MASTERARBEIT

SEGMENTIERUNG VON KUNDENDATEN IM E-COMMERCE KONTEXT DURCH DEN EINSATZ VON MACHINE LEARNING ALGORITHMEN ALS BASIS FÜR PERSONALISIERUNG

ausgeführt am

Studiengang
Informationstechnologien und Wirtschaftsinformatik

Von: Kevin Golob BSc
Personenkennzeichen: 1810320007

Graz, am 10. Juli 2020 .....

Unterschrift

EHRENWÖRTLICHE ERKLÄRUNG

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen nicht benützt und die benutzten Quellen wörtlich zitiert sowie inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

.....

Unterschrift

DANKSAGUNG

An dieser Stelle möchte ich mich bei all denjenigen bedanken, die mich während der Anfertigung dieser Masterarbeit unterstützt und motiviert haben.

Zuerst gebührt mein Dank Herrn Dipl.-Ing Hans-Peter Grahsl, der meine Masterarbeit betreut und begutachtet hat. Für die hilfreichen Anregungen und die konstruktive Kritik bei der Erstellung dieser Arbeit möchte ich mich herzlich bedanken.

Ich bedanke mich bei dem Experten Dipl.-Ing Christian Langmann, der mir geduldig auf meine Fragen während unseres Interviews geantwortet und als Ansprechpartner immer für mich Zeit gefunden hat.

Ein besonderer Dank geht an Dr. Vid Jelen, Univ. Dipl. Biotech., der mich in Machine Learning Themen beraten und unterstützt hat. Danke für die Vorschläge, die Anregungen und den Stoß in die richtige Richtung.

Abschließend möchte ich mich bei meiner Familie, vor allem meiner Frau und meinem Sohn bedanken, die Geduld und Verständnis für die zahlreichen Wochenenden, die ich dieser Arbeit gewidmet habe, aufgebracht haben.

KURZFASSUNG

Machine Learning ist seit vielen Jahren ein aufstrebender Begriff in der IT-Branche. Unabhängig von der jeweiligen Branche werden in nahezu allen Unternehmen, die Produkte oder Services an Kunden vertreiben, Daten über die Kunden im Rahmen des Verkaufsprozesses in diversen Formen gespeichert. Produktempfehlungen oder personalisierte Inhalte sind dabei ein Weg, die Kunden noch direkter und effektiver anzusprechen. Oft verwenden Unternehmen dazu Analysen aus Marketingabteilungen und legen statische Regeln für diese Inhalte fest. Diese Aufgabe kann jedoch auch Machine Learning übernehmen. Dabei sind vor allem bislang unbekannte Zusammenhänge innerhalb der verfügbaren Daten von Interesse. Welche neuen Informationen können mittels automatisierter Analysen gewonnen werden?

In dieser Arbeit wurde eine Clusteranalyse von 250.000 anonymisierten Onlinekunden des weltweit zweitgrößten Möbelhändlers, der auch im E-Commerce vertreten ist, mithilfe von Machine Learning Algorithmen vorgenommen. Dazu wurde ein Experteninterview durchgeführt, um die für das Unternehmen interessanten Aspekte der Kunden in die Analyse aufnehmen zu können. Es wurden statische Regeln für eine Einteilung der Kunden erstellt, die im Anschluss mit den Ergebnissen aus der automatisierten Clusteranalyse gegenübergestellt wurden. Nach dem Vergleich mehrerer Ansätze wurde festgestellt, dass der K-means Algorithmus die gestellten Anforderungen für eine anschließende Analyse der Clusterinhalte am besten erfüllen konnte. Die abschließende Analyse der erstellten Cluster hat gezeigt, dass eine Verwendung von automatisch erstellten Segmenten zur Anwendung von Personalisierung machbar ist. Die Gemeinsamkeiten von Kunden innerhalb einzelner Cluster konnten anhand der Daten belegt werden. Darüber hinaus konnten neue, interessante Zusammenhänge aus den Clustern gelesen werden. So hat sich zum Beispiel herausgestellt, dass es nicht die vom Experten genannten Zielgruppen waren, die häufig bestellt oder viel Geld ausgegeben haben.

ABSTRACT

The IT industry's attention is being increasingly drawn to machine learning. Every product- or service-provider collates customer data. Product recommendations or personalised content can help customers. Static rules based on marketing analyses are often used to provide this content. Machine learning algorithms can perform this duty. The conclusions which can be drawn from the data can be of great interest, potentially identifying information beyond the scope of automatic analysis.

This thesis examines data from 250,000 anonymised online customers, provided by the second biggest furniture retailer in the world, into a machine learning cluster analysis. An expert interview is conducted to gather all relevant customer attributes. These attributes are then analysed. To compare the result of the automatic clusters, static rules for segmentation are defined based on the chosen attributes. After comparing different approaches, the K-means algorithm is shown to fit the needs and requirements for an analysis of the cluster contents best. The final examination of the clusters shows that automatic created clusters can be used for personalisation. Common features of customers inside the clusters from the data can be shown. Beyond the cluster contents, there is interesting and unexpected information hidden in the data. The target groups defined by the expert are not the ones that had a high volume of orders or spent large sum of money.

INHALTSVERZEICHNIS

1 EINFÜHRUNG

In vielen Bereichen werden heutzutage Kundendaten angesammelt. Es gibt kaum noch Dienste oder Webseiten, bei denen man ohne ein Kundenkonto auf den vollen Serviceumfang zugreifen kann. Für Unternehmen sind Kundendaten dabei so wichtig wie nie zuvor. Anhand dieser Daten lassen sich Benutzerprofile generieren und sie sind Ausgangspunkt zahlreicher Analysen.

Es gibt viele unterschiedliche Herangehensweisen wie Unternehmen mit Kundendaten verfahren. Manche Unternehmen stehen erst am Anfang der Digitalisierung und beginnen damit ihre archaischen Unterlagen in Datenbanken und Managementsysteme zu überführen, andere wiederum haben die Wichtigkeit dieser Daten im Kampf um Wettbewerbsvorteile längst erkannt.

Ein Wesentlicher Vorteil, dessen sich Unternehmen dabei bedienen können ist, gezielt auf die Bedürfnisse Ihrer Kunden eingehen, oder auch Vorhersagen treffen zu können. Jeder der schon einmal auf Amazon unterwegs war, kennt die Produktvorschläge auf der Startseite, oder die „Andere Kunden kauften auch“ Slider auf den Produktdetailseiten. Doch wie können solche Unternehmen wissen was ihre Kunden interessiert? Wie funktioniert das?

1.1 Zielsetzung

Innerhalb dieser Arbeit werden Kundendaten aus dem SAP Commerce Cloud System eines der größten Onlinehändler im Möbelhandel anhand von Machine Learning (ML) analysiert und in verschiedene Kundensegmente eingeteilt. Dabei werden zwei unterschiedliche Algorithmen verwendet und verglichen. Die Ergebnisse werden dokumentiert, interpretiert und in Folge eine Erkenntnis abgeleitet.

Ziel der Arbeit ist es, mittels Machine Learning Kundendaten in Kundensegmente zu unterteilen, um diese in weiterer Folge für verschiedene Zwecke zur Verfügung zu stellen. Eine mögliche Anwendung wäre es, die Kundensegmente zu nutzen, um Kunden in verschiedenen Segmenten personalisierten Inhalt anzuzeigen. Machine Learning Algorithmen sollen dabei neue, bisher unbekannte Zusammenhänge in den Datensätzen erkennen können, die eine präzisere Anwendung von Personalisierung ermöglichen.

Es wird erwartet, dass der Kunde die Kundensegmente auch zu diesem Zweck nutzen kann. Das bedeutet, dass die entstandenen Segmente auf Gemeinsamkeiten hin untersucht werden müssen, um dem Kunden eine Personalisierung zu ermöglichen. Beispiele wären, ein Segment, in dem weibliche Kunden eines bestimmten Alters vertreten sind, oder ein Segment in dem Kunden Käufe von mindestens €500 getätigt haben.

Je nach Ergebnis der Literaturrecherche, werden dabei überwachte oder auch unüberwachte Machine Learning Algorithmen zur Segmentierung der Kundendaten verwendet. Es könnte sich

aus der Recherche auch ergeben, dass beide Ansätze zu verfolgen und gegenüberzustellen sind.

1.2 Motivation

Benutzerdaten lassen sich mithilfe statischer Regeln einfach in Segmente unterteilen. Wie eingangs kurz erwähnt, lassen sich Klassifizierungen anhand eines oder mehrerer Merkmale treffen. Diese Merkmale werden häufig von Menschen manuell festgelegt. Meist wird jemand aus dem Marketing oder ein Experte in der jeweiligen Branchendomäne diese Entscheidungen treffen.

Es gibt jedoch auch die Möglichkeit, dies komplett automatisiert und dynamisch mittels Machine Learning umzusetzen. Dabei werden die Segmente, denen die Kunden zugeteilt werden, aufgrund der angesammelten Daten bestimmt. Daraus resultiert eine Segmentierung, basierend auf Mustern, die von ML Verfahren erkannt wurden.

Die Fragen, die sich dabei stellen lauten: Sind die automatisch erstellten Segmente ähnlich? Gibt es Unterschiede in den Segmenten? Sind sie als Basis geeignet, um damit schlussendlich Produktvorschläge oder personalisierte Inhalte zu generieren?

1.3 Forschungsfrage

Die Forschungsfrage, die innerhalb dieser Arbeit beantwortet werden soll, lautet:

Welche neuen Zusammenhänge zwischen Daten lassen sich bei der Segmentierung von Kundendaten im E-Commerce Kontext durch den Einsatz von Machine Learning Algorithmen, im Vergleich zur Segmentierung aufgrund statischer Regeln, erkennen?

1.4 Methoden

Zur Beantwortung der Forschungsfrage werden zwei Wissenschaftliche Methoden verwendet.

Experteninterview mit dem Kunden

Um zu verstehen was der Kunde mit den Kundensegmenten erreichen will, bzw. anhand welcher Kriterien die Segmente entstehen sollen, wird ein Experteninterview geführt. Es wird in Erfahrung gebracht, welche Daten in das Segmentierungsmodell überführt werden sollen. Denkbar wäre auch das Segmentieren aufgrund unterschiedlich aufbereiteter Kundendaten. Einteilung nach demographischen oder bestellbezogenen Kundendaten als Basis für die Kundensegmente sind hier möglich. Weiters soll geklärt werden wie der Kunde mit den entstandenen Segmenten verfahren will. Mögliche Anwendungen wären hier die einmalige Einteilung in Segmente, die danach als Basis für personalisierte Inhalte verwendet wird, oder aber eine ständige automatische Auswertung der Daten, wodurch sich Segmente jederzeit

ändern können. Dabei könnten jederzeit neue Segmente entstehen, verschwinden oder sich die Zuteilung der Kunden zu den jeweiligen Segmenten ändern.

Auch der wirtschaftliche Faktor soll dabei eruiert werden. Als Beispiel wäre die Laufzeiten zu nennen, da dies Rechenzeit und somit Ressourcen benötigt. Je nachdem wie der entsprechende Service implementiert ist und wo er zur Verfügung gestellt werden soll (On-Premises vs. Cloud) ergeben sich unterschiedliche Auswirkungen auf die damit verbundenen Kosten.

Quantitative, statistische Analyse von Messwerten

Aus der Literatur werden gängige Modelle und Vorgehensweisen entnommen, mit deren Hilfe ein vorhandenes Datenset aus Kundendaten mittels ML segmentiert wird.

Die Ergebnisse aus den unterschiedlichen Modellen werden anhand von statistischen Merkmalen analysiert. Denkbare Messwerte wären unter anderem:

- Anzahl der Segmente
- Optimale Anzahl der Segmente (falls vorhanden)
- Statistische Analyse (Ähnlichkeiten) der Daten innerhalb der Segmente
- Genauigkeit der Vorhersagen/Zuteilung
- Vergleich zwischen statischer und automatischer Segmentierung

Aufbau der Arbeit

In den nachfolgenden Kapiteln werden wichtige Bestandteile erläutert, die für das Verständnis und den Kontext, in dem sich diese Arbeit bewegt, notwendig sind. Wie der Titel dieser Arbeit schon verrät, geht es um Anwendung von Machine Learning im E-Commerce Umfeld. Dazu wird ein Überblick über die Bereiche E-Business und E-Commerce in Kapitel 2 geboten. Es folgt eine Abgrenzung verschiedener Begriffe aus den Bereichen Computer- und Data Science in Kapitel 3, bevor es zum Hauptthema, Machine Learning, in Kapitel 4 übergeht. Vor der praktischen Umsetzung beginnend mit Kapitel 8, wird noch das Vorgehen und der Ablauf des Experteninterviews behandelt. Zum Abschluss werden die Ergebnisse präsentiert und die Forschungsfrage beantwortet.

2 E-BUSINESS UND E-COMMERCE

In diesem Kapitel werden die beiden Begriffe E-Business und E-Commerce erläutert und voneinander abgegrenzt. Zudem sollen typische Elemente eines Webshops und mögliche Einsatzbereiche von ML in E-Commerce Systemen identifiziert werden.

Der Begriff Electronic Business, oder E-Business beinhaltet die Anbahnung, Vereinbarung und Abwicklung elektronischer Geschäftsprozesse. Es findet also ein Leistungsaustausch über öffentliche oder private Kommunikationsnetze statt, um eine Wertschöpfung zu erzielen. Wie in Abbildung 2-1 zu sehen ist, gibt es drei Gruppen von Marktteilnehmern. Private Konsumenten (Consumer), Unternehmen (Business) und öffentliche Institutionen (Administration). Jede dieser Gruppen kann sowohl als Anbieter, als auch als Nachfrager einer Leistung auftreten. Durch die Kombination aus Anbietern und Nachfragern entstehen neun grundsätzliche Geschäftsbeziehungen (Meier & Stormer, 2012).

		Leistungsnachfrager		
		Consumer	**Business**	**Administration**
Leistungsanbieter	**Consumer**	**Consumer-to-Consumer (C2C)** z.B. Kleinanzeige auf einer persönlichen Homepage	**Consumer-to-Business (C2B)** z.B. Webseite mit persönlichem Fähigkeits-profil	**Consumer resp. Citizen-to-Administration (C2A)** z.B. Bürger bewertet öffentliches Umwelt-projekt
	Business	**Business-to-Consumer (B2C)** z.B. Produkte und Dienstleistungen in einem eShop	**Business-to-Business (B2B)** z.B. Bestellung bei Lieferanten (Supply Chain)	**Business-to-Administration (B2A)** z.B. elektronische Dienst-leistungen für öffentliche Verwaltungen
	Administration	**Administration-to-Consumer resp. Citizen (A2C)** z.B. Möglichkeit für elektronische Wahlen	**Administration-to-Business (A2B)** z.B. öffentliche Ausschreibung von Projektvorhaben	**Administration-to-Administration (A2A)** z.B. Zusammenarbeits-formen virtueller Gemeinden

Abbildung 2-1: Vielfalt elektronischer Geschäftsbeziehungen in Anlehnung (Meier & Stormer, 2012)

Die beiden Optionen Business-to-Consumer (B2C) und Business-to-Business (B2B) werden zu dem Begriff E-Commerce gezählt, da hier Unternehmen Produkte oder Dienstleistungen für Kunden oder andere Unternehmen bereitstellen (Meier & Stormer, 2012).

Eine Definition des Terms Electronic Commerce (EC) liefern beispielsweise (Turban, King, & Lang, 2010) wie folgt:

> *Electronic Commerce (EC) is the process of buying, selling, transferring, or exchanging products, services, and/or information via computer networks, mostly the Internet and intranets*

Der Unterschied zwischen E-Business und E-Commerce liegt also darin, dass der Begriff E-Business mehr umfasst als nur das Kaufen und Verkaufen von Produkten und Services. Hier geht es unter anderem auch um die Kollaboration zwischen Geschäftspartnern oder auch dem Anbieten von E-Learning, während die Definition von E-Commerce deutlich enger gefasst ist.

Wie oben bereits angeführt gehört vor allem B2C in den EC Kontext. Eine klassische Form von B2C bilden Webshops, Online-Handelsplattformen auf der Unternehmen Produkte und Services direkt an Endkunden vertreiben können.

Anwendung von Machine Learning im E-Commerce/Webshop Bereich

Nachdem der Begriff EC bereits abgesteckt wurde, werden nun typische Anwendungsbereiche von ML im EC-Kontext angeführt und kurz beschrieben. Dies soll einen Überblick über die vielen Möglichkeiten, die ML uns bietet, verschaffen. Bei (Singh, 2019) werden acht Anwendungszwecke beschrieben.

Kunden Segmentierung zur Personalisierung von Services und Inhalten

Kundensegmentierung teilt die Kunden verschiedenen Gruppen (Segmenten) zu. Dies kann aufgrund unterschiedlicher Unterscheidungsmerkmale geschehen. Dabei gibt es *demografische-* (Alter, Rasse, Geschlecht…), *geografische-* (Lebensraum, Arbeitsumgebung…), *psychografische-* (Soziale Stellung, Persönlichkeit…) und *Verhaltensmerkmale* (Kaufverhalten, Konsum…). Durch diese Einteilung können Unternehmen die Kunden gezielter ansprechen und Inhalte auf ihrer Webseite personalisieren. Werbung für bestimmte Produkte oder Services kann zum Beispiel nur Kunden gesendet oder angezeigt werden, die in der Vergangenheit Interesse daran gezeigt haben. Somit lassen sich Marketingressourcen effektiv einsetzen und es erlaubt Cross- und Up-Selling Potenziale besser zu nutzen. Nicht nur die Unternehmen, sondern auch die Kunden profitieren davon, da die Werbung relevanter und auf ihre Bedürfnisse zugeschnitten ist (Rouse, What is Customer Segmentation?, 2019).

Preisoptimierung

Hier geht es darum den besten Preis für sein Angebot zu finden. Was einfach klingt, ist in der Praxis allerdings alles andere als das. Der beste Preis hängt von vielen einzelnen Faktoren ab, die sich zudem noch ständig ändern können. Welchen Preis bietet ein Konkurrent? Wie sieht die Nachfrage derzeit am Markt aus? Diese und noch zahlreiche andere Überlegungen müssen

bei der Preisgestaltung miteinfließen. Mittels ML kann ein Modell erstellt werden, das sämtliche Daten, die für eine Preisberechnung notwendig sind, miteinbezieht. Die Daten müssen für diesen Zweck gesammelt und von Experten aufbereitet werden, um sicherzustellen, dass das Modell zuverlässig ist.

ML für den Zweck der Preisoptimierung zu nutzen bietet, abgesehen davon, dass es für Menschen unmöglich ist diese Berechnungen in kürzester Zeit anzustellen, einige Vorteile. Maschinen gehen an die Lösung eines Problems anders heran und finden damit Wege, an die ein Mensch nicht denken würde. Durch eine Berechnung des Preises für ein Portfolio an Produkten aus den Informationen mehrerer Datenquellen in Echtzeit kann man Trends noch früher erkennen. Außerdem lässt sich die Genauigkeit der Vorhersagen bestimmen. Reichen 90% oder müssen es 99% sein?

Fluglinien, Hotels und Modeketten sind nur einige Beispiele für Unternehmen, die auf diese Technik angewiesen sind, um im Preiskampf überleben zu können (Prize Optimization, 2019).

Schutz vor Betrug

Machine Learning kann benutzt werden, um sich vor Betrug (Fraud) zu schützen. Betrug ist eine Aneignung von Geld oder Besitz an einer Sache durch Vortäuschung falscher Tatsachen. Bei Banken gibt es unter anderem den Scheck oder Kreditkartenbetrug. Auch Versicherungsbetrug durch das absichtliche Herbeiführen von Unfällen wäre ein Beispiel. Menschen finden immer neue Wege zu betrügen, was es schwierig macht solche Fälle aufzudecken. Ein Weg sich dagegen zu Wehr zu setzen, bieten statistische Analysen und die Erkennung von Mustern. ML versucht hier bestimmte Charakteristiken in den Daten zu identifizieren, um so Betrugsfälle erkennen zu können (Rouse, What is fraud detection?, 2019).

Optimierte Suchergebnisse

Ein wichtiges Feature bei B2C Webseiten ist die Suche nach Produkten oder Services. Kunden müssen in der Lage sein zu finden wonach sie suchen. Aber nicht alle Kunden sind gleich und verwenden die gleichen Suchbegriffe. Die Suchfunktion sollte also mehr als nur Stichworte für die Suche heranziehen. Hier kann ML Muster in den Suchbegriffen, den Käufen der Kunden und deren Vorlieben extrahieren und dann die optimalen Suchergebnisse dazu ausliefern. Auch Produktvorschläge können auf diese Weise gemacht werden.

Produktempfehlungen

Empfehlungssysteme sollen sogenannte Items für einen aktiven Benutzer empfehlen. Bei diesen Items kann es sich zum Beispiel um Produkte, Dienstleistungen oder Orte handeln. Dazu werden vorhandene Informationen des Benutzers oder der Items herangezogen und ermittelt, ob das Item den Geschmack des Benutzers trifft. Dabei wird eine vorhergesagte Bewertung von Items für den Benutzer berechnet und eine Liste von empfehlenswerten Items für den Benutzer generiert. Organisationen wie Amazon oder Netflix verwenden diese Art von

Systemen, um den Kunden Produkte oder Filme aufgrund ihrer Präferenzen vorzuschlagen (Wörndl & Schlichter, 2019).

Customer Support

Machine Learning kann durch eine Automatisierung des Customer Supports dabei helfen die Kundenzufriedenheit zu erhöhen. Ein Beispiel hierfür sind Chatbots, die versuchen in einer Konversation mit den Kunden Probleme zu identifizieren und eine geeignete Lösung vorzuschlagen. Dies hilft den Kunden als auch dem Unternehmen selbst durch reduzierte Anfragen an den Support.

Managen von Angebot und Nachfrage

Unternehmen müssen auf die Nachfrage nach Produkten mit einem entsprechenden Angebot reagieren. Um dies zu erreichen, benötigt man eine Vorhersage, die auf Daten beruht. Diese Daten müssen dabei akkurat sein und auf die richtige Weise verarbeitet werden. Machine Learning kann große Mengen an Daten schnell verarbeiten und dabei aus ihnen lernen, um neue Einblicke zu gewähren. Das kann die Vorhersagen verbessern und das Unternehmen kann aus dem gelernten schöpfen, um seine Produkte oder Services zu verbessern.

Omnichannel Marketing

Das Ziel des Omnichannel Marketing ist, ein umfassendes Kundenerlebnis über alle Werbe- und Verkaufsplattformen eines Unternehmens zu schaffen. Kunden können dabei überall und jederzeit auf unterschiedlichen Wegen zu einer Kaufentscheidung gelangen. Es wird definiert als eine Strategie, bei der digitale, analoge sowie physische Kanäle zur Interaktion mit Kunden genutzt werden. Kundenrelevante Daten sind über alle Kanäle verfügbar. Die Customer Journey ist auf allen Plattformen einheitlich und auf die jeweilige Person abgestimmt. Machine Learning kann diese Strategie durch das Sammeln und Analysieren der Kundendaten erheblich unterstützen (Omnichannel Marketing Definition und Strategien, 2018).

3 DATA SCIENCE

Viele Artikel stützen sich auf Begriffe wie Data Science, Machine Learning, Künstliche Intelligenz (KI), Business Intelligence (BI) oder Data Mining (DM). Oft werden diese jedoch nicht voneinander abgegrenzt und dadurch tendenziell wie Synonyme behandelt. Dieses Kapitel soll diesem Umstand entgegenwirken und auf einige dieser Begriffe näher eingehen, um besser verstehen zu können wo Unterschiede und potenzielle Gemeinsamkeiten liegen.

Eine Definition von Data Science findet sich bei (Kelleher & Tierney, 2018):

> *Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting non-obvious and useful patterns from large data sets.*

Während sich Machine Learning auf Algorithmen und Mustererkennung fokussiert, beschäftigt sich Data Mining mit der Analyse strukturierter Daten. Data Science geht darüber hinaus und stellt sich auch anderen Problemen, wie der Gewinnung, der Bereinigung und der Transformation von unstrukturierten Daten. Auch die Handhabung von rechtlichen Verordnungen und der Ethik im Umgang mit Daten gehören in den Bereich der Data Science.

Der Term Data Science selbst stammt aus den 1990er Jahren im Rahmen von Gesprächen, in denen Statistiker und Informatiker sich einigten Rigor, die Mathematische Strenge, auf die Analyse von großen Datenbeständen anzuwenden. C.F. Jeff Wu hielt im Jahr 1997 einen Vortrag mit dem Titel „Statistics = Data Science“ in dem er vorschlug den Begriff der Statistik in Data Science umzubenennen.

In den vergangen beiden Jahrzehnten ist die Menge an Daten, die durch online Aktivitäten generiert wird, rasant gestiegen. Damit einhergehend, hat sich auch das Konzept von Data Science weiterentwickelt. Ausgehend von einem Synonym für den Begriff der Statistik bis zur Definition, die oben beschrieben ist. Heutzutage ist die Definition eines Data Scientists so breit gefächert, dass es in der Arbeitswelt keine Einigung darüber gibt, welche Ausbildung und Fähigkeiten zur Erfüllung dieser Rollenbeschreibung Voraussetzung sind. Einen Überblick über die verschiedenen Skills eines Data Scientists zeigt sich in Abbildung 3-1 (Kelleher & Tierney, 2018).

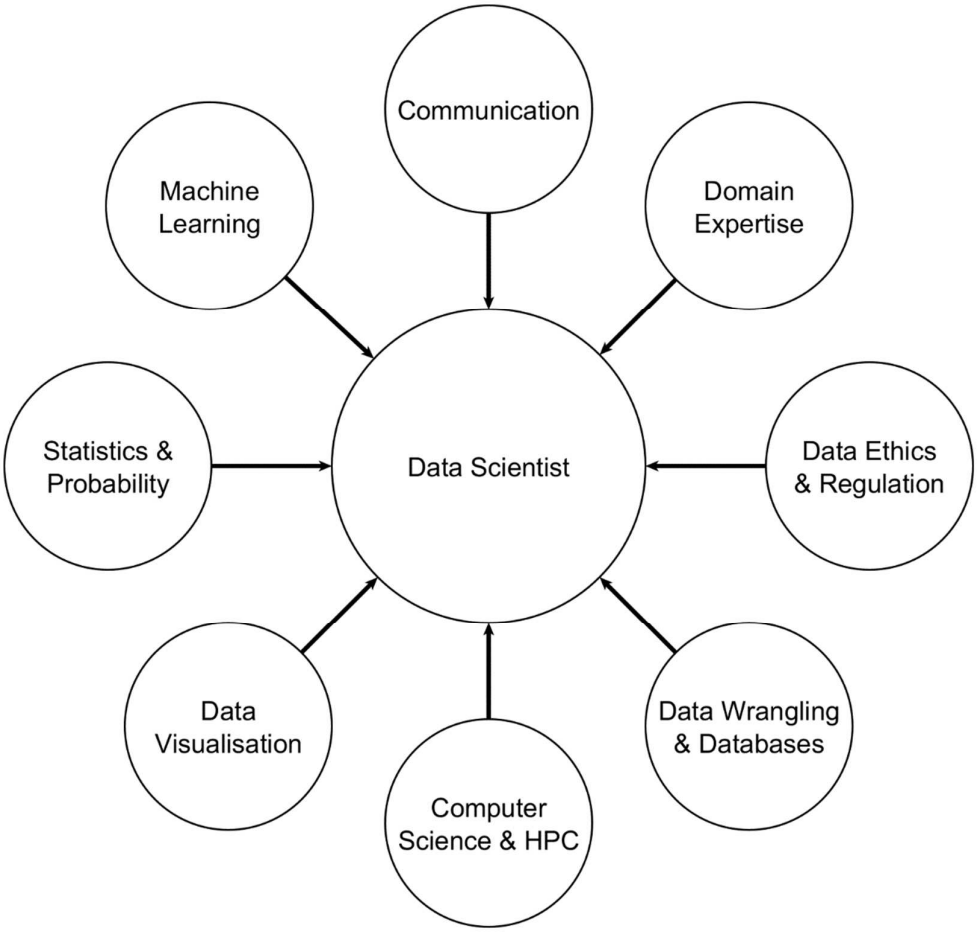

Abbildung 3-1: Erwünschtes Skill-set von Data Scientists in Anlehnung (Kelleher & Tierney, 2018)

3.1 Machine Learning

"Much of what we do with machine learning happens beneath the surface. Machine learning drives our algorithms for demand forecasting, product search ranking, product and deals recommendations, merchandising placements, fraud detection, translations, and much more. Though less visible, much of the impact of machine learning will be of this type — quietly but meaningfully improving core operations." (Bezos)

Machine Learning ist ein Begriff, der vielen Menschen heutzutage geläufig ist. Jeff Bezos, der Gründer von Amazon, nennt in seinem Zitat Beispiele für die Anwendung von ML, die bereits heute unseren Alltag maßgeblich bestimmen. Unternehmen setzen diese ML Verfahren ein, um den Zugang zu ihren Kunden neu zu gestalten und ihnen somit wesentlich mehr Nutzen zu stiften, als es in vergangen Jahrzehnten möglich gewesen wäre. Oft wird Machine Learning im alltäglichen Gebrauch mit dem Begriff der Künstlichen Intelligenz verwechselt oder aber gleichgesetzt. Was aber steckt hinter diesen Begriffen?

Machine Learning ist ein Teilbereich der Informatik, der sich mit Algorithmen beschäftigt. Computer benötigen Algorithmen, um praktische Probleme zu lösen. Sie sind Sequenzen von Befehlen, die ausgeführt werden, um einen bestimmten Input in einen Output zu transformieren. Dafür werden Datensets benötigt, die mit Hilfe eines Algorithmus in ein statistisches Modell überführt werden. Dieses Modell kann letztlich verwendet werden, um ein vorliegendes Problem zu lösen (Burkov, 2019).

What we lack in knowledge, we make up for in data. (Alpaydin, 2014)

Für manche Probleme haben wir allerdings keine Algorithmen, die uns bei deren Lösung helfen könnten. Beispiele hierfür wären etwa die Vorhersage von Konsumentenverhalten oder aber das Unterscheiden zwischen Spam und relevanten Mails. Wir haben Kenntnisse von Input (den Mails) und Output (Spam oder kein Spam), aber keine Ahnung wie der Input in passenden Output zu transformieren ist. Auch kann sich im Laufe der Zeit ändern, was wir unter Spam verstehen. Für solche Probleme, wo kein unmittelbarer, direkter Lösungsweg bekannt ist, benötigt man Daten, aus denen man durch das Erkennen bestimmter Zusammenhänge lernen kann. Ein anschauliches Beispiel dafür wäre folgende Aufgabenstellung: Man kann tausende Emails, bei denen man weiß, bei welchen es sich um Spam handelt und bei welchen nicht, als Input benutzen und den Computer aus diesen Daten einen Algorithmus zur Unterscheidung extrahieren lassen. Was dabei herauskommt, ist eine Annäherung einer Lösung. Auch wenn diese nicht alles erklärt, gehen wir davon aus, dass sie gut genug ist, um einen Teil unserer Daten zu klassifizieren. Auch wenn es nicht gelingt eine perfekte Lösung des Problems zu finden, lassen sich Muster und Regelmäßigkeiten in den Daten entdecken, die wir verwenden können, um in der Praxis sinnvolle Vorhersagen zu treffen (Alpaydin, 2014).

Das Hauptziel von ML ist mathematische Modelle, zu analysieren, zu entwickeln und zu verbessern. Es handelt sich dabei um Modelle die mithilfe kontextspezifischer Daten trainiert werden können, um Annahmen über die Zukunft zu treffen, ohne dabei alle externen Variablen zu kennen. Dabei spielen Agenten eine wichtige Rolle. Agenten lassen sich beschreiben als Instanzen in der Software, die Input ihrer Umgebung aufnehmen, die beste Aktion auswählen, um ein bestimmtes Ziel zu erreichen und das resultierende Ergebnis beobachten. Diese Agenten verwenden statistische Methoden um zu lernen und versuchen dabei Wahrscheinlichkeitsverteilungen zu bestimmen, um daraus die Aktion auszuwählen, die mit der größten Wahrscheinlichkeit zum Erfolg führt (Bonaccorso, Machine Learning Algorithms Popular algorithms for data science and machine learning, 2018).

3.2 Künstliche Intelligenz

Der Begriff der Künstlichen Intelligenz (KI), oder auch Artificial Intelligence (AI), wie es im Englischen heißt, wurde erstmals von John McCarthy im Jahre 1955 definiert. Seiner Ansicht nach war es das Ziel der KI Maschinen zu entwickeln, die sich verhalten, als ob sie über Intelligenz verfügten (Ertel, 2016).

Eine aktuellere Definition der KI bieten (Rich, Knight, & Shivashankar, 2017) wie folgt:

> *Artificial Intelligence (AI) is the study of how to make computers do things at which, at the moment, people are better.*

Während uns Computer in der Ausführung von vielen Berechnungen in kürzester Zeit bereits überlegen sind, gibt es andere Anwendungsbereiche, in denen der Mensch der Maschine noch einiges voraushat. Adaptivität ist eine Stärke der menschlichen Intelligenz. Menschen sind in der Lage sich innerhalb von Sekunden an unterschiedliche Umweltbedingungen anzupassen und dabei durch Lernen ihr Verhalten zu ändern. Gerade bei der Lernfähigkeit sind wir Menschen den Computern überlegen. Das obige Zitat ist zeigt also, dass gerade maschinelles Lernen ein wichtiges Teilgebiet der KI ist (Ertel, 2016).

(Neukart, Reverse Engineering the Mind, 2016) beschreibt in seinem Buch fünf Säulen der Künstlichen Intelligenz:

- Maschinelles Lernen
- Computer Vision
- Schließen und Entscheidungsfindung
- Sprache und Kommunikation
- Agenten und Aktionen

Das Maschinelle Lernen unterteilt sich in zwei Klassen - überwacht und nicht überwacht. Diese werden im nächsten Kapitel noch ausführlicher beschrieben. Bei Computer Vision handelt es sich um ein Forschungsfeld, das versucht von Kameras aufgenommene Bilder zu verarbeiten. Es geht darum den Inhalt zu verstehen und Vorgänge zu interpretieren. Anwendung findet dieses Gebiet vor allem in der autonomen Navigation von Robotern oder Fahrzeugen. Bei dem dritten Punkt handelt es sich um das Feld der Wissensrepräsentation (engl. knowledge representation & reasoning, KRR). Beim Schließen müssen datenbasierte Lösungen zu Problemen ohne menschliche Intervention oder Hilfe gefunden werden. Bei der Entscheidungsfindung steht die Beantwortung von Fragen über die Präferenzen zwischen Aktivitäten im Vordergrund. Beispielsweise wie ein autonomes System, etwa ein Fahrzeug, auf Veränderungen in der Umwelt reagieren muss. Die vierte Säule basiert auf der Verarbeitung von Sprache und teilt sich in zwei Felder. Komputationale Linguistik (Computational Linguistics, CL) und Verarbeitung von natürlicher Sprache (Natural Language Processing, NLP). CL fokussiert sich auf die Nutzung von Computern zur Sprachverarbeitung, während NLP sich auch anderer Ressourcen bedient. NLP erfordert eine konkrete Aufgabe und ist streng genommen keine eigene Forschungsdisziplin. Die letzte Säule bezieht sich auf die Agenten und deren Aktionen, wie sie schon oben beschreiben wurden.

3.3 Data Mining

Der Fortschritt in der Gewinnung von Daten und der Technologie der Speichermedien lässt Datenbanken immer weiter wachsen. Es ist kein Wunder, dass das Interesse daran steigt, diese Daten anzuzapfen und so darin schlummernde Informationen zu gewinnen. Diesen Prozess der Extraktion von Informationen aus riesigen Datenmengen nennt man Data Mining (DM). Eine Definition dieser Disziplin liefert (Hand, Mannila, & Smyth, 2001):

> *Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. (Hand, Mannila, & Smyth, 2001)*

Data Mining wird üblicherweise auf Daten angewandt, die bereits aus anderen Gründen als dem DM selbst, angesammelt wurden. DM hat also keinen Einfluss darauf wie die Daten gesammelt wurden, oder woher diese stammen. In diesem Punkt unterscheidet sich DM von der Statistik also in der Hinsicht, dass dort oft versucht wird Daten zur Beantwortung eines bestimmten Problems oder einer speziellen Fragestellung zu sammeln. Aus diesem Grund wird DM oft auch als Sekundäre Daten Analyse bezeichnet.

Wie im obigen Zitat zu lesen, geht es bei DM ebenfalls darum Neues zu entdecken. Neuartigkeit ist also Teil von DM. Es macht darum wenig Sinn, bereits bekannte Beziehungen und Muster in den Daten durch DM erneut entdecken zu lassen, außer man zielt darauf ab, eine bestehende Hypothese zu bestätigen. Das Konzept der Neuartigkeit ist allerding noch ein offenes Forschungsproblem, da es erfordert ein Vorwissen des oder der User miteinzubeziehen (Hand, Mannila, & Smyth, 2001).

3.4 Business Intelligence

Die Historie von Business Intelligence (BI) reicht zurück bis in die 60er Jahre. Damals wurde versucht, Führungskräfte durch den Einsatz von Informationssystemen, zu unterstützen. Die ersten Ansätze waren jedoch nicht erfolgreich und scheiterten. Im Laufe der Jahre entwickelten sich Einzelsysteme, die im Management erfolgreich eingesetzt werden konnten. In den 80er Jahren entstand der Sammelbegriff Management Support Systems (MSS) für diese Systeme. Der Begriff MSS findet auch heute noch Verwendung, insbesondere in der Wissenschaft. Mitte der 90er Jahre hat sich dann der Begriff der Business Intelligence, getrieben durch die Gartner Group, entwickelt (Kemper, Baars, & Mehanna, 2010).

Eine Definition von Business Intelligence wurde vom The Data Warehouse Institute (TDWI), einer Weiterbildungsplattform in den Themenbereichen Analytics, Big Data und BI, im Jahr 2002 in einem Newsletter veröffentlicht und von (Loshin, 2003) übernommen:

> *The processes, technologies, and tools needed to turn data into information, information into knowledge, and knowledge into plans that drive profitable business action. Business intelligence encompasses data warehousing, business analytic tools, and content/knowledge management. (Loshin, 2003)*

BI ist also nicht nur eine Ansammlung an Werkzeugen (Tools). Ohne die richtigen Prozesse und Menschen dahinter wird aus den Tools kaum Nutzen generiert. Zudem müssen gewonnene Informationen für einen Entscheidungsprozess genutzt werden, da der Einsatz von BI sonst wenig Sinn machen würde (Loshin, 2003).

