- Ludwig-Maximilians-Universität München -
Institut für Statistik

Segmentierung von Nutzern der autoscout24.de-Plattform unter Verwendung clusteranalytischer Methoden

BACHELORARBEIT

ZUR ERLANGUNG DES AKADEMISCHEN GRADES

BACHELOR OF SCIENCE (B. Sc.)

Autor: Yvonne Barth

Matrikelnummer: XXXXXXXX

Gutachter: Prof. Dr. Christian Heumann

Abgabedatum: 15. September 2015

Eidesstattliche Erklärung

Hiermit bestätige ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, wurden unter Angabe der Quelle kenntlich gemacht.

München, 15. September 2015 .....

(Unterschrift)

Inhaltsverzeichnis

1 Einleitung

1.1 Nutzersegmentierung und Intention der Arbeit

Das Internet bietet umfangreiche Möglichkeiten verhaltensbezogene Daten aufzuzeichnen. Jede Aktion kann unmittelbar in einer Datenbank abgespeichert werden. So werden beispielsweise Informationen wie Datum, Uhrzeit, Nutzernummer, unternehmensinterne Seiteninformationen oder die dem Nutzer angezeigten URLs gesammelt. Diese gesicherten Dateneinträge bilden die Grundlage zur Analyse des Surfverhaltens verschiedener Nutzer, welche sich in verschiedene Gruppen zusammenfassen lassen. Nutzersegmentierung hat das Ziel, Personen in Gruppen mit unterschiedlichen Verhaltensweisen und Bedürfnissen einzuteilen [Kopp (2014)].
Die Hauptmerkmale verschiedener Nutzer sollen innerhalb eines Segments möglichst einheitlich und zwischen den Segmenten so verschieden wie möglich sein. Um einzelne Personen individuell erreichen zu können muss klar sein, in welches Cluster man sie einordnen kann.

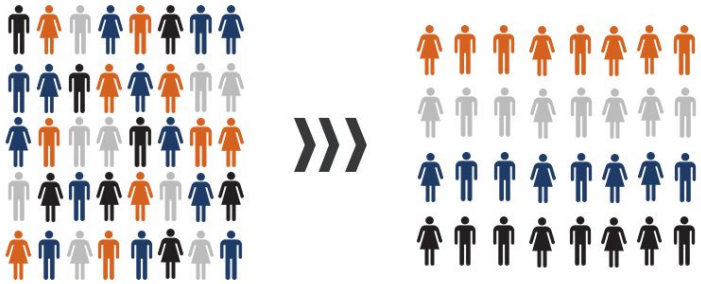


Abbildung 1: Grafische Darstellung der Nutzersegmentierung

Ziel dieser Arbeit ist es, circa 85.000 zufällig ausgewählte Nutzer, ausgehend von ihren aufgerufenen URLs der Internet-Plattform `www.autoscout24.de`, anhand zweier Clusteralgorithmen in verschiedene Nutzersegmente einzuteilen. Die Besucher wurden über einen Zeitraum von 15 Tagen beobachtet und alle von ihnen getätigten Aktionen aufgezeichnet. Die Ergebnisse beider Algorithmen werden im Anschluss verglichen, um überprüfen zu können, wie „einig“ sich beide Verfahren bezüglich der Klassifikation gleicher Nutzer sind. Auf Basis der Clustererkenntnisse soll es ermöglicht werden, neue Besucher diesen Gruppen zuzuordnen. Hierfür werden Entschei-

dungsbäume analysiert, um Kenntnisse über die Wichtigkeit einzelner Variablen für die Einteilung weiterer Nutzer zu erfahren. Zudem wird ein Random-Forest-Modell zur Klassifikation erstellt und dessen Vorhersagekraft zur Einteilung der Nutzer aufbauend auf den Ergebnissen der Clustermethoden beurteilt. Als weitergehende Maßnahme können die gewonnenen Erkenntnisse dann von AutoScout24 beispielsweise für gezielte Werbung genutzt werden.

1.2 Zu AutoScout24

Das münchner Unternehmen AutoScout24 ist einer der führenden Onlinemärkte sowohl für Neu- und Gebrauchtwagen, als auch für Nutzfahrzeuge und Motorräder. Zudem können Ersatzteile und Zubehör erworben werden.
Gegründet wurde die Tochter der Scout24-Gruppe im Jahr 1998 als MasterCarAG. Das Onlineportal ist mittlerweile in 17 europäischen Ländern vertreten und insgesamt gibt es mehr als zwei Millionen Angebote.
Seit 2011 bietet AutoScout24 auch ein Werkstattportal und Finanzierungsvorschläge auf der Webseite an. Des Weiteren kann man sich im AutoScout24 Magazin über neue Fahrzeugmodelle und Gebrauchtwagentests informieren.
Ruft man die Webseite auf, sieht man zunächst eine Box, in welche man Filterkriterien für das gewünschte Fahrzeug oder Motorrad eingeben kann. Klickt man weiterführend auf den Button zur Anzeige der entsprechenden Fortbewegungsmittel, erscheinen auf der Suchergebnisseite 20 Übersichtsfenster. Wählt man ein Fahrzeug der Suchergebnisseite aus, kann man auf der Detailseite nähere Informationen und Bilder betrachten.
Das Unternehmen finanziert sich primär über das B2B („Business-to-Business“) - Geschäftsmodell. Von den Händlern, welche auf der Plattform inserieren, wird eine Gebühr verlangt, damit diese mehr Kunden erreichen können. Außerdem können die Händler ergänzend noch ein gebührenpflichtiges Zusatzpaket erwerben, um ihre Inserate hervorzuheben. Eine zweite Einnahmequelle ist die Anzeige von Werbung.
Des Weiteren existiert zusätzlich noch das B2C („Business-to-Customer“) - Konzept. Hier haben private Inserenten, welche im Gegensatz zu Händlern kostenfrei inserieren können, ebenfalls die Möglichkeit, gegen Bezahlung ein Zusatzpaket zu erhalten, um so ihre Anzeigen auch auf einer der ersten Suchergebnisseiten zu platzieren.
Überdies besteht eine weitere Verdienstmöglichkeit durch die im Werkstattportal vertretenen Werkstätten, von welchen, ähnlich zu den Händlern, auch eine Gebühr erhoben wird. [vorliegende Informationen entstammen dem Unternehmensportrait von AutoScout24 (2015)]

2 Datengrundlage

2.1 Datenbeschaffung

Die Basisdaten zur späteren Nutzersegmentierung werden aus einem Hadoop Cluster abgefragt.
Unter Hadoop beziehungsweise dem „Hadoop ecosystem“ versteht man eine Sammlung von überwiegend in Java geschriebenen Software-Komponenten [White (2010)]. Diese Komponenten dienen der Speicherung und Verarbeitung großer Datenmengen. Die Daten werden unter Verwendung von „Hive“ aus dem Cluster exportiert. Bei Hive handelt es sich um eine Software, welche mit Hilfe einer SQL-ähnlichen Abfragesprache Map-Reduce-Programme erzeugt [Rutherglen et al. (2012)]. Hierzu greift Hive auf eigene Metadaten zu, welche im Hadoop Distributed File System (HDFS) gespeichert sind.
Die zwei wichtigsten zusammenarbeitenden Komponenten im Umfeld von Hive sind HDFS und die „Map-Reduce-Engine“, weshalb im Folgenden auf diese Komponenten eingegangen wird.

2.1.1 HDFS

Das „Hadoop Distributed File System“ (HDFS) ist nach White (2010) ein leistungsfähiger und ausfallsicherer Datenspeicher. Die Dateien werden von einem „HDFS Client“ mit Hilfe der „namenode“(Masterknoten) auf mehreren datanodes(Datenknoten) abgelegt. Hierzu werden große Dateien in einzelne Blöcke unterteilt um so eine bessere Verteilung innerhalb des Clusters sicherzustellen [White (2010)]. Zur Steigerung der Ausfallsicherheit und Leistungsfähigkeit des Clusters werden die datanodes auf „racks“, übersetzt Regale oder Gruppen, verteilt. Ist mehr als ein rack vorhanden, wird eine Kopie des Blocks auf ein anderes rack repliziert. Ansonsten wird innerhalb des racks bestmöglich auf die verschiedenen datanodes verteilt.
Folgende Grafik skizziert die eben beschriebene Architektur grundlegend:

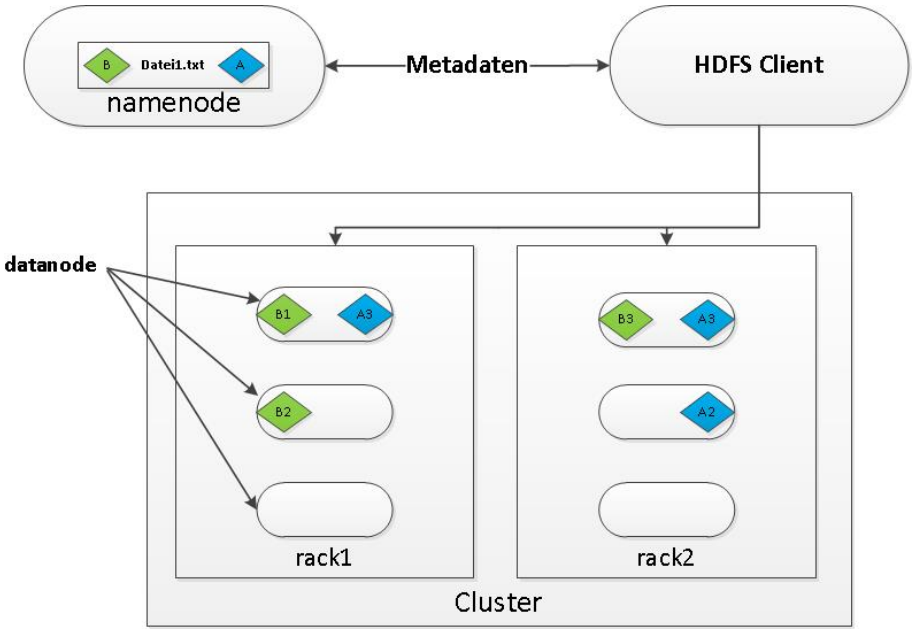


Abbildung 2: Grafische Darstellung eines Hadoop-Clusters und dessen Komponenten

Falls ein Block, ein Knoten oder gar ein ganzes rack ausfällt, stellt der HDFS Client die Datei wieder aus den Sicherungskopien anderer Blöcke, Knoten oder racks her. In dem Cluster können durch Tausende von Knoten Rechen- und Verarbeitungsprozesse mit Datenmengen im Petabytebereich parallel bearbeitet werden [White (2010)]. Wo genau der HDFS Client die Blöcke ablegen beziehungsweise abholen soll, entscheidet die namenode. Hinter jeder „node“, beziehungsweise jedem „Knoten“, steht ein Computer und jedes rack besteht aus n Computern beziehungsweise datanodes [White (2010)].