4 MACHINE LEARNING TYPEN

Nach Abgrenzung der verschiedenen Begriffe voneinander im vorhergehenden Kapitel, soll in diesem Kapitel nun Machine Learning, das Hauptthema dieser Arbeit, näher betrachtet werden. Es werden die verschiedenen Teilgebiete von ML beschrieben. Eine Auswahl der verwendeten Methoden für die praktische Umsetzung folgt in den nächsten Kapiteln.

Wie in (Burkov, 2019) angeführt, gibt es vier unterschiedliche Typen des Lernens, *Supervised Learning*, *Unsupervised Learning*, *Semi-Supervised Learning* und das *Reinforcement Learning*. Abbildung 4-1 verschafft einen Überblick über diese Typen.

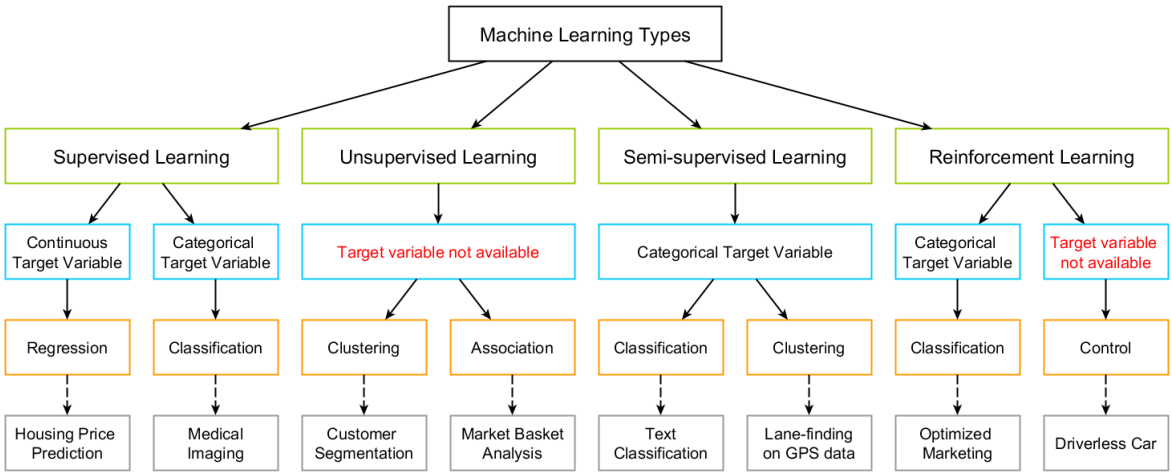

Abbildung 4-1: Machine Learning Typen nach (Fumo, 2015)

In der englischen Literatur werden Daten, die Ausgangsbasis für ML sind, unterteilt in *labeled* und *unlabeled* data. Labeled data bezeichnet Daten, bei denen man den Inhalt oder die Beschaffenheit kennt. Sie werden für Supervised Learning verwendet, da man hier Rückmeldung geben kann, ob das Lernen erfolgreich war und die Vorhersagen zutreffen. Als Beispiel dient hier wieder die Klassifizierung als Spam oder nicht-Spam. Wenn wir wissen welche Mails Spam sind, können wir klar zwischen Erfolg und Misserfolg der Klassifizierung unterscheiden. Bei Unsupervised Learning hingegen hat man kaum, teilweise auch gar keine Vorstellung, wie sich Daten verhalten oder anhand welcher Kriterien diese unterschieden werden sollten. Im Vergleich zu Supervised Learning werden bei diesem Typ für gewöhnlich mehr Daten als Input benötigt, um daraus Muster erkennen zu können.

4.1 Supervised Learning

Überwachte Lernverfahren erfordern neben den Eingangsvariablen (Prädiktoren) die bekannten Zielwerte (Labels) eines Problems. Für das Lernen wird ein Datenset zum Training und ein anderes Datenset zum Testen verwendet. Das Testset wird zur Überprüfung der vorhergesagten Daten nach dem Training verwendet. Mit diesen Daten kann ein Agent (siehe Abschnitt 3.1) schrittweise Fehler in der Vorhersage verringern, indem er Parameter anpasst und so das Ausmaß einer globalen Verlustfunktion reduziert. So wird nach jedem Durchlauf die Präzision erhöht und die Differenz zwischen vorhergesagtem und erwartetem Ergebnis minimiert. Um dem Problem der Überanpassung (engl. overfitting) entgegenzuwirken, muss das Modell in der Lage sein zu generalisieren, damit auch es auch mit nie zuvor gesehen Daten noch funktioniert (Bonaccorso, Machine Learning Algorithms Popular algorithms for data science and machine learning, 2018). Bei der Überanpassung, beschreibt das Modell Features, die sich aus Rauschen oder Abweichungen in den Trainingsdaten ergeben anstatt der zugrunde liegenden Verteilung dieser Daten. Dies führt zu einem Verlust der Genauigkeit bei Stichproben, die außerhalb der Trainingsdaten liegen (Webb, 2011).

Typische Anwendungsgebiete für Supervised Learning finden sich in folgenden Bereichen:

- Vorausschauende Analyse basierend auf Regression oder kategorialer Klassifizierung
- Erkennen von Spam
- Erkennen von Mustern
- Computerlinguistik
- Sentimentanalyse
- Automatische Bildklassifikation
- Automatische Sequenzverarbeitung (Musik oder Sprache)

Zu weit verbreiteten Algorithmen im Supervised Learning gehören nach (Fumo, 2015):

- K-Nächste Nachbarn
- Naive Bayes
- Entscheidungsbäume
- Lineare Regression
- Support Vector Maschinen (SVM)
- Neuronale Netzwerke

4.1.1 Regression

Regression ist ein Prozess, bei dem die Beziehung verschiedener Input Variablen zur Zielvariable durch Schätzungen ausgedrückt wird. Sie wird angewandt, wenn die Zielvariable, die vorhergesagt werden soll, beliebige Werte annehmen kann. Die Temperatur der nächsten Tage wäre ein Beispiel für eine Anwendung der Regression. Damit könnte man abschätzen, wie sich die Änderung von Variablen, wie etwa der Luftdruck, auf die Temperatur auswirken (Kaluza, 2016).

4.1.2 Classification

Wie oben schon erläutert, werden beim Supervised Learning labeled Datensets verwendet, die sich in Trainings- und Testdaten aufteilen. Das Ziel der Klassifikation ist es eine Zielvariable aus den Testdaten, nach dem Training vorherzusagen. Die möglichen Werte dieser Variable müssen sich dabei gegenseitig ausschließen, um zwischen einem Erfolg oder Misserfolg unterscheiden zu können. Ein Beispiel wäre die Kreditwürdigkeit einer Person. Als mögliche Werte gibt es hier ein „ja“ oder „nein, respektive „1“ oder „0“. Die am weitest verbreiteten Algorithmen beinhalten *Entscheidungsbäume* (engl. Decision Trees), *Naive Bayes, Support Vector Maschine, Neuronale Netzwerke*, und *Ensemble-Lernen* (Ensemble methods).

Entscheidungsbäume sind sogenannte „white box“ Verfahren. Das bedeutet, dass das Verfahren nachvollziehbar ist und man von der Ausgangssituation bis zur Lösung alle Schritte einsehen kann. Die Bäume bestehen aus Knoten und Kanten, wobei die Knoten einzelne Attribute und die Kanten deren mögliche Werte repräsentieren. Abbildung 4-2 zeigt als vereinfachtes Beispiel einen Baum, der mit einfachen „ja“ oder „nein“ Entscheidungen bei der Frage hilft, was man tun kann, wenn man Hunger hat. Wie man sieht, kann man die Problemlösung, also das Vorhersagemodell, auf diesem Weg einfach visualisieren (Kaluza, 2016).

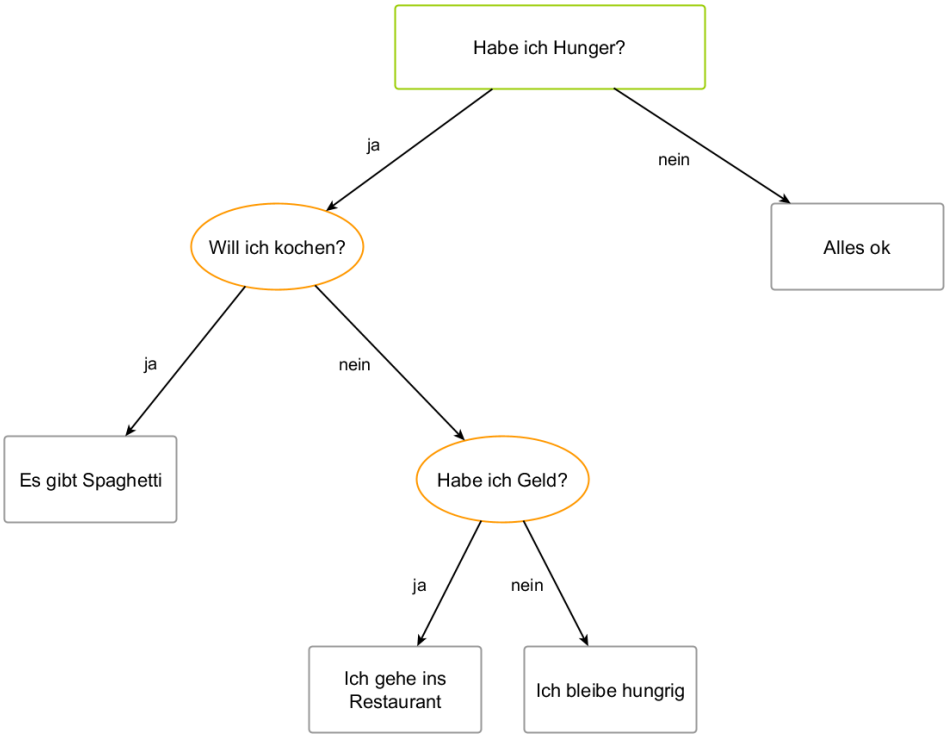

Abbildung 4-2: Entscheidungsbaum (eigene Darstellung)

4.2 Unsupervised Learning

Im Gegensatz zu Supervised Learning gibt es hier keinen „Aufpasser“ der über den Misserfolg des Ergebnisses Rückmeldung an den Agenten gibt. Auch die Aufteilung in Trainings- und Testdaten gibt es bei diesem Ansatz nicht. Hier gibt es nur die Ausgangsdaten die als Input dienen. Unsupervised Learning ist hilfreich um eine Klassifizierung, oder Gruppierung, der Daten anhand ähnlicher Merkmale zu erreichen. Dabei werden alle Eigenschaften der Elemente in dem Datensatz auf Gemeinsamkeiten hin untersucht. Die Algorithmen müssen Strukturen, Cluster, und Beziehungen zwischen den Daten selbständig ermitteln. Das Ergebnis des Lernens ist hier ein Modell, das die Daten den vorhandenen Clustern zuordnen kann. Die Anzahl und Art der Cluster kann sich dabei je nach verwendeten Daten ändern (Bonaccorso, Machine Learning Algorithms Popular algorithms for data science and machine learning, 2018). Abbildung 4-3 zeigt ein Beispiel mit zwei unterschiedlichen Clustern.

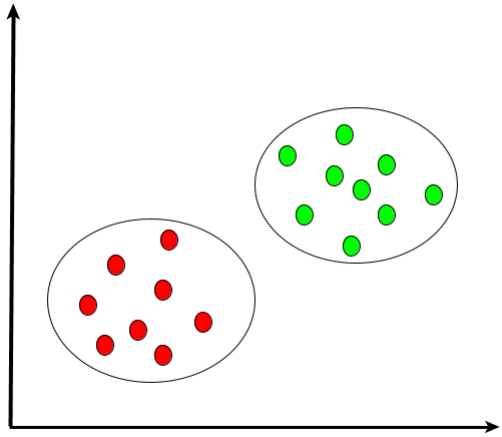

Abbildung 4-3: Beispiel für Cluster (eigene Darstellung)

Anwendungsgebiete von Unsupervised Learning umfassen zum Beispiel:

- Objektsegmentierung (Benutzer, Produkte, Filme…)
- Erkennen von Ähnlichkeiten
- Automatische Kennzeichnung
- Empfehlungsdienste

Zu weit verbreiteten Algorithmen im Unsupervised Learning gehören nach (Fumo, 2015):

- K-means Clustering
- Assoziationsanalyse

Unsupervised Learning wird vor allem für Probleme verwendet, die eine dynamische Lösungsstrategie erfordern. Alles was auf statischen Regeln beruht, wie zum Beispiel Prozesse, die sich in Flussdiagrammen abbilden lassen, sind keine Kandidaten für ML. Nach (Neukart, Hofmann, & Bäck, 2017) wird ML vor allem angewandt, wenn:

- Menschliche Expertise nicht existiert.
- Menschen ihre Expertise nicht ausdrücken können.
- Die Lösung sich über die Zeit verändert.
- Die Lösung auf spezifische Fälle angepasst werden muss.

4.2.1 Clustering

Das Ziel von Clustering Algorithmen ist es Gruppen, sogenannte Cluster, in unlabeled Datensets zu identifizieren. Es werden Daten, die sich in ein oder mehreren Aspekten ähnlich sind, gruppiert und Clustern zugeordnet. Für diese Einteilung werden die Distanzen zwischen den Datenpunkten berechnet. Punkte die weit auseinander liegen werden dann verschiedenen Clustern zugeteilt und jene die sich nahe sind, landen innerhalb eines Clusters. Bei dieser Methode gibt es zwei grundlegend verschiedenen Herangehensweisen.

Zum einen gibt es den *hierarchischen*, oder auch *agglomerativen* Ansatz genannt, bei dem jeder einzelne Datenpunkt als eigener Cluster angesehen wird. Ausgehend davon werden in einem iterativen Verfahren die Punkte, oder Cluster, die sich am ähnlichsten sind, zusammengefügt. Dies geschieht so lange, bis eine vordefinierte Anzahlt an Clustern erreicht ist, oder die Cluster so weit voneinander entfernt sind, dass das Zusammenlegen nicht mehr sinnvoll wäre.

Bei dem anderen Verfahren werden zunächst Mittelpunkte für Cluster bestimmt. Dies geschieht durch einfaches Schätzen. Danach werden die Punkte, die den Mittelpunkten am nächsten sind zu den jeweiligen Clustern zugeordnet. Dies wiederholt sich so lange bis alle Punkte einem Cluster angehören. Das bekannteste Verfahren nennt sich *K-means* clustering. Dabei werden die Clustermittelpunkte entweder so gewählt, dass sie am weitesten voneinander entfernt sind, oder der Algorithmus legt Cluster über verschiedene Punkte und nimmt den Punkt, der der Mitte des jeweiligen Clusters am nächsten ist (Kaluza, 2016).

4.2.2 Association Rules

Assoziation wird verwendet, wenn man interessante Beziehungen zwischen Klassen in einer Datenbank entdecken möchte. Meist wird dies im Handel verwendet, um herauszufinden welche Produkte häufig gemeinsam gekauft werden. Die Muster, die bei der Assoziation erkannt werden, werden in Regeln (engl. Rules) ausgedrückt. Folgende Regel lässt darauf schließen, dass ein Kunde der Zwiebeln und Kartoffeln kauft, sehr wahrscheinlich auch Burger kaufen wird:

$$\{Zwiebeln, Kartoffeln\} \rightarrow \{Burger\}$$

Ein klassisches Vorzeigebeispiel der Assoziation zeigt diese Regel:

$$\{Bier\} \rightarrow \{Windeln\}$$

Bei einer Analyse des Kundenverhaltens in einem Supermarkt zeigte sich, dass ein Kunde der Bier kauft, auch dazu neigt Windeln zu kaufen. Während das erste Beispiel nicht sonderlich verwundert, ist eine Assoziation zwischen Bier und Windeln ein unerwartetes Ergebnis, was auch die Popularität des Beispiels erklärt. Ein bekannter Algorithmus für dieses Verfahren nennt sich *Apriori* (Kaluza, 2016).

4.3 Semi-Supervised Learning

Bei diesem Ansatz besteht das Datenset aus labeled und unlabeled Daten. Ein Großteil der Daten gehört hier der Kategorie unlabeled an. Ein Grund dafür kann sein, dass es schwierig oder aufwändig ist die Daten zu benennen. Um Daten überhaupt mit einem Label versehen zu können, wird ein Experte auf dem jeweiligen Gebiet benötigt. Supervised Learning hier anzuwenden birgt die Gefahr, dass das resultierende Modell nicht repräsentativ für die gesamten Daten und somit unbrauchbar ist. Darum ist es notwendig einen Kompromiss zwischen Supervised- und Unsupervised Learning zu finden.

Das Vorgehen, bei dem ein Modell darauf abzielt, Labels für die unlabeled Werte zu finden, nennt man *Transductive Learning*. Wenn das Ziel ist, mehr über die unlabeled Werte herauszufinden, sollte diese Methode bevorzugt werden.

Inductive Learning versucht dagegen ein Modell zu finden, das das gesamte Datenset beschreibt und Proben beider Kategorien zum entsprechenden Label zuordnen kann. Diese Methode ist komplexer und erfordert mehr Rechenzeit (Bonaccorso, Mastering Machine Learning Algorithms, 2018) (Fumo, 2015).

4.4 Reinforcement Learning

Reinforcement Learning basiert, ähnlich wie Supervised Learning, auf Feedback. Dies kommt allerdings nicht von einem Supervisor, sondern der Umgebung (engl. Environment), in der die Maschine „lebt“. Der Agent erkennt den Zustand (engl. State), in dem sich das Environment befindet und kann dann Aktionen (Actions) ausführen. Diese Aktionen können wiederum eine Änderung in dem Zustand bewirken. Der Agent bekommt als Rückmeldung auf jede Aktion eine Belohnung (engl. Reward), wobei diese auch negativ sein kann. In solchen Fällen wird dann oft von Bestrafung (engl. Penalty) gesprochen. Diese helfen dem Agenten zu unterscheiden ob eine Aktion sinnvoll war oder nicht. Bei einer Belohnung wird der Agent in der Ausführung einer Aktion bestärkt. Eine Bestrafung bringt den Agenten dazu, sich in Zukunft anders zu entscheiden und eine andere Aktion auszuführen. So kann der Agent mit der Zeit herausfinden wie er zu einer maximalen Belohnung gelangt (Reinforcement Learning Definition & Erklärung, 2019) (Burkov, 2019). Abbildung 2-1 stellt die Interaktion des Agenten mit der Umgebung dar.

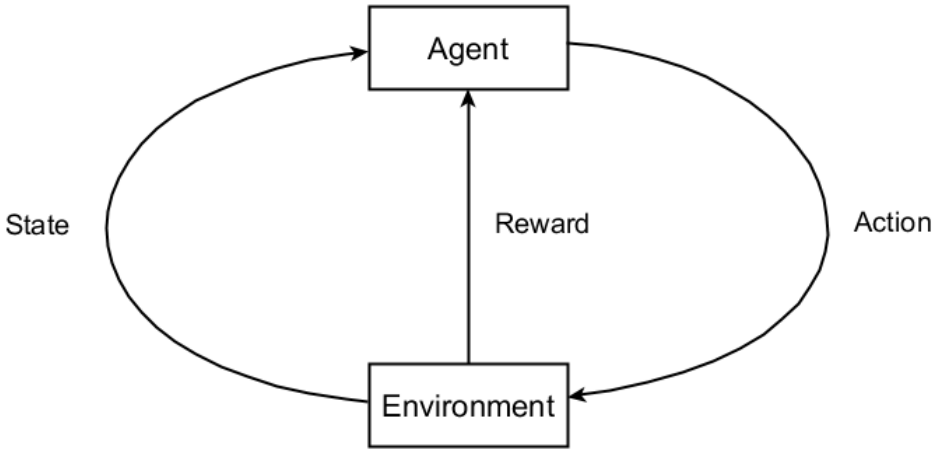

Abbildung 4-4: Reinforcement circle nach (Fumo, 2015)

5 CLUSTERING ALGORITHMEN

Wie in Kapitel 0 bereits beschrieben, gibt es zwei grundlegende Arten zu Clustern. Machine Learning bietet hierfür ein hierarchisches Modell, und das K-means Clustering als bekanntesten Vertreter in seinem Bereich. Nach (Tripathi, Bhardwaj, & Poovammal, 2018) sind dies auch die beiden Algorithmen, die im Bereich der Kundensegmentierung am häufigsten verwendet werden. Daher werden beide Vorgehen im Laufe dieser Arbeit verwendet und gegenübergestellt. Im Folgenden wird erklärt wie diese Algorithmen funktionieren, um zu verstehen wie Cluster entstehen.

5.1 Hierarchische Algorithmen

Bei hierarchischen Algorithmen wird versucht eine Hierarchie von Teilclustern zu finden. Es gibt zwei Herangehensweisen, die dafür verwendet und in (Bonaccorso, Machine Learning Algorithms Popular algorithms for data science and machine learning, 2018) beschrieben werden:

- **Agglomerativer Ansatz**: Wird auch bottom-up genannt. Hier beginnt der Algorithmus mit der Annahme, dass jedes Element im Datensatz ein eigener Cluster ist. Diese Cluster werden so lange zusammengelegt, bis ein Ende-Kriterium erreicht ist.
- **Divisiver Ansatz**: Bei diesem top-down Vorgehen, startet der Algorithmus mit einem einzigen Cluster, dem alle Elemente zugeordnet sind. Diese werden so lange aufgeteilt, bis jedes Element separiert ist. Danach werden die Elemente wieder aggregiert. Dies geschieht aufgrund eines Kriteriums, das auf der Unterschiedlichkeit der Elemente basiert. Es werden also nur ähnliche Elemente wieder zusammengefügt.

Bei dem agglomerativen Ansatz, gibt es unterschiedliche Metriken (engl. Affinity genannt), die zur Distanzberechnung zweier Datenpunkte eingesetzt werden können. Hierzu gehören:

- Euklidische Distanz
- Manhattan Distanz
- Kosinus Distanz

Nach der Festlegung auf eine Metrik, wird eine Strategie (engl. Linkage) zur Zusammenlegung von Clustern gewählt. Auch hier gibt es verschiedene Alternativen:

- Complete linkage
- Average linkage
- Ward's linkage

Die erste Strategie berechnet die maximale Distanz der am weitest entfernten Elemente innerhalb der Cluster und versucht diese Distanz zu minimieren. Average linkage funktioniert ähnlich, jedoch wir hier die durchschnittliche Distanz der Cluster berechnet, bevor fusioniert wird. Bei der ward's linkage Strategie wird die Summe der quadrierten Distanzen innerhalb jedes Clusters berechnet. Diese wird Summe wird dann durch das Zusammenlegen der Cluster minimiert. So erfolgt eine Reduktion der Varianz in den Clustern. Diese Methode kann nur in Verbindung mit der Euklidischen Distanz verwendet werden (Bonaccorso, Mastering Machine Learning Algorithms, 2018).

Im Rahmen dieser Arbeit wird der agglomerative Ansatz mit der ward's linkage Strategie und damit die euklidische Distanz zur Berechnung der Cluster verwendet.

$$d_{Euclidean}(\bar{x}_1, \bar{x}_2) = \|\bar{x}_1 - \bar{x}_2\|_2 = \sqrt{\sum_j (\bar{x}_1^{(j)} - \bar{x}_2^{(j)})^2}$$

Formel 1: Euklidische Distanz entnommen aus (Bonaccorso, Mastering Machine Learning Algorithms, 2018)

Formel 1 oben zeigt die Berechnung der euklidischen Distanz. Als Beispiel soll die Distanz zweier Vektoren berechnet werden:

$$x_1 = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, x_2 = \begin{pmatrix} 0 \\ 6 \end{pmatrix}$$

Nun wird die Differenz der Vektoren berechnet:

$$x_1 - x_2 = \begin{pmatrix} 3 \\ 4 \end{pmatrix} - \begin{pmatrix} 0 \\ 6 \end{pmatrix} = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$$

Die Differenz wird in die Formel eingesetzt:

$$|x_1 - x_2| = \sqrt{3^2 + \left(-2^2\right)} = \sqrt{9+4} = \sqrt{13} = 3{,}61$$

Die euklidische Distanz der zwei Vektoren beträgt also 3,61.

5.2 K-means

Der K-means, oder auch Lloyd's Algorithmus, gruppiert Daten indem er versucht die Stichproben in Gruppen mit gleicher Varianz zu trennen. Der Name entsteht durch die Anzahl (k) der definierten Zentroiden (engl. centroids oder means). Das Ziel ist es, optimale Zentroiden zu finden, durch die Maximierung zweier Kriterien:

- Ähnlichkeit innerhalb des Clusters
- Unterschiede zwischen den Clustern

Zu Beginn werden die Zentroiden geschätzt. Danach werden iterativ die Distanzen zwischen den Datensätzen und den Mittelpunkten berechnet und jeder Datensatz dem Cluster zugeteilt, zu dem die Distanz am geringsten ist. Dieser Vorgang wird als Minimierung der Trägheit (engl. Inertia) der Cluster bezeichnet. Wie die Berechnung dazu aussieht zeigt Formel 2 unten.

$$S = \sum_{j=1}^{k} \sum_{\bar{x}_i \in C_j} \left\| \bar{x}_i - \bar{\mu}_j \right\|^2$$

Formel 2: Trägheit, entnommen aus (Bonaccorso, Mastering Machine Learning Algorithms, 2018)

Wenn S einen großen Wert hat, ist die Ähnlichkeit der Daten innerhalb der Cluster gering und es wurden Daten mit großen euklidischen Abständen den Clustern zugeordnet. Sobald die Berechnung für alle Daten abgeschlossen ist, werden neue Zentroide für die jeweiligen Cluster berechnet. Im Anschluss werden wieder die Distanzen kalkuliert und die Datenpunkte werden den neuen Zentroiden zugeteilt. Dies wird so lange wiederholt, bis das Trägheitsmaß ein Minimum erreicht (unter einen Schwellenwert sinkt) und die Cluster stabil sind (Bonaccorso, Mastering Machine Learning Algorithms, 2018).

Die Berechnung der Trägheit hat allerdings auch Nachteile. Es wird davon ausgegangen, dass die Cluster konvex und isotrop sind, was nicht immer der Fall ist. Sind die Cluster in die Länge gezogen oder haben eine andere ungleichmäßige Form, ist das Ergebnis schlecht. Außerdem handelt es sich bei der Trägheit nicht um eine normalisierte Metrik. Es ist bekannt, dass kleinere Werte besser sind und Null das Optimum darstellt. Aber in höher dimensionierten Räumen tendiert die euklidische Distanz dazu sich aufzublähen. Hier kann eine Reduktion der Dimensionen, wie sie bei der Principal Component Analysis (PCA) gemacht wird, helfen diesen Effekt abzuschwächen und so das Ergebnis verbessern (scikit-learn Machine Learning in Python, n.d.).

Zusammengefasst, erfolgt die Einteilung also in drei Schritten.

- Initiale Zentroiden bestimmen
- Stichprobe dem nächsten Zentroid zuordnen
- Zentroiden neu berechnen

Die Schritte zwei und drei wechseln sich dabei ab, bis die Cluster stabil sind.

6 SPRACHEN, BIBLIOTHEKEN UND FRAMEWORKS

In diesem Kapitel werden verschiedene Angebote an Programmiersprachen, Frameworks oder Bibliotheken beschrieben, die Machine Learning Algorithmen anbieten. Eine Übersicht über verschiedene beliebte Sprachen und Frameworks bietet (Przystalski, 2019). Unter den Sprachen finden folgende Vertreter am meisten Verwendung:

- Python
- R
- Java
- Scala
- JavaScript

Nicht nur bei den Sprachen, sondern auch bei den Bibliotheken gibt es mehrere beliebte Alternativen:

- Tensorflow
- Keras
- Scikit-learn
- PyTorch
- H2O
- Caffe
- MLlib
- Mlr
- Deeplearning4J
- MXNet

Jede angeführte Software oder Bibliothek ist entweder Open Source oder unter einer freien Lizenz verfügbar als auch Plattform unabhängig. Für diese Arbeit werden jene Angebote untersucht, die aufgrund von Vorkenntnissen aus Beruf oder Studium oder den Interessen des Autors in Frage kommen.

6.1 JavaML

JavaML ist eine Machine Learning Bibliothek für die Sprache Java. Sie bietet verschiedene Techniken für Programmierer, wie etwa Classification oder auch Clustering. Die Bibliothek besteht aus Interfaces, die in der API (Application Programming Interface) Dokumentation ausführlich beschrieben werden. Unter den Algorithmen, die für Clustering zur Verfügung stehen, findet sich K-means, sowie eine Implementation des Weka Clustering Algorithmus. Hierarchische Modelle gibt es zum Zeitpunkt dieser Arbeit noch keine. In JavaML wird derzeit weder eine grafische Benutzeroberfläche noch grafische Darstellungen zur Visualisierung angeboten (Abeel, Van de Peer, & Saeys, 2009).

6.2 Apache Spark

Apache Spark ist eine Plattform für Datenverarbeitung. Mit MLlib, bietet Apache Spark eine Machine Learning Bibliothek, die in verschiedenen Sprachen genutzt werden kann, etwa in, Java, Scala, Python oder R. Im Bereich Clustering bietet die Bibliothek folgende Auswahl:

- K-means
- Gaussian mixture
- Power iteration clustering (PIC)
- Latent Dirichlet allocation (LDA)
- Bisecting k-means
- Streaming k-means

Hierarchische Clustering Algorithmen werden zu diesem Zeitpunkt keine unterstützt (The Apache Software Foundation, n.d.)

6.3 Weka

Weka ist eine weitere Software, die für statistische Analysen und Machine Learning genutzt werden kann. Das Paket kann samt grafischer Benutzeroberfläche unter https://waikato.github.io/weka-wiki/ heruntergeladen werden. Weka bietet zum Clustern eine Implementation des K-means Algorithmus, genannt SimpleKmeans. Auch bottom-up oder top-down Algorithmen für hierarchisches Clustering werden angeboten (Chaudari & Parikh, 2012).

6.4 Scikit-learn

Scikit-learn ist eine Machine Learning Bibliothek in der Programmier- und Skriptsprache Python. Es werden unterschiedliche Algorithmen mit dem Basispaket mitgeliefert. Was den Teilbereich des Clustering angeht, stehen folgende Implementationen zur Auswahl:

- K-means
- Affinity Propagation
- Mean Shift
- Spectral clustering
- Hierarchical clustering
- DBSCAN
- OPTICS
- Birch

Dank der Bibliothek matplotlib, die in Scikit-learn integriert ist, kann man Ergebnisse auch grafisch darstellen (scikit-learn Machine Learning in Python, n.d.).

6.5 R

R ist eine weitere gratis Software, die für statistische Berechnungen und grafische Ausgaben verwendet werden kann. Machine Learning ist in der Basisversion nicht inkludiert, kann aber durch Installation sogenannter Packages verwendet werden. Für R ist ebenfalls eine graphische Oberfläche vorhanden, das R-Studio, die man separat herunterladen und starten kann. Es gibt Packages für verschiedenste Machine Learning Algorithmen, auch K-means oder hierarchische Ansätze sind vorhanden. R, sowie alle verfügbaren Packages können über das Comprehensive R Archive Network (CRAN), einem Netzwerk aus Web- und FTP Servern, heruntergeladen werden.

6.6 Tensorflow

Tensorflow wurde ursprünglich von Google entwickelt und ist inzwischen unter der Apache-2.0 Open Source Lizenz verfügbar. Das Framework ist ausgelegt auf Deep Learning und die Verwendung von Neuronalen Netzen. Die API Dokumentation zeigt eine Implementierung des K-means Algorithmus, allerdings in einem experimentellen Zustand. Dies bedeutet, dass der Algorithmus zwar für die Sprache Python zur Verfügung steht, allerdings noch nicht auf seine Richtigkeit überprüft wurde. Hierarchische Algorithmen stehen momentan nicht zur Verfügung. Das Framework kann unter https://www.tensorflow.org/ heruntergeladen und installiert werden.

6.7 Auswahl

Für die Implementation der Clustering Algorithmen in dieser Arbeit wird Scikit-learn verwendet. Die Auswahl beruht darauf, dass Python eine Sprache ist, die für mich noch unbekannt ist und dies eine Gelegenheit ist Neues zu lernen. Außerdem bietet es eine willkommene Abwechslung zum Alltag als Java Entwickler. Weiters bietet Scikit-learn passende Algorithmen für die Umsetzung, wie sie in der Arbeit vorgesehen ist. Die Bibliothek ist bereits weit verbreitet und bietet eine große Community im Hintergrund. Zudem wurde die Verwendung dieser Bibliothek von Personen aus dem Arbeitsumfeld und Bekanntenkreis mehrfach empfohlen. R konnte ich während dem Studium kennen lernen. Der Umgang mit dem R Studio und der R spezifischen Syntax, war jedoch gewöhnungsbedürftig für mich, was mich dazu bewogen hat R für diese Arbeit auszuschließen. Apache Spark und Tensorflow scheiden aufgrund der fehlenden hierarchischen Clusteralgorithmen aus.

7 EXPERTENINTERVIEW

Um eine Segmentierung von Kundendaten vornehmen zu können, muss man wissen welche Daten man für eine Unterteilung in Segmente heranziehen soll. Um zu erfahren, was für das Unternehmen wichtig ist, wurde ein Experteninterview mit einem Vertreter des Unternehmens durchgeführt. Dabei wurden dem Experten Fragen zu dem Thema Kundendaten und Segmentierung gestellt, um zu erfahren welche Kunden für das Unternehmen interessant sind und wofür Segmente in weiterer Folge verwendet werden sollen. Als Ergebnis werden aus dem Interview Kriterien abgeleitet, die dann als sogenannte Features in das Machine Learning Modell mitaufgenommen werden. Es folgt eine kurze Zusammenfassung der wichtigsten Fragen und Antworten des Experten aus dem Transkript des Interviews:

Welche Pläne haben Sie für die Verwendung von verschiedenen Kundensegmenten?

Wir haben Online Kundendaten und wollen diese Daten auch zielgerichtet verwenden, um den Kunden in weiterer Folge auch besser ansprechen zu können. Cluster, die abgeleitet werden aus den Daten sind für uns in erster Linie ganz besonders relevant für Produktempfehlungsmodelle bzw. wo man dann entsprechend den Kunden auf Basis des Clusters, auf Basis der Segmentierung dann auch die richtigen Inhalte anzeigen kann.

Welche Kundengruppen sind für Sie interessant, um ein Targeting auszuführen?

Die ersten interessanten Informationen für uns wären ein Cluster z.b. wie bereits genannt, männliche Kunden, die über 40 Jahre alt sind, aber auch weibliche Kunden, die zwischen 20 und 35 Jahre alt sind. Das heißt das wäre auf jeden Fall entsprechend schon zwei relevante Cluster wo wir da das Geschlecht miteinbeziehen und auch die Altersgruppe miteinbeziehen. Was natürlich ganz essenziell ist, ist das Thema Kundenkarte ja/nein. Natürlich auch die Käufe, das heißt kommen die Kunden wieder? Hat der Kunde zumindest einmal gekauft? Hat der Kunde öfter gekauft? Gibt es da irgendwie Unterscheidungen? Wie kristallisieren sich diese Unterscheidungen heraus in den Daten? Wir bieten in unseren Onlineshops auch die Möglichkeit Produkte zu reservieren und auch das wäre für uns natürlich auch spannend welche Daten sich daraus ableiten und welche Käufe sich daraus ableiten lassen. Ob ein Kunde jetzt reserviert hat ja oder nein.

Wie würden Sie bewerten ob das Ergebnis im Ganzen sinnvoll ist, oder ein Cluster für sich genommen?

Im Detail ist da noch gar nichts geplant. Jetzt warten wir einmal grundsätzlich auf die Ergebnisse, sind da schon sehr gespannt drauf. Dann in weiterer Folge werden wir wahrscheinlich einzelne Cluster auch im Detail anschauen, bedeutet wir schauen was steckt da tatsächlich dahinter. Wir könnten dann auch in Richtung Detailanalyse gehen, bzw.: A/B testing gehen. Das heißt im Detail ist es noch nicht klar aber wir haben da schon unterschiedliche Werkzeuge auch im Unternehmen, um dann auch eben dieses Ergebnis noch weiter analysieren zu können.

7.1 Ergebnis des Interviews

Aus den Antworten des Experten lassen sich folgende Kriterien für Cluster ableiten:

- Männlich über 40 Jahre
- Weiblich zwischen 20-35 Jahren
- Kundenkarte ja/nein
- Käufe ja/nein
- Käufe mehrmals?
- Reservierungen ja/nein
- Umsatz

Diese Auflistung enthält statischen Regeln, die das Unternehmen durch manuelles Auswerten der Kundendaten bereits erkannt hat und derzeit für Marketingzwecke verwendet.

Die Unterscheidung zwischen Männern und Frauen und deren Einteilung in Altersgruppen liefert zwei Attribute, die aus den Kundendaten extrahiert werden müssen, nämlich das Geschlecht, sowie das Alter. Weiters wird ein Feature benötigt, das Auskunft über das Vorhandensein einer Kundenkarte gibt. Bei den Bestellungen und Reservierungen wird jeweils die Anzahl extrahiert. Dies erlaubt nicht nur eine Unterscheidung zwischen ja/nein, es gibt auch Auskunft darüber ob mehrmals bestellt oder reserviert wurde. Beim Umsatz müssen die Bestellwerte der Bestellungen summiert und in einem eigenen Feld extrahiert werden.

In die Analyse werden also folgende Merkmale aufgenommen:

- Geschlecht
- Alter
- Kundenkarte vorhanden
- Anzahl der Käufe
- Anzahl der Reservierungen
- Umsatz

In Kapitel 8 werden diese Merkmale der Kunden aus der Kundendatenbank extrahiert und für die weitere Verarbeitung vorbereitet.