2.1.2 Map-Reduce

Das Map-Reduce-Verfahren basiert ursprünglich auf einer Idee von Google und ermöglicht parallele Arbeit. Dieses Verfahren beinhaltet zentral die Mapping- und Reduce-Funktion, welche angelehnt an van Groningen (2009) in Abbildung 3 veranschaulicht werden.

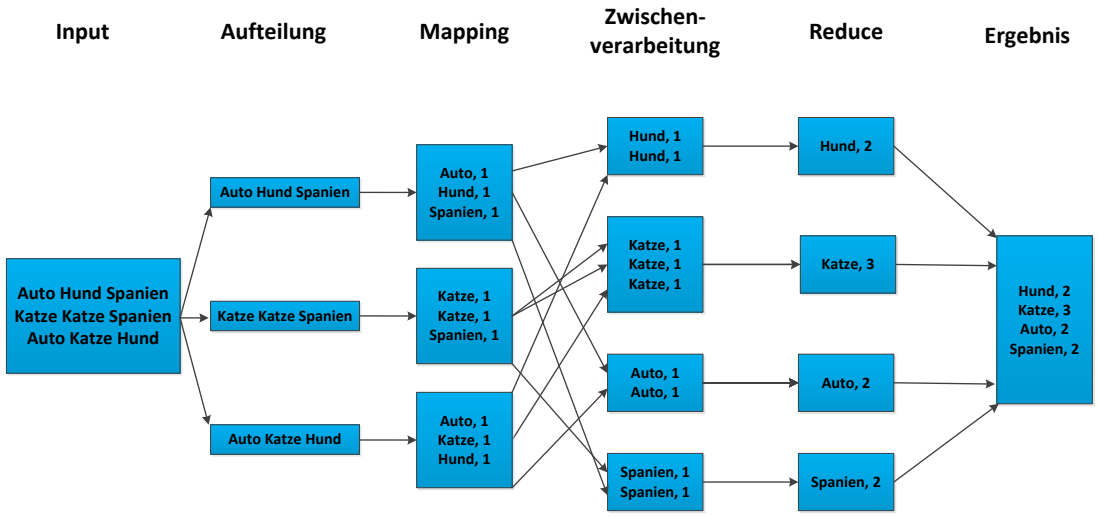


Abbildung 3: Map-Reduce Schritte [van Groningen (2009)]

Im Schritt „Aufteilung“ werden die Daten auf Blöcke verteilt.
Die Map-Funktion greift auf die Blöcke zu und erzeugt distinkte Schlüsselwert-Paare (engl.: key-value pairs).
In der Zwischenverarbeitung wird nach den einzelnen Schlüsseln, wie „Hund“ oder „Katze“ sortiert. Die Reduce-Funktion komprimiert die Daten zu einer Datei, verarbeitet die Ausgabe und gibt sie anschließend aggregiert in einer Ergebnisliste zurück. In diesem Beispiel werden die Werte summiert, jedoch kann auch beispielsweise das arithmetische Mittel, das Minimum oder Maximum berechnet werden [White (2010)].
Eine weitere Komponente der Hadoop Datenbank ist die Abfragesoftware Hive.
Hive erweitert Hadoop um eine Datenbank-Infrastruktur und ermöglicht das Abfragen von Daten im Hadoop System mit einer SQL-ähnlichen Sprache [Rutherglen et al. (2012)]. Die SQL-Abfragen werden im Hintergrund wieder in die Map-Reduce-

Funktion übersetzt, um somit die Daten zur Verfügung stellen zu können [Rutherglen et al. (2012)].
In dieser Arbeit werden die Variablen für die spätere Segmentierung unter Verwendung von Hive extrahiert. Der verwendete Befehl ist Anhang A zu entnehmen.
Die relevanten URLs, welche Detailseiten-, Suchergebnisseiten- und Filterseitenaufrufe beinhalten, werden im Vorfeld schon durch die Hive Abfrage gefiltert.
Gespeichert werden die Basisdaten mit circa 12 Millionen verschiedenen URLs von 89.821 Nutzern im csv-Format, um das spätere Einlesen mit R zu ermöglichen.

2.2 Datenaufbereitung

2.2.1 Basisdaten

Die Basisdaten liegen zunächst als Logfiles vor. Im Zusammenhang mit Datenbanken versteht man unter einem Logfile die Protokolldatei einer abgeschlossener Aktion, wie beispielsweise das Aufrufen einer Detailseite zu einem Fahrzeug. Die gewonnenen Protokolldateien enthalten anonymisierte Informationen über das Verhalten von circa 90.000 verschiedener Nutzer der Plattform `www.autoscout24.de`.
Die Zeitspanne erstreckt sich über 15 Tage, beginnend am Sonntag, 3. Mai 2015 und endend am Sonntag, 17. Mai 2015. Relevante Variablen für jede Aktion, die verschiedene Nutzer auf der Webseite tätigen, sind Datum, Uhrzeit, visitor-ID, csuriquery, csuristem und csreferer. Die visitor-ID bleibt pro IP-Adresse und Medium gleich, solange „Cookies“ gesetzt sind. „Cookies“ können ausgelesen werden, wenn im Browser des Nutzers die Skriptsprache Javascript aktiviert ist [Flanagan (2007)].
Csuriquery und csuristem sind unternehmensinterne Informationen, die protokolliert werden und dem Nutzer nicht sichtbar sind. Beispiele hierfür können Tabelle 1 entnommen werden.
Der csreferer zeichnet die URLs, welche dem Nutzer im Browserfenster angezeigt werden, auf. Bestandteil des csreferers ist die vorherige Seite, von der ein Nutzer kommt. Wenn jemand beispielsweise von der Startseite auf die Suchergebnisseite kommt, ist der csreferer `www.autoscout24.de`.
Während des Ladeprozesses einer Seite werden für csuriquery, csuristem und csreferer sekündlich Informationen weggeschrieben. Daher werden für jede Aktion, die der Nutzer tätigt, viele verschiedene Teilinformationen erfasst. Vorrangiges Ziel ist es, wiederkehrende Bestandteile relevanter Aktionen, wie das anfängliche Suchen der gewünschten Marke, herauszufiltern und als eigenständige Aktion zu erfassen.
Tabelle 1 beschreibt den Datensatz ausschnittsweise:

sdate	stime	as24visitorID	csuristem	csuriquery	csreferer
2015-05-10	16:49:01	000588f4-9758-ffe5-4cf2-b9d0-2162f44XXXXX[1]	/ArticleList /GetCounters	atype=C &make=60 &mmvmk0=1 &fregto=1985 &cy=D &ustate=N,U &sort=threetier, price&resu lts=20&pa ge=1&tabg =guidedfull	csreferer http://ww4. autoscout24 .de/fahrzeuge ?atype=C &make=60 &mmvmk0=60 &mmvco=1 &fregto=1985 &cy=D&ust ate=N,U &sort=price &results=20 &page=1 &event=sort &dtr=s
2015-05-10	16:49:10	000588f4-9758-ffe5-4cf2-b9d0-2162f44XXXXX	/classified /268648722	asrc=st|as &testvariant= list3tiers &tierlayer=st	http://ww4. autoscout24.de /fahrzeuge? atype=C &mmvmk0=21 &mmvmd0=185 45&mmvco=1 &make=21 &model=18545 &fuel=D &fregfrom=200 0&pricefrom=10 00&priceto=2500 &cy=D&ust ate=N,U &fromhome=1 &intcidm=HP Searchmask Button&dtr=s
2015-05-10	16:52:47	000588f4-9758-ffe5-4cf2-b9d0-2162f44XXXXX	/Parkdeck /Add /268648722	-	-

Tabelle 1: Rohdatensatz mit mehreren Zeilen pro Nutzer

[1] aus Gründen des Datenschutzes anonymisiert

Der in dieser Tabelle beispielhaft ausgewählte Nutzer hat am 10. Mai 2015 zwischen 16:49 Uhr und 16:52 Uhr drei relevante Aktionen auf der Webseite ausgeführt. Diese sind zunächst ein Aufruf der Suchergebnisseite (\ArticleList\GetCounters in csuristem) mit zuvor eingestellten Filtern, welche durch das „&“-Zeichen im csreferer und csuriquery getrennt sind. Wie aus der zweiten Zeile hervorgeht wurde ein Fahrzeug mit Artikelnummer „268648722“ im Detail angesehen (\classified\268648722 in csuristem) und die letzte Aktion war das Anlegen eines Merkzettels mit betreffendem Automobil (\Parkdeck\Add\268648722 in csuristem).

2.2.2 Transformation der Logfiles

Um das Nutzungsverhalten der Besucher besser zu beschreiben, werden sogenannte „Key Performance Indicators“, übersetzt Messwerte oder Kennzahlen, ermittelt. So wird der Rohdatensatz, welcher pro Nutzer mehrere Zeilen mit unterschiedlichen URLs enthält, in einen Datensatz mit jeweils einer Zeile pro Nutzer transformiert. Diese Zeile enthält vor allem individuelle Informationen über das Surf-Verhalten.
Für alle 89.821 Besucher werden sowohl Kennzahlen zum Nutzungsverhalten auf den verschiedenen Seitentypen als auch fahrzeugspezifische Messwerte ermittelt.
Für die Berechnung aller Kennzahlen wird die open-source Software R [R Core Team (2015)] verwendet.
Angefangen mit der Messung der Intensität eines Besuchs, wird das arithmetische Mittel aller relevanten Aktionen, die der Nutzer pro Tag auf der Webseite tätigte, berechnet.
Zu den relevanten Aktionen gehören unter anderem Seiten, auf welchen gefiltert wurde, sowie Detailseitenaufrufe. Suchergebnisseiten gehören dabei zu den Filterseiten, da beispielsweise beim Wechseln von der ersten Suchergebnisseiten auf die zweite im Hintergrund der AutoScout-Seite der Filter „page“ von Eins auf Zwei gesetzt wird. Außerdem beinhalten relevante Aktionen den Versand von Emails und Anfragen sowie das Anlegen von Merkzetteln.
Auf Basis dieses Datensatzes kann die durchschnittliche Anzahl der relevanten Aktionen pro Tag für jeden Nutzer berechnet werden.
Anschließend wird jeweils der prozentuale Anteil an den eben genannten Seiten im Verhältnis zu allen Aktionen ermittelt.
Des Weiteren wird berechnet, wieviele Filter durchschnittlich in jeder Initialsuche gesetzt werden, um die Phase der Kaufentscheidung eines Nutzers einordnen zu können. Wenn der Nutzer viele Filter setzt, liegt die Vermutung nahe, dass er schon eine klare Vorstellung vom gewünschten Auto hat.

Eine Initialsuche zeichnet sich dadurch aus, dass sich der Nutzer nach der Filtereingabe die erste Suchergebnisseite aller von ihm gefilterten Fahrzeuge ansieht.
Zudem wird bestimmt, ob sich ein Nutzer an mindestens einem Tag eingeloggt hat. Ein eingeloggter Besucher ist registriert und durch den Login kann man registrierte Nutzer von nicht registrierten unterscheiden. Weil der Beobachtungszeitraum über 15 Tage geht und es keine Auskunft darüber gibt, wie oft der Besucher zuvor schon auf der Webseite war, wird zusätzlich ein Score ermittelt, welcher wiedergibt, wie aktiv der Nutzer in den 15 Tagen war.
Der Score berechnet sich aus:

$$\frac{(Anzahl\ aktiver\ Tage)^2}{Zeitspanne}, \text{ Wertebereich } [\frac{4}{15}, 15]$$

Die Anzahl aktiver Tage wird quadriert, um zu vermeiden, dass ein User, der an zwei Tagen in Folge online war, in Bezug auf die Aktivität gleichgestellt wird mit einem Besucher, der an 15 Tagen jeden Tag auf der Webseite war. Ein kleiner Wert des Aktiv-Scores gibt an, dass ein Nutzer zwischen dem ersten und letzten Tag, an dem er die Webseite besucht hat, verhältnismäßig wenige Tage aktiv war. Ein hoher Score bezeichnet einen aktiveren Nutzer im Verhältnis zu seinem Zeitfenster. Wenn er zum Beispiel nur am 5. Mai und am 9.Mai online war, wäre die Zeitspanne 5 Tage (5., 6., 7., 8. und 9. Mai) und die Anzahl aktiver Tage ergibt sich aus 5.Mai + 9. Mai = 2 Tage.
Der Score ist somit $\frac{2^2}{5} = 0{,}8$.
Ein Nutzer, der in der gleichen Zeitspanne an vier Tagen aktiv war, erhält den Score $\frac{4^2}{5} = 3{,}2$. Demnach besitzt der zweite Nutzer im Vergleich zum ersten einen höheren Wert, da er im selben Zeitraum häufiger auf der Webseite war. Die auf dem Score basierenden Daten entstammen der Spalte „Datum“. Für alle Nutzer wurde ein dataframe folgender Form erstellt:

as24visitorID	03.05.	04.05.	05.05.	06.05.	07.05.	08.05.	09.05.	10.05.
00005274-0290-4843-b7e1-e5d6209XXXXX	0	0	1	1	1	0	0	1

11.05.	12.05.	13.05.	14.05.	15.05.	16.05	17.05	aktive_Tage	Zeitspanne	Score
1	1	1	1	0	0	0	8	10	6,4

Tabelle 2: Berechnung des Maßes für aktive Tage

Ist das betreffende Datum mit einer Eins vermerkt, war der Nutzer an dem Tag

online. Eine Null bedeutet, dass er die Plattform an dem Tag nicht besucht hat.
Die aktiven Tage ergeben sich aus der Summe der Einträge, die Zeitspanne ist die Differenz aus dem letzten und ersten aktiven Tag plus Eins. Der Score berechnet sich aus der oben genannten Formel.
Die letzte Kennzahl gibt Auskunft darüber, ob ein Nutzer mindestens ein Fahrzeug inseriert hat. Trifft dies zu, erhält der betreffende Nutzer eine Eins für dieses Variable, falls nicht, wird eine Null notiert.

Als nächstes werden fahrzeugspezifische Kennzahlen ermittelt.
Hier wird zuerst berechnet, wie hoch der Durchschnittspreis aller Fahrzeuge ist, die sich der Nutzer im Detail angesehen hat.
Um die Preisbereitschaft herauszufinden, wird sowohl das absolute Maximum aller Minimumpreise, als auch das absolute Minimum aller Maximumpreise jedes Nutzers notiert.
Zusätzlich wird noch die Information aufgenommen, ob ein Besucher spezifisch nach Fahrzeugen von Händlern oder privaten Anbietern gesucht hat, um die Präferenzen verschiedener Nutzer zu messen.
Für die beiden Verkaufstypen werden zwei Spalten mit Dummyvariablen erstellt. Hat ein Nutzer während seinen Suchen mindestens einmal den Filter „Fahrzeuge von privaten Anbietern“ gesetzt, dann erhält er in der Spalte „Privat“ eine Eins. Hat er den Filter „Händlerfahrzeuge“ nie ausgewählt, so steht in der entsprechenden Spalte „Händler“ eine Null.
Des Weiteren werden noch die zehn am häufigsten gesuchten Marken über alle Nutzer und eine zusätzliche Spalte mit den Restmarken betrachtet, um Markenpräferenzen als Information mit aufzunehmen. Die Variablenanzahl der verschiedenen Marken wird bewusst auf die Top 10 reduziert, da sonst der Großteil der Segmentierungsvariablen aus Marken besteht und die anderen Einflussvariablen wie Durchschnittspreis oder Anzahl aller Aktionen in den Hintergrund geraten.
In jeder Zeile steht nun der jeweilige prozentuale Anteil der einzelnen Marken- beziehungsweise Restmarkensuchen. Für den Fall, dass einzelne Nutzer einmal nicht speziell nach einer Marke gefiltert haben, sondern an anderen Filterkriterien interessiert waren, wird die Spalte „keine Marke“ eingeführt. Hier wird der relative Anteil der Suchen ohne Markenfilter eingetragen.
Durch die Kennzahlen für jeden Nutzer erhält man nun einen Datensatz, der die Nutzer mit verschiedenen Merkmalen und nicht mehr durch URLs beschreibt.
Die Merkmale sind zusammengefasst Messgrößen zu den unterschiedlichen Seitentypen und zu den Eigenschaften der betrachteten Detailseiten. Außerdem enthält der

transformierte Datensatz auch Informationen darüber, wieviele und welche Filtereinstellungen verwendet werden und wie aktiv der Nutzer ist.
Tabelle 3 enthält die Kennzahlen ausschnittsweise für einen Nutzer.

as24visitorID	alle_Aktionen	rel_suchen	rel_detail	rel_mails	rel_bookmarks
00002cb7-9758-44b2-95de-47dee09XXXXX	59,75	0,75	0,25	0,00	0,00

active_score	mind_1_Inserat	Haendler	Privat	Filter_Inital	Durchschnittspreis
1,77	0	1	0	2,54	7.663,16

Minumumpreis	Maximumpreis	keine_Marke	VW	Mercedes	BMW	Audi
1.000	12.000	0,30	0,00	0,00	0,24	0,08

Ford	Opel	Skoda	Toyota	Renault	Peugeot	uebrige_Marken
0,05	0,00	0,00	0,30	0,00	0,00	0,03

Tabelle 3: Transformierter Datensatz mit einer Zeile pro Nutzer

Diese Person hat durchschnittlich 59,75 Aktionen pro Tag, an dem er auf der Plattform war, getätigt.
Anteilig hat dieser Nutzer zu 75% Suchergebnis- und Filterseiten aufgerufen und zu 25% Detailseiten betrachtet. Es wurden weder Emails versendet, noch Lesezeichen angelegt.
Der Aktiv-Score liegt bei 1,77, was bedeutet, dass dieser Nutzer in dem Beobachtungszeitraum nicht oft auf der Plattform tätig war.
Er hat kein Fahrzeug inseriert und mindestens einmal explizit nach Händlerfahrzeugen gesucht. Am Anfang jeder Suche wurden im Mittel 2,54 Filter gesetzt.
Im Schnitt waren alle Fahrzeuge, die er sich im Detail angesehen hat, 7.663,16€ wert.
Für Fahrzeuge, die billiger als 1.000€ sind, hat er sich nicht interessiert und seine minimale Obergrenze lag bei 12.000€.
Betrachtet man alle seine Markenfilter hat er in 30% seiner Filtereinstellungen keine Marke angegeben. Zu 24% wurde die Marke BMW, zu 8% Audi, zu 5% Ford und zu 30% Toyota gesucht. 3% aller seiner Sucheinstellungen bestanden aus mindestens einer der 353 übrigen Marken.
Durch diese Kennzahlen wird der wichtigste Informationsbedarf, welcher für die Nutzersegmentierung interessant ist, abgedeckt.

3 Deskriptive Datenanalyse

Die deskriptive Analyse soll dazu dienen, Daten grafisch darzustellen und zu beschreiben [Fahrmeir et al. (2004)]. So kann ein Überblick über Strukturen und Besonderheiten gewonnen werden, was eine sehr wichtige Grundvoraussetzung für die weitere Arbeit mit den Daten ist.
Da der Datensatz bisher noch fehlende Werte enthält, bietet dieser noch keine Grundlage für eine Clusteranalyse. Fehlende Werte entstehen unter anderem für die beiden Spalten der Verkaufstypen an den Stellen, wo ein Nutzer nicht spezifisch nach Händler- oder Privatfahrzeugen gefiltert hat. Deshalb werden diese NA's durch eine Null ersetzt, da tatsächlich nicht nach diesen Fahrzeugen gesucht wurde. Wenn bei Minimum- und Maximumpreisen fehlende Werte erscheinen, liegt das daran, dass der Nutzer keine Preiseinschränkung gewählt hat. Hier werden beim Minimumpreis die NA's durch Null und beim Maximumpreis durch 100.000 (Obergrenze in den Suchfiltern auf der Webseite) ersetzt, da man davon ausgeht, dass im minimalen beziehungsweise maximalen Preisbereich gesucht wurde. Fehlende Werte beim Durchschnittspreis aller angesehenen Fahrzeuge haben den Hintergrund, dass ein Nutzer sich für kein Fahrzeug im Detail interessiert hat. Hier werden die NA's durch das arithmetische Mittel aller Durchschnittspreise imputiert [Enders (2010)].

Im Folgenden werden die Segmentierungsvariablen einzeln deskriptiv betrachtet. Alle Grafiken wurden unter Verwendung des `ggplot2`-Pakets für R [Wickham (2009)] erstellt.
Zunächst wird die Verteilung aller getätigten Aktionen betrachtet.