7.2 Ableiten von Use Cases

Aus der Befragung des Experten lassen sich nun auch Beispiele für eine Anwendung dieser Cluster für personalisierten Content ableiten. Dazu werden die oben festgestellten Merkmale zur Einteilung in Cluster herangezogen. Auf deren Basis können unterschiedliche statische Regeln erstellt, und ausgewählte Kunden aus dem Datensatz anschließend in verschiedene Cluster eingeteilt werden. Jeder dieser statischen Cluster bekommt dabei ein eigenes Produkt, symbolisch für den angepassten Inhalt, zugewiesen. Dies soll die unterschiedlichen Kundenerfahrungen, je nach Cluster, verdeutlichen.

Als Vergleich werden dieselben Produkte auch den Clustern zugeteilt, die nach dem Durchlauf des Machine Learning Programms entstehen. Dabei werden die automatischen Cluster aufgrund von statistischen Merkmalen analysiert. Nach dieser Analyse werden die statischen Regeln den Clustern zugeteilt, deren Inhalt am ehesten der jeweiligen Regeln entspricht. Im Anschluss wird festgestellt in welchen Clustern die ausgewählten Kunden gelandet sind, und ob sich ihr Produkt geändert hat.

Als erstes werden statische Regeln aufgestellt, und jeder Regel ein Produkt zugewiesen. Die Auswahl der Regeln orientiert sich an den Antworten des Experten. Durch unterschiedliche Kombination und Verknüpfung der Kriterien lässt sich eine Vielzahl an statischen Regeln erstellen. Werden die Regeln zu allgemein formuliert, wird ein Großteil der Kunden im selben Segment landen. Sind die Regeln zu speziell gewählt, kann es andererseits passieren, dass nur wenige Kunden in einem Segment landen. Als Zielgruppen wurden Männer über 40 Jahren, sowie Frauen zwischen 20 und 35 Jahren genannt. Anhand dieser Zielgruppen und der restlichen Merkmale werden insgesamt zehn Regeln erstellt. Diese Regeln und die zugehörigen Produkte können Tabelle 1 entnommen werden.

Cluster	Regel	Produkt
1	Männer über 40 mit Bestellungen	Grill
2	Männer über 40	Werkzeugset
3	Frauen zwischen 20-35 mit Bestellungen	Esstisch
4	Frauen zwischen 20-35	Wohnlandschaft
5	Kunden mit Ausgaben > €340	Wohnwand
6	Kunden mit Bestellungen	Barhocker
7	Kunden mit Bonuskarte	Bücherregal
8	Kunden mit Reservierungen	Boxspringbett
9	Männer	Fernsehsessel
10	Frauen	Sonnenliege

Tabelle 1: Statische Regeln und Produkte

Es wurden Regeln aufgestellt, wie sie im produktiven Betrieb zum Einsatz kommen könnten. Die Regel mit den Ausgaben über €340 soll Kunden, die mehr Geld als andere ausgegeben haben, von den anderen trennen. Diesen Kunden könnten teurere Produkte angezeigt werden. Die Schwelle von €340 ist nicht zufällig gewählt. Sie basiert auf der Tatsache, dass nur 25% der Kunden über diesem Wert liegen wie man in Abbildung 8-2 später noch sehen wird. Das obere Quantil der Ausgaben liefert diesen Wert.

Um eine statische Einteilung und später einen Vergleich vornehmen zu können, werden nun einzelne Kunden aus dem gesamten Datensatz herausgenommen. Die Kunden werden so gewählt, dass jede der oben definierten statischen Regeln mindestens einen Kunden enthält. Um den Umgang mit diesen Datensätzen zu erleichtern, werden für diese Kunden Personas mit Namen angelegt. Tabelle 2 zeigt die ausgewählten Kunden.

Name	ID	Alter	Bonuskarte	Bestellungen	Ausgaben in €	Reservierungen
Birgit	230227	24	Ja	1	103,85	0
Sarah	13203	45	Nein	1	83,91	0
Nicole	16734	52	Nein	0	0	0
Sandra	20498	22	Nein	2	388,00	0
Beate	219202	46	Ja	0	0	3
Gabi	189563	-	Nein	1	83,96	0
Martina	29105	25	Nein	0	0	0
Claudia	3955	26	Ja	0	0	0
Nina	63460	-	Nein	0	0	1
Peter	12644	53	Nein	1	349,00	0
Franz	128368	36	Ja	1	149,00	0
Jürgen	5807	51	Nein	1	128,90	0
Werner	197222	33	Ja	2	334,80	1
Karl	123167	-	Nein	1	544,95	0
Fabian	7135	38	Nein	0	0	0
Thomas	8755	-	Ja	2	210,20	0
Christian	78077	49	Nein	0	0	0
Simon	203957	41	Ja	0	0	0

Tabelle 2: Personas

Beim Aufruf der Startseite des Onlineshops werden diese Regeln nun angewandt. Sie werden der Reihe nach überprüft. Somit werden Regeln die weiter oben stehen priorisiert. Am Beispiel Birgit kann diese Priorisierung nachvollzogen werden. Birgit ist eine Frau zwischen 20-35 mit einer Bestellung. Sie hat außerdem eine Bonuskarte. Es würden als Regel 3 und Regel 7 aus Tabelle 1 zutreffen. Da Regel 3 weiter oben steht und somit eine höhere Priorität hat, wird Birgit dem Cluster 3 zugeordnet. Die beiden letzten Regeln beziehen sich nur auf das Geschlecht. Hier wird zwischen Männern und Frauen unterschieden, auf die keine der Regeln mit höherer Priorität zutrifft. So kann zumindest auf Basis des Geschlechtes noch unterschiedlicher Inhalt ausgespielt werden. Dies wird nun auf alle ausgewählten Kunden angewandt. Die folgende Tabelle 3 verschafft einen Überblick.

Name	Cluster	Produkt
Birgit	3	Esstisch
Sarah	6	Barhocker
Nicole	10	Sonnenliege
Sandra	3	Esstisch
Beate	7	Bücherregal
Gabi	6	Barhocker
Martina	4	Wohnlandschaft
Claudia	4	Wohnlandschaft
Nina	8	Boxspringbett
Peter	1	Grill
Franz	6	Barhocker
Jürgen	1	Grill
Werner	5	Wohnwand
Karl	5	Wohnwand
Fabian	9	Fernsehsessel
Thomas	6	Barhocker
Christian	2	Werkzeugset
Simon	2	Werkzeugset

Tabelle 3: Personas mit Produkten

Um einen Vergleich zwischen statischen Regeln und automatisiert erstellten Clustern anstellen zu können, müssen die notwendigen Daten erst aus der Kundendatenbank extrahiert werden. Auch ein geeigneter Algorithmus, der eine Interpretation der erstellten Cluster und deren Inhalt erlaubt muss noch ausgewählt werden. Die folgenden Kapitel beschäftigen sich mit dieser Thematik.

8 VORBEREITEN DER KUNDENDATEN

Die Daten, die als Ausgangsbasis für die Segmentierung dienen sollen, befinden sich in einer Datenbank beim Kunden und müssen von dort extrahiert werden. Die E-Commerce Shops des Kunden bauen auf einer SAP Commerce Cloud Lösung auf. Dies bedeutet, dass andere Unternehmen oder E-Commerce Shops, die ebenfalls auf dieser Lösung aufbauen, idente Datenstrukturen und auch Prozesse zur Verwaltung und Speicherung der Daten im Hintergrund verwenden. Die Datenstruktur bildet also ein Basis-Set, das potenziell bei zahlreichen anderen Kunden verschiedener Onlineshops ebenfalls extrahiert und zur Analyse verwendet werden kann.

Der Kunde betreibt E-Commerce Shops in mehreren Ländern, unter anderem Österreich und Deutschland. Die Daten aller Shops werden in derselben Datenbankinstanz verwaltet. Um Unterschiede zwischen den Ländern, Währungen oder Zeitzonen zu beseitigen und mögliche Einflüsse dieser Kriterien auf die Segmentierung auszuschließen, werden nur die Kunden jeweils eines Landes extrahiert und zur Analyse herangezogen. Da die Anzahl der Kunden und deren Umsatz des Shops in Deutschland am größten sind, wird mit diesen fortgefahren.

Im Folgenden wird der ETL-Prozess (Extraktion, Transformation, Laden) beschrieben sowie eine erste Übersicht zu den Daten präsentiert.

8.1 ETL – Prozess

Es gibt viele Wege Daten aus Systemen zu extrahieren. Im vorliegenden Fall werden zwei Methoden verwendet. Im ersten Schritt wird ein Datenbankdump der Kundendatenbank erstellt. Dieser wird in eine lokale Datenbank eingespielt und in einem weiteren Schritt werden die Daten mit einem Groovy-Skript aus dem lokalen System entnommen. Zwar wäre eine Extraktion grundsätzlich auch im Livebetrieb möglich, allerdings bietet die lokale Datenbank auch Vorteile. Zum einen wird damit die Datenbank zu einem bestimmten Zeitpunkt „eingefroren“, um sicherzustellen, dass keine neuen Kunden, Adressen oder Bestellungen hinzukommen. Dadurch ist garantiert, dass beim Extrahieren immer auf denselben Kundenstamm zugegriffen werden kann. Zum anderen wird der Livebetrieb des Kundensystems nicht beeinträchtigt. Ein direkter Zugriff auf alle Kunden des produktiv Systems, könnte dieses merkbar verlangsamen und im schlimmsten Fall sogar eine Verletzung der Service Level Agreements (SLAs) nach sich ziehen. Am Ende wird der Datensatz in Python geladen und zur Verarbeitung vorbereitet.

8.2 Datenbankdump

Die Daten befinden sich in einer MySQL Datenbank, welche auf einem Datenbankserver läuft. Verbindet man sich zu dem Server kann man durch einen Befehl, wie ihn Listing 8-1 zeigt, einen Datenbankdump erstellen und auf dem Server ablegen:

```
shell> mysqldump [arguments] | gzip -fc > dump.sql.gz
```

Listing 8-1 MySQL dump Befehl

Der komprimierte Datenbankdump wird anschließend mittels Secure Copy Protocol (SCP) auf den lokalen Rechner transferiert und in die lokale Datenbankinstanz eingespielt.

8.3 Groovy Cronjob

SAP Commerce Cloud bietet die Möglichkeit Skripte der Programmier- und Skriptsprache Groovy direkt in einer Konsole auszuführen. Dies eignet sich für kurze Befehle, die in wenigen Sekunden Ergebnisse liefern können, da es sonst zu einem Timeout im Browser kommt. Das Extrahieren von Daten geschieht darum über einen Cronjob, der direkt am Server ausgeführt werden kann. Der Cronjob wird nach dem Starten ein eigens erstelltes Groovy Skript ausführen, um die Extraktion der Kundendaten durchzuführen.

Das Skript wird die Daten dabei nicht nur extrahieren, sondern zugleich auch transformieren. Hierzu gehören folgende Transformationen:

- Syntaktische Transformationen
  - Umwandlung von Geburtsdatum in Alter
- Semantische Transformationen
  - Aggregation der Ausgaben von Bestellungen
  - Filtern von Bestellungen und Reservierungen sowie Entwickler und Test Accounts
  - Bereinigen von Kundeneingaben (Entfernen von Leerzeichen am Anfang und Ende)

Weiters ist das Skript so erstellt, dass die Daten in einem Comma-separated values (CSV) Format abgespeichert werden. Die erste Zeile besteht aus den Spaltenüberschriften und jede weitere Zeile in diesem Dokument repräsentiert einen Kundendatensatz.

8.4 Übersicht über die Daten

Die Datei besteht aus der Spaltenüberschrift und Daten von insgesamt 250.000 Kunden. Bei diesen Kunden handelt es sich um jene, die im Onlineshop ein Kundenkonto erstellt haben.

Anhand dieser Kundenkonten können die Bestellungen und Adressdaten abgerufen werden, die alle wichtigen Informationen für die spätere Segmentierung enthalten. Der Shop bietet auch die Möglichkeit anonyme Bestellungen zu tätigen. Dabei werden allerdings keine Kundenkontos angelegt, sondern bestellbezogene Daten lediglich mit einer Bestellung verknüpft. Eine Extraktion wäre auch hier grundsätzlich möglich. Im Rahmen dieser Arbeit wird darauf jedoch verzichtet.

Id	Gender	City	PostalCode	Age	BonusCardOwner	OrderCount	TotalOrderSum	ReservationCount
1	MALE	albstadt	72459	67	FALSE	1	1598	0
2	FEMALE	dülmen	48249	27	FALSE	1	4792	0
3	MALE	velbert	42553	NA	FALSE	0	0	0
4	MALE	walsrode	29664	43	FALSE	1	36335	0
5	FEMALE	tutzing	82327	46	FALSE	1	40895	0
6	FEMALE	offenburg	77652	46	FALSE	1	67439	0
7	MALE	hennef	53773	40	FALSE	0	0	0
8	MALE	lüdinghausen	59348	NA	FALSE	0	0	0
9	FEMALE	grenzach-wyhlen	79639	NA	FALSE	1	31875	0
10	FEMALE	bad soden	65812	26	TRUE	0	0	0
11	FEMALE	kalbach	36148	NA	FALSE	1	14295	0
12	MALE	herbertingen	88518	NA	FALSE	0	0	0
13	FEMALE	leimersheim	76774	NA	FALSE	2	25778	0
14	MALE	lingen	49809	NA	FALSE	1	298	0
15	FEMALE	kiel	24147	48	TRUE	0	0	0
16	FEMALE	speichersdorf	95469	41	FALSE	1	10789	0
17	FEMALE	sigmarszell	88138	49	FALSE	1	6996	0
18	MALE	fellbach	70736	55	TRUE	1	5385	0
19	MALE	mühldorf	84453	NA	FALSE	1	5493	0
20	MALE	heidenfeld	97520	NA	FALSE	1	23995	0
21	FEMALE	wangen	88239	NA	FALSE	0	0	0
22	FEMALE	würzburg	97076	27	TRUE	0	0	0
23	MALE	heilbronn	74078	NA	FALSE	1	8575	0
24	MALE	schnaittach	91220	NA	FALSE	1	398	0
25	MALE	ergersheim	91465	33	TRUE	0	0	0
26	FEMALE	dannenberg	29451	30	FALSE	2	16167	0
27	FEMALE	zell am main	97299	37	FALSE	1	42895	0
28	FEMALE	herxheim	76863	24	FALSE	1	11613	0
29	FEMALE	geldern	47608	NA	FALSE	0	0	0
30	MALE	kaiserslautern	67657	46	FALSE	0	0	0

Abbildung 8-1: Übersicht über die Kundendaten

Wie in Abbildung 8-1 zu sehen ist, besteht die Kopfzeile aus mehreren Spalten mit englischen Bezeichnungen, deren Inhalt im Folgenden kurz beschrieben wird.

Id

Die Id ist eine fortlaufende Nummer, die zum Zweck der Identifizierung eines Datensatzes generiert wurde. Der erste Kunde wurde mit der Id 1exportiert.

Gender

Die Spalte Gender enthält das Geschlecht des jeweiligen Kunden. Eine Besonderheit des Shops ist, dass es auch möglich ist als Firma zu bestellen. In solchen Fällen kann es sein, dass neben MALE und FEMALE auch noch das Geschlecht COMPANY aufscheint. Hat eine Person kein Geschlecht angegeben scheint dies als NA auf.

City

In dieser Spalte sind die Städte der Kunden, die diese als Lieferadresse angegeben haben gelistet. Bei den Städten handelt es sich um ein Feld das direkt vom Kunden ausgefüllt wird, daher sind Städte nicht immer gleich und können Umlaute oder Rechtschreibfehler enthalten. Manche Kunden schreiben auch den Stadtteil, aus dem sie stammen in dieses Feld. Beim

Extrahieren wurden hier Leerzeichen am Anfang und am Ende entfernt und sämtliche Städte in die Kleinschreibung transformiert.

PostalCode

Hierbei handelt es sich um die Postleitzahlen der Lieferadressen der Kunden. Deutsche Postleitzahlen bestehen aus einer fünfstelligen Zahl.

Age

Das Alter der Kunden wurde zum Zeitpunkt der Extraktion aus dem Geburtsdatum der Kunden errechnet. Im Shop des Kunden ist es möglich auch ohne Angabe eines Geburtsdatums einen Kauf zu tätigen. Hat ein Kunde kein Geburtsdatum angegeben, wird dies mit NA angeführt. Um für das Alter gültige und für die Segmentierung relevante Daten zu erhalten, wurden beim Extrahieren nur Kunden zwischen 15 und 100 Jahren berücksichtigt. Dies soll verhindern, dass fehlerhafte Datensätze oder bewusste Falschangaben der Kunden das Ergebnis beeinflussen.

BonuscardOwner

Der Shop bietet Kunden die Möglichkeit eine Art Freundschaftskarte im deutschen Shop, oder auch Bonuscard/Bonuskarte wie sie in anderen Ländern genannt wird, mit dem Kundenkonto zu verknüpfen. Hat ein Kunde eine solche Karte in seinem Account steht hier TRUE, andernfalls wird FALSE angegeben.

OrderCount

Die Spalte OrderCount gibt die Anzahl der Bestellungen an, die ein Kunde in dem jeweiligen Land getätigt hat. Damit eine Bestellung hier gezählt wird, muss sie bestimmte Kriterien erfüllen. Die Bestellung muss bezahlt und vom Onlineshop an das interne System des Händlers exportiert worden sein.

TotalOrderSum

Hier wird die Gesamtsumme der Ausgaben in Euro angegeben, die aus den Bestellungen der vorhergehenden Spalte errechnet werden. Die Gesamtsumme ist auf zwei Nachkommastellen gerundet.

ReservationCount

Ähnlich dem OrderCount wird hier die Anzahl der getätigten Reservierungen eines Kunden angegeben. Bei einer Reservierung kann der Kunde ein Produkt in einer Filiale für einen gewissen Zeitraum zur Abholung vormerken. Eine Reservierung ist kein Kauf und es erfolgt keine Bezahlung im Onlineshop. Das Einzige Kriterium, um hier gezählt zu werden ist, dass eine Reservierung an das händlerinterne System exportiert wurde.

8.5 Laden der Daten in Python

Listing 8-2 zeigt wie die Daten in Python geladen werden. Die CSV-Datei wird zuerst in ein Pandas-Dataframe eingelesen. Danach folgen noch notwendige Anpassungen, um das Dataframe und die Daten darin auf einen Cluster Algorithmus vorzubereiten.

```
# load data from csv
df = pd.read_csv('/media/backup/MasterThesis/customerData_{}_{}.csv'
   .format(mandant, sampleSize), delimiter=';')

# drop customers which do not have a gender assigned - they do not provide
any value for segmentation
df = df.dropna(subset=['Gender'])
# drop customers with negative orderSum - could have happend by a bug in
the shop
df.drop(df[ (df['TotalOrderSum'] < 0) ].index, inplace=True)
# drop company customers - its only 1 in 250k customers and this one has
no orders or anything else of value
df.drop(df[ (df['Gender'] == 'COMPANY') ].index, inplace=True)

# drop not relevant columns
df_metrics = df.drop(columns=['Id', 'City', 'PostalCode'])

# replace all NA cells with 0
df_metrics = df_metrics.fillna(0)

# make a copy of data frame
df_tr = df_metrics

# transform to dummies
df_tr = pd.get_dummies(df_tr, columns=['Gender',  'BonusCardOwner'])
```

Listing 8-2: Laden und vorbereiten der Daten in Python

Zunächst werden alle Kunden, die kein Geschlecht aufweisen aus den Daten entfernt. Kein Geschlecht bedeutet, dass diese Kunden keine Adressinformationen oder Bestellungen mit Ihrem Konto verknüpft haben, in denen das Geschlecht aufscheint. Sie sind daher für die Segmentierung uninteressant.

In einem weiteren Schritt wird ein Kunde, der eine negative TotalOrderSum aufweist aus den Daten entfernt. Ein negativer Wert hier lässt auf einen Fehler schließen, der im Shop zur Zeit der Bestellung vorhanden war. Ein Bestellwert kann unter normalen Umständen nie unter null liegen.

Ein weiterer Kunde mit dem Geschlecht COMPANY wird entfernt, da dieser wie auch die Kunden ohne Geschlecht zuvor, keine relevanten Informationen zur Segmentierung liefert.

Nicht relevante Spalten werden aus den Daten entfernt. Die Id ist eine fortlaufende Nummer und bietet für die Bildung von Segmenten keine relevanten Informationen. Cluster Algorithmen können nur mit numerischen Variablen arbeiten, weshalb die Städtenamen erst in eine solche Umgewandelt werden müssen. Die Postleitzahlen an sich, bestehen aus Nummern, jedoch gibt es tausende unterschiedliche Postleitzahlen. Eine Stadt allein kann mehrere unterschiedliche Postleitzahlen besitzen. Aus diesem Grund sind die Postleitzahlen allein nicht aussagekräftig. Es müsste erst ein Mapping von Postleitzahlen auf Städte oder Regionen gemacht werden, um hier Zusammenhänge erkennen zu können. Die Spalten City und PostalCode werden daher entfernt, da sie zum einen weitere Vorbehandlungen erfordern und zum anderen vom Kunden im Kontext dieser Arbeit nicht erwünscht sind. Eine Aufnahme dieser Merkmale soll erst in weiteren Analysen folgen, nachdem die Ergebnisse dieser Arbeit ausgewertet wurden. Dies ist auch der Grund warum diese Daten schon bei der Extraktion mitaufgenommen wurden.

Die beiden binären Variablen Gender und BonuscardOwner liegen als kategorische Variablen vor. Hier muss erst ein sogenanntes one-hot encoding gemacht werden, um diese Variablen nutzbar zu machen. Dabei werden die Kategorien in eigene Spalten aufgeteilt. So entstehen die Spalten Gender_Male und Gender_Female. Ist das Geschlecht des Kunden zutreffend auf die Spalte, wird eine 1, andernfalls wird eine 0 eingetragen. Dasselbe geschieht mit den Vorteilskarten. Der Grund warum hier das one-hot encoding verwendet wurde, anstatt die Werte einfach durch 0 und 1 zu ersetzen, ist die bessere Übersicht. Die Ergebnisse lassen sich so leichter aus dem Output des Programms lesen. Zudem kann diese Umwandlung für beide Spalten gleichzeitig in einem Schritt durchgeführt werden.

8.6 Beschreibung des Datensatzes

In Abbildung 8-2 unten findet sich eine Beschreibung des Datensatzes nach dem Laden und der Vorbehandlung der Daten in Python. Von ursprünglich 250.000 Kunden sind 249.989 Kunden in dem Datensatz verblieben, elf wurden entfernt. Im weiteren Verlauf werden die Quantilen, wie sie in der Abbildung zu sehen sind, für die Analyse verwendet. Diese sind im Rahmen der Arbeit ausreichend, um Aussagen über die Clusterinhalte treffen zu können. Das Unternehmen könnte, je nach Ergebnis allerdings entscheiden, dass andere Werte zur Interpretation herangezogen werden müssen. Ein Experte des Unternehmens könnte die Einteilung zum Beispiel aufgrund der 80% oder 90% Perzentile treffen, was das Resultat verändern könnte.

	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True
count	249989.000000	249989.000000	249989.000000	249989.000000	249989.000000	249989.000000	249989.000000	249989.000000
mean	26.440927	1.041718	278.363348	0.014473	0.504722	0.495278	0.734144	0.265856
std	23.496041	1.401581	715.040792	0.154686	0.499979	0.499979	0.441789	0.441789
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	30.000000	1.000000	103.940000	0.000000	1.000000	0.000000	1.000000	0.000000
75%	46.000000	1.000000	340.860000	0.000000	1.000000	1.000000	1.000000	1.000000
max	100.000000	257.000000	210416.760000	10.000000	1.000000	1.000000	1.000000	1.000000

Abbildung 8-2: Beschreibung der Daten

Das Alter reicht von 0 bis 100 Jahren. Das 75%-Perzentil zeigt, dass 75% der Kunden ein Alter von weniger als 46 Jahren aufweisen. Da jedoch viele Kunden ohne Altersangabe in dem Datensatz mit dem Alter 0 aufscheinen, sagt dies wenig aus. Ein Blick auf das Histogramm in Abbildung 8-3 zeigt, dass beinahe 100.000 Kunden kein Alter angegeben haben. Man erkennt auch, dass jene Kunden deren Alter angegeben wurde annähernd eine Normalverteilung aufweisen. Die höchste Anzahl dieser Kunden findet sich im Intervall zwischen 30 und 40 Jahren.

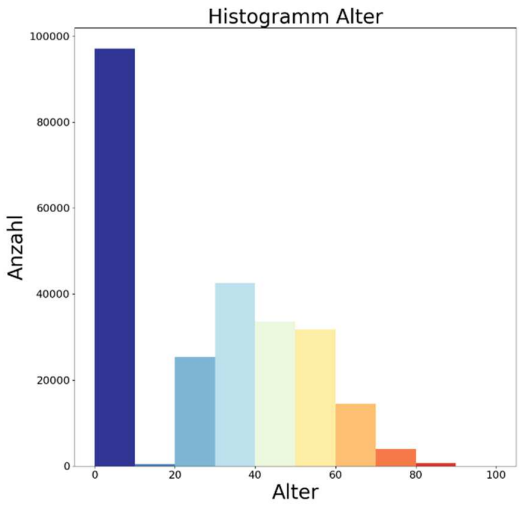

Abbildung 8-3: Histogramm Alter

Das 75%-Perzentil der Anzahl der Bestellungen zeigt, dass die meisten Kunden maximal 1-mal eine Bestellung getätigt haben. Die Standardabweichung liegt über dem Mittelwert, was sich durch den hohen Maximalwert erklären lässt, der bei 257 liegt. Die Histogramme in Abbildung 8-4 und Abbildung 8-5 zeigen die Verteilung der Bestellungen. Links erkennt man, dass mehr als 240.000 Kunden keine oder nur eine Bestellung aufgegeben haben. Wird herangezoomt, lässt sich in der rechten Abbildung feststellen, dass die Kunden kaum mehr als 50 Bestellungen abgeschlossen haben. Es lassen sich jedoch drei Ausreißer erkennen, deren Anzahl an Bestellungen weit über den Anderen liegt.

Abbildung 8-4: Histogramm Anzahl der Bestellungen

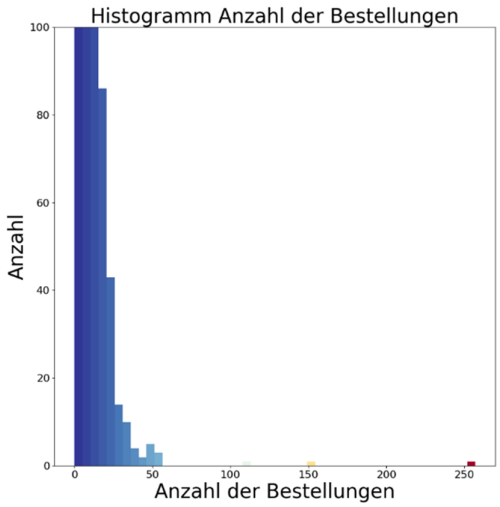

Abbildung 8-5: Histogramm Anzahl der Bestellungen 2

Bei den Reservierungen verhält es sich ähnlich wie bei den Bestellungen. Hier lässt sich anhand der Histogramme in Abbildung 8-6 und Abbildung 8-7 erkennen, dass die Anzahl der Reservierungen mit deren Häufigkeit stark abfällt. Die Mehrheit der Kunden haben noch nie reserviert. Eine Reservierung haben nur knapp mehr als 2000 Kunden und bei zwei Reservierungen sind es in etwa 340 Personen. Das Maximum an Reservierungen liegt bei 10.

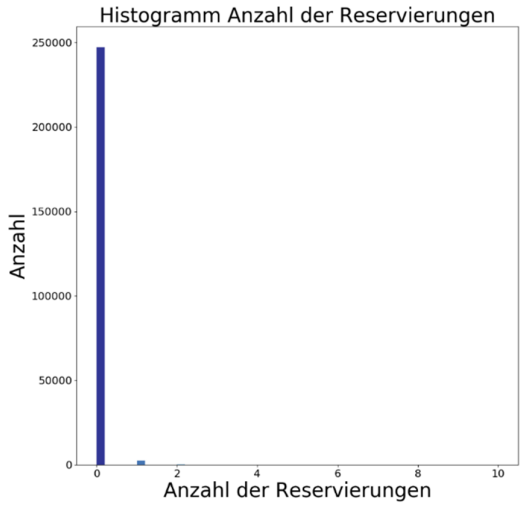

Abbildung 8-6: Histogramm Reservierungen

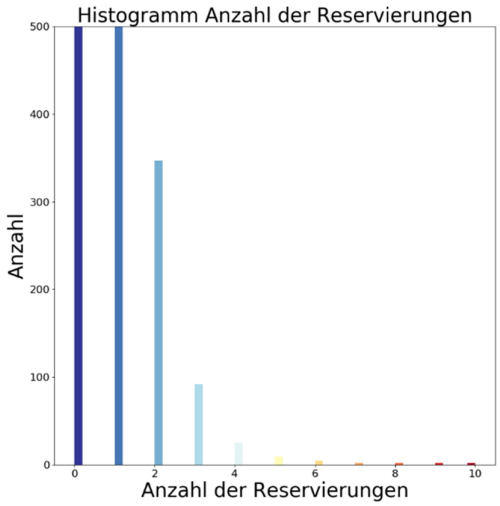

Abbildung 8-7: Histogramm Reservierungen 2

Bei den Ausgaben zeigt die Statistik, dass 75% der Kunden weniger als 340 Euro im Shop ausgegeben haben. Auch hier ist die Standardabweichung aufgrund des hohen Maximalwertes von 210.416 Euro über dem Mittelwert. Abbildung 8-8 und Abbildung 8-9 zeigen die Verteilung der Ausgaben. In der linken Abbildung kann man sehen, dass die Anzahl an Kunden, die nichts oder eher wenig Geld ausgegeben haben, nah an der totalen Anzahl der Kunden in dem

Datensatz liegt. Vergrößert man den Ausschnitt, lässt sich in der rechten Grafik ablesen, dass es kaum Kunden gibt die für mehr als 15.000 Euro bestellt haben. Ab rund 40.000 Euro gibt es nur noch drei Ausreißer, deren Bestellwert deutlich über jenem der anderen Kunden liegt.

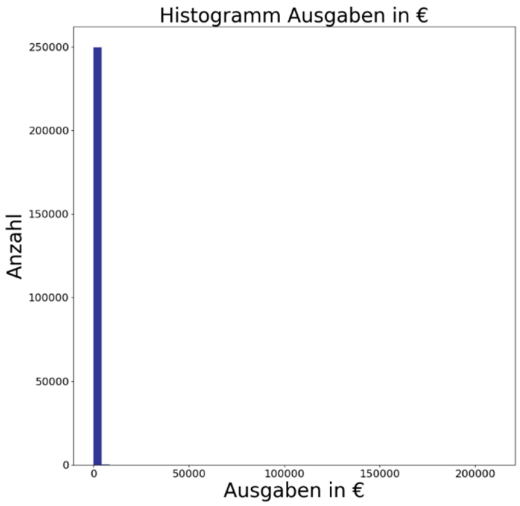

Abbildung 8-8: Histogramm Ausgaben in €

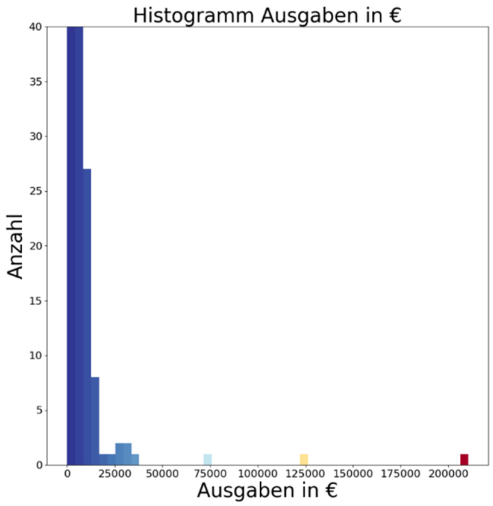

Abbildung 8-9: Histogramm Ausgaben in € 2

Bei den Binären Variablen, können die Mittelwerte direkt als Prozentangaben gelesen werden, da diese Variablen nur einen Wert von 1 oder 0 annehmen können. Addiert man die Werte der jeweiligen Spalten erhält man 1, was also 100% entspricht. Beim Geschlecht erkennt man, dass die Anzahl von Männern und Frauen beinahe gleich ist. Mit 50,5% Frauenanteil gibt es nur unwesentlich mehr Frauen als Männer, die mit 49,5% vertreten sind. Bei den Bonuskarten fällt der Unterschied deutlicher aus. 73,4% der Kunden besitzen keine Bonuskarte.

9 AUSWAHL EINES ALGORITHMUS

In diesem Kapitel werden die Ergebnisse eines ersten Durchlaufs der implementierten Algorithmen zusammengefasst. Die Anzahl an Clustern, die für die Analyse ausgewählt wurden, basiert auf der sogenannten Ellenbogen Methode (engl. elbow method). In (Kodinariya & Makwana, 2013) werden mehrere Methoden zur Feststellung der optimalen Clusteranzahl gelistet, wobei diese die älteste und bekannteste ist. Sie kann bei dem K-means Algorithmus berechnet werden und soll einen Anhaltspunkt für die optimale Anzahl an Clustern liefern. Die optimale Anzahl liegt an der Stelle, an der die Kurve beginnt abzuflachen. Abbildung 9-1 zeigt diese Kurve, die für den Datensatz und bis zu 20 Cluster berechnet wurde.

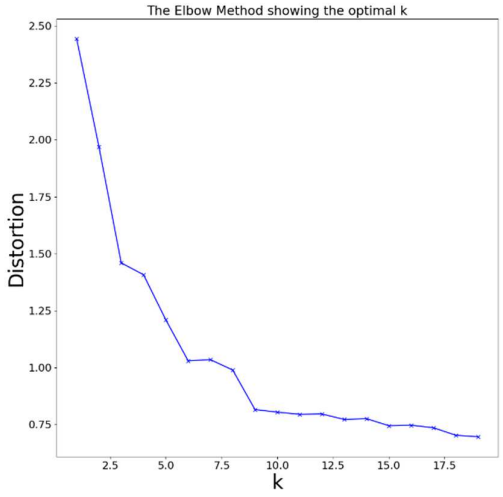

Abbildung 9-1: K-means Ellenbogen Methode

Auf der X-Achse wird die Anzahl der Cluster k dargestellt. Die Y-Achse zeigt die Summe der quadrierten Distanzen zu den Zentroiden der Cluster. Diese Summe fällt mit der Anzahl der Cluster, da sich diese Distanzen verringern. Bis zu dem Punkt wo jeder Datensatz seinen eigenen Cluster bildet und die quadrierten Distanzen auf 0 sinken. Die Kurve zeigt einen steilen Abfall, der bei neun Clustern beginnt abzuflachen. Ab diesem Punkt bringt die Erhöhung der Clusteranzahl nur mehr wenig Verbesserung, rein rechnerisch gesehen. Der Ausgangspunkt für Analysen liegt also bei neun Clustern. Um herauszufinden nach welchen Kriterien die Cluster entstehen, werden mehrere Durchläufe mit einer unterschiedlichen Anzahl an Clustern gemacht. Dieselbe Anzahl an Clustern wird auch für den hierarchischen Ansatz verwendet, um einen direkten Vergleich zu ermöglichen. Jedes Unterkapitel widmet sich einem Ansatz der Verfolgt wurde und enthält eine kurze Darstellung und Erläuterung der Implementation. Die Ergebnisse des jeweiligen Algorithmus werden der Reihe nach in ANHANG A -aufgeführt und anschließend in dem jeweiligen Kapitel gemeinsam analysiert. Bei der Analyse werden die

Kunden nach Clustern gruppiert und mithilfe von Mittelwert, Median, Quantilen, Minimum, Maximum, sowie Standardabweichung der Spalten interpretiert. Am Ende des Kapitels folgt ein Vergleich und es wird eine Auswahl getroffen, welcher der Ansätze für weitere Analysen verwendet wird.

Auch die Laufzeit der verschiedenen Ansätze wird angegeben, um hier einen Vergleich anstellen zu können. Um die Laufzeiten in Relation zu setzen wird die Hardware angegeben, die verwendet wurde, um die Berechnungen anzustellen:

- System: Dell Precision 5540
- Betriebssystem: Linux Mint 19.3 „Tricia“
- Intel Processor i7-9850H, 6 Cores, 12 MB Cache, 2,60 GHz bis zu 4,6 GHz Turbo, 45 W, vPro
- 32GB RAM (2x16GB RAM), DDR4-2.666 MHz SDRAM
- M.2 512 GB PCIe NVMe 40 Solid-State-Festplatte

9.1 Hierarchische Segmentierung

Scikit-learn bietet zur Segmentierung nur einen agglomerativen Ansatz. Hierbei werden alle Datenpunkte als eigenständige Cluster gesehen und in Iterationen so lange zusammengefügt, bis nur mehr ein einziger Cluster übrigbleibt, oder ein Abbruchkriterium erreicht ist. Ein Kriterium ist zum Beispiel die Anzahl an Endclustern, die dem Algorithmus übergeben werden können. Dieses Vorgehen wird auch Bottom-up Verfahren genannt.

```
cluster = AgglomerativeClustering(n_clusters=clusters,
affinity='euclidean', linkage='ward')

cluster.fit_predict(data_scaled)
```

Listing 9-1: Agglomeratives Clustering

Listing 9-1 zeigt die Anwendung des agglomerativen Algorithmus auf das Dataframe. Der Parameter *n_clusters* gibt an, wie viele Cluster am Ende entstehen sollen.

Bei der hierarchischen Segmentierung war der Datensatz zu groß, um ein Clustering mit allen 250.000 Kunden durchzuführen. Die Berechnung aller möglichen Kombinationen, die miteinander zusammengelegt werden könnten, erfordert erhebliche Ressourcen. Abbildung 9-2 zeigt, dass zur Berechnung insgesamt 233 GiB (Gibibyte, entspricht ~250 Gigabyte) benötigt würden. Darum wurde die Anzahl der Kunden auf 50.000 reduziert, um die Segmentierung erfolgreich durchführen zu können. Die Ergebnisse für die Durchläufe mit fünf, neun und 15 Clustern, sind in den Abbildung 12-1 bis Abbildung 12-11 dokumentiert.