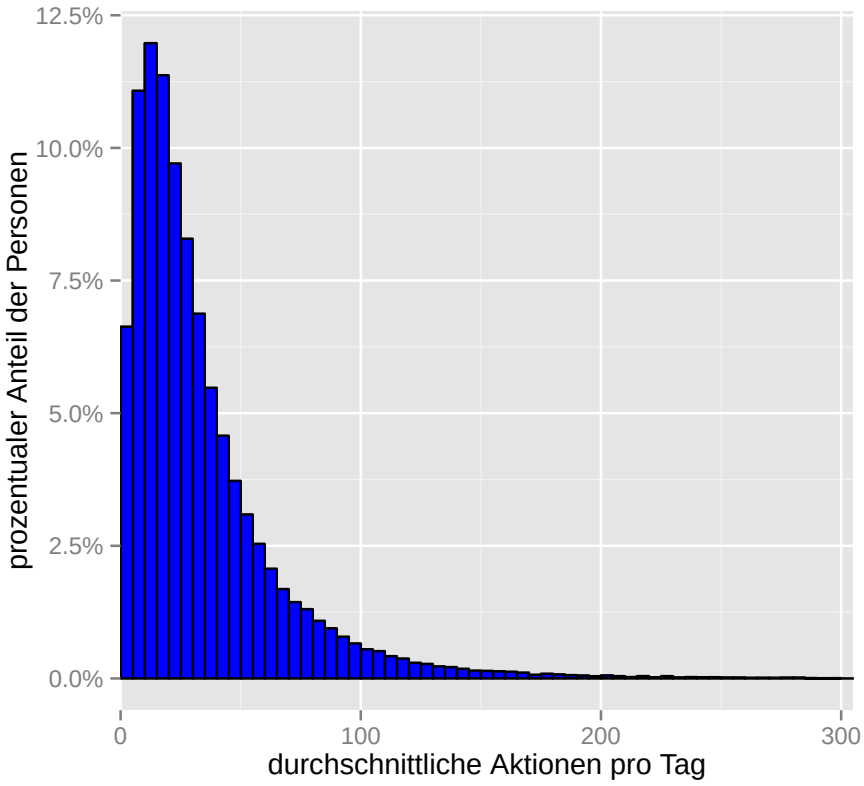


Abbildung 4: Histogramm der Variable „alle Aktionen“

Man erkennt, dass die meisten Personen durchschnittlich zwischen fünf und 25 Aktionen täglich ausführen. Der prozentuale Anteil der Personen, die mehr als 25 Aktionen tätigen, geht konstant mit der Anzahl der Aktionen zurück und ab einer durchschnittlichen Anzahl von 150 sind kaum noch Beobachtungen vertreten.
Um die 7% tätigen durchschnittlich nur weniger als fünf Aktionen pro Tag. Diese Zielgruppe kann man als „bouncer“ oder auch „Abspringer“ bezeichnen, da sie die Seite sofort wieder verlassen.
Insgesamt 103 Nutzer führen im Schnitt weit mehr als 300 Aktionen pro Tag durch, diese sind jedoch nicht in der Grafik enthalten.
Nutzer, die mehr als 300 Aktionen pro Tag tätigen, fallen vermutlich in die Gruppe „grabber“ oder „crawler“. Für diese Gruppe wird davon ausgegangen, dass automatisiert oder computergestützt gezielt Informationen der Webseite gesammelt werden.

Als nächstes wird der prozentuale Anteil der vier Seitentypen „Suchseiten“, „Detailseiten“, „Emailversand“ und „Anlegen von Merkzetteln“ analysiert.
Zunächst werden Such- und Detailseiten betrachtet.

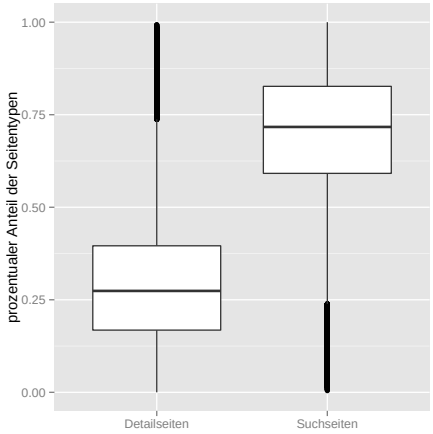


Abbildung 5: Boxplots der Variablen „Detailseiten“ und „Suchseiten“

Der Median der Detailseiten liegt bei circa 25% während sich das 50%-Quantil für Suchseiten knapp unter 75% befindet. Die Interquartilsabstände beider Boxplots sind nahezu gleich breit. Die Verteilung der Detailseiten ist linkssteil, die der Suchseiten rechtssteil. Für die Hälfte der Grundgesamtheit machen Detailseiten zwischen 15% und 40% aller Aktionen aus, wohingegen Suchseiten mit 60% bis 80% öfter aufgerufen werden. Seiten, auf welchen Emails versendet oder Merkzettel angelegt werden, machen einen wesentlich kleineren prozentualen Anteil aus, wie man Abbildung sechs und sieben entnehmen kann. Diese Seitentypen wurden anhand zweier Histogramme dargestellt, da ein Boxplot vergleichsweise keinen anschaulichen Interquartilsabstand liefert und primär Ausreißer zu erkennen sind.

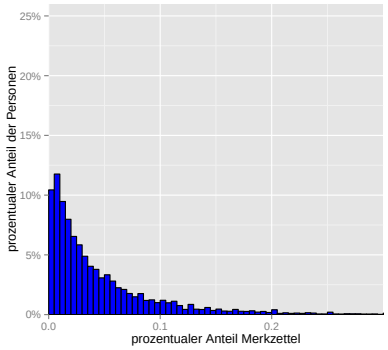


Abbildung 6: Histogramm der Variable „Merkzettel“

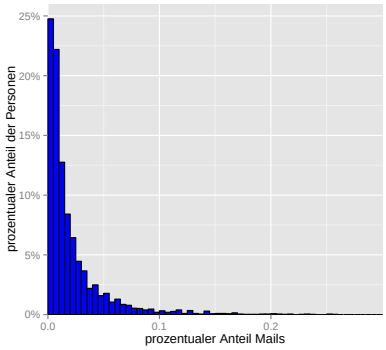


Abbildung 7: Histogramm der Variable „Mails“

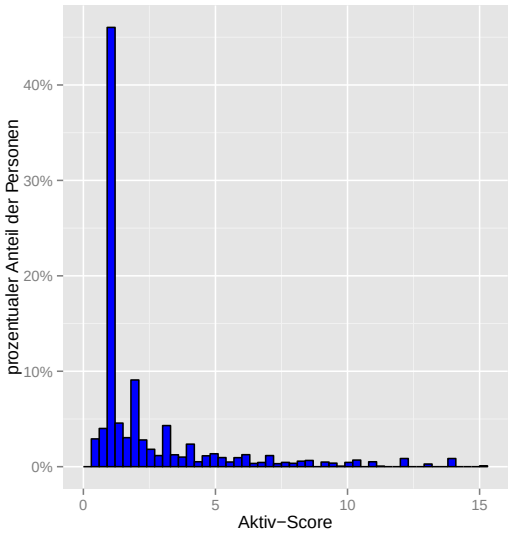


Abbildung 8: Histogramm der Variable „Aktiv-Score“

In Abbildung Acht wird die Verteilung der Aktivität betrachtet.
Es ist deutlich zu erkennen dass fast die Hälfte aller Nutzer eine Aktivitätszahl im Bereich von Eins haben.
Ein Wert zwischen $\frac{4}{15}$ und Fünf tritt bei mehr Besuchern auf, als ein Wert über Fünf.
Insgesamt besitzt die Mehrheit der Nutzer einen eher niedrigen Aktiv-Score, da die „Bouncer“ bereits einen großen Anteil ausmachen. Insgesamt lässt sich sagen, dass auf der Plattform mehr Gelegenheitsnutzer als intensive Besucher vertreten sind.

In der Grafik rechts wird die Verteilung der durchschnittlich gesetzten Initialfilter dargestellt[1]. Hierbei handelt es sich um eine bimodale Verteilung. Die erste Gruppe charakterisiert sich dadurch, wenige Filter in der Initialsuche zu setzen. Die zweite Gruppe stellt im Schnitt sehr viele Filter ein und scheint daher schon im Vergleich zur ersten Gruppe genauere Vorstellungen vom Fahrzeug zu besitzen.

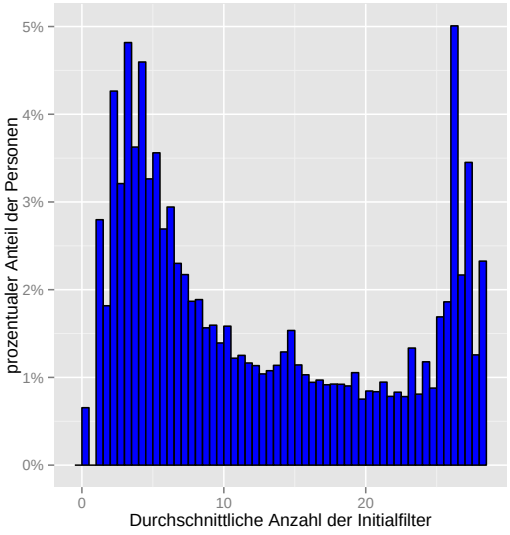


Abbildung 9: Histogramm der Variable „Initialfilter“

[1] Wählt man im Filterfenster zu Beginn keine Filter aus und sucht alle Fahrzeuge, sind automatisch drei Filter eingestellt: Ein Minimumpreis von 1.000€, Fahrzeuge des Landes Deutschland und Neu- & Gebrauchtwagen

Zu den Durchschnittspreisen aller angesehenen Fahrzeuge ist zu bemerken, dass der Wert, welcher bei der Datenaufbereitung zu Beginn durch den Mittelwert aller Preise ersetzt wurde, ausgeschlossen wird, da nur die tatsächlich angesehenen Durchschnittspreise aller Nutzer betrachtet werden sollen. Damit die Grafik anschaulich wird, werden die 4.000 Nutzer mit einem Durchschnittspreis über 50.000€ ebenfalls entfernt.