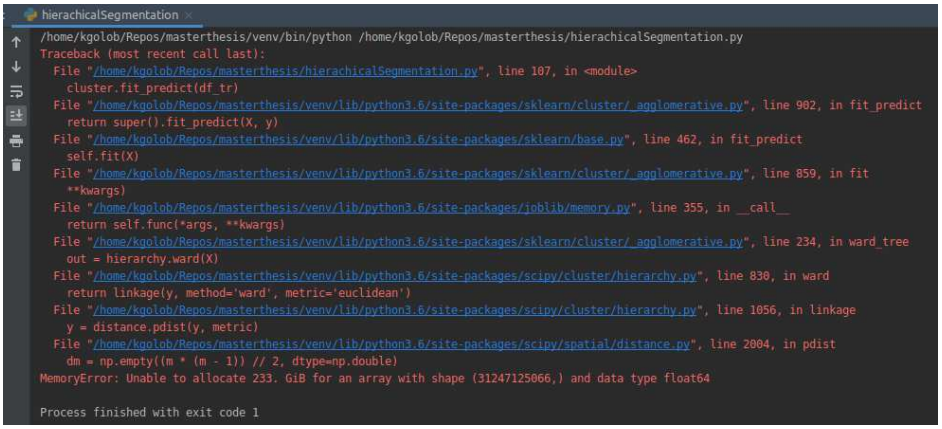

Abbildung 9-2: Fehler - zu wenige Ressourcen:

9.1.1 Laufzeit

Die Laufzeit über 50.000 Kunden für die Hierarchische Segmentierung sind in Tabelle 4 unten festgehalten. Obwohl die Kundenanzahl reduziert wurde, liegt die Laufzeit bei etwas über einer Minute. Das klingt zunächst nicht nach viel, ist aber im Vergleich zum K-means Algorithmus erheblich langsamer, wie später noch zu sehen sein wird.

Clusteranzahl	Laufzeit in Sekunden
5	85,21
9	85,33
15	85,56

Tabelle 4: Laufzeit Hierarchisches Clustering

9.1.2 Analyse

Für den Durchlauf mit 50.000 Kunden wird keine eigene Analyse erstellt. Der hierarchische Ansatz wird nach einer Anpassung noch einmal auf den gesamten Datensatz angewandt.

9.2 Hierarchische Segmentierung mit Connectivity Matrix

Da der hierarchische Algorithmus im ersten Versuch nur in der Lage war 50.000 Kunden zu verarbeiten, wurde eine Optimierung vorgenommen, um eine Verarbeitung aller Kunden zu ermöglichen und die Ergebnisse mit dem K-means Algorithmus zu vergleichen.

```
connectivity = kneighbors_graph(data_scaled, n_neighbors=300,
include_self=False)

cluster = AgglomerativeClustering(n_clusters=clusters,
connectivity=connectivity, linkage='ward')
cluster.fit_predict(data_scaled)
```

Listing 9-2: Hierarchisches Clustering mit Connectivitx Matrix

Listing 9-2 zeigt die Anpassung, die vorgenommen wurde. Es handelt sich um einen kneighbors Graph, der aus den Daten erstellt wird. Jeder Kunde wird mit seinen n_neighbors verknüpft. Der Algorithmus basiert auf der Annahme, dass Datensätze, die eine ähnliche Struktur aufweisen auch nahe beieinander liegen. Sind zwei Datensätze also in unmittelbarer Nähe zueinander, weisen sie vermutlich die gleichen Merkmale auf. In diesem Graphen müssen alle Kunden miteinander verknüpft sein, damit der hierarchische Algorithmus später seine Berechnungen anstellen kann. Um die Rechenpower und die Laufzeit gering zu halten, sollte der Graph so klein wie möglich sein. Nach einigen Versuchen hat sich gezeigt, dass für 250.000 Kunden mindestens die 300 nächsten Nachbarn gebraucht werden, damit kein Kunde in dem Graph isoliert ist. Die Anzahl kann und wird sich mit anderen Daten aber ändern. Im vorhandenen Datensatz finden sich einige Ausreißer, wie ein Blick auf die Anzahl der Bestellungen oder die Ausgaben zeigt.

Dieser Graph, wird anschließend als Connectivity-Parameter an den hierarchischen Algorithmus übergeben. Als Kandidaten für denselben Cluster werden nun nur noch mit dem Kunden verknüpfte Nachbarn herangezogen. Dies verringert die nötige Rechenpower zur Berechnung der Cluster. Der Parameter affinity=“euclidean“ fehlt hier, ist aber bei dem Parameter linkage=“ward“ eine Standardeinstellung und muss nicht extra angegeben werden.

9.2.1 Laufzeit

Auch bei der Nutzung aller Kundendaten, wurde wieder eine Laufzeitanalyse gemacht und in der Tabelle 5 angeführt. Die Gesamtzeit setzt sich aus zwei Messungen zusammen. Zum einen, der Aufbau des kneighbors Graphen, oder auch Connectivity Matrix genannt, zum anderen die Zeit, die der hierarchische Algorithmus benötigt um zu Clustern. Im Vergleich zum reduzierten Datensatz, hat sich die Laufzeit nun beinahe verachtfacht.

Clusteranzahl	Aufbau der Matrix in Sekunden	Laufzeit in Sekunden	Gesamt
5	31,25	639,64	670,89
9	28,46	641,33	669,79
15	30,16	637,38	667,54

Tabelle 5: Laufzeit Hierarchisches Clustering mit Connectivity Matrix

9.2.2 Analyse

Bei dem Durchlauf mit 5 Clustern lässt sich in Abbildung 12-12 ff. erkennen, dass es einen Cluster gibt, in dem sich Personen ohne Altersangabe befinden, nämlich Cluster 3. Sowohl der Mittelwert als auch das Minimum und Maximum liegen beim Alter bei 0. Mit 96% besitzen die meisten Kunden in dem Cluster keine Bonuskarte. Das untere Quantil zeigt, dass 75% der Kunden zumindest eine Bestellung haben. Unter den 83.256 Kunden, befinden sich 52% Männer.

Cluster 0 sind nur drei Personen zugewiesen. Es handelt sich um zwei Männer und eine Frau, die aufgrund der hohen Anzahl an Bestellungen und Ausgaben in dem Cluster gelandet sind. Um welche Kunden es sich handelt, zeigt Abbildung 9-3 unten. Nur die Frau aus diesem Cluster ist im Besitz einer Bonuskarte und hat kein Alter angegeben.

	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True	clusters
99286	40.0	152	210416.76	0	0	1	1	0	0
127005	0.0	23	73312.18	0	1	0	0	1	0
130276	44.0	48	122695.53	0	0	1	1	0	0

Abbildung 9-3: Cluster 0

In dem Cluster mit der Nummer 1 befinden sich offenbar Kunden, die mehrere Bestellungen abgeschlossen haben. Der Mittelwert liegt bei über 4, der Median zeigt immerhin noch 3 Bestellungen. Auch die Ausgaben liegen im Median bei €2.874,55, was im Vergleich zu anderen Clustern sehr hoch ist. Das Verhältnis von Frauen zu Männern liegt bei 40% zu 60%. Nur 35% der Kunden besitzen eine Bonuskarte.

Cluster 2 enthält augenscheinlich etwas ältere Kunden. Der Mittelwert liegt hier bei 43,48 Jahren. Der Minimalwert von 15 zeigt, dass sich nur Kunden mit Altersangaben in diesem Cluster befinden. Der Median liegt bei 42 Jahren. Die Quantilen zeigen, dass der Prozentsatz der Kunden, die bestellt haben, irgendwo zwischen 25% und 50% liegt. Der Mittelwert liegt aber bei 0,76 Bestellungen. Der Grund dafür liegt darin, dass der Maximalwert bei 20 liegt und somit den Mittelwert nach oben zieht. Die Frauen sind mit 54% in der Überzahl und auch die Kunden ohne Bonuskarte überwiegen mit 57% in diesem Cluster.

In Nummer 4 finden sich im Durchschnitt jüngere Kunden wieder. Auch wenn das Maximum beim Alter 100 Jahre anzeigt, zeigt der Median 32 Jahre und liegt somit 10 Jahre unter dem von Cluster 2. Es befinden sich auch Kunden ohne Alter in dem Cluster. Wie das Minimum zeigt, haben alle Kunden zumindest eine Bestellung. Die Ausgaben liegen im unteren Quantil bereits bei €638,76. Dies weist darauf hin, dass sich eher jüngere Kunden mit Bestellungen und höheren Ausgaben in diesem Cluster zusammengefunden haben. Das Geschlecht, sowie der Bonuskartenbesitz sind wieder gemischt. Frauen sowie Bonuskartenbesitzer sind in der Unterzahl.

Die Reservierungen verteilen sich auf vier Cluster. Cluster 1 hat noch den größten Anteil, der allerdings nur bei 0,04 Reservierungen liegt. Somit lassen sich hier keine Aussagen über Kunden mit Reservierungen machen.

Auffällig ist die Tatsache, dass die Ausgaben auch in jenen Clustern, in denen die Männer stärker vertreten sind, höher liegen. Hier könnte ein Zusammenhang bestehen, der noch genauer untersucht werden könnte.

Werden dem Algorithmus neun Cluster als Ziel übergeben, zeigen die Abbildung 12-15 ff. ein interessantes Ergebnis. Es gibt wieder nur einen Cluster, der nur Kunden ohne Alter enthält. Bei genauerem Hinsehen fällt auf, dass dieser Cluster identisch ist, mit jenem der auch bei dem Durchlauf mit 5 Clustern entstanden ist. Wieder trägt er die Nummer 3. Er enthält die exakt gleiche Anzahl von 83.256 Kunden. Auch alle anderen Werte sind identisch. Die fehlende Altersangabe ist für den hierarchischen Algorithmus offenbar eines der Hauptkriterien zur Einteilung.

Die drei Kunden aus dem vorherigen Cluster 0 wurden nun aufgeteilt in die Cluster 1 und 5. Der Mann mit den Ausgaben von €210.416 wurde von den beiden anderen getrennt in Cluster 5 untergebracht.

Mit Cluster 8 ist ein neues Segment mit nur sechs Kunden entstanden, das es vorher nicht gab. Der Minimalwert von €21.183,47 bei den Ausgaben verrät, dass es sich um Kunden handelt, die viel Geld ausgegeben haben. Zwei Drittel davon sind Männer und nur ein Drittel Bonuskartenbesitzer.

Auch in Cluster 6 befinden sich im Verhältnis zur Gesamtanzahlt nur wenige Kunden. Es sind 78. Alle Kunden haben mindestens eine Bestellung. Der Median liegt hier sogar bei 9,5 Bestellungen. Die Ausgaben sind ebenfalls überdurchschnittlich hoch. Das untere Quantil liegt bereits bei € 7.328,86.

Der Cluster 2 entspricht beinahe dem vorherigen Cluster 1. Die Werte sind sehr ähnlich. Bei einem Blick auf die Anzahl der Kunden im Cluster lässt sich erkennen, dass 84 Kunden von diesem Cluster abgewandert sind und neu verteilt wurden.

Als weitere Cluster, die Kunden mit Bestellungen beinhalten, lassen sich die Nummer 4 und 7 identifizieren. Die Kunden aus dem vorherigen Cluster 4 haben sich also in diese zwei aufgeteilt. In Cluster 7, der weniger Kunden enthält, wurden die Kunden ausgelagert, deren Bestell- und Ausgabenwerte höher ausfallen. Dies lässt sich an den Medianen gut erkennen.

Es bleibt noch Cluster 0. Auch hier lässt sich feststellen, dass dieser Cluster identisch ist mit Cluster 2 aus dem vorherigen Durchlauf. Somit gibt es zwei Cluster, die vollkommen identisch sind. Für insgesamt 209.028 Kunden hat sich also nichts geändert. Die Einteilung scheint vor allem aufgrund von Bestellungen und Ausgaben zu geschehen. Darum sind auch Geschlecht, Bonuskarten und Reservierungen durchwegs vermischt und auf die Cluster aufgeteilt.

Werden die Kunden in 15 Cluster aufgeteilt, zeigt sich, im Vergleich zu neun Clustern, ein ähnliches Bild. Cluster 6 ist jetzt Cluster 0, Cluster 8 bleibt bei Nummer 8. Der Kunde aus Cluster 5 findet sich nun in Cluster 11. Die beiden verbleibenden Kunden aus Cluster 1 wurden nun auf die Cluster 9 und 10 aufgeteilt. Drei Cluster sind also identisch und einer hat sich geteilt.

Auch der ursprüngliche Cluster 3 hat sich aufgeteilt. Die Kunden in Cluster 2 und 5 weisen keine Altersangaben auf. Wird die Anzahl der Kunden in den Clustern zusammengezählt, erhält man wieder die 83.256 Kunden ohne Alter. In Cluster 2 befinden sich diejenigen die zumindest einmal bestellt haben. Auch die Ausgaben liegen hier höher, was die Quantilen anzeigen.

Die Cluster 3 und 4 bestehen aus Kunden, die ihr Alter angegeben haben, wobei Kunden in Nummer 3 wenigstens eine Bestellung abgeschlossen haben. Der Mittelwert des Alters liegt bei beiden jenseits der 40 Jahre.

Die restlichen Cluster 1, 6, 7, 12, 13 und 14 sind ebenfalls Cluster, in denen die Kunden mindestens eine Bestellung aufweisen. Diese Segmente unterscheiden sich durch eine Kombination aus Alter, Bestellanzahl und Höhe der Ausgaben.

Reservierungen, Geschlecht und Bonuskarten sind auf die Cluster verteilt. Eine Interpretation ist hier nicht ohne weiteres möglich.

9.3 K-means Segmentierung

Bei dem K-means Algorithmus ist an sich keine weitere Vorbehandlung notwendig. Mehrere Versuche, haben allerdings gezeigt, dass eine z-Transformation der Daten zu Clustern führt, die in der Beschaffenheit näher an jenen statischen Regeln aus Kapitel 7.2 liegen. Ohne diese Transformation, zeigen sich ähnliche Ergebnisse wie beim hierarchischen Algorithmus, bei dem dies aufgrund der benötigten Rechenleistung zur Erstellung der Connectivity Matrix der transformierten Daten nicht möglich war. Diese Transformation wird in Listing 9-3 gezeigt.

```
# standardize
df_tr_std = stats.zscore(df_tr[clmns])

# cluster the data
kmeans = KMeans(n_clusters=clusters, random_state=0, init='k-means++',
n_jobs=-1).fit(df_tr_std)
```

Listing 9-3: K-means Clustering mit z-Transformation

Die z-Transformation ist eine Form der Standardisierung, bei der die Werte nicht mehr in der originalen Maßeinheit gemessen werden, sondern in Vielfachen der Standardabweichung. Die Mittelwerte der transformierten Daten liegen immer bei null. Dies wird gemacht um Stichproben, die unterschiedliche Maßeinheiten verwenden, vergleichbar zu machen.

9.3.1 Laufzeit

Tabelle 6 zeigt die Laufzeit des K-means Algorithmus über den Datensatz. Die Laufzeit bei fünf Clustern beträgt nur 1,37 Sekunden. 15 Cluster benötigen lediglich 5,29 Sekunden Rechenzeit. Der Algorithmus ist anscheinend besser geeignet, um große Datensätze zu verarbeiten als das hierarchische Modell, zumindest was die Laufzeit angeht.

Clusteranzahl	Laufzeit in Sekunden
5	1,37
9	2,18
15	5,29

Tabelle 6: Laufzeit K-means Clustering

9.3.2 Analyse

Beginnend mit der Analyse von fünf Clustern, zu sehen in Abbildung 12-23 ff., zeigen die Clustermittelwerte, dass der Algorithmus zunächst die Binären Variablen Gender und BonusCardOwner in einzelne Cluster aufgeteilt hat. Es gibt also die Unterteilung in vier grundlegende Cluster:

- Frauen ohne Bonuskarte – Cluster 0
- Frauen mit Bonuskarte – Cluster 4
- Männer ohne Bonuskarte – Cluster 1
- Männer mit Bonuskarte – Cluster 3

Dies lässt sich daran erkennen, dass in diesen Spalten die Mittelwerte entweder bei 0 oder 1 liegen. Auch Minimum und Maximum stimmen mit dem Mittelwert überein, nur Cluster 1 enthält offenbar einige Frauen, hier ist der Mittelwert knapp unter 1. Der Minimalwert der Spalte Gender_Male liegt hier bei 0. Kundin 127006 wurde diesem Cluster zugeteilt.

In Cluster 2 finden sich fast ausschließlich jene Kunden, die eine Reservierung getätigt haben. Wie der Minimalwert von 0 zeigt, gibt es allerdings auch einige Kunden ohne Reservierung in diesem Cluster. Bei Durchsicht der Daten, trifft dies auf drei der insgesamt 2890 Kunden in dem Cluster zu. Die Kunden mit den Ids 99287, 130277 und 182207 sind aufgrund der Distanzberechnung in diesem Cluster gelandet.

Erhöht sich die Anzahl der Cluster auf neun, zeigt sich in Abbildung 12-26 ff. wieder, wie der Algorithmus die Cluster in die vier oben erwähnten Kategorien aufteilt. Im Gegensatz zu vorher befinden sich nun aber nur noch etwas mehr als die Hälfte der Frauen ohne Bonuskarte in Cluster 0. Auch bei den anderen Clustern verhält es sich ähnlich. Der Cluster 1, also Männer ohne Bonuskarten, ist nun rein männlich. Keine Frau befindet sich mehr in diesem Cluster, wie sich an der glatten 1 in der Spalte Gender_Male erkennen lässt.

Auch hier findet sich wieder ein Cluster mit Reservierungen. Diesmal ist es Cluster 5. Die Kunden ohne Reservierung sind nun auch von diesem Cluster in einen anderen gewandert. Der Minimalwert an Reservierungen liegt in diesem Cluster nun bei 1.

Interessant ist weiters, dass sowohl Frauen als auch Männer ohne Bonuskarte sich jeweils in zwei Cluster aufteilen. Bei den Frauen sind das Cluster 0 und 3, bei den Männern Cluster 1 und

6. Auffällig ist hier der Unterschied beim Alter zwischen den Clustern. Die Cluster 1 und 3 haben einen Mittelwert, der unter 1 liegt. Dies lässt sich dadurch erklären, dass sich dort Kunden befinden, die entweder keine Altersangabe gemacht haben, das Alter also mit 0 übernommen wurde, oder aber Kunden, die eher jung sind. Der Maximalwert verrät, dass sich keine Personen, die älter als 23 Jahre sind in diesen Clustern befinden.

Ein weiterer interessanter Cluster ist der mit der Nummer 8. In Abbildung 12-28 wird gezeigt, dass nur zwei von insgesamt 249.989 Kunden in diesem Cluster gelandet sind. Abbildung 9-4 verrät, um welche Kunden es sich hierbei handelt. Addiert man 1 zu der Nummer in der ersten Spalte, erhält man die Id der Kunden. Die Unterschiedlichen Nummern gehen auf die verschiedenen Extraktionsmethoden zurück. Python beginnt mit 0 beim Generieren des Outputs, das Groovy Skript zur Extraktion hingegen verwendet 1 als Beginn.

	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True	clusters
99286	40.0	152	210416.76	0	0	1	1	0	8
130276	44.0	48	122695.53	0	0	1	1	0	8

Abbildung 9-4: K-means Clustering mit 9 Clustern - Cluster 8

Somit kommt man zu den Ids 99287 und 130277. Diese Ids wurden bereits weiter oben erwähnt. Sie wurden in den Reservierungscluster beim Durchlauf mit nur fünf Clustern aufgenommen. Mit erhöhter Anzahl an Clustern war die Distanz zu anderen Kunden also groß genug, um einen eigenen Cluster zu bilden. Der Grund hierfür lässt sich auch aus der Abbildung herauslesen. Die TotalOrderSum der beiden Männer ist mit Abstand höher als die anderer Kunden.

In Cluster 7 befinden sich Kunden, die mindestens einmal im Shop bestellt haben. Dabei handelt es sich um Männer und Frauen - mit und ohne Bonuskarte. Auch einige Kunden mit Reservierungen sind in diesem Cluster gelandet, wie ein Blick auf den ReservationCount verrät. Der Höchstwert an Bestellungen eines Kunden liegt bei 257. Dieser Wert ist extrem hoch im Vergleich zu anderen, was vermuten lässt, dass dieser Kunde den Cluster wechseln wird, wenn die Clusteranzahl erhöht wird.

Die Abbildung 12-30 ff., zeigen die Aufteilung der Kunden in 15 verschiedene Cluster. Bei dieser Einteilung fallen sofort zwei Cluster auf, in denen sich Personen mit Reservierungen befinden. Das Geschlecht und die Bonuskartenbesitzer sind in beiden Clustern gemischt. Kunden in Cluster 4 haben genau eine Reservierung getätigt, wie der Minimal- und Maximalwert anzeigen. In Cluster 14 hingegen, befinden sich Personen, die mehrmals reserviert haben. Das Minimum liegt hier bei 2 Reservierungen. Ein paar Personen mit Reservierungen finden sich auch in Cluster 7 wieder. Ihre Anzahlt ist allerdings so gering, dass der Cluster nicht zu jenen mit Reservierungen gezählt werden kann. Eine Erhöhung der Clusteranzahl hat also dazu geführt, dass sich der Reservierungscluster teilt.

In den Clustern 5 und 12 befindet sich jeweils nur ein Kunde. In Cluster 11 sind es zwei. Der Mittelwert bei Nummer 5 zeigt, dass der Kunde aufgrund der hohen Ausgaben und der hohen Bestellanzahl von den anderen getrennt wurde. Aus diesem Grund sind auch die zwei Personen in Cluster 11 isoliert. Deren Ausgaben waren nicht so hoch, allerdings doch um einiges höher als bei der Masse der Kunden. Bei Cluster 12 scheint es die Anzahl der Bestellungen zu sein, die ausschlaggebend war.

Auch hier findet sich wieder eine Trennung in die grundlegenden Cluster, wie schon oben erwähnt. Es sind allerdings mehr als vier. Die Cluster 0, 2, 6 und 9 sind Frauen vorbehalten. Von diesen Frauen besitzen lediglich jene in Cluster 2 eine Bonuskarte. Cluster 6 unterscheidet sich von den zwei übrig gebliebenen dadurch, dass die Bestellanzahl im Median bei 3 liegt. Bei 0 und 9 liegt der Median bei 1. Diese beiden Cluster wiederum grenzen sich voneinander aufgrund des Alters ab. Der Median in Cluster 9 zeigt 42 Jahre, jener von Cluster 0 liegt bei 0.

Die Männer verteilen sich auf die Cluster 1, 3, 8 und 10. Auch hier gibt es wieder nur einen Cluster für Männer mit Bonuskarte, nämlich den mit der Nummer 3. Bei Cluster 1 haben sich einige wenige Frauen untergemischt. Der Mittelwert von Gender_MALE zeigt rund 99,84% Männeranteil an. Dieser hohe Anteil genügt jedoch um den Cluster den Männern zuzusprechen. Auch bei den Bonuskarten gibt es eine Unreinheit, die aber ignoriert werden kann. Der Wert bei BonusCardOwner_FALSE liegt bei 99,98%. Zudem unterscheidet sich Cluster 1 von den anderen durch einen Wert von 3 beim Median der Bestellungen. In den Clustern 8 und 10 liegt dieser bei einer 1. Wie auch schon bei den Frauen, liegt der Unterschied bei den zwei verbleibenden Clustern im Alter. So finden sich bei Nummer 8 keine Personen unter 24 Jahren, während dies das Maximum in Cluster 8 darstellt.

Nun gibt es noch Cluster 7 und 13. In diesen Clustern befinden sich Kunden, die bereits ein oder mehrmals bestellt haben. Mit einem Median von 13 Bestellungen in Cluster 7, befinden sich dort Kunden, die öfter bestellt haben als jene in Cluster 13 wo der Maximalwert bei 11 Bestellungen liegt. Auch die Ausgaben sind dort höher. Zusammengenommen, befinden sich allerdings weniger als 5000 Kunden in diesen Clustern. Die Anzahl der Bestellungen scheint also kein Hauptkriterium zur Auftrennung zu sein.

Der Grund weshalb Reservierungen eigene Cluster bilden ist vermutlich, dass dieses Feature weniger Menschen benutzen. In den Reservierungsclustern befinden sich weniger als 3000 Kunden.

Die Abbildungen Abbildung 9-5 und Abbildung 9-6 zeigen eine graphische Darstellung der Clusteraufteilung nach K-means im dreidimensionalen Raum. Als Beispiel dient die Einteilung in 5 Cluster. Links ist die Gesamtverteilung aller Daten sichtbar, während rechts die Achsen limitiert wurden auf €2500 bei den Ausgaben und 10 bei den Bestellungen. Die unterschiedlichen Farben stehen für die verschiedenen Cluster. Die Säulen im Vordergrund der rechten Grafik zeigen die Bestellungen der Kunden ohne Altersangaben. Der Abstand zu den anderen Punkten erklärt sich durch das Mindestalter von 15 Jahren mit dem die restlichen Kunden in den Datensatz aufgenommen wurden.

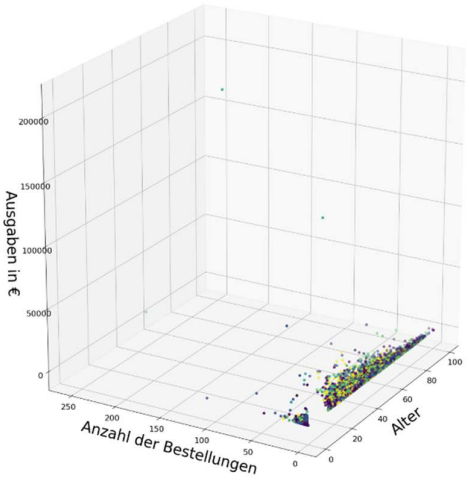

Abbildung 9-5: 3D Plot 5 Cluster

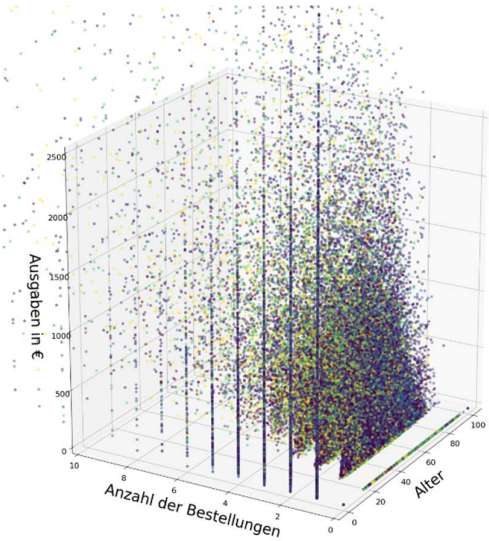

Abbildung 9-6: 3D Plot 5 Cluster vergrößert

9.4 PCA reduzierte K-means Segmentierung

Bei der PCA reduzierten Segmentierung, werden zunächst die Features auf eine bestimmte Anzahl an Dimensionen reduziert. Der K-means Algorithmus wird danach auf das reduzierte Datenset angewandt. Eine Reduktion der Variablen kann möglicherweise das Modell verbessern. Es werden zwei, drei und vier Dimensionen des K-means Durchlaufs mit neun Clustern untersucht. Listing 9-4 zeigt die Anwendung der PCA Reduktion.

```
# standardize
df_tr_std = stats.zscore(df_tr[clmns])
reduced_data = PCA(n_components=2).fit_transform(df_tr_std)

# cluster the data
kmeans = KMeans(n_clusters=clusters, random_state=0, init='k-means++',
n_jobs=-1)
kmeans.fit(reduced_data)
```

Listing 9-4: PCA reduziertes K-means Clustering

9.4.1 Laufzeit

Da die Clusteranzahl sich nicht ändert und der Unterschied nur in den reduzierten Dimensionen liegt, ändert sich die Laufzeit nicht wesentlich wie Tabelle 7 unten zeigt. Vergleicht man die Zeiten mit der Laufzeit des 9 Cluster K-means ohne PCA Reduktion, gibt es auch hier keinen wesentlichen Unterschied. Ein Reduzieren der Dimensionen hat für den vorliegenden Datensatz also kaum Einfluss auf die Laufzeit des Algorithmus.

Clusteranzahl	Dimensionen	Laufzeit in Sekunden
9	2	1,44
9	3	2,27
9	4	2,04

Tabelle 7: Laufzeit PCA K-means Clustering

9.4.2 Analyse

Bei der Analyse mit zwei Dimensionen in Abbildung 12-34 ff. fällt sofort auf, dass es keine Cluster mit weniger als 10.000 Kunden gibt. Die Kunden, die bei K-means als Ausreißer identifiziert wurden, sind hier mit den anderen vermischt und wurden nicht von anderen separiert. Auch die Reservierungen bilden diesmal keinen eigenen Cluster. Die beiden ersten Cluster haben einen niedrigen Altersdurchschnitt gemein. Das obere Quantil zeigt jeweils ein Alter von 0 an. Die Maximalwerte liegen jedoch bei 74 und 78 Jahren. Auch bei der Bestellanzahl reicht die Spanne von 0 bis 50, bzw. 0 bis 257 in Cluster 1. Die Einteilung der Cluster basiert also auf der Trennung zwischen Männern und Frauen ohne Bonuskarte, von denen der Großteil kein Alter angegeben hat. Ganz sauber ist diese Trennung allerdings nicht, wie die Werte in diesen Spalten erkennen lassen.

Cluster 2 und 6 enthalten Männer mit Bonuskarten. Nummer 2 besteht aus Männern, die eher älter sind und wenige bis gar keine Bestellungen vorweisen können. In Cluster 6 haben zumindest 50% eine Bestellung abgeschlossen. Auch das Alter ist hier mit einem Median von 31 Jahren geringer als das von Cluster 2. Die restlichen Männer, ohne Bonuskarte, finden sich in Cluster 5.

Bei den Frauen sind mit 3 und 8 zwei Cluster mit Kundinnen entstanden, die eine Bonuskarte besitzen. Auch hier gibt es wieder einen Altersunterschied zu vermerken. Cluster 3 enthält die älteren Frauen mit eher wenig Bestellungen. Bei Cluster 8, den jüngeren Frauen, hat zumindest jede zweite schon einmal bestellt. Die Cluster 4 und 7, Frauen mit Bonuskarten, sind ähnlich aufgebaut. Auch hier sind es die älteren Frauen, die im Schnitt weniger bestellt haben.

Die Reduktion der Dimensionen hat also funktioniert und das Ergebnis unterscheidet sich deutlich von dem K-means Ergebnis ohne die Verwendung von PCA. Durch das Reduzieren sind allerdings Cluster entstanden, deren Inhalt sich nicht für den Zweck der Personalisierung eignet. Schon der Blick auf die Tabelle der Minimalwerte zeigt, dass das Minimum bei jedem Cluster bei 0 liegt. Der Informationsgehalt der Cluster, also die Qualität der Cluster selbst war ohne PCA besser. Das Unternehmen muss wissen welche Kunden sich in den Segmenten befinden.

Die Abbildung 9-7 unten zeigt eine graphische Repräsentation der Cluster im zweidimensionalen Raum. Jeder Cluster hat eine andere Farbe. Die Nummern in weiß zeigen die Clusterzentroiden.

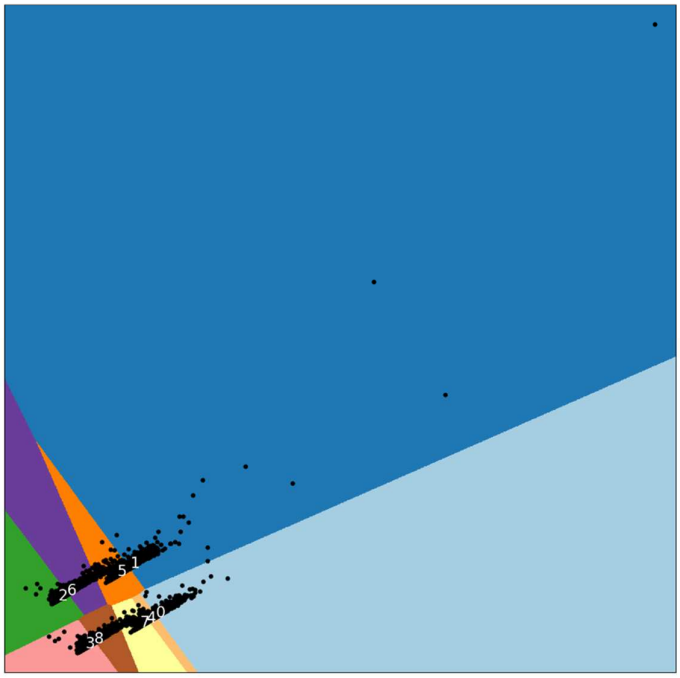

Abbildung 9-7: PCA K-means mit 2 Dimensionen

Eine Verbesserung zeigt der Durchlauf mit drei Dimensionen. Die Abbildung 12-38 ff. zeigen, dass die Ausreißer nun erkannt wurden und in eigenen Clustern gelandet sind. Der Kunde mit den höchsten Ausgaben ist Cluster 4 zugewiesen. Cluster 8 enthält drei Personen. Abbildung 9-8 zeigt diese drei Kunden.

	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True	clusters
127005	0.0	23	73312.18	0	1	0	0	1	8
130276	44.0	48	122695.53	0	0	1	1	0	8
182206	50.0	257	9532.66	0	0	1	0	1	8

Abbildung 9-8: PCA 3 Dimensionen - Cluster 8

Es lässt sich erkennen, dass die ersten beiden hohe Ausgaben gemein haben, während der dritte Kunde aufgrund seiner hohen Anzahl an Bestellungen in dem Cluster gelandet ist.

Cluster 5 und 7 fallen dadurch auf, dass sich Männer und Frauen diese Cluster teilen. Auch das Verhältnis der Bonuskarten ist gemischt. Nummer 5 zeigt Kunden, deren Bestellungen und Ausgaben höher ausfallen. Das Minimum liegt bei einer Bestellung, der Median liegt bei 16. Kunden in Cluster 7 haben weniger Bestellungen, aber mit einem Median von 5 immer noch mehr als jene in anderen Clustern.

Die restlichen Segmente zeigen einen ähnlichen Aufbau wie zuvor mit zwei Dimensionen, die Unreinheiten beim Geschlecht und den Bonuskarten sind jedoch verschwunden. Die Unterteilung erfolgt hier vor allem aufgrund des Alters und der Anzahl an Bestellungen. Auch Reservierungen sind nach wie vor auf alle Cluster verteilt.

Erhöht sich die Dimensionsanzahl auf vier, zeigt sich in den Abbildung 12-42 ff., dass sich nun eigene Cluster mit Reservierungen gebildet haben. Bis auf die Cluster 5,7 und 8 sind alle frei

von Reservierungen. Cluster 7 enthält die Kunden, die mindestens drei Reservierungen abgeschlossen haben. In Cluster 5 befinden sich jene die ein oder zwei Mal reserviert haben. Einige wenige sind aus anderen Gründen in Cluster 8 eingeteilt worden.

Die Cluster mit den Nummern 2 und 8 sind wieder jene, bei denen Geschlecht und Bonuskarten gemischt ausfallen. Diesmal ist es die 8, die Kunden mit sehr hohen Ausgaben und vielen Bestellungen enthält.

Die Ausreißer finden sich in Cluster 6 wieder. Es sind drei Personen, wie zuvor. Allerdings gibt es einen Unterschied. Wie Abbildung 9-9 zeigt, befindet sich der Kunde, der zuvor in einem eigenen Cluster gelandet ist unter den Kunden. Er hat die Kundin mit 23 Bestellungen aus Abbildung 9-8 verdrängt.

	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True	clusters
99286	40.0	152	210416.76	0	0	1	1	0	6
130276	44.0	48	122695.53	0	0	1	1	0	6
182206	50.0	257	9532.66	0	0	1	0	1	6

Abbildung 9-9: PCA 4 Dimesnionen - Cluster 6

Die verbleibenden vier Cluster teilen sich in Männer und Frauen, mit und ohne Bonuskarte auf. Auch beim Alter gibt es Unterschiede. Die Minimal- und Maximalwerte sind jedoch wieder breit gestreut. Die Höchstzahl an Bestellungen liegt jeweils bei sechs.

Durch die Erhöhung der Dimensionen, nähert sich das Ergebnis immer weiter an das Standard-K-means Verfahren an. Es wird leichter die Cluster zu interpretieren und Aussagen über die Kunden darin zu treffen.

9.5 Zusammenfassung der Algorithmen

Wie sich in den Versuchen gezeigt hat, eignet sich K-means ohne PCA reduzierte Dimensionen am besten für den, in dieser Arbeit, verfolgten Zweck. Sowohl bei dem hierarchischen Ansatz als auch bei der Anwendung der PCA Reduktion waren die Clusterinhalte nicht ausreichend geeignet für eine Interpretation. Die Kriterien, nach denen die Cluster eingeteilt wurden, waren andere als bei dem K-means Algorithmus. Es gab keine eindeutige Unterscheidung nach Geschlecht oder Bonuskartenbesitz. Auch Kunden mit Reservierungen konnten nicht von anderen ausreichend getrennt werden. Da diese Merkmale sich über die Cluster verteilt und sogar innerhalb der Cluster gemischt waren, kann keine brauchbare Aussage über Kunden innerhalb der Cluster gemacht werden, wenn das Ziel ist diese Cluster für personalisierten Inhalt zu verwenden. Zumindest was den Kontext dieser Arbeit betrifft. Ziel der Analyse war es, Cluster zu erhalten die eine möglichst genaue Aussage über deren Inhalt ermöglichen. Der Inhalt der Cluster sollte sich dabei an den Merkmalen, die der Experte genannt hat, orientieren.

Die Laufzeit der Algorithmen spricht hingegen eindeutig für die Verwendung von K-means. Abhängig davon, wo und wie oft eine Analyse durchgeführt wird, können die Kosten erheblich gesenkt werden. Vorstellbar wäre eine solche Analyse durch ein Microservice, das in einer Cloudumgebung läuft. Sowohl der Bedarf an Speicher als auch die Dauer der Berechnungen würden sich in zusätzlichen Kosten niederschlagen, wird der hierarchische Algorithmus

gewählt, der für seine Analysen über zehn Minuten Rechenzeit benötigt. Zudem muss bedacht werden, dass die Anzahl der Kunden im Laufe der Zeit wächst. Dieses Wachstum stellt für den K-means Algorithmus kein Problem dar. Auch eine Z-Transformation konnte beim hierarchischen Algorithmus aufgrund fehlender Ressourcen nicht durchgeführt werden.