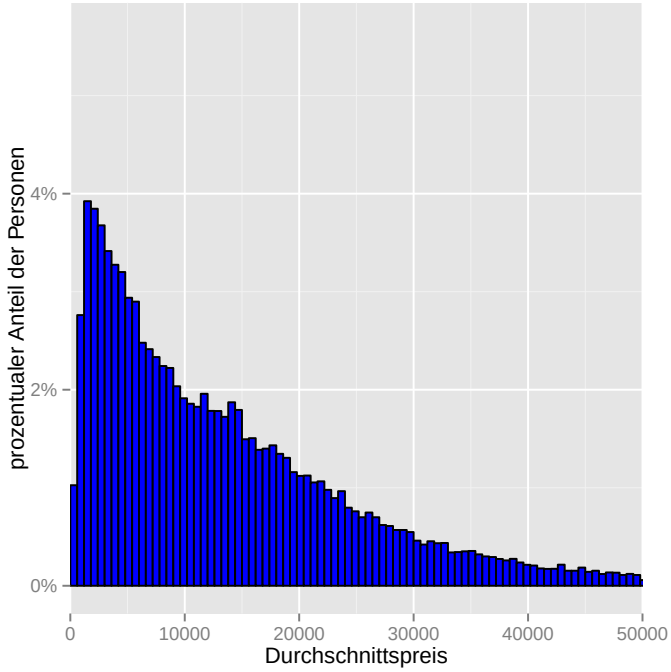


Abbildung 10: Histogramm der Variable „Durchschnittspreis“

Aus der Grafik kann man ablesen, dass um die 4% aller Nutzer günstige Fahrzeuge mit einem Durchschnittspreis zwischen 2.000€ und 4.000€ betrachtet haben.
Die Verteilung fällt mit steigendem Preis konstant, was bedeutet, dass die Nachfrage nach billigeren Fahrzeugen höher ist.

Zur weiteren Variable „Logins“ ist zu bemerken, dass sich von den insgesamt 89.821 betrachteten Nutzern 719 mindestens einmal im Laufe des Aktivitätszeitraums eingeloggt haben. Somit sind die meisten Nutzer der Webseite unregistrierte Besucher.

Ferner werden die Preisfilter betrachtet. Beim Minimumpreis ist standardmäßig 1.000€ eingestellt, um Unfallfahrzeuge und nicht fahrbereite Automobile auszuschließen. Beim Maximumpreis ist keine obere Grenze voreingestellt, weshalb die fehlenden Werte durch 100.000€ ersetzt wurden. Sowohl 1.000€ als auch 100.000€ werden für folgende Grafiken ausgeschlossen, um besser erkennen zu können, wonach explizit gefiltert wurde. Des Weiteren sind 984 Nutzer, welche einen Minimumpreis gesetzt haben, der 25.000€ überschreitet, zur besseren Anschaulichkeit nicht in der Grafik enthalten.

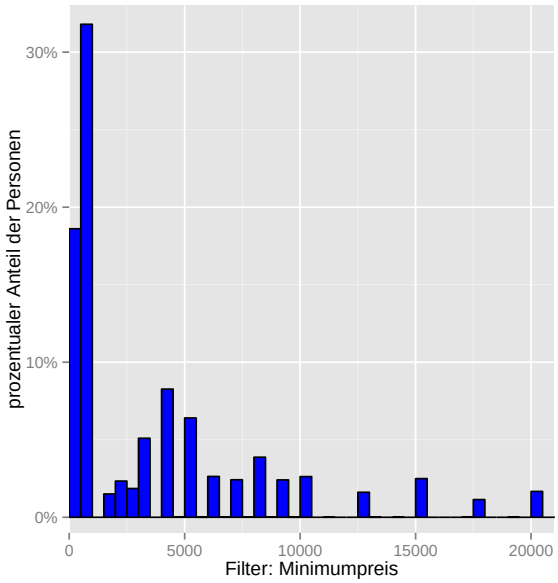


Abbildung 11: Histogramm der Variable „Minimumpreis“

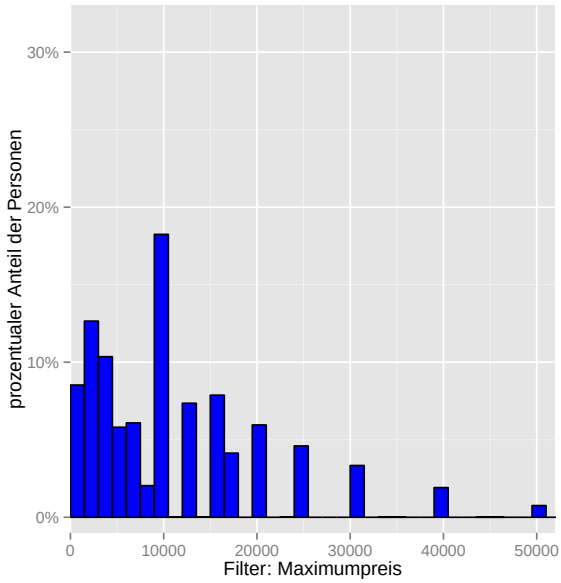


Abbildung 12: Histogramm der Variable „Maximumpreis“

Der Standard-Minimumpreis wurde von über 30% der Nutzer auf 500€ reduziert. An zweiter Stelle beschränken circa 20% aller Personen den Filter auf ein Minimum von 0€. Weniger als jeweils 5% setzen den Mindestpreis höher als 7.500€ und ein kleiner Teil der beobachteten Grundgesamtheit sucht nach Fahrzeugen ab 10.000€. Zusammenfassend liegt das 50%-Quantil des Minimumpreises bei 500€, was bedeutet, dass die Hälfte der in der Grafik betrachteten Nutzer den Minimumpreis auf einen Preis unter oder gleich 500€ setzt, während die andere Hälfte höhere Minimumpreise zwischen]500, 25.000[€ eingibt.
Als Obergrenze für den Fahrzeugpreis wurde am häufigsten von 20% aller Personen ein Wert um die 9.000€ gesetzt. An zweiter Stelle befindet sich ein kleiner Maximalpreis zwischen 2.000€ und 4.000€, welcher von circa 10% der Nutzer eingestellt wurde. Die meisten Personen filtern nach verschiedenen Maximalpreisen im Inter-

vall bis 30.000€. Insgesamt suchen die unteren 50% aller Besucher nach Fahrzeugen mit einem Preis unter 10.000€, wohingegen die oberen 50% nach Automobilen im Intervall zwischen]10.000, 100.000[€ filtern.
Nachfolgend wird der explizite Filter nach Händlern und privaten Verkäufern betrachtet.

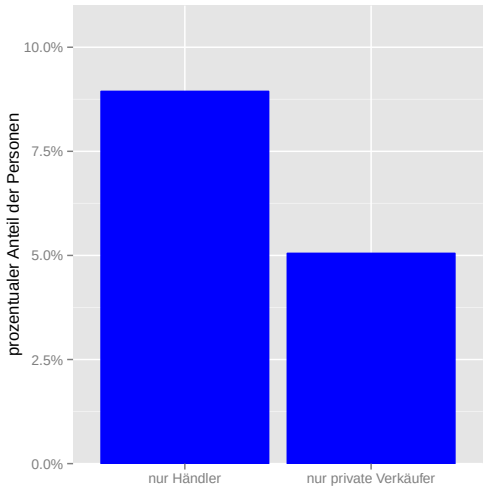


Abbildung 13: Balkendiagramm der Verkaufstypen

Circa Neun Prozent aller Nutzer haben im Laufe ihrer Suche mindestens einmal spezifisch nach Automobilen von Händlern gesucht, während im Vergleich ungefähr Fünf Prozent den Filter „private Fahrzeuge“ gesetzt haben. Insgesamt nutzen 14% aller Nutzer einen der beiden Filter, was bedeutet, dass 86% der Grundgesamtheit nicht explizit nach Verkaufstypen unterscheidet.
Anknüpfend wird die relative Suchhäufigkeit der Marken analysiert. Bei der Suche nach den Top 10 Marken, allen übrigen Marken und keiner Marke ist der prozentuale Anteil an Suchen nach den jeweiligen Marken gegeben.
In Abbildung 14 wird visualisiert, welchen Anteil die unterschiedlichen Personen für die Suche nach der Marke VW aufweisen.
Es ist deutlich zu erkennen, dass bei einer Mehrheit von 75% VW nie in den Filterkriterien nach Marken vorkommt.
Lediglich 5% der Nutzer sind markentreu und suchen ausschließlich nach VW.
Bei den übrigen 20% wird nicht vollständig nach VW gefiltert, was auf eine geringe Markenpräferenz und -treue hinweist.
Die Verteilung für die Top 2-10, als auch für keine und übrige Marken ist sehr ähnlich zu der von VW, was in Anhang B nachgesehen werden kann.

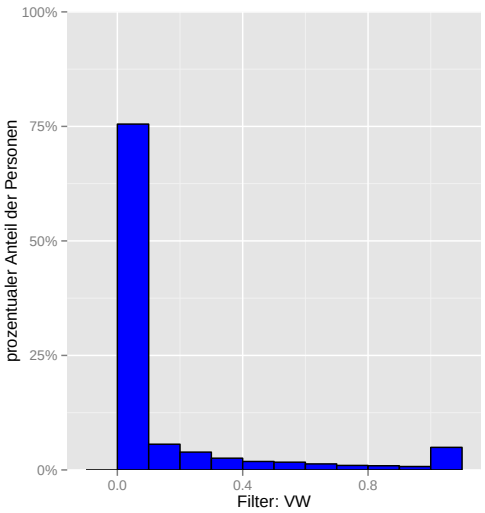


Abbildung 14: Balkendiagramm der Variable „VW“

Ergänzend zu der deskriptiven Analyse kann gesagt werden, dass 0,05% aller Nutzer mindestens ein Fahrzeug inserieren.
Da dies ein sehr kleiner prozentualer Anteil der Grundgesamtheit ist, wird diese Variable nicht mit in die Clusteranalyse aufgenommen.

Bei der Verteilung der durchschnittlichen Aktionen pro Tag, des Minimumpreises, des Durchschnittspreises und der relativen Häufigkeit von Mails und Merkzetteln treten im oberen Bereich Ausreißer auf. Da ohne Ausreißerbehandlung die Clusterschätzung verzerrt wird, werden im folgenden alle Werte der Variablen, welche größer als das jeweilige 99%-Quantil sind, entfernt. Wie Tabelle 4 zu entnehmen ist, werden beispielsweise für die Variable „Minimumpreis“ alle Beobachtungen, welche für diese Variable Werte größer als 25.000 enthalten, beseitigt.

Variable	Wert
durchschnittliche Aktionen pro Tag	168,86
Minimumpreis	25.000€
Durchschnittspreis	161.036, 24€
relativer Anteil Mails	2,90%
relativer Anteil Merkzettel	14,96%

Tabelle 4: Werte des 99%-Quantils ausgewählter Variablen

Durch die Ausreißerbehandlung fallen ungefähr 5.000 der ursprünglich circa 90.000 Nutzer weg und rund 85.000 Personen bilden somit die Basis für eine Clusteranalyse.

4 Clusteranalyse

Die Clusteranalyse ist ein multivariates statistisches Verfahren. Ziel einer Clusteranalyse beziehungsweise Datensegmentierung ist es, Objekte aufgrund vorgegebener Merkmalsvariablen in disjunkte Gruppen $\{C_1, ..., C_k\}$ zu unterteilen [Steinhausen & Langer (1977)]. Die Clusterbeobachtungen sollen innerhalb einer Gruppe möglichst homogen und untereinander so heterogen wie möglich sein. Darüberhinaus dient das Clustering der deskriptiven Analyse, also der Beantwortung der Frage, ob in den Daten überhaupt verschiedene Untergruppen, welche sich im Wesentlichen voneinander unterscheiden, zu finden sind [Handl (2010)].
Bei der Clusteranalyse handelt es sich um „unsupervised learning“ [Hastie et al. (2009)]. Hier ist die Gruppenzugehörigkeit, im Gegensatz zu „supervised learning“, einzelner Beobachtungen und die Anzahl der verschiedenen Cluster a priori nicht bekannt.
Die Qualität einer Gruppe wird durch ein Gütekriterium bestimmt. „Gesucht ist eine Partition, die hinsichtlich des Gütekriteriums optimal ist“ [Fahrmeir et al. (1996)]. Zentrale Merkmale der Güte sind nach Everitt (1993) die Varianz- und Determinantenreduktion. Die „besten“ Einteilungen variieren jedoch abhängig vom Kriterium und da die Clusterzugehörigkeit im Vorfeld nicht bekannt ist, gibt es auch kein ideales Gütekriterium [Everitt (1993)]. Der zentrale Aspekt ist die „Brauchbarkeit der erhaltenen Klassifikation für das Untersuchungsziel“ [Fahrmeir et al. (1996)]. Daher eignet sich eine Clusteranalyse zur Bestimmung der Klasseneinteilung beispielsweise im Vergleich zur Diskriminanzanalyse, wo die Klassenzugehörigkeit bekannt ist, besser.

In diesem Abschnitt werden zwei Clusteralgorithmen behandelt. Der erste Algorithmus, K-Means, wird unter Verwendung des Pakets `stats` in R auf die Daten angewendet, wobei die zweite Systematik, Two-Step-Cluster, aus dem Softwarepaket SPSS stammt. Beide Algorithmen setzen verschiedene Skalierungen voraus. Während K-Means mit dem euklidischen Distanzmaß arbeitet und deshalb nur mit skalierten Daten umgehen kann, ist der Two-Step-Cluster Algorithmus auf Daten, die sowohl metrische als auch kategorielle Variablen beinhalten, ausgelegt. Zum besseren Vergleich der Variablenunterschiede werden die Ergebnisse aus der SPSS Software anschließend wieder in R eingelesen. Der K-Means Algorithmus wird verwendet, da er aufgrund seiner Anschaulichkeit (das Verfahren besteht lediglich aus Abstandsberechnungen und Neuzuordnungen) und geringen Anzahl an Iterationsschritten zu den beliebtesten und am häufigsten angewandten Clusterverfahren gehört [Cleve

& Lämmel (2014)]. Die Wahl für den zu vergleichenden Algorithmus fiel auf den Two-Step-Cluster Algorithmus, da er laut SPSS (2006) zur „effizienten Kundensegmentierung speziell für große Datenmengen“ empfohlen wird und somit sehr gut auf den vorhandenen Fall anzuwenden ist.

4.1 Die Funktionsweise des K-Means Algorithmus

K-Means ist ein partitionierendes Clusterverfahren und gehört zu den iterativen Algorithmen. Im Gegensatz zu hierarchischen Verfahren kann sich bei dieser Methode die Klassenzugehörigkeit einzelner Beobachtungen im Iterationsverlauf ändern [Handl (2010)]. Grundlegend für diesen Algorithmus sind metrische Variablen und als Distanzmaß wird die quadrierte euklidische Distanz verwendet,welche sich nach Hastie et al. (2009) wie folgt berechnet:

$$d(x_i, x_{i'}) = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2 \tag{1}$$

Da bei K-Means zu Beginn die Anzahl der gewünschten Cluster k festgelegt werden sollen, müssen diese zunächst durch ein geeignetes Gütekriterium bestimmt werden. Abhängig von dieser Anzahl werden zu Beginn k zufällige Startpartitionen festgelegt.
Jeder Unterteilung k wird ein Mittelwert beziehungsweise Zentroid $\bar{x_k}$ zugeordnet. Alle Individuen werden iterativ dem Clusterzentroid, zu dem die kleinste Distanz besteht, zugewiesen.
Ziel ist es hier unter Verwendung der quadratischen euklidischen Distanz die Streuung zwischen den Clustern zu minimieren.
Die Streuung zwischen den Clustern ergibt sich aus folgender Formel:

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(x_i, x_{i'})$$

$$= \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} ||x_i - x_{i'}||^2$$

$$= \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \bar{x_k}||^2 \tag{2}$$

wobei $\bar{x_k} = (\bar{x_{1k}}, ..., \bar{x_{pk}})$ der Mittelwertsvektor des k-ten Cluster ist und $N_k = \sum_{i=1}^{N} I(C(i) = k)$

Somit wird im ersten Schritt versucht, die Clustervarianz zu minimieren:

$$\min_{C,m_k} \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - m_k||^2 \tag{3}$$

wobei $\{m_1, ..., m_K\}$ die Menge der Zentroide darstellt.
Die Streuung wird kleiner, je ähnlicher die Werte der Beobachtungen den Zentroiden sind. Nach diesem Vorgang werden im zweiten Schritt auf Basis der neu angeordneten Datenpunkte neue Zentroide wie folgt berechnet:

$$\bar{x_S} = \underset{m}{\operatorname{argmin}} \sum_{i \in S} ||x_i - m||^2 \tag{4}$$

wobei m der derzeitige Mittelwert des zugeteilten Cluster ist.
Anschließend werden die Beobachtungen wieder neuen Clustern C(i) unter Anwendung von (5) zugeteilt.

$$C(i) = \underset{1 \leq k \leq K}{\operatorname{argmin}} ||x_i - m_k||^2 \tag{5}$$

Diese beiden Schritte werden solange wiederholt, bis sich die Klassenzugehörigkeit der einzelnen Datenpunkte nicht mehr ändert. Durch diesen iterativen Prozess soll die Streuung zwischen den Clustern zu einem lokalen Minimum konvergieren. Folgende Grafik beschreibt den K-Means Iterationsprozess für simulierte Daten und 20 Wiederholungen.

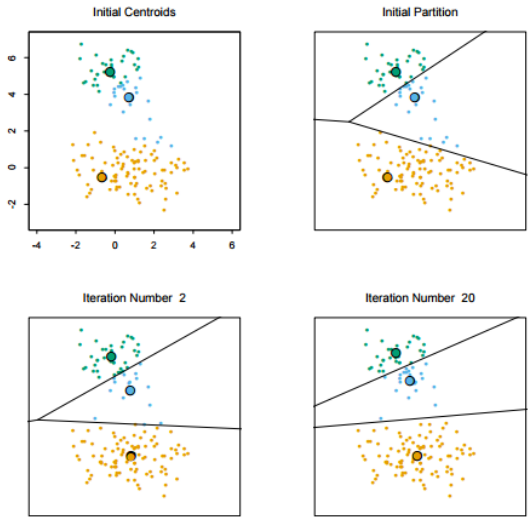


Abbildung 15: sukzessive Iteration des K-Means Clusteralgorithmus für simulierte Daten [Hastie et al. (2009)]

Die Zentroide werden durch durch die drei größeren schwarz umrandeten Kreise dargestellt. Die geraden Linien trennen die Cluster zur Veranschaulichung voneinander. Formeln (1) bis (5) und Erläuterungen stammen aus Hastie et al. (2009).
Der Nachteil des K-Means Algorithmus ist, dass die Gruppeneinteilung von der Startpartition, welche für jeden Algorithmusdurchlauf unterschiedlich ist, abhängt. Somit ist die Segmentierung einzelner Nutzer in immer gleiche Gruppen bei verschiedenen Startpartitionen nicht für alle Wiederholungen gegeben.

4.2 K-Means in der Anwendung

Bevor der K-Means Algorithmus angewandt werden kann, ist eine Skalierung der Daten notwendig. K-Means verwendet die euklidische Distanz als Distanzmaß, welches nicht skaleninvariant ist. Hierzu werden die Daten Z-Standardisiert. Die dummykodierten kategoriellen Variablen werden durch diese Standardisierung in quasimetrischen Variablen transformiert [Müller (2010)]. Dies erlaubt auch die Durchführung des K-Means Algorithmus, da dieser nur metrische Inputvariablen nutzen kann.
Nachdem die Daten aufbereitet wurden, bleibt noch zu klären, wieviele Clustergruppen erstellt werden sollen. Zur Bestimmung der Anzahl der Segmente wird das Gütekriterium der Varianzreduktion verwendet [Everitt (1993)]. Hier wird versucht für die Streuung zwischen den Clustern ein lokales Minimum zu erreichen. Da die Streuung jedoch abhängig von den Startpartitionen verschieden ist, wird in Abbildung 16 für neun wiederholte Durchläufe die Summe der Streuung innerhalb der Cluster für unterschiedliche Clustergruppen Eins bis Fünfzehn veranschaulicht um ein globales Optimum zu finden. Die Clusteranzahl wird durch das Ellbogenkriterium bestimmt [Everitt (1993)]. Bei acht bis dreizehn Clustern ist in den neun Grafiken, welche in Abbildung 16 zu sehen sind, ein Ellbogen beziehungsweise der erste herausstechende Knick zu sehen. Die Streuung innerhalb der Gruppen wird an der Stelle nicht mehr reduziert, sondern teilweise leicht erhöht. Da sich die optimale Clusteranzahl auf Basis der neun Plots zwischen acht und dreizehn bewegt, erscheinen neun Cluster als akzeptabel und brauchbar hinsichtlich des Ziels, unterschiedliche Nutzergruppen zu finden. Der prozentuale Anteil der Beobachtungen in allen neun Segmenten kann Tabelle 5 entnommen werden.
Zusammenfassend lässt sich sagen, dass in drei große Segmente, Cluster fünf, sechs und sieben, drei mittelgroße, Cluster eins, vier und acht, und drei kleine Gruppen, Cluster zwei, drei und neun, aufgeteilt wurde.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
8,89%	4,64%	3,71%	12,63%	16,14%	21,48%

Cluster 7	Cluster 8	Cluster 9
17,11%	12,74%	2,66%

Tabelle 5: Resultat Clustering: Verteilung der Nutzer auf alle neun Cluster / K-Means