Die Anwendung der PCA hat gezeigt, dass die Bestellungen und Ausgaben in den Clustern ähnliche Verteilungen aufweisen. Vor allem die Reduktion auf 2 Dimensionen ließ wenig Schlüsse über Bestellungen und Ausgaben zu, die Mittelwerte lagen hier bei knapp über bzw. unter null. Diese Informationen sind allerdings entscheidend, will das Unternehmen neue Erkenntnisse aus den Daten gewinnen. Das Ergebnis verbessert sich allerdings, wenn die Anzahl der Dimensionen erhöht wird. Darum wird eine Reduktion der Dimensionen für das weitere Verfahren ausgeschlossen.

Nach der Analyse lässt sich auch behaupten, dass neun Cluster nur rein rechnerisch das Optimum darstellen. Wie sich gezeigt hat, kann sich das Ergebnis durch die Erhöhung oder Reduzierung der Clusteranzahl ändern. Je höher die Clusteranzahl, desto mehr Kriterien werden bei der Aufteilung berücksichtigt. Es bleibt dem Unternehmen überlassen, welche Anzahl an Clustern am sinnvollsten für eine Interpretation scheint. Für die weitere Analyse in der Arbeit wird nicht mehr von der Ellenbogenmethode Gebrauch gemacht, sondern eine für eine Interpretation geeignete Anzahl an Clustern durch mehrere Versuche ermittelt. Eine Analyse wird nur mehr für die jeweilig ermittelte Anzahl an Clustern vorgenommen.

10 VORBEREITUNG FÜR ENDANALYSE

Nachdem mit K-means ein, für diese Arbeit, geeigneter Algorithmus für das Clustern des Datensatzes gefunden wurde, werden die Kundendaten nun noch einmal einer Vorbehandlung unterzogen. Ziel ist es, durch zusätzliche Vorverarbeitung der Daten, das auf K-means basierte automatische Clustering so nahe wie möglich an die Kundensegmente basierend auf den statischen Regeln aus Kapitel 7.2 heranzubringen.

Vorgenommene Änderungen

Der Datensatz wird in Männer und Frauen aufgeteilt um in weiterer Folge das Alter in, zu dem Geschlecht passende, Kategorien aufteilen zu können.

Folgende Einteilungen werden dabei getroffen:

- Männer unter 40
- Männer über 40 (Zielgruppe)
- Frauen zwischen 20-35 (Zielgruppe)
- Frauen unter 20 oder über 35

Die Geschlechter werden somit schon im Vorbereitungsschritt getrennt und werden dem Algorithmus darum nicht mehr als Feature übergeben. Somit können Männer und Frauen getrennt voneinander untersucht werden. Es kommen neue Kategorien beim Alter hinzu, das zuvor eine rein metrische Variable war. Die Unterteilung des Alters in eine binäre Variable, soll es dem K-means Algorithmus besser ermöglichen die Kunden der Zielgruppen vom Rest der Kunden zu trennen. Kunden ohne Altersangabe werden ebenfalls aus dem Datensatz entfernt. Diese Kunden werden getrennt analysiert. Dadurch soll gewährleistet werden, dass sich die Zielgruppe nicht mit diesen Kunden vermischt und Zusammenhänge aufgrund der Zielgruppeneinteilung erkannt werden können. Listing 10-1 bis Listing 10-3 zeigen die Änderungen, die vorgenommen wurden, um Männer und Frauen getrennt voneinander in Zielgruppen einzuteilen.

```
# drop female gender, to get male only data
df.drop(df[ (df['Gender'] == 'FEMALE') ].index, inplace=True)
…
…
# drop all customers with no age given
df_metrics.drop(df_metrics[ (df_metrics['Age'] == 0) ].index,
inplace=True)

# create category for males of age 40+
df_metrics['Targetgroup'] = 'FALSE'
df_metrics.loc[df_metrics['Age'] >= 40, 'Targetgroup'] = 'TRUE'
```

Listing 10-1: Vorbereitung für männnliche Zielgruppe

Listing 10-1 oben zeigt die Vorbereitungen, die für eine Analyse der Männer durchgeführt wurde. Es werden zuerst alle Frauen aus dem Datensatz entfernt. Im Anschluss werden alle verbleibenden Männer ohne Altersangaben verworfen. In einem weiteren Schritt werden die Männer über 40 Jahre der Zielgruppe zugeteilt.

```
# drop male gender, to get female only data
df.drop(df[ (df['Gender'] == 'MALE') ].index, inplace=True)
…
…
# drop all customers with no age given
df_metrics.drop(df_metrics[ (df_metrics['Age'] == 0) ].index,
inplace=True)

# create category for females of age between 20-35
df_metrics['Targetgroup'] = 'FALSE'
df_metrics.loc[(df_metrics['Age'] >= 20) & (df_metrics['Age'] <= 35),
'Targetgroup'] = 'TRUE'
```

Listing 10-2: Vorbereitung für weibliche Zielgruppe

Die Vorbereitung bei den Frauen geschieht auf die gleiche Weise. Erst werden alle Männer und dann alle Kundinnen ohne Alter aus dem Datensatz genommen. Die letzte Zeile in Listing 10-2 zeigt die Einteilung der Frauen zwischen 20 und 35 Jahren in die Zielgruppe.

```
# replace all NA cells with 0
df_metrics = df_metrics.fillna(0)

# drop all customers with no age given
df_metrics.drop(df_metrics[ (df_metrics['Age'] != 0) ].index,
inplace=True)
```

Listing 10-3: Vorbereitung für Kunden ohne Altersangabe

In Listing 10-3 wird gezeigt, wie alle Kunden deren Alter nicht 0 ist aus dem Datensatz entfernt werden. Es verbleiben somit nur die Kunden ohne Altersangaben für die Clusteranalyse.

11 PRÄSENTATION UND INTERPRETATION DER ERGEBNISSE

Nach der erweiterten Datenvorverarbeitung werden in diesem Kapitel die finalen Ergebnisse der Durchläufe präsentiert und interpretiert. Männer und Frauen, sowie Personen ohne Altersangaben werden jeweils getrennt voneinander untersucht. Für jede dieser Gruppen wird eine Clusteranzahl gewählt, die eine Interpretation und Zuteilung der Cluster zu den statischen Regeln aus Kapitel 7.2 möglichst einfach macht. Dabei wird nicht von der Ellenbogen Methode gebraucht gemacht, um diese Clusteranzahl zu finden. Dieser Schritt erfolgt manuell, indem mehrere Versuche mit unterschiedlicher Clusteranzahl generiert und verglichen werden. Die Kunden in den jeweiligen Clustern werden auf Gemeinsamkeiten untersucht und der Cluster anschließend jener Regel aus Tabelle 1 zugeordnet, die am besten auf den Clusterinhalt zutrifft. In Abschnitt 11.4 werden die Cluster identifiziert, in denen die ausgewählten Personen nach der automatischen Segmentierung gelandet sind und welche Produkte sie zu sehen bekommen. Abschießend werden Zusammenhänge besprochen, die nach der automatischen Segmentierung aus den Daten ablesbar sind. Sämtliche Inhalte können den Abbildungen über den Output des Programms in Anhang B nachgeschlagen werden.

Während der Analyse werden die Männer und Frauen in den Clustern oft als Zielgruppe bezeichnet. Um noch einmal einen Überblick über die Zielgruppen zu verschaffen werden diese hier erneut angeführt:

- Männer über 40 Jahre
- Frauen zwischen 20-35 Jahren

11.1 Analyse Männer mit Altersangaben

Zur Analyse wird ein Durchlauf mit sieben Clustern verwendet, in die sich die Männer aufteilen, zu sehen in Abbildung 12-46 ff. Unter den Männern mit Altersangabe finden sich insgesamt 71794 Kunden. Der Cluster mit Nummer 4 ist auf den ersten Blick als Cluster mit Reservierungen zu erkennen. Das 25% Quantil zeigt, dass 75% der Männer in diesem Cluster eine oder mehr als eine Reservierung abgeschlossen haben. Auch das Minium zeigt, dass es in dem Cluster nur Männer mit Reservierungen gibt. Diese 907 Kunden werden als solche mit Reservierungen eingestuft.

Die Cluster 3 und 6 enthalten Ausreißer in den Daten und diese sind der Grund warum bei den Männern insgesamt sieben Cluster in die finale Analyse eingehen. Cluster 3 enthält zwei Kunden, Cluster 6 nur einen einzigen. Abbildung 11-1 und Abbildung 11-2 unten zeigen die

Werte der Kunden in den Clustern. Cluster 3 wurde aufgrund der hohen Bestellsumme erstellt, während der Kunde in Cluster 6 vor allem durch die Anzahl der Bestellungen hervorsticht.

	Age	OrderCount	TotalOrderSum
99286	40.0	152	210416.76
130276	44.0	48	122695.53

Abbildung 11-1: Männer - Cluster 3

	Age	OrderCount	TotalOrderSum
182206	50.0	257	9532.66

Abbildung 11-2: Männer - Cluster 6

Es stehen nun mehrere statische Regeln zur Auswahl. Zum einen Männer über 40 mit Bestellungen, zum anderen die Kunden mit Ausgaben >€340. Zwar sind die Männer beider Cluster ein Teil der Zielgruppe, jedoch waren die Gründe der Entstehung der Cluster andere. Darum werden diese Cluster der Ausgaben orientierten Regel zugeteilt.

Nun sollen die restlichen Männer der Zielgruppe analysiert werden. Sie verteilen sich auf die Cluster 1 und 2. In Cluster 1 befinden sich vor allem Männer ohne Bonuskarte, während Cluster 2 ausschließlich Bonuskartenbesitzer über 40 Jahren beinhaltet. Ein weiterer Unterschied scheint die Anzahl der Bestellungen zu sein. Cluster 1 zeigt im unteren Quantil bereits eine 1, also sind 75% der Kunden im Cluster mit mindestens einer Bestellung. Bei Cluster 2 hingegen lässt der Mittelwert von 0,74 bereits vermuten, dass hier die meisten Kunden keine Bestellung getätigt haben. Dies bestätigen auch die Quantilen, der Median liegt bei 0, erst das obere Quantil liegt bei einer Bestellung. Somit können die Cluster 1 und 2 den statischen Regeln 1 und 2 zugeordnet werden.

Nun fehlt noch der Rest, die Männer, die nicht der Zielgruppe angehören. Sie finden sich in den Clustern 0 und 5 wieder. Gleich wie bei den Männern in der Zielgruppe, kann hier zwischen jenen unterschieden werden, die bestellt haben und jenen die dies nicht getan haben. Cluster 0 wäre in diesem Fall jener ohne Bestellungen. Dafür besitzen alle Männer in diesem Cluster eine Bonuskarte. Die statische Regel hierfür wäre jene, die Kunden mit Bonuskarte definiert. Bei Cluster 5 gibt es zwei mögliche Optionen. Kunden mit Bestellungen ist aber in diesem Fall die zutreffendere Wahl. Auch wenn der Mittelwert der Bestellsumme mit €398,82 über dem Schwellenwert von €340 liegt, zeigt der Median nur €199,00. Aus diesem Grund scheidet die Regel mit den Ausgaben aus.

Tabelle 8 unten zeigt nun die Clusternummern und deren Zuweisung zu den statischen Regeln.

Cluster	Regel
0	Kunden mit Bonuskarte
1	Männer über 40 mit Bestellungen
2	Männer über 40
3	Kunden mit Ausgaben > €340
4	Kunden mit Reservierungen
5	Kunden mit Bestellungen
6	Kunden mit Ausgaben > €340

Tabelle 8: Analyse der Cluster – Männer

11.2 Analyse Frauen mit Altersangaben

Bei den Frauen erfolgt die Analyse anhand von sechs Clustern. Hier gibt es keine Extremwerte, die aufgrund der Distanz zu anderen Kundinnen in eigenen Clustern landen würden. In Abbildung 12-51 ff. finden sich die Daten der erstellten Cluster. Es gibt insgesamt 81099 Frauen, die ihr Alter angegeben haben. Gleich wie bei den Männern, lässt sich hier der Cluster 2, nach einem Blick auf die Mittelwerte, als Reservierungscluster identifizieren. Das Minimum zeigt, dass sich keine Frauen ohne Reservierungen in dem Cluster befinden, darum wird die Regel für Kunden mit Reservierung direkt zugewiesen.

Die Frauen, die der Zielgruppe angehören befinden sich in den Clustern 0 und 5. Die Quantilen der Cluster weisen darauf hin, dass ein Großteil der Kundinnen in Cluster 0 eine Bestellung vorweisen können. In Cluster 5 hingegen befinden sich hauptsächlich Kundinnen ohne Bestellung, hier liegt nur das obere Quantil bei 1. Die Cluster werden also den Regeln für Frauen in der Zielgruppe zugeordnet.

Die Kundinnen, die nicht der Zielgruppe zugehörig sind, teilen sich auf die Cluster 3 und 4 auf. Hier ist die Nummer 3 jener Cluster der als Bestellcluster identifiziert werden kann. Der Mittelwert, sowie das obere Quantil liegen beide unter der €340 Schwelle, weshalb die Regel für Kunden mit Bestellungen zum Zug kommt. Cluster 4 enthält Kundinnen außerhalb der Zielgruppe, großteils ohne Bestellungen. Allerdings haben alle Kundinnen in diesem Cluster eine Bonuskarte. Die Regel für Kunden mit Bonuskarte fällt darum diesem Cluster zu.

Der letzte verbleibende Cluster mit der Nummer 1 enthält augenscheinlich Kundinnen, die mehrmals bestellt und dabei viel Geld ausgegeben haben. Der Median der Bestellanzahl liegt bereits bei 3 Bestellungen. Die Ausgaben zeigen im unteren Quantil bereits €1189, was bedeutet, dass 75% der Kundinnen über diesem Wert liegen. Die Regel für Ausgaben über €340 wird auf diesen Cluster angewandt.

Tabelle 9 verschafft einen Überblick über die Regeln, die den Clustern zugewiesen wurden.

Cluster	Regel
0	Frauen zwischen 20-35 mit Bestellungen
1	Kunden mit Ausgaben > €340
2	Kunden mit Reservierungen
3	Kunden mit Bestellungen
4	Kunden mit Bonuskarte
5	Frauen zwischen 20-35

Tabelle 9: Analyse der Cluster - Frauen

11.3 Analyse Kunden ohne Altersangaben

Die Ergebnisse für Kunden ohne Altersangabe wurden mit fünf Clustern erstellt. Die Einteilung, zu sehen in Abbildung 12-54 ff., weist wieder einen reinen Cluster auf, in dem Kunden mit Reservierungen zu finden sind. Das Minimum an Reservierungen liegt hier bei 1. Die Regel für Kunden mit Reservierung wird somit Cluster 4 zugewiesen.

Die beiden Cluster 0 und 1 wurden nach Geschlecht getrennt. Frauen befinden sich Cluster 0 und Männer in Cluster 1. Beide enthalten ausschließlich Kunden ohne Bonuskarte, der Mittelwert für die Bestellanzahl liegt knapp unter 1 und das untere Quantil zeigt jeweils eine 1. Somit haben mindestens 75% der Kunden in diesen Clustern eine Bestellung getätigt. Beide Cluster werden darum der Regel für Kunden mit Bestellungen zugeordnet. Es ließe sich auch argumentieren, diese beiden Cluster den letzten beiden statischen Regeln für Männer und Frauen zuzuweisen. Dies sollte geschehen, wenn gänzlich vermieden werden soll, dass auch Kunden ohne Bestellungen, wie sie in diesen Clustern vorkommen, der Bestellorientierten Regel zugeteilt werden sollen. Im Rahmen der Arbeit wird jedoch davon ausgegangen, dass 75% der Kunden mit Bestellungen genug sind, für eine Zuweisung zu der Bestellregel.

Der Cluster mit der Nummer 2 ist auf den ersten Blick schwierig einzuteilen. Der Mittelwert der Bestellungen liegt bei über 1. Allerdings liegt das untere Quantil bei 0 und erst der Median zeigt eine 1. Potenziell haben also nur 50% der Kunden eine Bestellung abgeschlossen. Bei einem Blick auf die Bonuskarten, fällt auf, dass alle Kunden in diesem Cluster eine Bonuskarte besitzen. Die Regel für Kunden mit Bonuskarte scheint somit besser geeignet, als die für Bestellungen.

Cluster 3 hingegen ist nicht schwer zu deuten. Hier befinden sich Kunden, die im Schnitt über 3 Bestellungen vorweisen und das untere Quantil der Bestellsumme zeigt, dass 75% der Kunden mehr als €836,90 ausgegeben haben. Dieser Cluster wird der Ausgabenorientierten Regel zugeteilt.

Tabelle 10 zeigt die Zuteilung der statischen Regeln zu den einzelnen Clustern.

Cluster	Regel
0	Kunden mit Bestellungen
1	Kunden mit Bestellungen
2	Kunden mit Bonuskarte
3	Kunden mit Ausgaben > €340
4	Kunden mit Reservierungen

Tabelle 10: Analyse der Cluster - Kunden ohne Altersangabe

11.4 Vergleich mit statischen Regeln

Nachdem nun alle Kunden in Cluster eingeteilt und den passenden Regeln zugeteilt wurden, werden die ausgewählten Personen aus den Clustern herausgesucht. Eine Übersicht der Personen und ihrer Clusterzuteilung wird am Ende von Anhang B bereitgestellt. In weiterer Folge wird verglichen, welche Produkte diese Personen nun zu sehen bekommen. Auch wird untersucht, ob sich diese Produkte im Vergleich zur Einteilung aufgrund der statischen Regeln geändert haben.

Die Resultate des K-means Clustering sind in Tabelle 11 zusammengefasst. Die unterschiedlichen Farben weisen auf die jeweilige Analyse hin. Bei orangem Hintergrund wurde die Clusteranalyse der Frauen für die Bestimmung der Produkte herangezogen. Blau weist auf die Männer hin und Grau sind die Kunden ohne Altersangabe. Aus diesem Grund hat zum Beispiel ein grauer Cluster 0 eine andere statische Regel hinterlegt als ein Cluster 0 bei den Frauen.

Es lässt sich erkennen, dass einige der statischen Regeln gar nicht aufscheinen. Die Ausgaben über €340, sowie die letzten beiden Regeln für Männer und Frauen fehlen in der Tabelle. Dies liegt unter anderem an den ausgewählten Personen, den aufgestellten statischen Regeln, sowie der Clusteranzahl der Analysen.

Name	ID	Cluster	Regel	Produkt
Birgit	230227	5	Frauen zwischen 20-35	Wohnlandschaft
Sarah	13203	3	Kunden mit Bestellungen	Barhocker
Nicole	16734	3	Kunden mit Bestellungen	Barhocker
Sandra	20498	0	Frauen zwischen 20-35 mit Bestellungen	Esstisch
Beate	219202	2	Kunden mit Reservierungen	Boxspringbett
Gabi	189563	0	Kunden mit Bestellungen	Barhocker
Martina	29105	0	Frauen zwischen 20-35 mit Bestellungen	Esstisch
Claudia	3955	5	Frauen zwischen 20-35	Wohnlandschaft
Nina	63460	4	Kunden mit Reservierungen	Boxspringbett
Peter	12644	0	Kunden mit Bonuskarte	Bücherregal
Franz	128368	0	Kunden mit Bonuskarte	Bücherregal
Jürgen	5807	1	Männer über 40 mit Bestellungen	Grill
Werner	197222	4	Kunden mit Reservierungen	Boxspringbett
Karl	123167	1	Kunden mit Bestellungen	Barhocker
Fabian	7135	5	Kunden mit Bestellungen	Barhocker
Thomas	8755	2	Kunden mit Bonuskarte	Bücherregal
Christian	78077	1	Männer über 40 mit Bestellungen	Grill
Simon	203957	2	Männer über 40	Werkzeugset

Tabelle 11: Produkte nach automatischem Clustering

Nachdem die Produkte bestimmt sind, die die ausgewählten Personen zu sehen bekommen, können diese mit der Zuteilung aus Kapitel 7.2 und der Tabelle 3 verglichen werden. Die Details dieses Vergleichs sind in der Tabelle 12 veranschaulicht. Auch hier gibt es eine Farblegende, wobei grün bedeutet, dass sich die Produkte nicht geändert haben, während rot signalisiert, dass ein Wechsel zu einem anderen Produkt stattgefunden hat.

Name	Produkt (statisch)	Produkt (ML)
Birgit	Esstisch	Wohnlandschaft
Sarah	Barhocker	Barhocker
Nicole	Sonnenliege	Barhocker
Sandra	Esstisch	Esstisch
Beate	Bücherregal	Boxspringbett
Gabi	Barhocker	Barhocker
Martina	Wohnlandschaft	Esstisch
Claudia	Wohnlandschaft	Wohnlandschaft
Nina	Boxspringbett	Boxspringbett
Peter	Grill	Bücherregal
Franz	Barhocker	Bücherregal
Jürgen	Grill	Grill
Werner	Wohnwand	Boxspringbett
Karl	Wohnwand	Barhocker
Fabian	Fernsehsessel	Barhocker
Thomas	Barhocker	Bücherregal
Christian	Werkzeugset	Grill
Simon	Werkzeugset	Werkzeugset

Tabelle 12: Vergleich der Produkte

Sieben der 18 Personen sehen bei beiden Einteilungen das gleiche Produkt. Bei den restlichen elf Kunden hat sich das Produkt geändert. Dies ist das Ergebnis, das aufgrund der Auswahl der statischen Regeln, sowie der Vor- und Aufbereitung der Daten für das K-means Clustering, entstanden ist. Es zeigt, dass der Einsatz von Machine Learning für die Erstellung von Kundensegmenten als Basis für personalisierten Inhalt, verwendet werden kann. Aus den Daten die der K-means Algorithmus liefert, lässt sich gut herauslesen welche Kunden sich in den Clustern befinden. Auch wenn bei dem Vergleich hier in der Arbeit nur knapp 40% das gleiche Produkt zu sehen bekommen, muss dieses Ergebnis nicht als schlecht angesehen werden. Es bedeutet, dass sich die Art und Kriterien der Clustereinteilung unterscheiden. Die statischen Regeln sind von oben nach unten priorisiert. Eine solche Priorisierung gibt es bei der automatischen Einteilung nicht. Am Beispiel Birgit, lässt sich dieser Unterschied erkennen. Birgit ist eine Frau zwischen 20 und 35 Jahren mit einer Bestellung. Nach den statischen Regeln wurde sie eben diesem Cluster zugeteilt. In Tabelle 11 zeigt die Zuteilung nach automatischem Clustering, dass sie der Regel für Frau zwischen 20 und 35 Jahren zugewiesen

wurde. Dies geschah aufgrund der Tatsache, dass der Cluster in dem Birgit gelandet ist überwiegend aus Frauen ohne Bestellungen besteht. Die Einteilung war nicht komplett falsch, nur hatte diese eine Bestellung zu wenig Gewicht und andere Kriterien hatten Vorrang. Eine Änderung der Clusteranzahl bei den Frauen könnte hier vielleicht dazu führen, dass Birgit in dem Cluster mit Bestellungen landet.

Ein Faktor, der bei der Umlegung der automatisierten Cluster auf statische Regeln noch berücksichtigt werden sollte, ist die Tatsache, dass es auch vorkommen kann, dass Datensätze dadurch falsch zugeteilt werden. Bei Nicole zeigt Tabelle 11 die Zuteilung zu der Regel für Kunden mit Bestellungen. Ein Vergleich mit Tabelle 2, lässt erkennen, dass diese Kategorisierung nicht auf Nicole zutrifft. Sie hat keine Bestellungen vorzuweisen. Die Einteilung geschah aufgrund der Tatsache, dass ein Großteil der Kundinnen in ihrem Cluster Bestellungen getätigt haben. Aber eben nicht alle. Sie wurde diesem Cluster vermutlich zugeteilt, weil sie keine Bonuskarte besitzt. Dieses Merkmal hatte für den Algorithmus offenbar mehr Gewicht als die fehlende Bestellung. Hier könnte das Unternehmen einen Schwellenwert definieren, wie viele dieser falsch positiven Kunden erlaubt sind und den Cluster einer anderen statischen Regel zuteilen, sofern dieser überschritten wird.

Das Ergebnis, das erreicht wurde, kann also weder als gut noch als schlecht bezeichnet werden. Es soll den Unterschied zwischen der Verwendung von statischen Regeln und einem automatisierten Ansatz veranschaulichen und zeigen was mit Machine Learning möglich ist. Die statischen Regeln, wie sie in dieser Arbeit aufgestellt wurden, könnten natürlich von jenen abweichen, die das Unternehmen am Ende im produktiven Einsatz verwendet. Auch die Anzahl der verwendeten Cluster könnte unterschiedlich sein. Welchen Ansatz das Unternehmen in Zukunft verfolgen wird, hängt von dessen Prioritäten und Ressourcen ab. Auch eine Kombination aus statischen und automatisiert erstellten Regeln wäre denkbar.

11.5 Abschließende Analyse der automatischen Cluster

In den Daten verstecken sich noch mehr Informationen als nur die Einteilung in die Cluster. Die Analyse kann noch weiter gehen, als nur zu bestimmen welche Kunden in einem Cluster gelandet sind. Dazu wird wieder derselbe Output wie in den vorherigen Kapiteln herangezogen und die Analyse vertieft.

Männer

Bei den Männern sind, wie in Abbildung 12-46 ff. zu sehen ist, in Cluster 3 und 6 jene drei Kunden die als Ausreißer bezeichnet werden können. Ihre Bestellanzahl und Bestellsumme liegen weit über den anderen. Außerdem sind sie Teil der Zielgruppe. Geben Männer in der Zielgruppe also mehr aus als die anderen? Hierzu werden Cluster 1 und 5 verglichen. Beide Cluster haben einen ähnlichen Aufbau. Die Quantilen der Bestellungen zeigen alle eine Bestellung. Auch die Bonuskarten Verteilung ist beinahe identisch und weist auf Kunden ohne Bonuskarte. Cluster 1 beinhaltet die Zielgruppe. Bei näherer Betrachtung der Mittelwerte der

Ausgaben ist zu erkennt, dass die Zielgruppe jedoch unter den anderen Kunden liegt. Der Median der TotalOrderSum zeigt dies auch. Ebenso ist der Aufbau der beiden Cluster 0 und 2 ähnlich. Diesmal befinden sich nur Kunden ohne Bonuskarte in den Clustern. Die Mediane der Ausgaben stehen hier bei 0. Das obere Quantil in Cluster 0 liegt bei €258,85, jenes von Cluster 2 bei €192,95. Es schein also auch hier der Fall zu sein, dass Kunden in Cluster 0, die nicht der Zielgruppe angehören, mehr ausgegeben haben. Vielleicht wurde das Alter der Zielgruppe zu hoch angesetzt? Warum Männer über 40 die Zielgruppe bilden wurde vom befragten Experten nicht genannt. Möglicherweise kaufen diese Kunden vor allem Produkte, bei denen die Marge besonders hoch ist. Dieser Faktor ist bei der Erstellung der Cluster nicht berücksichtigt, kann aber in Zukunft aufgenommen werden.

Es lässt sich noch ein Vergleich für die Männer anstellen. Diesmal geht es darum, ob eine Bonuskarte Einfluss auf die Bestellungen oder Ausgaben hat. Es werden die Männer in der Zielgruppe vergleichen. Sie befinden sich in Cluster 1 und 2, wobei Bonuskatenbesitzer in Cluster 2 vertreten sind. An den Mittelwerten lässt sich bereits erahnen, dass jene Männer ohne Bonuskarte mehr Bestellungen und Ausgaben aufweisen. Dies bestätigen auch die Quantilen. Sowohl der Median der Bestellungen als auch jener der Ausgaben liegen in Cluster 2 bei 0, Die Bestellungen für Kunden ohne Karte zeigen 1 und die Ausgaben 161,86 im Median. Nun werden die Männer, die nicht der Zielgruppe angehören verglichen. Cluster 0, die Bonuskartenbesitzer weisen auch hier weniger Bestellungen und Ausgaben vor als jene Kunden ohne Karte in Cluster 5. Männer ohne Karte haben also öfter gekauft und mehr Geld im Onlineshop ausgegeben. Die Karte bietet nicht nur Vorteile für eine Bestellung. Wie von Möbelhändlern bekannt, gibt es noch andere Zusatz Services, die angeboten werden, wie etwa das Ausleihen eines Lieferwagens. Möglicherweise ist dies einer der Hauptgründe für den Besitz einer Karte. Hier stellt sich die Frage wie sinnvoll eine Bonuskarte dann für die Auswertung ist. Das Unternehmen könnte die Bonuskarte stärker bewerben und über einen längeren Zeitraum verfolgen ob diese Maßnahmen auch zu mehr Verkäufen geführt haben.

Bei dem Reservierungscluster lässt sich erkennen, dass etwas mehr Männer der Zielgruppe diese Funktion bereits benutzt haben. Das Verhältnis liegt hier bei 55% zu 45%. 65% der Männer, die reserviert haben, besitzen auch eine Bonuskarte. Vielleicht gibt es also für die Reservierung Vorteile, beim Besitz einer Bonuskarte. Der Median der Bestellanzahl zeigt an, dass zumindest die Hälfte der Männer in diesem Cluster auch zumindest eine Bestellung aufgegeben haben.

Frauen

Die Analyse wird mit den Frauen fortgeführt. Hierzu werden wieder Abbildung 12-51 ff. verwendet. Bei den Frauen in der Zielgruppe, welche sich in Cluster 0 und 5 befinden, zeigt sich, dass Cluster 0 Frauen enthält, die keine Bonuskarte besitzen. Cluster 5 hingegen, besteht aus Frauen zwischen 20-35 Jahren mit Karte. Gleich wie bei den Männern lassen die Mittelwerte den Schluss zu, dass Frauen ohne Karte mehr Bestellungen und Ausgaben vorweisen. Der Median bei Frauen mit Karte liegt bei den Bestellungen bei 0 und die TotalOrderSum ebenfalls. Dem stehen ein Median von 1 bei Bestellungen und ein Wert von

€133,87 bei den Frauen ohne Karte gegenüber. Bei Frauen außerhalb der Zielgruppe zeigt sich das gleiche Muster. Cluster 3 mit jenen ohne Karte hat mehr Bestellungen und Ausgaben als die Frauen in Cluster 4. Auch hier lässt sich herauslesen, dass der Besitz einer Bonuskarte sich nicht in der Bestellanzahl oder den Ausgaben niederschlägt.

Es folgt der Vergleich der Zielgruppe mit den restlichen, noch nicht betrachteten Frauen. Cluster 0 und 3 sind die Cluster mit Bestellungen, wobei in Cluster 0 Frauen aus der Zielgruppe enthalten sind. Der Unterschied bei den Mittelwerten fällt minimal aus. Die Mediane zeigen ein ähnliches Bild. Sie sind bei den Bestellungen identisch. Bei der TotalOrderSum ist das untere Quantil bei den Frauen in Cluster 3 etwas höher. Beim Median dreht sich dieses Bild zugunsten der Frauen in der Zielgruppe. Das obere Quantil zeigt einen Abstand, der ein wenig deutlicher ausfällt. Mit €363,93 haben die Frauen aus der Zielgruppe die Nase vor den anderen Frauen mit einem Wert von €338,32. Die Differenz bei den Clustern 4 und 5, fällt wesentlich knapper aus. Doch auch hier haben Frauen der Zielgruppe geringfügig mehr Ausgaben. Dies zeigt das obere Quantil. Der Abstand beträgt allerdings nur €0,2. Die Mediane beider Gruppen liegen bei 0. Diese beiden Cluster können folglich als gleichwertig gesehen werden, was die Bestellungen und Ausgaben betrifft. Kann im Gegensatz zu den Männern, kann also behauptet werden, dass die Zielgruppeneinteilung bei den Frauen zutrifft? Die Daten zeigen, dass die Differenzen gering ausfallen und es wahrscheinlich eine Verbesserung gibt, wenn sich die Altersgrenzen der Zielgruppe ein wenig verschieben.

Der Reservierungscluster mit der Nummer 2 verrät, dass mehr Frauen außerhalb der Zielgruppe, der Prozentsatz liegt bei 53%, eine Reservierung vorweisen. Wie auch schon bei den Männern, liegt der Anteil der Kundinnen mit Bonuskarte in diesem Cluster über jenem ohne Karte. Beinahe 60% der Kundinnen besitzen eine Karte. Interessant an diesem Cluster ist ebenfalls die Tatsachen, dass der Mittelwert der Bestellungen bei über 2 liegt. Auch die Ausgaben liegen höher als bei Frauen in anderen Clustern. Der Median bei OrderCount zeigt, dass mindestens die Hälfte der Frauen zumindest eine Bestellung vorweisen können. 25% haben sogar 3 oder mehr Bestellungen. Der Median bei den Ausgaben zeigt einen Wert von €161,93 an und liegt damit sogar über dem Wert der Cluster 0 und 3 die als Cluster mit Bestellungen identifiziert wurden. Es dürften also viele Frauen, die reserviert haben, auch im Shop bestellt haben. Frauen, die ihr Alter angegeben haben und mindestens eine Reservierung gemacht haben, scheinen also kauffreudiger zu sein als ihr männliches Gegenstück.

Im letzten Cluster befinden sich offenbar Kundinnen die oft bestellt und viel ausgegeben haben. Alle Kundinnen in Cluster 1 haben mindestens eine Bestellung abgeschlossen. Die Ausgaben liegen bei wenigstens € 94,69. Das untere Quantil liegt bereits bei €1189,30. 75% der Kundinnen in dem Cluster haben also mehr ausgegeben. Unter diesen kaufkräftigen Kundinnen finden sich allerdings nur 32% Frauen der Zielgruppe. Eine Anpassung der Altersgrenzen scheint also empfehlenswert zu sein. Auch in diesem Cluster befinden sich mit rund 60% mehr Frauen ohne Bonuskarte. Nach Analyse dieses Clusters lässt sich die Frage von oben nun beantworten. Auch bei den Frauen haben jene aus der Zielgruppe weniger bestellt und ausgegeben.

Kunden ohne Altersangabe

Bei den Kunden ohne Altersangabe befinden sich nun wieder Männer und Frauen gemeinsam. Zur Analyse dienen wieder die Abbildung 12-54 ff. Bei den Clustern 0 und 1 springt die Unterteilung der Geschlechter ins Auge. Cluster 0 enthält die Frauen, Cluster 1 die Männer. Die Kunden beider Cluster besitzen keine Bonuskarten. Die Mediane zeigen, dass sich hauptsächlich Kunden mit einer Bestellung in den Clustern befinden. Die Höchstwerte liegen bei 3 bzw. 4 Bestellungen, die Mittelwerte befinden sich knapp unter 1. Bei den Ausgaben liegt Cluster 1 durchwegs vor den Frauen. Damit lässt sich folgende Aussage treffen: Von den Kunden ohne Altersangabe, die keine Bonuskarte besitzen und die einmal im Shop bestellt haben, haben Männer mehr Geld bei ihrer Bestellung ausgegeben.

Die Bonuskartenbesitzer befinden sich fast ausschließlich in Cluster 2. Einige wenige Kunden sind auch in Cluster 3 und 4 abgewandert da andere Attribute stärker ausgeprägt waren. Bei diesen Kunden sind die Männer mit 56% stärker vertreten sind als die Frauen. Unter den Kunden ohne Alter befinden sich also mehr Bonuskarten im Besitz von Männern. Das untere Quantil zeigt an, dass wenigstens 25% dieser Kunden keine Bestellung abgeschlossen haben. Der Median der Ausgaben liegt bei €117,41.

Der Cluster mit den Reservierungen trägt die Nummer 4. Ähnlich wie bei den Bonuskarten sind einige Kunden mit einer Reservierung in andere Cluster gewandert. In dem Cluster befinden sich fast nur Kunden ohne Bonuskarte. Mit 60% haben mehr Männer als Frauen reserviert. Der Median der Bestellungen zeigt an, dass zumindest 50% keine Bestellung aufweisen. Der echte Prozentsatz liegt irgendwo zwischen 50% und 75%, da das obere Quantil bei einer 1 liegt.

Zu guter Letzt, bleibt der Cluster 3 zu analysieren. Auch hier befinden sich fast ausschließlich Kunden ohne Bonuskarte. Die Mittelwerte zeigen, dass es sich um kauffreudige Kunden handelt. Die Bestellungen reichen von einer bis zu 109 als Maximalwert. Im Median haben Kunden 3 Bestellungen. Das untere Quantil der Ausgaben liegt bei €836,90. Interessant ist hier die Tatsache, dass die Bestellungen im Minimum 1 sind und die Ausgaben aber bei 0. Dieser Umstand lässt auf einen Einsatz von Gutscheinen schließen. Bei diesen kauffreudigen Kunden handelt es sich mit rund 72% Wahrscheinlichkeit um einen Mann.

Ein weiterer Faktor, der noch nicht besprochen wurde, ist die Anzahl der Kunden ohne Altersangabe selbst. Unter den Kunden waren insgesamt 97096 die kein Alter angegeben haben. Das sind beinahe 40% aller Onlinekunden in diesem Land. Um die Analysen in Zukunft noch genauer durchführen zu können, kann hier angesetzt und eine Altersangabe für das Erstellen eines Kundenkontos erforderlich gemacht werden.

11.6 Zusammenfassung neuer Erkenntnisse

Da die abschießende Analyse ausführlich beschrieben ist und dabei der Überblick verloren gehen kann, werden hier noch einmal die gewonnenen Erkenntnisse zusammengefasst.

Geben Männer in der Zielgruppe mehr aus als die anderen Männer?

Nein. Männer in der Zielgruppe bestellen weniger oft und geben insgesamt weniger aus. Dies gilt sowohl für Männer mit und ohne Bonuskarte.

Hat der Besitz einer Bonuskarte bei Männern Einfluss auf die Bestellungen und Ausgaben?

Ja. Allerdings einen negativen. Männer ohne Karte haben öfter gekauft und mehr Geld im Onlineshop ausgegeben, egal ob sie der Zielgruppe angehören oder nicht.

Reservieren Männer in der Zielgruppe öfter als andere Männer?

Ja. Der Unterschied ist allerdings nicht sehr deutlich. 55% der Reservierungen wurden von Männern in der Zielgruppe abgeschlossen.

Geben Frauen in der Zielgruppe mehr aus als die anderen Frauen?

Nein. Auch bei den Frauen haben jene in aus Zielgruppe weniger bestellt und ausgegeben. Zwar gab es beinahe einen Gleichstand bei Clustern mit wenigen Bestellungen, bei dem Cluster mit kauffreudigen Kundinnen befanden sich allerdings nur 32% in der Zielgruppe.

Hat der Besitz einer Bonuskarte bei Frauen Einfluss auf die Bestellungen und Ausgaben?

Ja. Auch hier lässt sich herauslesen, dass der Besitz einer Bonuskarte sich nicht in der Bestellanzahl oder den Ausgaben niederschlägt. Auch hier ist der Effekt der Bonuskarte negativ.

Reservieren Frauen in der Zielgruppe öfter als andere Frauen?

Nein. Der Anteil an Frauen in der Zielgruppe bei den Reservierungen liegt bei 47%.

Weitere Erkenntnisse

- Frauen, die ihr Alter angegeben haben und mindestens eine Reservierung gemacht haben, scheinen kauffreudiger zu sein als ihr männliches Gegenstück.
- Von den Kunden ohne Altersangabe, die keine Bonuskarte besitzen und die einmal im Shop bestellt haben, haben Männer mehr Geld bei ihrer Bestellung ausgegeben.
- Unter den Kunden ohne Altersangabe befinden sich mit 56% mehr Bonuskarten im Besitz von Männern.
- Bei den Kunden ohne Altersangabe haben mit einem Anteil von 60% mehr Männer als Frauen reserviert.
- Bei kauffreudigen Kunden (3 Bestellungen im Median) ohne Altersangabe handelt es sich mit rund 72% Wahrscheinlichkeit um einen Mann.

12 EVALUATION DER ERGEBNISSE

Nach der Analyse und Interpretation der Ergebnisse, wird nun auf die Beantwortung der Forschungsfrage eingegangen. Weiters wird die Relevanz der Erkenntnis für die Anwendung in der Praxis hervorgehoben. Im Anschluss wird ein Ausblick auf mögliche fortführende Forschung in diesem Gebiet gegeben. Den Abschluss bilden die Worte und Gedanken des Autors.

12.1 Beantwortung der Forschungsfrage

Die Forschungsfrage wird anhand der Analysen aus dem vorhergehenden Kapitel beantwortet. Zur Erinnerung, die Frage, die beantwortet werden soll, lautet:

Welche neuen Zusammenhänge zwischen Daten lassen sich bei der Segmentierung von Kundendaten im E-Commerce Kontext durch den Einsatz von Machine Learning Algorithmen, im Vergleich zur Segmentierung aufgrund statischer Regeln, erkennen?

Abgesehen von der Tatsache, dass es möglich ist durch Machine Learning sinnvolle Cluster für eine Verwendung im E-Commerce Kontext zu erhalten, zeigt vor allem Kapitel 11.6 einige Beispiele für Schlüsse, welche aus den Daten gezogen werden können. In dem vorliegenden Fall ließ sich erkennen, dass die Zielgruppen des Unternehmens nicht diejenigen waren, die auch am meisten Käufe getätigt oder besonders hohe Ausgaben haben. Das Unternehmen kennt seine Kunden mit Sicherheit gut, jedoch ist das Betreiben eines Onlineshops ein relativ neues Betätigungsfeld. Es kann also durchaus sein, dass die Kunden die online bestellen, nicht mit Kunden die direkt in Möbelhäusern einkaufen zu vergleichen sind und dieser Umstand ein Umdenken sowie eine Anpassung erfordert. Auch der Effekt, der bei den Bonuskarten beobachtet werden konnte, stellt eine neue, so nicht zu erwartende Information dar. Möchte das Unternehmen die Bonuskarte dazu nutzen mehr Bestellungen zu generieren, müssten den Kunden mehr Anreize geboten werden. Die Forschungsfrage ist also bereits mit der abschließenden Clusteranalyse in Kapitel 11.5 und der Zusammenfassung neuer Erkenntnisse in Kapitel 11.6 beantwortet worden.

Hier bleibt nur noch anzumerken, dass sich die Zusammenhänge mit der Art der Daten und deren Aufbereitung ändern werden. Wichtig ist vor allem an dieser Stelle festzuhalten, dass solche Zusammenhänge durch den Einsatz von Machine Learning aus den Daten herausgelesen und für weitere Zwecke verwendet werden können.

12.2 Relevanz der Ergebnisse

Diese Arbeit zeigt, wie eine Kundensegmentierung mithilfe von Machine Learning Algorithmen in einem SAP Commerce Cloud Framework umgesetzt werden kann. Von der Extraktion der Daten bis hin zur automatischen Einteilung der Cluster und der anschließenden Analyse der Clusterinhalte. Es wurden Daten für das Clustern verwendet, wie sie, auch außerhalb dieses Frameworks, in vielen anderen Onlineshops ebenfalls vorliegen. Diese Daten werden von Unternehmen bei der Abwicklung ihrer Bestellungen gesammelt. Die Arbeit zeigt auf, wie aus den Daten zusätzliche Informationen gewonnen werden können, auch wenn bereits statische Regeln zur Segmentierung verwendet werden. Diese Informationen kann ein Unternehmen nutzen um seine Kundenbeziehungen noch effektiver zu Nutzen. Dabei müssen statische Regeln nicht durch die automatisierten Cluster ersetzt werden. Sie können auch als Ergänzung oder einfach nur als Quelle von Informationen und Zusammenhängen gesehen werden, um das bestehende System zu verbessern. Ein wichtiger Faktor ist die Zeit, in der diese Informationen gewonnen werden. Wie der K-means Algorithmus gezeigt hat, lassen sich Daten von 250.000 Kunden in wenigen Sekunden zu wertvollen Informationen verarbeiten, nachdem alle vorbereitenden Schritte dafür gemacht wurden. Eine händische Analyse sämtlicher Daten, auch wenn dazu verschiedene andere Tools verwendet würden, würde wesentlich länger dauern. Auch bei der Anwendung von Machine Learning bedarf es einer Vorbehandlung und Aufbereitung der Daten. Ist dies aber einmal abgeschlossen, lässt sich die Analyse beliebig oft wiederholen. Ein Beispiel wäre die Analyse für Kunden anderer Länder in dem das Unternehmen vertreten ist. Innerhalb der Arbeit wurden Kunden aus Deutschland verwendet. Kundenverhalten aus Österreich oder der Schweiz zu analysieren würde so nur einen Augenblick dauern. Die Kriterien bleiben dabei dieselben und auch das Ergebnis wäre auf die gleiche Weise zu Interpretieren.

Natürlich muss jedes Unternehmen selbst bestimmen, welche Kundendaten für eine Segmentierung interessant sind. Auch die Vorbehandlung und Aufbereitung wird von Fall zu Fall verschieden sein. Die Resultate sprechen jedoch dafür, sich dieser Herausforderung zu stellen. Unternehmen, die ihre Kunden besser verstehen, werden im Wettbewerb um diese Kunden auch immer einen Vorteil gegenüber den anderen haben.

12.3 Weiteres Vorgehen

Im Rahmen dieser Arbeit wurden der hierarchische, sowie der K-means Algorithmus untersucht. Sie sind aber nicht die einzigen Algorithmen, die zum Clustern verwendet werden können. Es wäre also denkbar noch andere Algorithmen auf den Datensatz anzuwenden. Auch bei den angewandten Algorithmen könnten noch weitere Wege verfolgt werden. Durch das Hinzufügen oder Weglassen von Features, einer Änderung in der Vorbehandlung, oder einer Gewichtung der Attribute können sich die Ergebnisse und die Informationen, die sich daraus lesen lassen, entscheidend verändern. Es gibt eine schier unendliche Anzahl an Möglichkeiten und Kombinationen von Schritten, die in diesem Betätigungsfeld verfolgt werden können. Manche Änderungen können dabei zu einem, für das Unternehmen subjektiv „besseren“, Ergebnis

führen, andere zu „schlechteren“ Ergebnissen. Eine dieser Änderungen, war innerhalb der Arbeit die Kunden in Männer, Frauen sowie jene ohne Alter aufzuteilen, um sich besser an den statischen Regeln orientieren zu können. Auch dies muss nicht gemacht werden, sollte das Interesse nicht vorrangig bei den Zielgruppen liegen.

Bei der Extraktion der Daten wurden die Postleitzahlen und Städtenamen in den Datensatz mitaufgenommen. Ein Regionales Clustering könnte damit erstellt werden, um zu sehen in welchen Regionen die Kunden mehr bestellt haben. Wie der Experte erwähnt hat, wäre das in einem weiteren Schritt für das Unternehmen in Zukunft von Interesse. Auch welche Produkte die Kunden gekauft haben sollte in eine Analyse miteinfließen. Zudem könnten auch Daten aus anderen Quellen gesammelt und aufbereitet werden. Eine Analyse von Kunden in Verbindung mit Trackingdaten wäre vorstellbar. Hier könnte untersucht werden, welche Produkte angesehen und dann auch tatsächlich gekauft wurden, oder welche Kunden besonders häufig vor einem Kauf noch abgesprungen sind.

Um diese erstellten Cluster automatisiert nutzbar zu machen, wäre es möglich die Clusterzuteilung wieder in die Kundendatenbank der SAP Commerce Cloud Lösung zu übertragen und bereits bestehende Regeln für personalisierten Inhalt damit anzureichern, oder diese sogar zu ersetzen. Die der Extraktion der Daten, die Clustereinteilung und der Import der Cluster in die Datenbank lassen sich dabei vollkommen automatisierten. Ist dieser Prozess einmal aufgesetzt, könnte eine Kundensegmentierung jeden Tag erneut durchgeführt werden, für beliebig viele Kunden aus unterschiedlichen Ländern.

12.4 Abschließende Worte

Ich arbeite ich nun schon seit einigen Jahren mit dem Unternehmen, für das diese Analyse zur Kundensegmentierung gemacht wurde, zusammen. Als Software Engineer habe ich bereits viel Erfahrung mit der Entwicklung von Software. So war die Programmierung an sich keine Herausforderung. Der Bereich Machine Learning und Data Science im Allgemeinen waren jedoch Neuland für mich. Diese Forschungsthemen interessieren mich persönlich schon seit längerem. Aus diesem Grund entstand auch diese Arbeit. Ich wollte mich in diesem Bereich fortbilden und dabei etwas erschaffen, das einen geschäftlichen Nutzen stiftet. Selbst wenn die Algorithmen relativ schnell erste Ergebnisse lieferten, war ich doch überwältigt von den zahlreichen Optionen und Richtungen, die dabei eingeschlagen werden können. Es war nicht möglich einzuschätzen, wie meine Entscheidungen am Ende das Ergebnis beeinflussen werden. Es hat Monate gedauert das Programm immer weiter anzupassen und dabei verschiedene Dinge auszuprobieren. Leider ist es kaum möglich in so kurzer Zeit, in der diese Arbeit entstanden ist, alles zu berücksichtigen. Dennoch war ich erstaunt über die Ergebnisse und wie gut sich die entstandenen Cluster deuten ließen. Ich hoffe das Unternehmen kann von den gewonnenen Einsichten profitieren und wird diesen Weg weiterverfolgen.

Mit den Erfahrungen, die ich im Laufe dieser Arbeit gewonnen habe, freue ich mich schon auf mein nächstes Machine Learning Projekt.

ANHANG A - 1. Anhang

Ergebnisse - Hierarchische Segmentierung

```
Compute unstructured hierarchical clustering...
Elapsed time: 85.21s
Cluster means
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         26.189533    1.606945     577.406569          0.014700       0.492402     0.507598              0.800312             0.199688
1         40.451613   12.612903    8287.052581          0.161290       0.290323     0.709677              0.677419             0.322581
2          0.000000   23.000000   73312.180000          0.000000       1.000000     0.000000              0.000000             1.000000
3         26.190224    0.699011      65.592997          0.013675       0.513704     0.486296              0.703579             0.296421
4         31.231683    3.219142    2034.482706          0.019142       0.423102     0.576898              0.699670             0.300330
##################################################
Cluster min
          Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         0.0           1         246.08                 0              0            0                     0                    0
1         0.0           1        5047.22                 0              0            0                     0                    0
2         0.0          23       73312.18                 0              1            0                     0                    1
3         0.0           0           0.00                 0              0            0                     0                    0
4         0.0           1        1362.45                 0              0            0                     0                    0
##################################################
Cluster max
            Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         100.0          25        1373.95                 9              1            1                     1                    1
1          66.0          80       32605.89                 2              1            1                     1                    1
2           0.0          23       73312.18                 0              1            0                     0                    1
3         100.0          12         264.61                 6              1            1                     1                    1
4          85.0         109        4922.13                 3              1            1                     1                    1
##################################################
Cluster std
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         22.896917    1.243331     272.954729          0.162965       0.499960     0.499960              0.399780             0.399780
1         18.602578   15.217703    5285.909335          0.522607       0.461414     0.461414              0.475191             0.475191
2               NaN         NaN            NaN               NaN            NaN          NaN                   NaN                  NaN
3         23.662149    0.671930      74.934213          0.141235       0.499819     0.499819              0.456686             0.456686
4         20.842012    4.122084     664.582827          0.155151       0.494214     0.494214              0.458553             0.458553
```

Abbildung 12-1: Hierarchisches Clustering mit 5 Clustern Output

	Age								OrderCount							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	14082.0	26.189533	22.896917	0.0	0.0	30.0	45.0	100.0	14082.0	1.606945	1.243331	1.0	1.0	1.0	2.0	25.0
1	31.0	40.451613	18.602578	0.0	35.0	44.0	53.5	66.0	31.0	12.612903	15.217703	1.0	3.5	10.0	14.0	80.0
2	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0	1.0	23.000000	NaN	23.0	23.0	23.0	23.0	23.0
3	34370.0	26.190224	23.662149	0.0	0.0	29.0	46.0	100.0	34370.0	0.699011	0.671930	0.0	0.0	1.0	1.0	12.0
4	1515.0	31.231683	20.842012	0.0	0.0	35.0	47.0	85.0	1515.0	3.219142	4.122084	1.0	1.0	2.0	4.0	109.0

Abbildung 12-2: Hierarchisches Clustering mit 5 Clustern Output 2

TotalOrderSum								ReservationCount							
count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
14082.0	577.406569	272.954729	246.08	353.850	493.83	742.5925	1373.95	14082.0	0.014700	0.162965	0.0	0.0	0.0	0.0	9.0
31.0	8287.052581	5285.909335	5047.22	5550.615	6399.90	9124.2000	32605.89	31.0	0.161290	0.522607	0.0	0.0	0.0	0.0	2.0
1.0	73312.180000	NaN	73312.18	73312.180	73312.18	73312.1800	73312.18	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0
34370.0	65.592997	74.934213	0.00	0.000	35.94	114.2225	264.61	34370.0	0.013675	0.141235	0.0	0.0	0.0	0.0	6.0
1515.0	2034.482706	664.582827	1362.45	1565.720	1802.29	2279.4900	4922.13	1515.0	0.019142	0.155151	0.0	0.0	0.0	0.0	3.0

Abbildung 12-3: Hierarchisches Clustering mit 5 Clustern Output 3

```
Compute unstructured hierarchical clustering...
Elapsed time: 85.33s
Cluster means
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         27.171060    1.866997     888.949724          0.015064       0.472745     0.527255              0.781169             0.218831
1         26.190224    0.699011      65.592997          0.013675       0.513704     0.486296              0.703579             0.296421
2         25.641585    1.461768     403.484337          0.014496       0.503375     0.496625              0.810999             0.189001
3         30.670509    2.842459    1670.984284          0.017291       0.425552     0.574448              0.705091             0.294909
4         32.464135    4.046414    2832.798861          0.023207       0.417722     0.582278              0.687764             0.312236
5          0.000000   23.000000   73312.180000          0.000000       1.000000     0.000000              0.000000             1.000000
6         41.000000   41.000000   32605.890000          0.000000       0.000000     1.000000              1.000000             0.000000
7         32.285714    7.000000   11825.388571          0.000000       0.285714     0.714286              0.285714             0.714286
8         42.913043   13.086957    6152.826957          0.217391       0.304348     0.695652              0.782609             0.217391
##################################################
Cluster min
           Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0          0.0           1         607.65                 0              0            0                     0                    0
1          0.0           0           0.00                 0              0            0                     0                    0
2          0.0           1         246.08                 0              0            0                     0                    0
3          0.0           1        1362.45                 0              0            0                     0                    0
4          0.0           1        2119.26                 0              0            0                     0                    0
5          0.0          23       73312.18                 0              1            0                     0                    1
6         41.0          41       32605.89                 0              0            1                     1                    0
7          0.0           1        9695.30                 0              0            0                     0                    0
8          0.0           1        5047.22                 0              0            0                     0                    0
```

Abbildung 12-4: Hierarchisches Clustering mit 9 Clustern Output

Cluster max

clusters	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True
0	89.0	25	1373.95	4	1	1	1	1
1	100.0	12	264.61	6	1	1	1	1
2	100.0	12	621.28	9	1	1	1	1
3	85.0	109	2111.49	1	1	1	1	1
4	72.0	37	4922.13	3	1	1	1	1
5	0.0	23	73312.18	0	1	0	0	1
6	41.0	41	32605.89	0	0	1	1	0
7	57.0	21	16691.92	0	1	1	1	1
8	66.0	80	8553.10	2	1	1	1	1

##

Cluster std

clusters	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True
0	22.251310	1.572196	197.391740	0.146905	0.499306	0.499306	0.413494	0.413494
1	23.662149	0.671930	74.934213	0.141235	0.499819	0.499819	0.456686	0.456686
2	23.232723	0.985061	99.545164	0.171284	0.500016	0.500016	0.391531	0.391531
3	20.978881	4.029478	202.214107	0.130416	0.494664	0.494664	0.456221	0.456221
4	20.506271	4.205850	627.976928	0.199071	0.493705	0.493705	0.463895	0.463895
5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	24.212354	7.164728	2301.373023	0.000000	0.487950	0.487950	0.487950	0.487950
8	16.865546	15.965625	1014.429447	0.599736	0.470472	0.470472	0.421741	0.421741

Abbildung 12-5: Hierarchisches Clustering mit 9 Clustern Output 2

	Age								OrderCount							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	5045.0	27.171060	22.251310	0.0	0.0	31.0	45.0	89.0	5045.0	1.866997	1.572196	1.0	1.0	1.0	2.0	25.0
1	34370.0	26.190224	23.662149	0.0	0.0	29.0	46.0	100.0	34370.0	0.699011	0.671930	0.0	0.0	1.0	1.0	12.0
2	9037.0	25.641585	23.232723	0.0	0.0	29.0	45.0	100.0	9037.0	1.461768	0.985061	1.0	1.0	1.0	2.0	12.0
3	1041.0	30.670509	20.978881	0.0	0.0	35.0	47.0	85.0	1041.0	2.842459	4.029478	1.0	1.0	2.0	3.0	109.0
4	474.0	32.464135	20.506271	0.0	24.0	36.0	49.0	72.0	474.0	4.046414	4.205850	1.0	1.0	3.0	5.0	37.0
5	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0	1.0	23.000000	NaN	23.0	23.0	23.0	23.0	23.0
6	1.0	41.000000	NaN	41.0	41.0	41.0	41.0	41.0	1.0	41.000000	NaN	41.0	41.0	41.0	41.0	41.0
7	7.0	32.285714	24.212354	0.0	13.5	42.0	50.0	57.0	7.0	7.000000	7.164728	1.0	2.0	4.0	9.5	21.0
8	23.0	42.913043	16.865546	0.0	37.5	46.0	53.5	66.0	23.0	13.086957	15.965625	1.0	5.0	11.0	14.0	80.0

Abbildung 12-6: Hierarchisches Clustering mit 9 Clustern Output 3

TotalOrderSum								ReservationCount							
count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
5045.0	888.949724	197.391740	607.65	721.6500	841.91	1028.9500	1373.95	5045.0	0.015064	0.146905	0.0	0.0	0.0	0.0	4.0
34370.0	65.592997	74.934213	0.00	0.0000	35.94	114.2225	264.61	34370.0	0.013675	0.141235	0.0	0.0	0.0	0.0	6.0
9037.0	403.484337	99.545164	246.08	318.9500	392.89	479.9500	621.28	9037.0	0.014496	0.171284	0.0	0.0	0.0	0.0	9.0
1041.0	1670.984284	202.214107	1362.45	1498.5800	1637.95	1828.9500	2111.49	1041.0	0.017291	0.130416	0.0	0.0	0.0	0.0	1.0
474.0	2832.798861	627.976928	2119.26	2343.4075	2652.50	3160.8775	4922.13	474.0	0.023207	0.199071	0.0	0.0	0.0	0.0	3.0
1.0	73312.180000	NaN	73312.18	73312.1800	73312.18	73312.1800	73312.18	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0
1.0	32605.890000	NaN	32605.89	32605.8900	32605.89	32605.8900	32605.89	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0
7.0	11825.388571	2301.373023	9695.30	10547.5700	11501.07	11897.1450	16691.92	7.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
23.0	6152.826957	1014.429447	5047.22	5423.1400	5680.43	6487.9200	8553.10	23.0	0.217391	0.599736	0.0	0.0	0.0	0.0	2.0

Abbildung 12-7: Hierarchisches Clustering mit 9 Clustern Output 4

```
Compute unstructured hierarchical clustering...
Elapsed time: 85.56s
Cluster means
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         30.670509    2.842459    1670.984284          0.017291       0.425552     0.574448              0.705091             0.294909
1         32.285714    7.000000   11825.388571          0.000000       0.285714     0.714286              0.285714             0.714286
2         25.484841    1.561285     502.048878          0.016066       0.489764     0.510236              0.805649             0.194351
3         42.913043   13.086957    6152.826957          0.217391       0.304348     0.695652              0.782609             0.217391
4         26.594015    1.770859     766.688791          0.016324       0.485187     0.514813              0.794740             0.205260
5         33.059028    3.659722    2422.265347          0.024306       0.434028     0.565972              0.684028             0.315972
6         21.358777    1.135607      97.067462          0.006902       0.556232     0.443768              0.873867             0.126133
7         34.142857    5.675325    3979.331039          0.012987       0.350649     0.649351              0.636364             0.363636
8         24.468022    1.259652     191.809372          0.008400       0.532714     0.467286              0.839080             0.160920
9         28.270006    2.050086    1121.787473          0.012666       0.449050     0.550950              0.755325             0.244675
10        28.536667    0.350879      11.665334          0.017925       0.491755     0.508245              0.595742             0.404258
11         0.000000   23.000000   73312.180000          0.000000       1.000000     0.000000              0.000000             1.000000
12        25.758401    1.387601     330.027294          0.013326       0.513519     0.486481              0.814986             0.185014
13        41.000000   41.000000   32605.890000          0.000000       0.000000     1.000000              1.000000             0.000000
14        29.706422    3.917431    3107.575688          0.027523       0.422018     0.577982              0.733945             0.266055
##################################################
Cluster min
           Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0          0.0           1        1362.45                 0              0            0                     0                    0
1          0.0           1        9695.30                 0              0            0                     0                    0
2          0.0           1         404.85                 0              0            0                     0                    0
3          0.0           1        5047.22                 0              0            0                     0                    0
4          0.0           1         607.65                 0              0            0                     0                    0
5          0.0           1        2119.26                 0              0            0                     0                    0
6          0.0           1          54.99                 0              0            0                     0                    0
7          0.0           1        3453.67                 0              0            0                     0                    0
8          0.0           1         122.78                 0              0            0                     0                    0
9          0.0           1         942.48                 0              0            0                     0                    0
10         0.0           0           0.00                 0              0            0                     0                    0
11         0.0          23       73312.18                 0              1            0                     0                    1
12         0.0           1         246.08                 0              0            0                     0                    0
13        41.0          41       32605.89                 0              0            1                     1                    0
14         0.0           1        2862.77                 0              0            0                     0                    0
```

Abbildung 12-8: Hierarchisches Clustering mit 15 Clustern Output

Cluster max

clusters	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True
0	85.0	109	2111.49	1	1	1	1	1
1	57.0	21	16691.92	0	1	1	1	1
2	100.0	12	621.28	9	1	1	1	1
3	66.0	80	8553.10	2	1	1	1	1
4	87.0	13	971.35	4	1	1	1	1
5	72.0	37	2842.82	3	1	1	1	1
6	89.0	11	145.21	3	1	1	1	1
7	71.0	22	4922.13	1	1	1	1	1
8	92.0	12	264.61	5	1	1	1	1
9	89.0	25	1373.95	2	1	1	1	1
10	100.0	5	69.00	6	1	1	1	1
11	0.0	23	73312.18	0	1	0	0	1
12	92.0	12	421.45	4	1	1	1	1
13	41.0	41	32605.89	0	0	1	1	0
14	64.0	18	3432.73	2	1	1	1	1

##

Cluster std

clusters	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True
0	20.978881	4.029478	202.214107	0.130416	0.494664	0.494664	0.456221	0.456221
1	24.212354	7.164728	2301.373023	0.000000	0.487950	0.487950	0.487950	0.487950
2	23.030553	1.101187	59.213269	0.201732	0.499960	0.499960	0.395751	0.395751
3	16.865546	15.965625	1014.429447	0.599736	0.470472	0.470472	0.421741	0.421741
4	22.443072	1.391398	95.101703	0.160433	0.499856	0.499856	0.403953	0.403953
5	20.791817	4.199989	204.932330	0.211432	0.496491	0.496491	0.465711	0.465711
6	23.962664	0.451509	22.517898	0.097788	0.496861	0.496861	0.332022	0.332022
7	19.332080	4.913573	426.627651	0.113961	0.480302	0.480302	0.484200	0.484200
8	23.864360	0.649300	34.709922	0.124116	0.498965	0.498965	0.367484	0.367484
9	21.845741	1.855469	115.119807	0.116895	0.497541	0.497541	0.430018	0.430018
10	23.161256	0.505573	18.498656	0.158842	0.499944	0.499944	0.490760	0.490760
11	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
12	23.383793	0.881537	45.212567	0.144489	0.499865	0.499865	0.388346	0.388346
13	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
14	20.475506	3.361350	156.819123	0.213382	0.496163	0.496163	0.443934	0.443934

Abbildung 12-9: Hierarchisches Clustering mit 15 Clustern Output 2

	Age								OrderCount							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	1041.0	30.670509	20.978881	0.0	0.0	35.0	47.0	85.0	1041.0	2.842459	4.029478	1.0	1.0	2.0	3.0	109.0
1	7.0	32.285714	24.212354	0.0	13.5	42.0	50.0	57.0	7.0	7.000000	7.164728	1.0	2.0	4.0	9.5	21.0
2	3859.0	25.484841	23.030553	0.0	0.0	29.0	44.0	100.0	3859.0	1.561285	1.101187	1.0	1.0	1.0	2.0	12.0
3	23.0	42.913043	16.865546	0.0	37.5	46.0	53.5	66.0	23.0	13.086957	15.965625	1.0	5.0	11.0	14.0	80.0
4	3308.0	26.594015	22.443072	0.0	0.0	30.0	44.0	87.0	3308.0	1.770859	1.391398	1.0	1.0	1.0	2.0	13.0
5	288.0	33.059028	20.791817	0.0	25.0	36.0	50.0	72.0	288.0	3.659722	4.199989	1.0	1.0	2.0	5.0	37.0
6	7389.0	21.358777	23.962664	0.0	0.0	0.0	43.0	89.0	7389.0	1.135607	0.451509	1.0	1.0	1.0	1.0	11.0
7	77.0	34.142857	19.332080	0.0	27.0	37.0	50.0	71.0	77.0	5.675325	4.913573	1.0	2.0	4.0	8.0	22.0
8	6786.0	24.468022	23.864360	0.0	0.0	27.0	45.0	92.0	6786.0	1.259652	0.649300	1.0	1.0	1.0	1.0	12.0
9	1737.0	28.270006	21.845741	0.0	0.0	32.0	46.0	89.0	1737.0	2.050086	1.855469	1.0	1.0	1.0	2.0	25.0
10	20195.0	28.536667	23.161256	0.0	0.0	32.0	47.0	100.0	20195.0	0.350879	0.505573	0.0	0.0	0.0	1.0	5.0
11	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0	1.0	23.000000	NaN	23.0	23.0	23.0	23.0	23.0
12	5178.0	25.758401	23.383793	0.0	0.0	29.0	46.0	92.0	5178.0	1.387601	0.881537	1.0	1.0	1.0	1.0	12.0
13	1.0	41.000000	NaN	41.0	41.0	41.0	41.0	41.0	1.0	41.000000	NaN	41.0	41.0	41.0	41.0	41.0
14	109.0	29.706422	20.475506	0.0	0.0	35.0	44.0	64.0	109.0	3.917431	3.361350	1.0	1.0	3.0	5.0	18.0

Abbildung 12-10: Hierarchisches Clustering mit 15 Clustern Output 3

TotalOrderSum								ReservationCount							
count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
1041.0	1670.984284	202.214107	1362.45	1498.5800	1637.950	1828.9500	2111.49	1041.0	0.017291	0.130416	0.0	0.0	0.0	0.0	1.0
7.0	11825.388571	2301.373023	9695.30	10547.5700	11501.070	11897.1450	16691.92	7.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
3859.0	502.048878	59.213269	404.85	444.0000	498.820	548.2600	621.28	3859.0	0.016066	0.201732	0.0	0.0	0.0	0.0	9.0
23.0	6152.826957	1014.429447	5047.22	5423.1400	5680.430	6487.9200	8553.10	23.0	0.217391	0.599736	0.0	0.0	0.0	0.0	2.0
3308.0	766.688791	95.101703	607.65	686.4225	758.895	838.9500	971.35	3308.0	0.016324	0.160433	0.0	0.0	0.0	0.0	4.0
288.0	2422.265347	204.932330	2119.26	2237.6050	2390.730	2593.0775	2842.82	288.0	0.024306	0.211432	0.0	0.0	0.0	0.0	3.0
7389.0	97.067462	22.517898	54.99	78.9000	95.970	113.8500	145.21	7389.0	0.006902	0.097788	0.0	0.0	0.0	0.0	3.0
77.0	3979.331039	426.627651	3453.67	3590.5900	3909.070	4236.1500	4922.13	77.0	0.012987	0.113961	0.0	0.0	0.0	0.0	1.0
6786.0	191.809372	34.709922	122.78	161.9500	189.550	219.7100	264.61	6786.0	0.008400	0.124116	0.0	0.0	0.0	0.0	5.0
1737.0	1121.787473	115.119807	942.48	1027.0000	1099.000	1215.9100	1373.95	1737.0	0.012666	0.116895	0.0	0.0	0.0	0.0	2.0
20195.0	11.665334	18.498656	0.00	0.0000	0.000	23.9300	69.00	20195.0	0.017925	0.158842	0.0	0.0	0.0	0.0	6.0
1.0	73312.180000	NaN	73312.18	73312.1800	73312.180	73312.1800	73312.18	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0
5178.0	330.027294	45.212567	246.08	293.8500	328.950	368.3875	421.45	5178.0	0.013326	0.144489	0.0	0.0	0.0	0.0	4.0
1.0	32605.890000	NaN	32605.89	32605.8900	32605.890	32605.8900	32605.89	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0
109.0	3107.575688	156.819123	2862.77	2973.8000	3094.800	3228.3500	3432.73	109.0	0.027523	0.213382	0.0	0.0	0.0	0.0	2.0

Abbildung 12-11: Hierarchisches Clustering mit 15 Clustern Output 4

Ergebnisse - Hierarchische Segmentierung mit Connectivity Matrix

```
Compute connectivity matrix..
Elapsed time: 31.25s
Compute structured hierarchical clustering...
Elapsed time: 639.64s
Cluster means
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         28.000000   74.333333  135474.823333          0.000000       0.333333     0.666667              0.666667             0.333333
1         31.973262    4.731729    3412.777701          0.039216       0.400624     0.599376              0.651515             0.348485
2         43.482333    0.756575     111.669340          0.012698       0.537035     0.462965              0.574563             0.425437
3          0.000000    0.939308     131.305716          0.016035       0.477407     0.522593              0.960099             0.039901
4         27.619182    1.968745     944.005073          0.015447       0.464535     0.535465              0.771452             0.228548
##################################################
Cluster min
            Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0          0.0          23       73312.18                 0              0            0                     0                    0
1          0.0           1        2263.39                 0              0            0                     0                    0
2         15.0           0           0.00                 0              0            0                     0                    0
3          0.0           0           0.00                 0              0            0                     0                    0
4          0.0           1         506.60                 0              0            0                     0                    0
##################################################
Cluster max
             Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0          44.0         152      210416.76                 0              1            1                     1                    1
1          91.0         257       37338.31                 4              1            1                     1                    1
2         100.0          20         516.54                10              1            1                     1                    1
3           0.0          26         507.00                 9              1            1                     1                    1
4         100.0         109        2283.96                10              1            1                     1                    1
##################################################
Cluster std
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         24.331050   68.412962   69439.896442          0.000000       0.577350     0.577350              0.577350             0.577350
1         21.008717    7.568634    1988.178277          0.274125       0.490134     0.490134              0.476597             0.476597
2         13.479000    0.846117     139.829597          0.146538       0.498628     0.498628              0.494411             0.494411
3          0.000000    0.622775     133.996709          0.158577       0.499492     0.499492              0.195728             0.195728
4         22.335950    1.900749     394.387480          0.162062       0.498747     0.498747              0.419903             0.419903
```

Abbildung 12-12: Hierarchisches Clustering mit Matrix mit 5 Clustern Output

	Age								OrderCount							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	3.0	28.000000	24.331050	0.0	20.00	40.0	42.0	44.0	3.0	74.333333	68.412962	23.0	35.5	48.0	100.0	152.0
1	2244.0	31.973262	21.008717	0.0	20.75	36.0	48.0	91.0	2244.0	4.731729	7.568634	1.0	2.0	3.0	5.0	257.0
2	125772.0	43.482333	13.479000	15.0	32.00	42.0	54.0	100.0	125772.0	0.756575	0.846117	0.0	0.0	1.0	1.0	20.0
3	83256.0	0.000000	0.000000	0.0	0.00	0.0	0.0	0.0	83256.0	0.939308	0.622775	0.0	1.0	1.0	1.0	26.0
4	38714.0	27.619182	22.335950	0.0	0.00	32.0	45.0	100.0	38714.0	1.968745	1.900749	1.0	1.0	1.0	2.0	109.0

Abbildung 12-13: Hierarchisches Clustering mit Matrix mit 5 Clustern Output 2

TotalOrderSum								ReservationCount							
count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
3.0	135474.823333	69439.896442	73312.18	98003.8550	122695.53	166556.1450	210416.76	3.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
2244.0	3412.777701	1988.178277	2263.39	2513.8675	2874.55	3589.5325	37338.31	2244.0	0.039216	0.274125	0.0	0.0	0.0	0.0	4.0
125772.0	111.669340	139.829597	0.00	0.0000	51.53	185.5725	516.54	125772.0	0.012698	0.146538	0.0	0.0	0.0	0.0	10.0
83256.0	131.305716	133.996709	0.00	22.2000	85.70	203.6500	507.00	83256.0	0.016035	0.158577	0.0	0.0	0.0	0.0	9.0
38714.0	944.005073	394.387480	506.60	638.7550	828.15	1127.2675	2283.96	38714.0	0.015447	0.162062	0.0	0.0	0.0	0.0	10.0

Abbildung 12-14: Hierarchisches Clustering mit Matrix mit 5 Clustern Output 3

```
Compute connectivity matrix..
Elapsed time: 28.46s
Compute structured hierarchical clustering...
Elapsed time: 641.33s
Cluster means
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         43.482333    0.756575     111.669340          0.012698       0.537035     0.462965              0.574563             0.425437
1         22.000000   35.500000   98003.855000          0.000000       0.500000     0.500000              0.500000             0.500000
2         31.868056    4.311574    3135.877287          0.037963       0.402778     0.597222              0.655556             0.344444
3          0.000000    0.939308     131.305716          0.016035       0.477407     0.522593              0.960099             0.039901
4         27.004208    1.776935     764.437329          0.013894       0.473449     0.526551              0.786053             0.213947
5         40.000000  152.000000  210416.760000          0.000000       0.000000     1.000000              1.000000             0.000000
6         35.307692   16.064103    9126.820641          0.076923       0.346154     0.653846              0.538462             0.461538
7         29.718308    2.623461    1556.934552          0.020748       0.434109     0.565891              0.721614             0.278386
8         26.500000    8.666667   28814.368333          0.000000       0.333333     0.666667              0.666667             0.333333
##################################################
Cluster min
           Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         15.0           0           0.00                 0              0            0                     0                    0
1          0.0          23       73312.18                 0              0            0                     0                    0
2          0.0           1        2263.39                 0              0            0                     0                    0
3          0.0           0           0.00                 0              0            0                     0                    0
4          0.0           1         506.60                 0              0            0                     0                    0
5         40.0         152      210416.76                 0              0            1                     1                    0
6          0.0           1        6555.48                 0              0            0                     0                    0
7          0.0           1        1154.41                 0              0            0                     0                    0
8          0.0           1       21183.47                 0              0            0                     0                    0
```

Abbildung 12-15: Hierarchisches Clustering mit Matrix mit 9 Clustern Output

Cluster max

clusters	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True
0	100.0	20	516.54	10	1	1	1	1
1	44.0	48	122695.53	0	1	1	1	1
2	91.0	53	6485.25	4	1	1	1	1
3	0.0	26	507.00	9	1	1	1	1
4	100.0	41	1181.41	6	1	1	1	1
5	40.0	152	210416.76	0	0	1	1	0
6	85.0	257	17571.80	3	1	1	1	1
7	95.0	109	2283.96	10	1	1	1	1
8	59.0	20	37338.31	0	1	1	1	1

##

Cluster std

clusters	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True
0	13.479000	0.846117	139.829597	0.146538	0.498628	0.498628	0.494411	0.494411
1	31.112698	17.677670	34919.301663	0.000000	0.707107	0.707107	0.707107	0.707107
2	20.941762	4.742431	856.774291	0.267844	0.490570	0.490570	0.475297	0.475297
3	0.000000	0.622775	133.996709	0.158577	0.499492	0.499492	0.195728	0.195728
4	22.446445	1.468905	181.625891	0.145082	0.499303	0.499303	0.410097	0.410097
5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	22.119104	29.974595	2634.435994	0.419314	0.478822	0.478822	0.501745	0.501745
7	21.825724	2.833085	296.836679	0.209843	0.495668	0.495668	0.448230	0.448230
8	29.649621	7.174027	5713.607043	0.000000	0.516398	0.516398	0.516398	0.516398