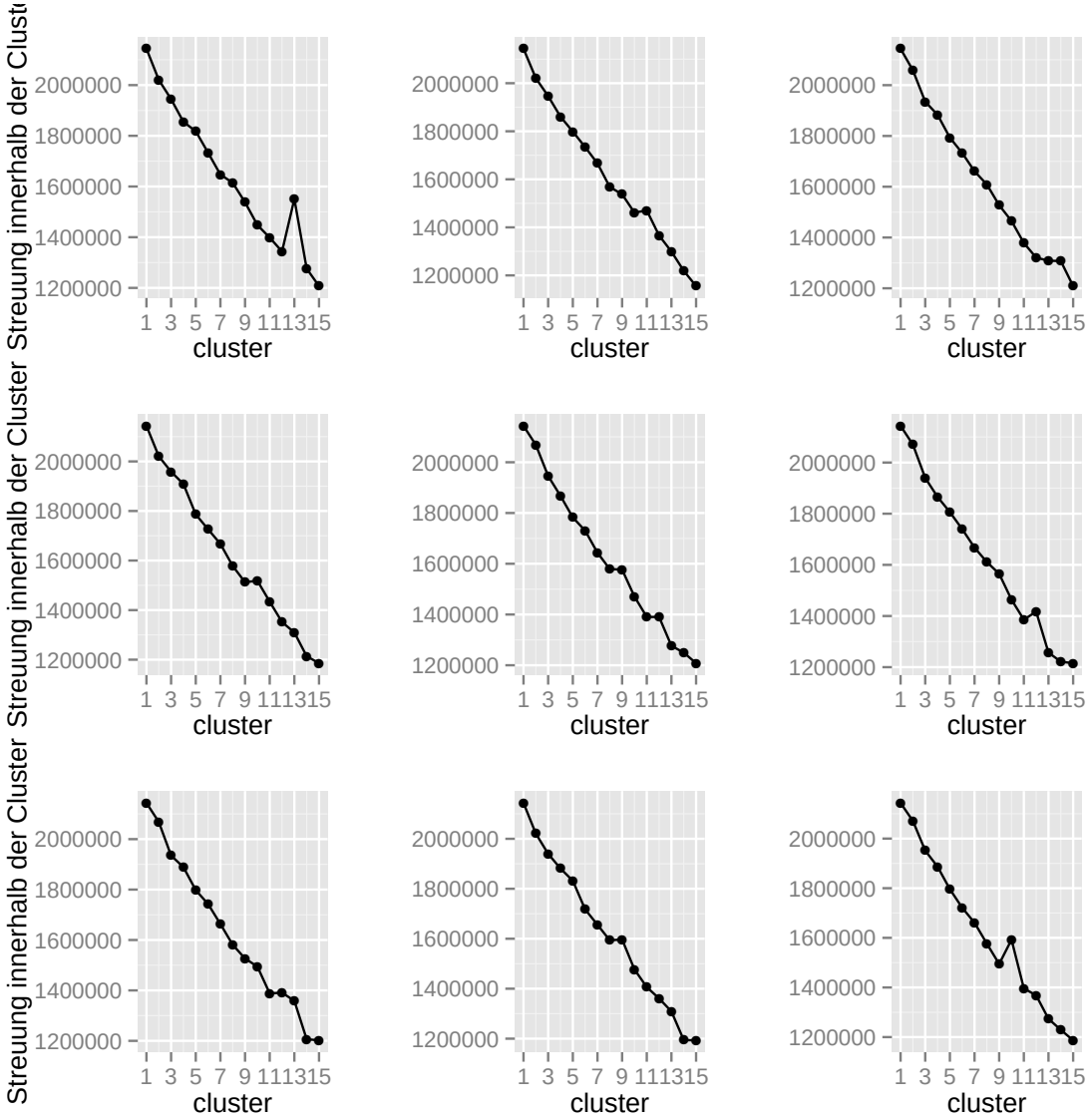


Abbildung 16: Bestimmung der Clusteranzahl anhand des Ellbogenkriteriums

Nachdem die Daten segmentiert wurden, wird nun auf die Unterschiede zwischen den einzelnen Clustern eingegangen. Hierfür werden Boxplots verwendet, welche die z-standardisierten Werte der einzelnen Variablen in jedem Cluster veranschaulichen. Die Boxplots werden ohne Ausreißer dargestellt, damit der Wertebereich der y-Achse übersichtlicher wird und die einzelnen Interquartilsabstände somit besser interpretierbar sind.

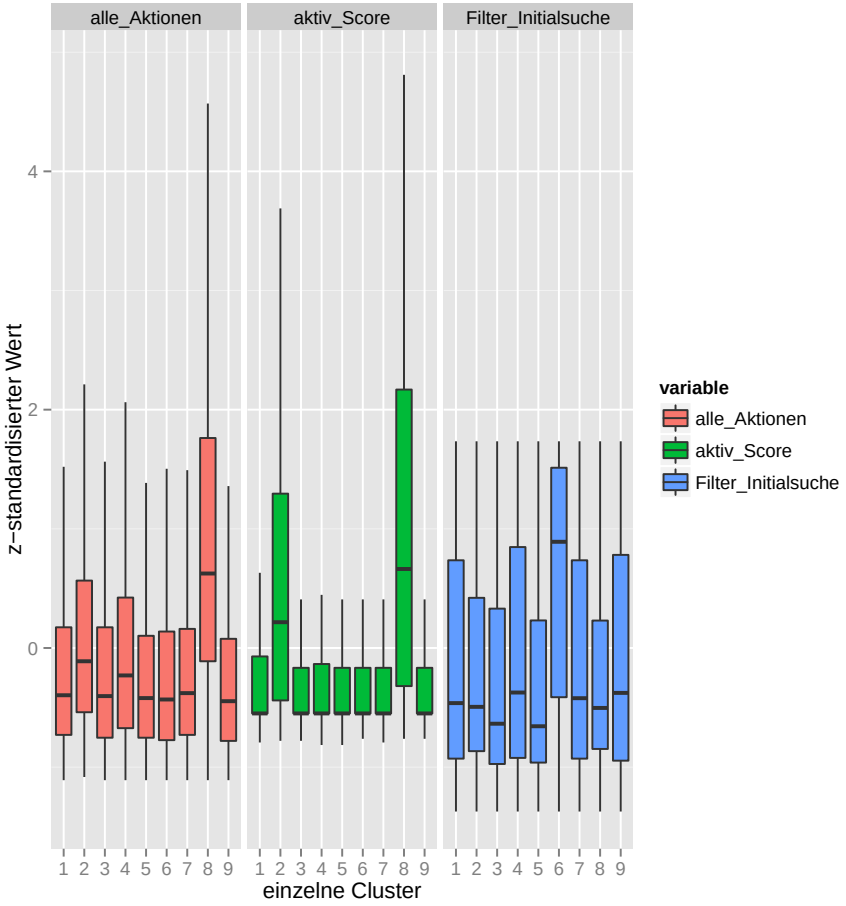


Abbildung 17: Boxplots der Variablen „alle_Aktionen“, „aktiv_Score“ & „Filter_Initialsuche“ / K-Means

Hier ist zu erkennen, dass Nutzer in Cluster zwei und acht durchschnittlich wesentlich mehr Aktionen pro Tag haben und eine höhere Aktivität aufweisen als die Übrigen. Personen aus Gruppe vier führen vergleichsweise auch mehr Aktionen täglich aus als die restlichen Gruppen. Bei der Variable „Initialfilter“ ist zu erkennen, dass der Median für Cluster sechs wesentlich höher liegt, als bei den übrigen Segmenten. Nutzer dieser Gruppe setzen im Vergleich zu den übrigen Clustern sehr viele Filter zu Beginn jeder Suche.

Anknüpfend werden die relativen Anteile der Seitentypen betrachtet.

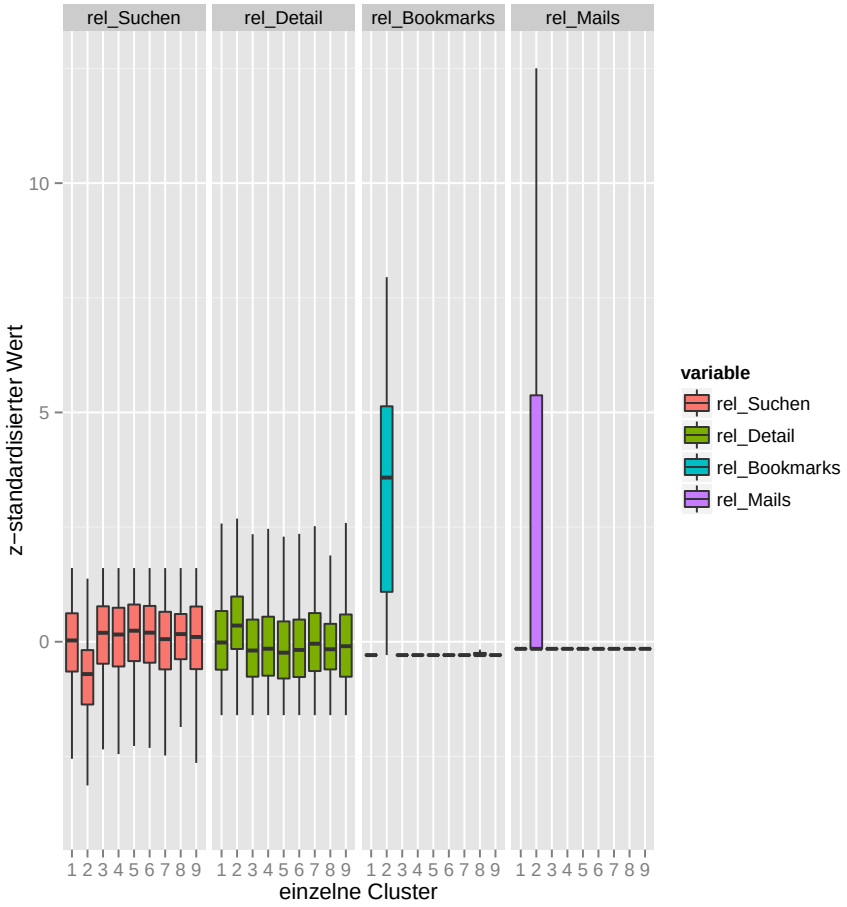


Abbildung 18: Boxplots der vier Seitentypen / K-Means

Übergreifend ist zur Interpretation der Boxplots für die prozentualen Anteile der einzelnen Seitentypen zu sagen, dass es keine großen Unterschiede zwischen den Clustern gibt und lediglich Gruppe zwei heraussticht.
Dieses Segment besitzt eine verhältnismäßig kleine Quote an Filter- und Suchergebnisseiten und hebt sich durch einen hohen Anteil an Detailseiten von den restlichen Nutzergruppen ab. Des Weiteren legen Nutzer des zweiten Clusters viele Lesezeichen an und versenden mehr Mails als die übrigen Nutzer, was darauf hinweist, dass diese schon sehr fortgeschritten sind in der Suche nach einem Fahrzeug.
Bei den übrigen Gruppen ist das Verhältnis zwischen Such- und Detailseiten ausgeglichen. Der Emailversand ist im Vergleich zum zweiten Segment sehr gering und Lesezeichen werden kaum angelegt.