Abbildung 12-16: Hierarchisches Clustering mit Matrix mit 9 Clustern Output 2

	Age								OrderCount							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	125772.0	43.482333	13.479000	15.0	32.00	42.0	54.00	100.0	125772.0	0.756575	0.846117	0.0	0.00	1.0	1.00	20.0
1	2.0	22.000000	31.112698	0.0	11.00	22.0	33.00	44.0	2.0	35.500000	17.677670	23.0	29.25	35.5	41.75	48.0
2	2160.0	31.868056	20.941762	0.0	19.75	36.0	48.00	91.0	2160.0	4.311574	4.742431	1.0	2.00	3.0	5.00	53.0
3	83256.0	0.000000	0.000000	0.0	0.00	0.0	0.00	0.0	83256.0	0.939308	0.622775	0.0	1.00	1.0	1.00	26.0
4	29942.0	27.004208	22.446445	0.0	0.00	31.0	45.00	100.0	29942.0	1.776935	1.468905	1.0	1.00	1.0	2.00	41.0
5	1.0	40.000000	NaN	40.0	40.00	40.0	40.00	40.0	1.0	152.000000	NaN	152.0	152.00	152.0	152.00	152.0
6	78.0	35.307692	22.119104	0.0	27.00	37.5	52.75	85.0	78.0	16.064103	29.974595	1.0	4.00	9.5	18.00	257.0
7	8772.0	29.718308	21.825724	0.0	0.00	34.0	47.00	95.0	8772.0	2.623461	2.833085	1.0	1.00	2.0	3.00	109.0
8	6.0	26.500000	29.649621	0.0	0.00	21.0	54.00	59.0	6.0	8.666667	7.174027	1.0	3.00	9.0	11.25	20.0

Abbildung 12-17: Hierarchisches Clustering mit Matrix mit 9 Clustern Output 3

TotalOrderSum								ReservationCount							
count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
125772.0	111.669340	139.829597	0.00	0.0000	51.530	185.5725	516.54	125772.0	0.012698	0.146538	0.0	0.0	0.0	0.0	10.0
2.0	98003.855000	34919.301663	73312.18	85658.0175	98003.855	110349.6925	122695.53	2.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
2160.0	3135.877287	856.774291	2263.39	2504.1275	2838.190	3480.1525	6485.25	2160.0	0.037963	0.267844	0.0	0.0	0.0	0.0	4.0
83256.0	131.305716	133.996709	0.00	22.2000	85.700	203.6500	507.00	83256.0	0.016035	0.158577	0.0	0.0	0.0	0.0	9.0
29942.0	764.437329	181.625891	506.60	604.0375	732.200	899.0000	1181.41	29942.0	0.013894	0.145082	0.0	0.0	0.0	0.0	6.0
1.0	210416.760000	NaN	210416.76	210416.7600	210416.760	210416.7600	210416.76	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0
78.0	9126.820641	2634.435994	6555.48	7328.8625	8196.155	10041.7050	17571.80	78.0	0.076923	0.419314	0.0	0.0	0.0	0.0	3.0
8772.0	1556.934552	296.836679	1154.41	1304.7500	1490.415	1759.8825	2283.96	8772.0	0.020748	0.209843	0.0	0.0	0.0	0.0	10.0
6.0	28814.368333	5713.607043	21183.47	25597.1925	28745.420	31430.7100	37338.31	6.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0

Abbildung 12-18: Hierarchisches Clustering mit Matrix mit 9 Clustern Output 4

```
Compute connectivity matrix..
Elapsed time: 30.16s
Compute structured hierarchical clustering...
Elapsed time: 637.38s
Cluster means
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         35.307692   16.064103    9126.820641          0.076923       0.346154     0.653846              0.538462             0.461538
1         31.293578    5.718654    4203.437798          0.056575       0.392966     0.607034              0.623853             0.376147
2          0.000000    1.223296     307.916832          0.008548       0.442079     0.557921              0.960334             0.039666
3         42.525627    1.481307     321.784574          0.014680       0.559359     0.440641              0.722554             0.277446
4         43.809045    0.509082      39.915732          0.012021       0.529412     0.470588              0.524025             0.475975
5          0.000000    0.817200      55.367365          0.019254       0.492597     0.507403              0.959998             0.040002
6         31.523218    2.950324    1852.613998          0.024298       0.435475     0.564525              0.694924             0.305076
7         28.055133    1.897556     936.846172          0.014826       0.455819     0.544181              0.773432             0.226568
8         26.500000    8.666667   28814.368333          0.000000       0.333333     0.666667              0.666667             0.333333
9         44.000000   48.000000  122695.530000          0.000000       0.000000     1.000000              1.000000             0.000000
10         0.000000   23.000000   73312.180000          0.000000       1.000000     0.000000              0.000000             1.000000
11        40.000000  152.000000  210416.760000          0.000000       0.000000     1.000000              1.000000             0.000000
12        32.117530    3.700531    2672.275312          0.029880       0.407039     0.592961              0.669323             0.330677
13        26.150363    1.678935     624.360361          0.013136       0.487772     0.512228              0.796308             0.203692
14        28.399171    2.384570    1340.834183          0.018153       0.433110     0.566890              0.741121             0.258879
################################################
Cluster min
          Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         0.0           1        6555.48                 0              0            0                     0                    0
1         0.0           1        3292.26                 0              0            0                     0                    0
2         0.0           1         174.39                 0              0            0                     0                    0
3        16.0           1         174.98                 0              0            0                     0                    0
4        15.0           0           0.00                 0              0            0                     0                    0
5         0.0           0           0.00                 0              0            0                     0                    0
6         0.0           1        1554.56                 0              0            0                     0                    0
7         0.0           1         761.80                 0              0            0                     0                    0
8         0.0           1       21183.47                 0              0            0                     0                    0
9        44.0          48      122695.53                 0              0            1                     1                    0
10        0.0          23       73312.18                 0              1            0                     0                    1
11       40.0         152      210416.76                 0              0            1                     1                    0
12        0.0           1        2263.39                 0              0            0                     0                    0
13        0.0           1         506.60                 0              0            0                     0                    0
14        0.0           1        1154.41                 0              0            0                     0                    0
```

Abbildung 12-19: Hierarchisches Clustering mit Matrix mit 15 Clustern Output

```
Cluster max
             Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0           85.0         257       17571.80                 3              1            1                     1                    1
1           79.0          53        6485.25                 4              1            1                     1                    1
2            0.0          26         507.00                 5              1            1                     1                    1
3           93.0          20         516.54                 9              1            1                     1                    1
4          100.0          10         197.70                10              1            1                     1                    1
5            0.0          11         174.50                 9              1            1                     1                    1
6           89.0         109        2283.96                 6              1            1                     1                    1
7           91.0          25        1181.41                 6              1            1                     1                    1
8           59.0          20       37338.31                 0              1            1                     1                    1
9           44.0          48      122695.53                 0              0            1                     1                    0
10           0.0          23       73312.18                 0              1            0                     0                    1
11          40.0         152      210416.76                 0              0            1                     1                    0
12          91.0          50        3305.45                 4              1            1                     1                    1
13         100.0          41         770.16                 5              1            1                     1                    1
14          95.0          47        1572.55                10              1            1                     1                    1
##################################################
Cluster std
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         22.119104   29.974595    2634.435994          0.419314       0.478822     0.478822              0.501745             0.501745
1         21.034124    6.039461     779.707383          0.338691       0.488783     0.488783              0.484788             0.484788
2          0.000000    0.681189      94.077063          0.120944       0.496644     0.496644              0.195177             0.195177
3         12.847075    0.945561      94.464973          0.160924       0.496472     0.496472              0.447746             0.447746
4         13.672861    0.643783      54.376340          0.141285       0.499137     0.499137              0.499425             0.499425
5          0.000000    0.552735      51.880419          0.172150       0.499949     0.499949              0.195966             0.195966
6         21.208416    3.449448     199.967340          0.224007       0.495886     0.495886              0.460502             0.460502
7         22.255453    1.621197     114.846368          0.148019       0.498063     0.498063              0.418627             0.418627
8         29.649621    7.174027    5713.607043          0.000000       0.516398     0.516398              0.516398             0.516398
9               NaN         NaN            NaN               NaN            NaN          NaN                   NaN                  NaN
10              NaN         NaN            NaN               NaN            NaN          NaN                   NaN                  NaN
11              NaN         NaN            NaN               NaN            NaN          NaN                   NaN                  NaN
12        20.903610    3.899224     280.979610          0.230059       0.491445     0.491445              0.470613             0.470613
13        22.565092    1.324404      72.804685          0.142651       0.499866     0.499866              0.402755             0.402755
14        22.175402    2.250078     112.614041          0.198837       0.495554     0.495554              0.438062             0.438062
```

Abbildung 12-20: Hierarchisches Clustering mit Matrix mit 15 Clustern Output 2

	Age								OrderCount							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	78.0	35.307692	22.119104	0.0	27.0	37.5	52.75	85.0	78.0	16.064103	29.974595	1.0	4.0	9.5	18.00	257.0
1	654.0	31.293578	21.034124	0.0	0.0	35.0	47.00	79.0	654.0	5.718654	6.039461	1.0	2.0	4.0	7.00	53.0
2	25034.0	0.000000	0.000000	0.0	0.0	0.0	0.00	0.0	25034.0	1.223296	0.681189	1.0	1.0	1.0	1.00	26.0
3	32017.0	42.525627	12.847075	16.0	32.0	41.0	52.00	93.0	32017.0	1.481307	0.945561	1.0	1.0	1.0	2.00	20.0
4	93755.0	43.809045	13.672861	15.0	32.0	42.0	54.00	100.0	93755.0	0.509082	0.643783	0.0	0.0	0.0	1.00	10.0
5	58222.0	0.000000	0.000000	0.0	0.0	0.0	0.00	0.0	58222.0	0.817200	0.552735	0.0	1.0	1.0	1.00	11.0
6	3704.0	31.523218	21.208416	0.0	0.0	35.0	48.00	89.0	3704.0	2.950324	3.449448	1.0	1.0	2.0	4.00	109.0
7	13422.0	28.055133	22.255453	0.0	0.0	32.0	46.00	91.0	13422.0	1.897556	1.621197	1.0	1.0	1.0	2.00	25.0
8	6.0	26.500000	29.649621	0.0	0.0	21.0	54.00	59.0	6.0	8.666667	7.174027	1.0	3.0	9.0	11.25	20.0
9	1.0	44.000000	NaN	44.0	44.0	44.0	44.00	44.0	1.0	48.000000	NaN	48.0	48.0	48.0	48.00	48.0
10	1.0	0.000000	NaN	0.0	0.0	0.0	0.00	0.0	1.0	23.000000	NaN	23.0	23.0	23.0	23.00	23.0
11	1.0	40.000000	NaN	40.0	40.0	40.0	40.00	40.0	1.0	152.000000	NaN	152.0	152.0	152.0	152.00	152.0
12	1506.0	32.117530	20.903610	0.0	22.0	36.0	48.00	91.0	1506.0	3.700531	3.899224	1.0	1.0	3.0	5.00	50.0
13	16520.0	26.150363	22.565092	0.0	0.0	30.0	44.00	100.0	16520.0	1.678935	1.324404	1.0	1.0	1.0	2.00	41.0
14	5068.0	28.399171	22.175402	0.0	0.0	33.0	46.00	95.0	5068.0	2.384570	2.250078	1.0	1.0	2.0	3.00	47.0

Abbildung 12-21: Hierarchisches Clustering mit Matrix mit 15 Clustern Output 3

TotalOrderSum								ReservationCount							
count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
78.0	9126.820641	2634.435994	6555.48	7328.8625	8196.155	10041.7050	17571.80	78.0	0.076923	0.419314	0.0	0.0	0.0	0.0	3.0
654.0	4203.437798	779.707383	3292.26	3566.2625	3977.785	4618.4300	6485.25	654.0	0.056575	0.338691	0.0	0.0	0.0	0.0	4.0
25034.0	307.916832	94.077063	174.39	224.1000	298.950	387.9500	507.00	25034.0	0.008548	0.120944	0.0	0.0	0.0	0.0	5.0
32017.0	321.784574	94.464973	174.98	238.9500	309.000	399.0000	516.54	32017.0	0.014680	0.160924	0.0	0.0	0.0	0.0	9.0
93755.0	39.915732	54.376340	0.00	0.0000	0.000	75.7500	197.70	93755.0	0.012021	0.141285	0.0	0.0	0.0	0.0	10.0
58222.0	55.367365	51.880419	0.00	4.9000	40.880	95.6575	174.50	58222.0	0.019254	0.172150	0.0	0.0	0.0	0.0	9.0
3704.0	1852.613998	199.967340	1554.56	1679.7225	1820.395	2009.0450	2283.96	3704.0	0.024298	0.224007	0.0	0.0	0.0	0.0	6.0
13422.0	936.846172	114.846368	761.80	836.9500	923.530	1028.9500	1181.41	13422.0	0.014826	0.148019	0.0	0.0	0.0	0.0	6.0
6.0	28814.368333	5713.607043	21183.47	25597.1925	28745.420	31430.7100	37338.31	6.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
1.0	122695.530000	NaN	122695.53	122695.5300	122695.530	122695.5300	122695.53	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0
1.0	73312.180000	NaN	73312.18	73312.1800	73312.180	73312.1800	73312.18	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0
1.0	210416.760000	NaN	210416.76	210416.7600	210416.760	210416.7600	210416.76	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0
1506.0	2672.275312	280.979610	2263.39	2433.7575	2620.405	2884.8500	3305.45	1506.0	0.029880	0.230059	0.0	0.0	0.0	0.0	4.0
16520.0	624.360361	72.804685	506.60	558.8875	618.850	688.9500	770.16	16520.0	0.013136	0.142651	0.0	0.0	0.0	0.0	5.0
5068.0	1340.834183	112.614041	1154.41	1238.9500	1328.950	1434.4700	1572.55	5068.0	0.018153	0.198837	0.0	0.0	0.0	0.0	10.0

Abbildung 12-22: Hierarchisches Clustering mit Matrix mit 15 Clustern Output 4

Ergebnisse - K-means Segmentierung

```
Compute kmeans clustering...
Elapsed time: 1.37s
Cluster means
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         23.447600    1.110434     265.121792          0.000000       1.000000     0.000000              1.000000             0.000000
1         19.409573    1.092530     322.657504          0.000000       0.000012     0.999988              0.999849             0.000151
2         24.015225    1.551557     435.068401          1.251903       0.430796     0.569204              0.584775             0.415225
3         41.610086    0.798289     231.520105          0.000000       0.000000     1.000000              0.000000             1.000000
4         38.431434    0.917679     234.072722          0.000000       1.000000     0.000000              0.000000             1.000000
##################################################
Cluster min
          Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         0.0           0            0.0                 0              1            0                     1                    0
1         0.0           0            0.0                 0              0            0                     0                    0
2         0.0           0            0.0                 0              0            0                     0                    0
3         0.0           0            0.0                 0              0            1                     0                    1
4         0.0           0            0.0                 0              1            0                     0                    1
##################################################
Cluster max
            Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         100.0          47       31446.00                 0              1            0                     1                    0
1         100.0         109       73312.18                 0              1            1                     1                    1
2          91.0         257      210416.76                10              1            1                     1                    1
3          95.0          31       26106.00                 0              0            1                     0                    1
4          91.0          50       16025.13                 0              1            0                     0                    1
##################################################
Cluster std
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         23.209554    1.039584     427.836225           0.00000       0.000000     0.000000              0.000000             0.000000
1         24.090742    1.159411     595.372981           0.00000       0.003414     0.003414              0.012307             0.012307
2         22.582117    6.243884    4580.327692           0.72167       0.495273     0.495273              0.492846             0.492846
3         16.719595    1.454711     555.907853           0.00000       0.000000     0.000000              0.000000             0.000000
4         15.784377    1.646750     512.214159           0.00000       0.000000     0.000000              0.000000             0.000000
```

Abbildung 12-23: K-means Clustering mit 5 Clustern Output

	Age								OrderCount							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	96030.0	23.447600	23.209554	0.0	0.0	26.0	43.0	100.0	96030.0	1.110434	1.039584	0.0	1.0	1.0	1.0	47.0
1	85821.0	19.409573	24.090742	0.0	0.0	0.0	41.0	100.0	85821.0	1.092530	1.159411	0.0	1.0	1.0	1.0	109.0
2	2890.0	24.015225	22.582117	0.0	0.0	27.5	42.0	91.0	2890.0	1.551557	6.243884	0.0	0.0	1.0	2.0	257.0
3	36349.0	41.610086	16.719595	0.0	32.0	41.0	53.0	95.0	36349.0	0.798289	1.454711	0.0	0.0	0.0	1.0	31.0
4	28899.0	38.431434	15.784377	0.0	29.0	37.0	50.0	91.0	28899.0	0.917679	1.646750	0.0	0.0	0.0	1.0	50.0

Abbildung 12-24: K-means Clustering mit 5 Clustern Output 2

TotalOrderSum								ReservationCount							
count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
96030.0	265.121792	427.836225	0.0	34.93	123.75	328.9500	31446.00	96030.0	0.000000	0.00000	0.0	0.0	0.0	0.0	0.0
85821.0	322.657504	595.372981	0.0	33.85	149.00	403.8500	73312.18	85821.0	0.000000	0.00000	0.0	0.0	0.0	0.0	0.0
2890.0	435.068401	4580.327692	0.0	0.00	40.69	359.2725	210416.76	2890.0	1.251903	0.72167	0.0	1.0	1.0	1.0	10.0
36349.0	231.520105	555.907853	0.0	0.00	0.00	237.4200	26106.00	36349.0	0.000000	0.00000	0.0	0.0	0.0	0.0	0.0
28899.0	234.072722	512.214159	0.0	0.00	0.00	255.8150	16025.13	28899.0	0.000000	0.00000	0.0	0.0	0.0	0.0	0.0

Abbildung 12-25: K-means Clustering mit 5 Clustern Output 3

```
Compute kmeans clustering...
Elapsed time: 2.18s
Cluster means
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         42.922689    1.106148     280.387073          0.000000       1.000000     0.000000              1.000000             0.000000
1          0.389733    1.026855     268.843224          0.000000       0.000000     1.000000              1.000000             0.000000
2         41.675646    0.655513     188.510637          0.000000       0.000000     1.000000              0.000000             1.000000
3          0.675817    0.981436     212.192965          0.000000       1.000000     0.000000              1.000000             0.000000
4         38.428318    0.740018     189.621723          0.000000       1.000000     0.000000              0.000000             1.000000
5         23.821090    1.186643     276.661033          1.255536       0.428822     0.571178              0.589104             0.410896
6         45.927969    1.036710     345.364141          0.000000       0.000000     1.000000              1.000000             0.000000
7         36.000000    7.642663    2173.485354          0.013860       0.515215     0.484785              0.524857             0.475143
8         42.000000  100.000000  166556.145000          0.000000       0.000000     1.000000              1.000000             0.000000
##################################################
Cluster min
           Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         18.0           0           0.00                 0              1            0                     1                    0
1          0.0           0           0.00                 0              0            1                     1                    0
2          0.0           0           0.00                 0              0            1                     0                    1
3          0.0           0           0.00                 0              1            0                     1                    0
4          0.0           0           0.00                 0              1            0                     0                    1
5          0.0           0           0.00                 1              0            0                     0                    0
6         21.0           0           0.00                 0              0            1                     1                    0
7          0.0           1          30.41                 0              0            0                     0                    0
8         40.0          48      122695.53                 0              0            1                     1                    0
```

Abbildung 12-26: K-means Clustering mit 9 Clustern Output

Cluster max

clusters	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True
0	100.0	6	4136.60	0	1	0	1	0
1	23.0	6	5139.88	0	0	1	1	0
2	95.0	6	4596.90	0	0	1	0	1
3	23.0	6	4955.92	0	1	0	1	0
4	91.0	6	3961.04	0	1	0	0	1
5	91.0	53	9499.18	10	1	1	1	1
6	100.0	6	4826.95	0	0	1	1	0
7	87.0	257	73312.18	3	1	1	1	1
8	44.0	152	210416.76	0	0	1	1	0

##

Cluster std

clusters	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True
0	12.863758	0.828296	377.643980	0.000000	0.000000	0.000000	0.000000	0.000000
1	2.864138	0.640627	385.998564	0.000000	0.000000	0.000000	0.000000	0.000000
2	16.731079	0.947362	384.781951	0.000000	0.000000	0.000000	0.000000	0.000000
3	3.708113	0.599434	315.559244	0.000000	0.000000	0.000000	0.000000	0.000000
4	15.809786	1.011289	365.389729	0.000000	0.000000	0.000000	0.000000	0.000000
5	22.656069	2.029831	550.977310	0.724599	0.494995	0.494995	0.492083	0.492083
6	13.137380	0.849133	465.028888	0.000000	0.000000	0.000000	0.000000	0.000000
7	18.334150	6.589171	2411.458159	0.129172	0.499844	0.499844	0.499457	0.499457
8	2.828427	73.539105	62028.276587	0.000000	0.000000	0.000000	0.000000	0.000000

Abbildung 12-27: K-means Clustering mit 9 Clustern Output 2

	Age								OrderCount							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	50976.0	42.922689	12.863758	18.0	32.0	42.0	52.0	100.0	50976.0	1.106148	0.828296	0.0	1.0	1.0	1.0	6.0
1	49711.0	0.389733	2.864138	0.0	0.0	0.0	0.0	23.0	49711.0	1.026855	0.640627	0.0	1.0	1.0	1.0	6.0
2	35563.0	41.675646	16.731079	0.0	32.0	41.0	54.0	95.0	35563.0	0.655513	0.947362	0.0	0.0	0.0	1.0	6.0
3	44117.0	0.675817	3.708113	0.0	0.0	0.0	0.0	23.0	44117.0	0.981436	0.599434	0.0	1.0	1.0	1.0	6.0
4	28152.0	38.428318	15.809786	0.0	29.0	37.0	50.0	91.0	28152.0	0.740018	1.011289	0.0	0.0	0.0	1.0	6.0
5	2845.0	23.821090	22.656069	0.0	0.0	27.0	42.0	91.0	2845.0	1.186643	2.029831	0.0	0.0	1.0	2.0	53.0
6	35304.0	45.927969	13.137380	21.0	35.0	45.0	55.0	100.0	35304.0	1.036710	0.849133	0.0	1.0	1.0	1.0	6.0
7	3319.0	36.000000	18.334150	0.0	28.0	38.0	49.0	87.0	3319.0	7.642663	6.589171	1.0	5.0	6.0	8.0	257.0
8	2.0	42.000000	2.828427	40.0	41.0	42.0	43.0	44.0	2.0	100.000000	73.539105	48.0	74.0	100.0	126.0	152.0

Abbildung 12-28: K-means Clustering mit 9 Clustern Output 3

TotalOrderSum								ReservationCount							
count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
50976.0	280.387073	377.643980	0.00	43.9400	142.950	370.0150	4136.60	50976.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
49711.0	268.843224	385.998564	0.00	32.9250	132.950	349.0000	5139.88	49711.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
35563.0	188.510637	384.781951	0.00	0.0000	0.000	204.6250	4596.90	35563.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
44117.0	212.192965	315.559244	0.00	26.3900	103.340	267.2900	4955.92	44117.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
28152.0	189.621723	365.389729	0.00	0.0000	0.000	222.8500	3961.04	28152.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
2845.0	276.661033	550.977310	0.00	0.0000	33.850	337.5200	9499.18	2845.0	1.255536	0.724599	1.0	1.0	1.0	1.0	10.0
35304.0	345.364141	465.028888	0.00	30.7400	169.000	471.2850	4826.95	35304.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
3319.0	2173.485354	2411.458159	30.41	982.5700	1654.140	2720.6150	73312.18	3319.0	0.013860	0.129172	0.0	0.0	0.0	0.0	3.0
2.0	166556.145000	62028.276587	122695.53	144625.8375	166556.145	188486.4525	210416.76	2.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0

Abbildung 12-29: K-means Clustering mit 9 Clustern Output 4

```
Compute kmeans clustering...
Elapsed time: 5.29s
Cluster means
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0          0.429945    0.949882     195.032560          0.000000       1.000000     0.000000              1.000000             0.000000
1         27.342953    2.856763    1541.870457          0.000000       0.001593     0.998407              0.999823             0.000177
2         38.421049    0.613625     150.587512          0.000000       1.000000     0.000000              0.000000             1.000000
3         41.895034    0.523501     137.629286          0.000000       0.000000     1.000000              0.000000             1.000000
4         23.561418    1.116927     257.549156          1.000000       0.427184     0.572816              0.593499             0.406501
5         40.000000  152.000000  210416.760000          0.000000       0.000000     1.000000              1.000000             0.000000
6         35.223657    3.330761    1076.282126          0.000000       1.000000     0.000000              1.000000             0.000000
7         37.257143   15.420952    3420.929276          0.080000       0.500952     0.499048              0.514286             0.485714
8         46.463229    0.889702     241.843348          0.000000       0.000000     1.000000              1.000000             0.000000
9         42.938880    0.917393     209.714118          0.000000       1.000000     0.000000              1.000000             0.000000
10         0.424225    0.976191     222.961975          0.000000       0.000000     1.000000              1.000000             0.000000
11        22.000000   35.500000   98003.855000          0.000000       0.500000     0.500000              0.500000             0.500000
12        50.000000  257.000000    9532.660000          0.000000       0.000000     1.000000              0.000000             1.000000
13        38.030620    4.267980    1358.004318          0.000000       0.432708     0.567292              0.000000             1.000000
14        25.312629    1.573499     380.996004          2.498965       0.443064     0.556936              0.561077             0.438923
##################################################
Cluster min
          Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         0.0           0           0.00                 0              1            0                     1                    0
1         0.0           1           5.00                 0              0            0                     0                    0
2         0.0           0           0.00                 0              1            0                     0                    1
3         0.0           0           0.00                 0              0            1                     0                    1
4         0.0           0           0.00                 1              0            0                     0                    0
5        40.0         152      210416.76                 0              0            1                     1                    0
6         0.0           1           0.00                 0              1            0                     1                    0
7         0.0           1         123.79                 0              0            0                     0                    0
8        24.0           0           0.00                 0              0            1                     1                    0
9        22.0           0           0.00                 0              1            0                     1                    0
10        0.0           0           0.00                 0              0            1                     1                    0
11        0.0          23       73312.18                 0              0            0                     0                    0
12       50.0         257        9532.66                 0              0            1                     0                    1
13        0.0           1          18.04                 0              0            0                     0                    1
14        0.0           0           0.00                 2              0            0                     0                    0
```

Abbildung 12-30: K-means Clustering mit 15 Clustern Output

Cluster max

clusters	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True
0	21.0	3	2086.20	0	1	0	1	0
1	91.0	10	12711.85	0	1	1	1	1
2	91.0	4	2433.90	0	1	0	0	1
3	95.0	3	2196.95	0	0	1	0	1
4	91.0	11	5041.98	1	1	1	1	1
5	40.0	152	210416.76	0	0	1	1	0
6	100.0	11	5906.82	0	1	0	1	0
7	85.0	109	37338.31	4	1	1	1	1
8	100.0	4	1767.46	0	0	1	1	0
9	100.0	3	1637.95	0	1	0	1	0
10	24.0	4	1484.53	0	0	1	1	0
11	44.0	48	122695.53	0	1	1	1	1
12	50.0	257	9532.66	0	0	1	0	1
13	85.0	11	9695.30	0	1	1	0	1
14	76.0	24	9499.18	10	1	1	1	1

##

Cluster std

clusters	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True
0	2.917410	0.503206	258.904360	0.000000	0.000000	0.000000	0.000000	0.000000
1	21.775155	1.676995	962.304630	0.000000	0.039890	0.039890	0.013306	0.013306
2	15.846610	0.804709	285.401838	0.000000	0.000000	0.000000	0.000000	0.000000
3	16.705269	0.729270	273.758103	0.000000	0.000000	0.000000	0.000000	0.000000
4	22.733634	1.682407	468.077369	0.000000	0.494774	0.494774	0.491284	0.491284
5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	17.950917	1.614531	795.595332	0.000000	0.000000	0.000000	0.000000	0.000000
7	18.066474	8.324750	3847.829149	0.340184	0.500476	0.500476	0.500273	0.500273
8	13.134904	0.648466	282.237669	0.000000	0.000000	0.000000	0.000000	0.000000
9	13.021498	0.541192	251.934046	0.000000	0.000000	0.000000	0.000000	0.000000
10	2.997542	0.539685	266.362336	0.000000	0.000000	0.000000	0.000000	0.000000
11	31.112698	17.677670	34919.301663	0.000000	0.707107	0.707107	0.707107	0.707107
12	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13	15.831699	1.810231	960.625510	0.000000	0.495510	0.495510	0.000000	0.000000
14	22.175731	2.463064	832.954223	1.101441	0.497263	0.497263	0.496770	0.496770

Abbildung 12-31: K-means Clustering mit 15 Clustern Output 2

	Age								OrderCount							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	42859.0	0.429945	2.917410	0.0	0.0	0.0	0.0	21.0	42859.0	0.949882	0.503206	0.0	1.00	1.0	1.00	3.0
1	5648.0	27.342953	21.775155	0.0	0.0	32.0	44.0	91.0	5648.0	2.856763	1.676995	1.0	2.00	3.0	4.00	10.0
2	26966.0	38.421049	15.846610	0.0	29.0	37.0	50.0	91.0	26966.0	0.613625	0.804709	0.0	0.00	0.0	1.00	4.0
3	33849.0	41.895034	16.705269	0.0	32.0	42.0	54.0	95.0	33849.0	0.523501	0.729270	0.0	0.00	0.0	1.00	3.0
4	2369.0	23.561418	22.733634	0.0	0.0	27.0	42.0	91.0	2369.0	1.116927	1.682407	0.0	0.00	1.0	2.00	11.0
5	1.0	40.000000	NaN	40.0	40.0	40.0	40.0	40.0	1.0	152.000000	NaN	152.0	152.00	152.0	152.00	152.0
6	6349.0	35.223657	17.950917	0.0	27.0	36.0	48.0	100.0	6349.0	3.330761	1.614531	1.0	2.00	3.0	4.00	11.0
7	525.0	37.257143	18.066474	0.0	30.0	39.0	50.0	85.0	525.0	15.420952	8.324750	1.0	11.00	13.0	17.00	109.0
8	31995.0	46.463229	13.134904	24.0	35.0	46.0	56.0	100.0	31995.0	0.889702	0.648466	0.0	1.00	1.0	1.00	4.0
9	46679.0	42.938880	13.021498	22.0	32.0	42.0	53.0	100.0	46679.0	0.917393	0.541192	0.0	1.00	1.0	1.00	3.0
10	48050.0	0.424225	2.997542	0.0	0.0	0.0	0.0	24.0	48050.0	0.976191	0.539685	0.0	1.00	1.0	1.00	4.0
11	2.0	22.000000	31.112698	0.0	11.0	22.0	33.0	44.0	2.0	35.500000	17.677670	23.0	29.25	35.5	41.75	48.0
12	1.0	50.000000	NaN	50.0	50.0	50.0	50.0	50.0	1.0	257.000000	NaN	257.0	257.00	257.0	257.00	257.0
13	4213.0	38.030620	15.831699	0.0	30.0	38.0	49.0	85.0	4213.0	4.267980	1.810231	1.0	3.00	4.0	5.00	11.0
14	483.0	25.312629	22.175731	0.0	0.0	29.0	45.0	76.0	483.0	1.573499	2.463064	0.0	0.00	1.0	2.00	24.0

Abbildung 12-32: K-means Clustering mit 15 Clustern Output 3

TotalOrderSum								ReservationCount							
count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
42859.0	195.032560	258.904360	0.00	26.3800	101.160	256.3900	2086.20	42859.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
5648.0	1541.870457	962.304630	5.00	947.7275	1393.055	1878.5250	12711.85	5648.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
26966.0	150.587512	285.401838	0.00	0.0000	0.000	182.9500	2433.90	26966.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
33849.0	137.629286	273.758103	0.00	0.0000	0.000	156.3900	2196.95	33849.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
2369.0	257.549156	468.077369	0.00	0.0000	19.810	327.4900	5041.98	2369.0	1.000000	0.000000	1.0	1.0	1.0	1.0	1.0
1.0	210416.760000	NaN	210416.76	210416.7600	210416.760	210416.7600	210416.76	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0
6349.0	1076.282126	795.595332	0.00	480.9700	888.960	1482.9600	5906.82	6349.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
525.0	3420.929276	3847.829149	123.79	1365.5200	2386.080	4008.6200	37338.31	525.0	0.080000	0.340184	0.0	0.0	0.0	0.0	4.0
31995.0	241.843348	282.237669	0.00	4.9550	138.470	366.7900	1767.46	31995.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
46679.0	209.714118	251.934046	0.00	37.5600	116.920	298.7000	1637.95	46679.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
48050.0	222.961975	266.362336	0.00	29.7500	123.645	323.1500	1484.53	48050.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
2.0	98003.855000	34919.301663	73312.18	85658.0175	98003.855	110349.6925	122695.53	2.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
1.0	9532.660000	NaN	9532.66	9532.6600	9532.660	9532.6600	9532.66	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0
4213.0	1358.004318	960.625510	18.04	674.4600	1136.170	1795.4800	9695.30	4213.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
483.0	380.996004	832.954223	0.00	0.0000	91.380	416.8700	9499.18	483.0	2.498965	1.101441	2.0	2.0	2.0	3.0	10.0

Abbildung 12-33: K-means Clustering mit 15 Clustern Output 4

Ergebnisse - PCA reduzierte K-means Segmentierung

```
Compute PCA kmeans clustering...
Elapsed time: 1.44s
Cluster means
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0          1.310505    1.168300     269.730848          0.008935       1.000000     0.000000              0.999867             0.000133
1          1.918916    1.187955     347.206724          0.012082       0.000057     0.999943              0.999792             0.000208
2         52.617276    0.386604      87.208592          0.028677       0.000000     1.000000              0.000000             1.000000
3         46.153595    0.478329     106.339239          0.025473       1.000000     0.000000              0.000000             1.000000
4         33.927281    1.203935     316.786131          0.008534       1.000000     0.000000              0.999933             0.000067
5         46.638624    0.977233     301.166601          0.016684       0.000000     1.000000              0.998494             0.001506
6         28.708810    1.257001     389.715673          0.016103       0.000000     1.000000              0.000000             1.000000
7         54.669372    0.907232     189.401124          0.011143       1.000000     0.000000              0.998578             0.001422
8         25.339564    1.633993     443.765529          0.018054       1.000000     0.000000              0.000091             0.999909
##################################################
Cluster min
          Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         0.0           0            0.0                 0              1            0                     0                    0
1         0.0           0            0.0                 0              0            0                     0                    0
2         0.0           0            0.0                 0              0            1                     0                    1
3         0.0           0            0.0                 0              1            0                     0                    1
4         0.0           0            0.0                 0              1            0                     0                    0
5         0.0           0            0.0                 0              0            1                     0                    0
6         0.0           0            0.0                 0              0            1                     0                    1
7         0.0           0            0.0                 0              1            0                     0                    0
8         0.0           0            0.0                 0              1            0                     0                    0
```

Abbildung 12-34: PCA K-means 9 Cluster 2 Dimensionen Output

```
Cluster max
             Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0           74.0          50       16025.13                 3              1            0                     1                    1
1           78.0         257      210416.76                 4              1            1                     1                    1
2           95.0           7        2798.00                10              0            1                     0                    1
3           91.0           9        3181.35                10              1            0                     0                    1
4           79.0          28       11607.03                 4              1            0                     1                    1
5          100.0          31       26106.00                 7              0            1                     1                    1
6           78.0          22        9499.18                 4              0            1                     0                    1
7          100.0          28        8304.26                 6              1            0                     1                    1
8           76.0          19       11937.18                 6              1            0                     1                    1
##################################################
Cluster std
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0          6.498196    1.346672     484.925743          0.097117       0.000000     0.000000              0.011533             0.011533
1          7.673375    1.881060    1271.949171          0.117898       0.007526     0.007526              0.014409             0.014409
2         10.379308    0.669877     216.875485          0.231865       0.000000     0.000000              0.000000             0.000000
3         11.662303    0.807276     248.870926          0.228235       0.000000     0.000000              0.000000             0.000000
4          7.898292    0.929048     404.160632          0.125525       0.000000     0.000000              0.008181             0.008181
5         13.179660    1.057013     483.171211          0.187739       0.000000     0.000000              0.038778             0.038778
6         13.188596    1.735684     663.716755          0.142403       0.000000     0.000000              0.000000             0.000000
7          8.794567    0.966449     291.807385          0.149027       0.000000     0.000000              0.037678             0.037678
8         13.074135    1.983813     673.501691          0.162178       0.000000     0.000000              0.009549             0.009549
```

Abbildung 12-35: PCA K-means 9 Cluster 2 Dimensionen Output 2

	Age								OrderCount							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	45104.0	1.310505	6.498196	0.0	0.0	0.0	0.0	74.0	45104.0	1.168300	1.346672	0.0	1.0	1.0	1.0	50.0
1	52970.0	1.918916	7.673375	0.0	0.0	0.0	0.0	78.0	52970.0	1.187955	1.881060	0.0	1.0	1.0	1.0	257.0
2	19842.0	52.617276	10.379308	0.0	44.0	52.0	60.0	95.0	19842.0	0.386604	0.669877	0.0	0.0	0.0	1.0	7.0
3	18412.0	46.153595	11.662303	0.0	36.0	45.0	55.0	91.0	18412.0	0.478329	0.807276	0.0	0.0	0.0	1.0	9.0
4	29882.0	33.927281	7.898292	0.0	28.0	33.0	39.0	79.0	29882.0	1.203935	0.929048	0.0	1.0	1.0	1.0	28.0
5	33865.0	46.638624	13.179660	0.0	36.0	46.0	56.0	100.0	33865.0	0.977233	1.057013	0.0	1.0	1.0	1.0	31.0
6	17140.0	28.708810	13.188596	0.0	26.0	31.0	35.0	78.0	17140.0	1.257001	1.735684	0.0	0.0	1.0	2.0	22.0
7	21807.0	54.669372	8.794567	0.0	49.0	54.0	60.0	100.0	21807.0	0.907232	0.966449	0.0	1.0	1.0	1.0	28.0
8	10967.0	25.339564	13.074135	0.0	23.0	27.0	32.0	76.0	10967.0	1.633993	1.983813	0.0	0.0	1.0	2.0	19.0

Abbildung 12-36: PCA K-means 9 Cluster 2 Dimensionen Output 3

TotalOrderSum								ReservationCount							
count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
45104.0	269.730848	484.925743	0.0	29.8275	109.85	308.9500	16025.13	45104.0	0.008935	0.097117	0.0	0.0	0.0	0.0	3.0
52970.0	347.206724	1271.949171	0.0	34.9425	149.76	402.9500	210416.76	52970.0	0.012082	0.117898	0.0	0.0	0.0	0.0	4.0
19842.0	87.208592	216.875485	0.0	0.0000	0.00	58.8500	2798.00	19842.0	0.028677	0.231865	0.0	0.0	0.0	0.0	10.0
18412.0	106.339239	248.870926	0.0	0.0000	0.00	93.4025	3181.35	18412.0	0.025473	0.228235	0.0	0.0	0.0	0.0	10.0
29882.0	316.786131	404.160632	0.0	54.9725	170.39	431.8250	11607.03	29882.0	0.008534	0.125525	0.0	0.0	0.0	0.0	4.0
33865.0	301.166601	483.171211	0.0	9.9900	144.81	404.9200	26106.00	33865.0	0.016684	0.187739	0.0	0.0	0.0	0.0	7.0
17140.0	389.715673	663.716755	0.0	0.0000	105.85	512.8450	9499.18	17140.0	0.016103	0.142403	0.0	0.0	0.0	0.0	4.0
21807.0	189.401124	291.807385	0.0	26.9300	97.49	243.1500	8304.26	21807.0	0.011143	0.149027	0.0	0.0	0.0	0.0	6.0
10967.0	443.765529	673.501691	0.0	0.0000	191.67	599.0000	11937.18	10967.0	0.018054	0.162178	0.0	0.0	0.0	0.0	6.0

Abbildung 12-37: PCA K-means 9 Cluster 2 Dimensionen Output 4

```
Compute PCA kmeans clustering...
Elapsed time: 2.27s
Cluster means
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         23.133151    1.023689     233.876412          0.008360       1.000000     0.000000              1.000000             0.000000
1         27.930014    1.973441     900.047110          0.026182       0.000000     1.000000              1.000000             0.000000
2         41.636861    0.633038     171.365715          0.020652       0.000000     1.000000              0.000000             1.000000
3         38.402096    0.699485     168.499806          0.018511       1.000000     0.000000              0.000000             1.000000
4         40.000000  152.000000  210416.760000          0.000000       0.000000     1.000000              1.000000             0.000000
5         37.971831   18.024648    5303.054718          0.158451       0.440141     0.559859              0.457746             0.542254
6         17.234359    0.822027     160.667004          0.010213       0.000000     1.000000              1.000000             0.000000
7         36.921830    5.408699    1853.514633          0.079401       0.625564     0.374436              0.522364             0.477636
8         31.333333  109.333333   68513.456667          0.000000       0.333333     0.666667              0.333333             0.666667
##################################################
Cluster min
           Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0          0.0           0           0.00                 0              1            0                     1                    0
1          0.0           0           0.00                 0              0            1                     1                    0
2          0.0           0           0.00                 0              0            1                     0                    1
3          0.0           0           0.00                 0              1            0                     0                    1
4         40.0         152      210416.76                 0              0            1                     1                    0
5          0.0           1         323.09                 0              0            0                     0                    0
6          0.0           0           0.00                 0              0            1                     1                    0
7          0.0           0           0.00                 0              0            0                     0                    0
8          0.0          23        9532.66                 0              0            0                     0                    0
```

Abbildung 12-38: PCA K-means 9 Cluster 3 Dimensionen Output

```
Cluster max
             Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0          100.0           6        2831.07                 6              1            0                     1                    0
1           93.0           8        4204.72                 7              0            1                     1                    0
2           95.0           6        2873.95                 7              0            1                     0                    1
3           91.0           5        2560.00                 8              1            0                     0                    1
4           40.0         152      210416.76                 0              0            1                     1                    0
5           85.0         109       37338.31                 4              1            1                     1                    1
6          100.0           2         847.14                 3              0            1                     1                    0
7          100.0          16        8121.41                10              1            1                     1                    1
8           50.0         257      122695.53                 0              1            1                     1                    1
##################################################
Cluster std
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         23.197879    0.689295     310.264835          0.111836       0.000000     0.000000              0.000000             0.000000
1         24.057109    1.062759     568.876018          0.231722       0.000000     0.000000              0.000000             0.000000
2         16.777360    0.911902     334.561251          0.175589       0.000000     0.000000              0.000000             0.000000
3         15.865580    0.945131     311.274874          0.164043       0.000000     0.000000              0.000000             0.000000
4               NaN         NaN            NaN               NaN            NaN          NaN                   NaN                  NaN
5         17.467362   10.754377    4774.393170          0.537859       0.497280     0.497280              0.499091             0.499091
6         23.600904    0.469548     180.706270          0.111700       0.000000     0.000000              0.000000             0.000000
7         17.967874    2.586801    1126.843090          0.472575       0.484027     0.484027              0.499551             0.499551
8         27.300794  128.492542   56733.848767          0.000000       0.577350     0.577350              0.577350             0.577350
```

Abbildung 12-39: PCA K-means 9 Cluster 3 Dimensionen Output 2

	Age								OrderCount							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	94855.0	23.133151	23.197879	0.0	0.0	25.0	43.00	100.0	94855.0	1.023689	0.689295	0.0	1.0	1.0	1.0	6.0
1	15889.0	27.930014	24.057109	0.0	0.0	32.0	49.00	93.0	15889.0	1.973441	1.062759	0.0	1.0	2.0	2.0	8.0
2	35832.0	41.636861	16.777360	0.0	32.0	41.0	54.00	95.0	35832.0	0.633038	0.911902	0.0	0.0	0.0	1.0	6.0
3	28145.0	38.402096	15.865580	0.0	29.0	37.0	50.00	91.0	28145.0	0.699485	0.945131	0.0	0.0	0.0	1.0	5.0
4	1.0	40.000000	NaN	40.0	40.0	40.0	40.00	40.0	1.0	152.000000	NaN	152.0	152.0	152.0	152.0	152.0
5	284.0	37.971831	17.467362	0.0	31.0	40.0	50.25	85.0	284.0	18.024648	10.754377	1.0	12.0	16.0	21.0	109.0
6	70106.0	17.234359	23.600904	0.0	0.0	0.0	38.00	100.0	70106.0	0.822027	0.469548	0.0	1.0	1.0	1.0	2.0
7	4874.0	36.921830	17.967874	0.0	29.0	38.0	50.00	100.0	4874.0	5.408699	2.586801	0.0	4.0	5.0	7.0	16.0
8	3.0	31.333333	27.300794	0.0	22.0	44.0	47.00	50.0	3.0	109.333333	128.492542	23.0	35.5	48.0	152.5	257.0

Abbildung 12-40: PCA K-means 9 Cluster 3 Dimensionen Output 3

TotalOrderSum								ReservationCount							
count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
94855.0	233.876412	310.264835	0.00	33.940	119.000	311.265	2831.07	94855.0	0.008360	0.111836	0.0	0.0	0.0	0.0	6.0
15889.0	900.047110	568.876018	0.00	489.580	835.910	1148.790	4204.72	15889.0	0.026182	0.231722	0.0	0.0	0.0	0.0	7.0
35832.0	171.365715	334.561251	0.00	0.000	0.000	199.000	2873.95	35832.0	0.020652	0.175589	0.0	0.0	0.0	0.0	7.0
28145.0	168.499806	311.274874	0.00	0.000	0.000	206.250	2560.00	28145.0	0.018511	0.164043	0.0	0.0	0.0	0.0	8.0
1.0	210416.760000	NaN	210416.76	210416.760	210416.760	210416.760	210416.76	1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0
284.0	5303.054718	4774.393170	323.09	2484.185	4025.520	6559.055	37338.31	284.0	0.158451	0.537859	0.0	0.0	0.0	0.0	4.0
70106.0	160.667004	180.706270	0.00	16.940	99.000	243.750	847.14	70106.0	0.010213	0.111700	0.0	0.0	0.0	0.0	3.0
4874.0	1853.514633	1126.843090	0.00	1011.575	1639.205	2454.615	8121.41	4874.0	0.079401	0.472575	0.0	0.0	0.0	0.0	10.0
3.0	68513.456667	56733.848767	9532.66	41422.420	73312.180	98003.855	122695.53	3.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0

Abbildung 12-41: PCA K-means 9 Cluster 3 Dimensionen Output 4

```
Compute PCA kmeans clustering...
Elapsed time: 2.04s
Cluster means
                Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0         38.466872    0.716379     175.238743          0.000000       1.000000     0.000000              0.000000             1.000000
1         23.166127    1.019053     231.620399          0.000000       1.000000     0.000000              1.000000             0.000000
2         35.528254    4.814104    1825.251854          0.000000       0.465292     0.534708              0.675901             0.324099
3         19.015538    0.994214     271.462204          0.000000       0.000000     1.000000              1.000000             0.000000
4         41.744859    0.624143     169.092237          0.000000       0.000000     1.000000              0.000000             1.000000
5         23.788603    1.185294     268.459875          1.126103       0.430882     0.569118              0.586765             0.413235
6         44.666667  152.333333  114214.983333          0.000000       0.000000     1.000000              0.666667             0.333333
7         25.352518    2.215827     499.373453          3.748201       0.417266     0.582734              0.625899             0.374101
8         38.010239   17.699659    5445.345119          0.116041       0.464164     0.535836              0.450512             0.549488
##################################################
Cluster min
           Age  OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0          0.0           0           0.00                 0              1            0                     0                    1
1          0.0           0           0.00                 0              1            0                     1                    0
2          0.0           1           6.32                 0              0            0                     0                    0
3          0.0           0           0.00                 0              0            1                     1                    0
4          0.0           0           0.00                 0              0            1                     0                    1
5          0.0           0           0.00                 1              0            0                     0                    0
6         40.0          48        9532.66                 0              0            1                     0                    0
7          0.0           0           0.00                 3              0            0                     0                    0
8          0.0           1         323.09                 0              0            0                     0                    0
```

Abbildung 12-42: PCA K-means 9 Cluster 4 Dimensionen Output

Cluster max

clusters	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True
0	91.0	6	2827.30	0	1	0	0	1
1	100.0	6	2608.20	0	1	0	1	0
2	100.0	16	8121.41	0	1	1	1	1
3	100.0	6	2671.91	0	0	1	1	0
4	95.0	6	2873.95	0	0	1	0	1
5	91.0	14	5041.98	2	1	1	1	1
6	50.0	257	210416.76	0	0	1	1	1
7	76.0	53	9499.18	10	1	1	1	1
8	85.0	109	73312.18	3	1	1	1	1

##

Cluster std

clusters	Age	OrderCount	TotalOrderSum	ReservationCount	Gender_FEMALE	Gender_MALE	BonusCardOwner_False	BonusCardOwner_True
0	15.810112	0.978632	326.828823	0.000000	0.000000	0.000000	0.000000	0.000000
1	23.204699	0.672806	303.986042	0.000000	0.000000	0.000000	0.000000	0.000000
2	19.605240	2.504136	1055.470304	0.000000	0.498833	0.498833	0.468074	0.468074
3	24.020910	0.665357	347.790016	0.000000	0.000000	0.000000	0.000000	0.000000
4	16.718478	0.898309	330.771072	0.000000	0.000000	0.000000	0.000000	0.000000
5	22.655141	1.820828	487.010612	0.332026	0.495291	0.495291	0.492505	0.492505
6	5.033223	104.500399	100710.203863	0.000000	0.000000	0.000000	0.577350	0.577350
7	22.071315	5.363271	1116.919753	1.435035	0.494891	0.494891	0.485640	0.485640
8	17.678926	10.453021	6173.766584	0.388422	0.499567	0.499567	0.498396	0.498396

Abbildung 12-43: PCA K-means 9 Cluster 4 Dimensionen Output 2

	Age								OrderCount							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	27907.0	38.466872	15.810112	0.0	29.0	37.0	50.0	91.0	27907.0	0.716379	0.978632	0.0	0.0	0.0	1.0	6.0
1	93946.0	23.166127	23.204699	0.0	0.0	25.0	43.0	100.0	93946.0	1.019053	0.672806	0.0	1.0	1.0	1.0	6.0
2	6353.0	35.528254	19.605240	0.0	28.0	38.0	50.0	100.0	6353.0	4.814104	2.504136	1.0	3.0	5.0	6.0	16.0
3	83471.0	19.015538	24.020910	0.0	0.0	0.0	41.0	100.0	83471.0	0.994214	0.665357	0.0	1.0	1.0	1.0	6.0
4	35157.0	41.744859	16.718478	0.0	32.0	41.0	54.0	95.0	35157.0	0.624143	0.898309	0.0	0.0	0.0	1.0	6.0
5	2720.0	23.788603	22.655141	0.0	0.0	27.0	42.0	91.0	2720.0	1.185294	1.820828	0.0	0.0	1.0	2.0	14.0
6	3.0	44.666667	5.033223	40.0	42.0	44.0	47.0	50.0	3.0	152.333333	104.500399	48.0	100.0	152.0	204.5	257.0
7	139.0	25.352518	22.071315	0.0	0.0	28.0	44.0	76.0	139.0	2.215827	5.363271	0.0	0.0	1.0	2.0	53.0
8	293.0	38.010239	17.678926	0.0	31.0	40.0	51.0	85.0	293.0	17.699659	10.453021	1.0	12.0	16.0	21.0	109.0

Abbildung 12-44: PCA K-means 9 Cluster 4 Dimensionen Output 3

TotalOrderSum								ReservationCount							
count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
27907.0	175.238743	326.828823	0.00	0.000	0.00	212.515	2827.30	27907.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
93946.0	231.620399	303.986042	0.00	33.940	119.00	308.950	2608.20	93946.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
6353.0	1825.251854	1055.470304	6.32	1036.150	1671.21	2401.650	8121.41	6353.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
83471.0	271.462204	347.790016	0.00	29.900	140.95	379.000	2671.91	83471.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
35157.0	169.092237	330.771072	0.00	0.000	0.00	198.450	2873.95	35157.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
2720.0	268.459875	487.010612	0.00	0.000	32.14	338.130	5041.98	2720.0	1.126103	0.332026	1.0	1.0	1.0	1.0	2.0
3.0	114214.983333	100710.203863	9532.66	66114.095	122695.53	166556.145	210416.76	3.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
139.0	499.373453	1116.919753	0.00	0.000	109.83	443.755	9499.18	139.0	3.748201	1.435035	3.0	3.0	3.0	4.0	10.0
293.0	5445.345119	6173.766584	323.09	2430.080	3934.33	6485.250	73312.18	293.0	0.116041	0.388422	0.0	0.0	0.0	0.0	3.0

Abbildung 12-45: PCA K-means 9 Cluster 4 Dimensionen Output 4

ANHANG B - 2. Anhang

Ergebnisse - Männer

```
Compute kmeans clustering...
Elapsed time: 0.82s
Cluster means
          OrderCount  TotalOrderSum  ReservationCount  Targetgroup_FALSE  Targetgroup_TRUE  BonusCardOwner_False  BonusCardOwner_True
clusters
0           0.821270     243.091304          0.000000            1.00000           0.00000              0.000000             1.000000
1           1.116439     363.090618          0.000000            0.00000           1.00000              0.999780             0.000220
2           0.736306     207.848279          0.000000            0.00000           1.00000              0.000000             1.000000
3         100.000000  166556.145000          0.000000            0.00000           1.00000              1.000000             0.000000
4           1.656009     412.059757          1.262404            0.45204           0.54796              0.345094             0.654906
5           1.162931     398.823210          0.000000            1.00000           0.00000              0.999787             0.000213
6         257.000000    9532.660000          0.000000            0.00000           1.00000              0.000000             1.000000
################################################
Cluster min
          OrderCount  TotalOrderSum  ReservationCount  Targetgroup_FALSE  Targetgroup_TRUE  BonusCardOwner_False  BonusCardOwner_True
clusters
0                  0           0.00                 0                  1                 0                     0                    1
1                  0           0.00                 0                  0                 1                     0                    0
2                  0           0.00                 0                  0                 1                     0                    1
3                 48      122695.53                 0                  0                 1                     1                    0
4                  0           0.00                 1                  0                 0                     0                    0
5                  0           0.00                 0                  1                 0                     0                    0
6                257        9532.66                 0                  0                 1                     0                    1
```

Abbildung 12-46: Männer 7 Cluster Output

Cluster max

clusters	OrderCount	TotalOrderSum	ReservationCount	Targetgroup_FALSE	Targetgroup_TRUE	BonusCardOwner_False	BonusCardOwner_True
0	31	16691.92	0	1	0	0	1
1	53	31384.84	0	0	1	1	1
2	24	26106.00	0	0	1	0	1
3	152	210416.76	0	0	1	1	0
4	53	9499.18	10	1	1	1	1
5	43	9277.90	0	1	0	1	1
6	257	9532.66	0	0	1	0	1

##

Cluster std

clusters	OrderCount	TotalOrderSum	ReservationCount	Targetgroup_FALSE	Targetgroup_TRUE	BonusCardOwner_False	BonusCardOwner_True
0	1.518172	550.047345	0.000000	0.000000	0.000000	0.000000	0.000000
1	1.360997	643.738827	0.000000	0.000000	0.000000	0.014829	0.014829
2	1.386476	541.705986	0.000000	0.000000	0.000000	0.000000	0.000000
3	73.539105	62028.276587	0.000000	0.000000	0.000000	0.000000	0.000000
4	3.238952	833.689158	0.754346	0.497969	0.497969	0.475661	0.475661
5	1.365811	577.551177	0.000000	0.000000	0.000000	0.014609	0.014609
6	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Abbildung 12-47: Männer 7 Cluster Output 2

clusters	OrderCount							
	count	mean	std	min	25%	50%	75%	max
0	14452.0	0.821270	1.518172	0.0	0.0	0.0	1.0	31.0
1	22733.0	1.116439	1.360997	0.0	1.0	1.0	1.0	53.0
2	19644.0	0.736306	1.386476	0.0	0.0	0.0	1.0	24.0
3	2.0	100.000000	73.539105	48.0	74.0	100.0	126.0	152.0
4	907.0	1.656009	3.238952	0.0	0.0	1.0	2.0	53.0
5	14055.0	1.162931	1.365811	0.0	1.0	1.0	1.0	43.0
6	1.0	257.000000	NaN	257.0	257.0	257.0	257.0	257.0

Abbildung 12-48: Männer 7 Cluster Output 3

TotalOrderSum							
count	mean	std	min	25%	50%	75%	max
14452.0	243.091304	550.047345	0.00	0.0000	0.000	258.8525	16691.92
22733.0	363.090618	643.738827	0.00	33.8500	161.860	446.7300	31384.84
19644.0	207.848279	541.705986	0.00	0.0000	0.000	192.9500	26106.00
2.0	166556.145000	62028.276587	122695.53	144625.8375	166556.145	188486.4525	210416.76
907.0	412.059757	833.689158	0.00	0.0000	93.850	441.8250	9499.18
14055.0	398.823210	577.551177	0.00	30.1100	199.000	538.9500	9277.90
1.0	9532.660000	NaN	9532.66	9532.6600	9532.660	9532.6600	9532.66

Abbildung 12-49: Männer 7 Cluster Output 4

ReservationCount							
count	mean	std	min	25%	50%	75%	max
14452.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
22733.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
19644.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
2.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
907.0	1.262404	0.754346	1.0	1.0	1.0	1.0	10.0
14055.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
1.0	0.000000	NaN	0.0	0.0	0.0	0.0	0.0

Abbildung 12-50: Männer 7 Cluster Output 5

Ergebnisse - Frauen

```
Compute kmeans clustering...
Elapsed time: 0.96s
Cluster means
          OrderCount  TotalOrderSum  ReservationCount  Targetgroup_FALSE  Targetgroup_TRUE  BonusCardOwner_False  BonusCardOwner_True
clusters
0           1.077540     254.115472          0.000000           0.000000          1.000000              1.000000             0.000000
1           5.337473    1896.202793          0.003194           0.679560          0.320440              0.599006             0.400994
2           2.051451     394.031675          1.282322           0.531662          0.468338              0.403694             0.596306
3           1.062843     244.960998          0.000000           1.000000          0.000000              1.000000             0.000000
4           0.705969     157.343379          0.000000           1.000000          0.000000              0.000000             1.000000
5           0.676008     158.732355          0.000000           0.000000          1.000000              0.000000             1.000000
##################################################
Cluster min
          OrderCount  TotalOrderSum  ReservationCount  Targetgroup_FALSE  Targetgroup_TRUE  BonusCardOwner_False  BonusCardOwner_True
clusters
0                  0           0.00                 0                  0                 1                     1                    0
1                  1          94.69                 0                  0                 0                     0                    0
2                  0           0.00                 1                  0                 0                     0                    0
3                  0           0.00                 0                  1                 0                     1                    0
4                  0           0.00                 0                  1                 0                     0                    1
5                  0           0.00                 0                  0                 1                     0                    1
##################################################
Cluster max
          OrderCount  TotalOrderSum  ReservationCount  Targetgroup_FALSE  Targetgroup_TRUE  BonusCardOwner_False  BonusCardOwner_True
clusters
0                  6        2072.10                 0                  0                 1                     1                    0
1                 50       16025.13                 1                  1                 1                     1                    1
2                 47        7459.66                10                  1                 1                     1                    1
3                  6        1857.66                 0                  1                 0                     1                    0
4                  6        1857.35                 0                  1                 0                     0                    1
5                  7        2089.89                 0                  0                 1                     0                    1
##################################################
Cluster std
          OrderCount  TotalOrderSum  ReservationCount  Targetgroup_FALSE  Targetgroup_TRUE  BonusCardOwner_False  BonusCardOwner_True
clusters
0           0.834352     311.079074          0.000000           0.000000          0.000000              0.000000             0.000000
1           4.078484    1125.051298          0.056433           0.466729          0.466729              0.490187             0.490187
2           3.179150     635.127290          0.763762           0.499326          0.499326              0.490961             0.490961
3           0.767507     297.078008          0.000000           0.000000          0.000000              0.000000             0.000000
4           0.984462     281.778935          0.000000           0.000000          0.000000              0.000000             0.000000
5           0.987514     294.154976          0.000000           0.000000          0.000000              0.000000             0.000000
```

Abbildung 12-51: Frauen 6 Cluster Output

	OrderCount								TotalOrderSum							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	18326.0	1.077540	0.834352	0.0	1.0	1.0	1.0	6.0	18326.0	254.115472	311.079074	0.00	37.9225	133.870	363.9325	2072.10
1	2818.0	5.337473	4.078484	1.0	3.0	5.0	7.0	50.0	2818.0	1896.202793	1125.051298	94.69	1189.3025	1731.555	2283.4275	16025.13
2	758.0	2.051451	3.179150	0.0	0.0	1.0	3.0	47.0	758.0	394.031675	635.127290	0.00	0.0000	167.925	543.1950	7459.66
3	33194.0	1.062843	0.767507	0.0	1.0	1.0	1.0	6.0	33194.0	244.960998	297.078008	0.00	43.9300	132.950	338.3225	1857.66
4	15012.0	0.705969	0.984462	0.0	0.0	0.0	1.0	6.0	15012.0	157.343379	281.778935	0.00	0.0000	0.000	198.8075	1857.35
5	10991.0	0.676008	0.987514	0.0	0.0	0.0	1.0	7.0	10991.0	158.732355	294.154976	0.00	0.0000	0.000	199.0000	2089.89

Abbildung 12-52: Frauen 6 Cluster Output 2

ReservationCount							
count	mean	std	min	25%	50%	75%	max
18326.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
2818.0	0.003194	0.056433	0.0	0.0	0.0	0.0	1.0
758.0	1.282322	0.763762	1.0	1.0	1.0	1.0	10.0
33194.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
15012.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
10991.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0

Abbildung 12-53: Frauen 6 Cluster Output 3

Ergebnisse - Kunden ohne Altersangabe

```
Cluster means
          OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0           0.957938     193.942747          0.000000       1.000000     0.000000              1.000000             0.000000
1           0.966004     223.288685          0.000000       0.000000     1.000000              1.000000             0.000000
2           1.100969     309.686511          0.029782       0.439952     0.560048              0.000000             1.000000
3           3.334331    1605.243413          0.001331       0.281437     0.718563              0.992016             0.007984
4           0.613260     149.043554          1.256906       0.393186     0.606814              0.978821             0.021179
##################################################
Cluster min
          OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0                  0            0.0                 0              1            0                     1                    0
1                  0            0.0                 0              0            1                     1                    0
2                  0            0.0                 0              0            0                     0                    1
3                  1            0.0                 0              0            0                     0                    0
4                  0            0.0                 1              0            0                     0                    0
##################################################
Cluster max
          OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0                  4        1847.00                 0              1            0                     1                    0
1                  3        1571.80                 0              0            1                     1                    0
2                 12        4782.95                 1              1            1                     0                    1
3                109       73312.18                 1              1            1                     1                    1
4                  8        3855.53                 9              1            1                     1                    1
##################################################
Cluster std
          OrderCount  TotalOrderSum  ReservationCount  Gender_FEMALE  Gender_MALE  BonusCardOwner_False  BonusCardOwner_True
clusters
0           0.516345     253.338675          0.000000       0.000000     0.000000              0.000000             0.000000
1           0.507453     270.209574          0.000000       0.000000     0.000000              0.000000             0.000000
2           1.214558     484.955762          0.170006       0.496441     0.496441              0.000000             0.000000
3           3.327143    1974.039917          0.036460       0.449775     0.449775              0.089011             0.089011
4           1.055846     345.161052          0.701101       0.488683     0.488683              0.144046             0.144046
```

Abbildung 12-54: Kunden ohne Alter 5 Cluster Output

	OrderCount								TotalOrderSum							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
clusters																
0	41986.0	0.957938	0.516345	0.0	1.0	1.0	1.0	4.0	41986.0	193.942747	253.338675	0.0	26.5675	102.105	256.2800	1847.00
1	46888.0	0.966004	0.507453	0.0	1.0	1.0	1.0	3.0	46888.0	223.288685	270.209574	0.0	29.8000	123.330	319.6000	1571.80
2	4130.0	1.100969	1.214558	0.0	0.0	1.0	1.0	12.0	4130.0	309.686511	484.955762	0.0	0.0000	117.415	403.5500	4782.95
3	3006.0	3.334331	3.327143	1.0	2.0	3.0	4.0	109.0	3006.0	1605.243413	1974.039917	0.0	836.9025	1370.895	1987.5075	73312.18
4	1086.0	0.613260	1.055846	0.0	0.0	0.0	1.0	8.0	1086.0	149.043554	345.161052	0.0	0.0000	0.000	138.9375	3855.53

Abbildung 12-55: Kunden ohne Alter 5 Cluster Output 2

	ReservationCount							
	count	mean	std	min	25%	50%	75%	max
	41986.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
	46888.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
	4130.0	0.029782	0.170006	0.0	0.0	0.0	0.0	1.0
	3006.0	0.001331	0.036460	0.0	0.0	0.0	0.0	1.0
	1086.0	1.256906	0.701101	1.0	1.0	1.0	1.0	9.0

Abbildung 12-56: Kunden ohne Alter 5 Cluster Output 3

Clusterzuteilung der Personas

Birgit

	Age	OrderCou	TotalOrde	Reservati	Gender_F	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
230226	24.0	1	103.85	0	1	0	1	0	1	5	230227

Sarah

	Age	OrderCou	TotalOrde	Reservati	Gender_F	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
13202	45.0	1	83.91	0	1	1	0	1	0	3	13203

Nicole

	Age	OrderCou	TotalOrde	Reservati	Gender_F	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
16733	52.0	0	0.0	0	1	1	0	1	0	3	16734

Sandra

	Age	OrderCou	TotalOrde	Reservati	Gender_F	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
20497	22.0	2	388.0	0	1	1	0	0	1	0	20498

Beate

	Age	OrderCou	TotalOrde	Reservati	Gender_F	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
219201	46.0	0	0.0	3	1	0	1	1	0	2	219202

Gabi

	Age	OrderCou	TotalOrde	Reservati	Gender_F	Gender_M	BonusCar	BonusCar	clusters	Id
189562	0.0	1	83.96	0	1	0	1	0	0	189563

Martina

	Age	OrderCou	TotalOrde	Reservati	Gender_F	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
29104	25.0	0	0.0	0	1	1	0	0	1	0	29105

Claudia

	Age	OrderCou	TotalOrde	Reservati	Gender_F	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
3954	26.0	0	0.0	0	1	0	1	0	1	5	3955

Nina

	Age	OrderCou	TotalOrde	Reservati	Gender_F	Gender_M	BonusCar	BonusCar	clusters	Id
63459	0.0	0	0.0	1	1	0	1	0	4	63460

Peter

	Age	OrderCou	TotalOrde	Reservati	Gender_M	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
12643	31.0	0	0.0	0	1	0	1	1	0	0	12644

Franz

	Age	OrderCou	TotalOrde	Reservati	Gender_M	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
128367	36.0	1	149.0	0	1	0	1	1	0	0	128368

Jürgen

	A	B	C	D	E	F	G	H	I	J	K	L
		Age	OrderCou	TotalOrde	Reservati	Gender_M	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
5806		51.0	1	128.9	0	1	1	0	0	1	1	5807

Werner

	A	B	C	D	E	F	G	H	I	J	K	L
		Age	OrderCou	TotalOrde	Reservati	Gender_M	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
197221		33.0	2	334.8	1	1	0	1	1	0	4	197222

Karl

	A	B	C	D	E	F	G	H	I	J	K
		Age	OrderCou	TotalOrde	Reservati	Gender_F	Gender_M	BonusCar	BonusCar	clusters	Id
123166		0.0	1	544.95	0	0	1	1	0	1	123167

Fabian

	A	B	C	D	E	F	G	H	I	J	K	L
		Age	OrderCou	TotalOrde	Reservati	Gender_M	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
7134		38.0	0	0.0	0	1	1	0	1	0	5	7135

Thomas

	A	B	C	D	E	F	G	H	I	J	K
		Age	OrderCou	TotalOrde	Reservati	Gender_F	Gender_M	BonusCar	BonusCar	clusters	Id
8754		0.0	2	210.2	0	0	1	0	1	2	8755

Christian

	A	B	C	D	E	F	G	H	I	J	K	L
		Age	OrderCou	TotalOrde	Reservati	Gender_M	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
78076		49.0	0	0.0	0	1	1	0	0	1	1	78077

Simon

	A	B	C	D	E	F	G	H	I	J	K	L
		Age	OrderCou	TotalOrde	Reservati	Gender_M	BonusCar	BonusCar	Targetgro	Targetgro	clusters	Id
203956		41.0	0	0.0	0	1	0	1	0	1	2	203957

ABKÜRZUNGSVERZEICHNIS

AI ..... *Artificial Intelligence*
API ..... *Application Programming Interface*
B2B ..... *Business-to-Business*
B2C ..... *Business-to-Consumer*
BI ..... *Business Intelligence*
CL ..... *Computational Linguistics*
CRAN ..... *Comprehensive R Archive Network*
CSV ..... *Comma Seperated Values*
DM ..... *Data Mining*
EC ..... *Electronic Commerce*
ETL ..... *Extraktion, Transformation, Laden*
GiB ..... *Gibibyte*
KI ..... *Künstliche Intelligenz*
KRR ..... *Knowledge Representation & Reasoning*
LDA ..... *Latent Dirichlet Allocation*
ML ..... *Machine Learning*
MSS ..... *Management Support Systeme*
NLP ..... *Natural Language Processing*
PCA ..... *Principal Component Analysis*
PIC ..... *Power Iteration Clustering*
SCP ..... *Secure Copy Protocol*
SLA ..... *Service Level Agreement*
SVM ..... *Support Vector Maschinen*
TDWI ..... *The Data Warehouse Institute*

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

LISTINGS

LITERATURVERZEICHNIS

Abeel, T., Van de Peer, Y., & Saeys, Y. (10. April 2009). *Java Machine Learning Library (Java-ML)*. Von Java-ML: A Machine Learning Library, Journal of Machine Learning Research: http://java-ml.sourceforge.net/ abgerufen

Alpaydin, E. (2014). *Introduction to Machine Learning.* The MIT Press.

Bezos, J. P. (kein Datum). Jeff Bezos Compilation of Amazon shareholder letters 1997-2016. Abgerufen am 22. Juni 2020 von https://wordsofward.files.wordpress.com/2017/04/jeff-bezos-compilation-of-amazon-shareholder-letters-1997-2016-final.pdf

Bonaccorso, G. (2018). *Machine Learning Algorithms Popular algorithms for data science and machine learning*. Packt Publishing.

Bonaccorso, G. (2018). *Mastering Machine Learning Algorithms.* Packt Publishing.

Burkov, A. (2019). *The Hundred-Page Machine Learning Book.* Andriy Burkov.

Chaudari, B., & Parikh, M. (2012, Oktober). A Comparative Study of clustering algorithms Using weka tools. *International Journal of Application or Innovation in Engineering & Management*. Retrieved from https://pdfs.semanticscholar.org/8589/47586e2e41e38ba2448d5b149c7d0f5d6b86.pdf

Ertel, W. (2016). *Grundkurs Künstliche Intelligenz: Eine praxinsorientierte Einführung 4.Auflage.* Springer Vieweg.

Fumo, D. (15. Juni 2015). *Types of Machine Learning Algorithms You Should Know*. Von Towards Data Science: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861 abgerufen

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining.* MIT Press.

Kaluza, B. (2016). *Machine Learning in Java.* Packt Publishing.

Kelleher, J. D., & Tierney, B. (2018). *Data Science.* The MIT Press.

Kemper, H.-G., Baars, H., & Mehanna, W. (2010). *Business Intelligence - Grundlagen und praktische Anwendungen Auflage 3.* Vieweg+Teubner Verlag.

Kodinariya, T. M., & Makwana, P. R. (2013). *Review on determining number of Cluster in K-Means Clustering*. International Journal of Advance Research in Computer Science and Management Studies. Abgerufen am 24. Juni 2020 von https://d1wqtxts1xzle7.cloudfront.net/34194098/V1I6-0015.pdf?1405312820=&response-content-disposition=inline%3B+filename%3DReview_on_determining_number_of_Cluster.pdf&Expires=1

593018677&Signature=PVPO7wZrsAJwkWZRsDi2RYZZBWCZB68V2vwLN~XTOd2jW4UGfPtgeQpvEWG

Loshin, D. (2003). *Business Intelligence: The Savvy Manager's Guide.* Morgan Kaufmann.

Meier, A., & Stormer, H. (2012). *eBusiness & eCommerce Management der digitalen Wertschöpfungskette.* Springer-Verlag Berlin Heidelberg.

Neukart, F. (2016). *Reverse Engineering the Mind.* (W. Springer, Hrsg.) doi:https://doi.org/10.1007/978-3-658-16176-7

Neukart, F., Hofmann, M., & Bäck, T. (6. September 2017). Artificial Intelligence and Data Science in the Automotive Industry. Abgerufen am 26. Juni 2020 von https://arxiv.org/ftp/arxiv/papers/1709/1709.01989.pdf

Omnichannel Marketing Definition und Strategien. (4. Mai 2018). Abgerufen am 20. Juni 2020 von Ionos: https://www.ionos.at/digitalguide/online-marketing/verkaufen-im-internet/omnichannel-marketing-definition-und-strategien/

Prize Optimization. (20. Oktober 2019). Von RapidMiner: https://rapidminer.com/glossary/price-optimization/ abgerufen

Przystalski, K. (20. August 2019). *Machine Learning Libraries Overview: Top 10 Libraries and Frameworks You Should Know*. Abgerufen am 20. Juni 2020 von Codete: https://codete.com/blog/machine-learning-libraries-overview-top-10-libraries-and-frameworks-you-should-know/

Reinforcement Learning Definition & Erklärung. (13. Oktober 2019). Von Datenbanken verstehen: https://www.datenbanken-verstehen.de/lexikon/reinforcement-learning/ abgerufen

Rich, E., Knight, K., & Shivashankar, N. B. (2017). *Artificial Intelligence.* Mcgraw Hill Higher Education; Third Edition.

Rouse, M. (20. Oktober 2019). *What is Customer Segmentation?* Von TechTarget: https://searchcustomerexperience.techtarget.com/definition/customer-segmentation abgerufen

Rouse, M. (20. Oktober 2019). *What is fraud detection?* Von TechTarget: https://searchsecurity.techtarget.com/definition/fraud-detection abgerufen

scikit-learn Machine Learning in Python. (n.d.). Retrieved from scikit-learn: https://scikit-learn.org/stable/

Singh, H. (14. Juni 2019). *8 Ways to Use Machine Learning for Ecommerc*. Abgerufen am 20. Juni 2020 von Omnisend: https://www.omnisend.com/blog/machine-learning/

The Apache Software Foundation. (n.d.). *Apache Spark*. Retrieved from Apache Spark: https://spark.apache.org/

Tripathi, S., Bhardwaj, A., & Poovammal, E. (2018). Approaches to Clustering in Customer Segmentation. *Approaches to Clustering in Customer Segmentation*. (I. j. technology, Hrsg.) Science Publishing Corporation. doi:10.14419/ijet.v7i3.12.16505

Turban, E., King, D., & Lang, J. (2010). *Introduction to Electronic Commerce 3rd Edition.* United States: Pearson Education (US).

Webb, G. I. (2011). *Overfitting*. (B. M. Springer, Herausgeber) doi:https://doi.org/10.1007/978-0-387-30164-8

Wörndl, W., & Schlichter, J. (19. Februar 2019). *Empfehlungssysteme*. Abgerufen am 20. Juni 2020 von Enzyklopädie der Wirtschaftsinformatik: https://enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/daten-wissen/Business-Intelligence/Analytische-Informationssysteme--Methoden-der-/empfehlungssysteme