Im weiteren Verlauf wird auf die Unterschiede der Variablen „Durchschnittspreis“, „Minimumpreis“ und „Maximumpreis“ eingegangen.

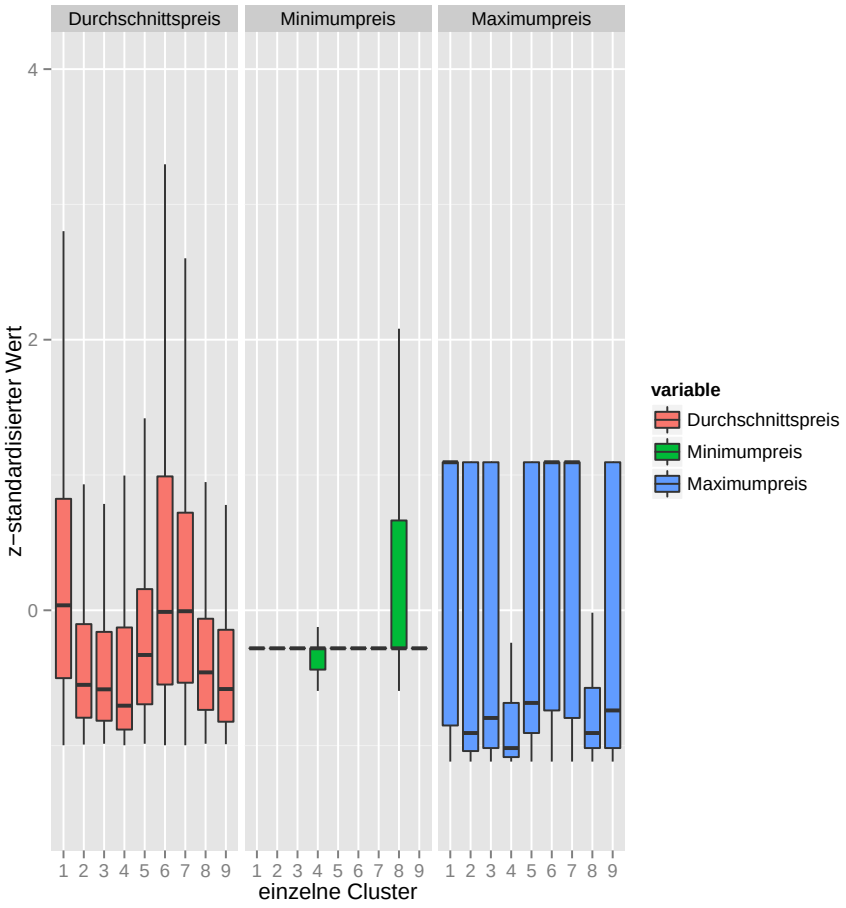


Abbildung 19: Boxplots der Variablen „Durchschnitts-“, „Minimum-“ und „Maximumpreis“ / K-Means

Bei der Betrachtung des Durchschnittspreises aller angesehenen Fahrzeuge fallen die Gruppen eins, sechs und sieben auf. Diese Segmente suchen im Vergleich zu den anderen Gruppen eher Fahrzeuge im höheren Preisbereich. Die niedrigsten Durchschnittspreise sind in Gruppe vier zu finden.
Wirft man einen Blick auf den Minimumpreis, fällt auf, dass der Median in allen Segmenten identisch ist, was darauf zurückzuführen ist, dass der voreingestellte Standardpreis von 1.000€ in allen Segmenten oft vorkommt. In Cluster acht ist der Abstand zwischen dem 75% und 25% Quantil am größten, in Gruppe vier überschneidet sich das dritte Quantil mit dem Median.
Beim Vergleich der Maximumpreise differenzieren sich die Gruppen vier und acht von den übrigen, da der Interquartilsabstand hier wesentlich kleiner ist. In Segment

eins, sechs und sieben ist das 50% Quantil gleich dem 75% Quantil. Diese Gruppen filtern nach Fahrzeugen mit hohem Maximalpreis. In den Segmenten zwei, drei, vier, fünf, acht und neun liegt der Median eher im unteren Bereich, was bedeutet, dass Nutzer dieser Gruppe einen niedrigen Maximalpreis setzen.
Insgesamt kann man über Gruppe eins, sechs und sieben sagen, dass sich Nutzer dieser Cluster für Fahrzeuge im durchschnittlich oberen Preisbereich interessieren, einen zu anderen Segmenten ähnlichen Minimumpreis angeben und nach einem hohen Maximumpreis filtern. Gruppe vier charakterisiert sich durch die Betrachtung von Fahrzeugen im unteren Preisniveau und sucht im unteren Preisbereich, da nach einem niedrigen Minimumpreis sowie einem niedrigen Maximumpreis gefiltert wird. Nutzer des Clusters acht sehen sich Fahrzeuge mit einem niedrigen Durchschnittswert an, setzen eine hohe Grenze für den Minimumpreis und einen niedrigen Maximumpreis fest.
In den nachfolgenden Grafiken, welche gestapelte Balkendiagramme zeigen, wird auf die Verteilung der Dummyvariablen „Filter: Händler“, „Filter: Privat“ und „Login“ eingegangen.

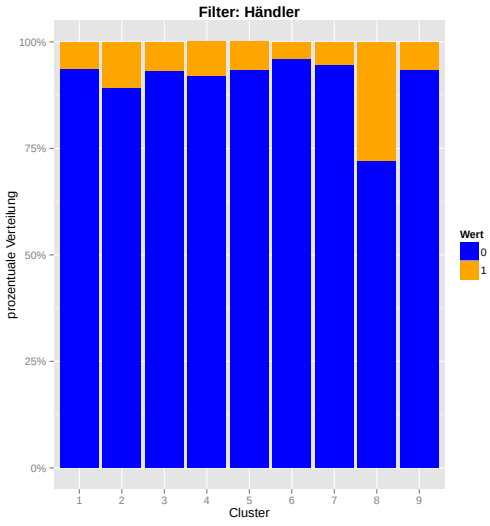


Abbildung 20: gestapeltes Balkendiagramm der Variable „Händler“ / K-Means

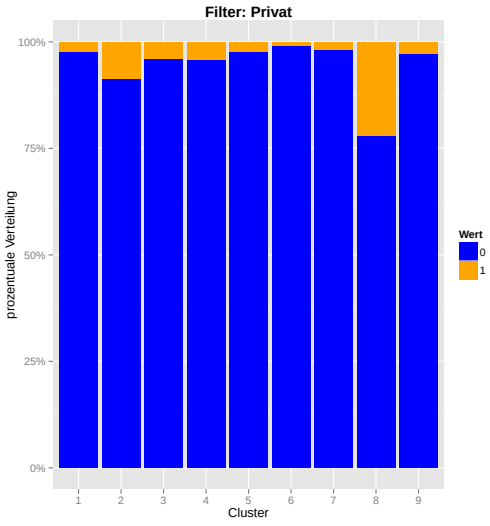


Abbildung 21: gestapeltes Balkendiagramm der Variable „Privat“ / K-Means

Abbildung 20 zeigt, aufgeteilt nach Segmenten, den prozentualen Anteil der Nutzer, welche im Laufe ihrer Suche mindestens einmal explizit nach Händlerfahrzeugen gesucht haben (Wert 1) im Vergleich zu denjenigen, welche nicht danach gefiltert haben (Wert 0). Am häufigsten suchen über 25% der Nutzer aus Gruppe acht nach Händlerfahrzeugen. Im Gegensatz dazu steht die übrigen Cluster, in denen nur zwischen

6% und 10% aller Personen mindestens einmal ausschließlich Interesse an Händlerfahrzeugen zeigen.
Abbildung 21 zeigt eine ähnliche Verteilung der „Privatfahrzeuge“ wie die der „Händlerfahrzeuge“.
Hier suchen auch wieder um die 25% der Nutzer, welche sich in Cluster acht befinden, nach Privatfahrzeugen, während verglichen mit Segment sechs am wenigsten Nutzer nach privaten Automobilen filtern.
Insgesamt wird also nicht nach den Filterkriterien „Händler“ und „Privat“ unterschieden, sondern es werden diejenigen Nutzer, welche generell den Filter setzen, in ähnliche Cluster eingeteilt.

Abbildung 22 visualisiert, analog zu den beiden vorherigen Grafiken, den prozentualen Anteil der Nutzer, welche sich im Laufe ihrer Suche mindestens einmal eingeloggt haben.

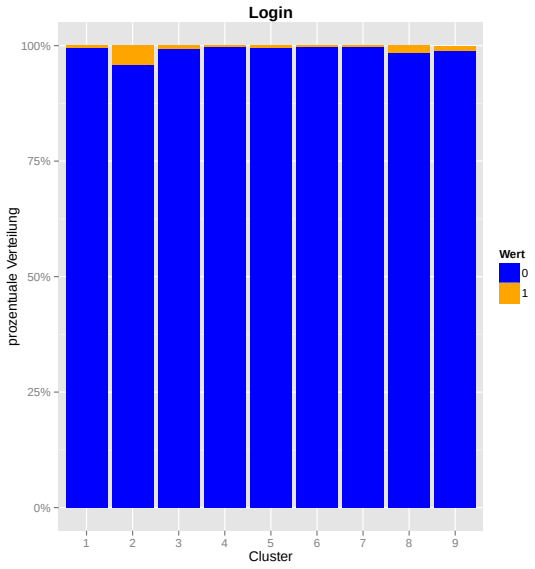


Abbildung 22: gestapeltes Balkendiagramm - Logins / K-Means

Hier sind kaum Unterschiede zwischen den einzelnen Clustern zu erkennen. Das zweite Segment hat mit circa 5% am meisten Personen, die sich mindestens einmal eingeloggt haben. In den anderen Segmenten sind vergleichweise wenige Nutzer vertreten.

Als Resultat der Clusterunterschiede kann man sagen, dass kaum Unterschiede für die drei quasi-stetigen Variablen zu finden sind. Ursache hierfür ist, dass generell wenige Nutzer nach Händler- oder Privatfahrzeugen filtern und sich auch nur ein

kleiner Anteil mindestens einmal einloggt, was bereits in der deskriptiven Analyse der Daten gezeigt wurde. Diese Variablen wurden dennoch mit in die Segmentierung aufgenommen, da sie in der Analyse der Entscheidungsbäume (Abschnitt 5.4.1 und 5.4.2) als ausschlaggebend erachtet werden.

Abschließend wird auf Unterschiede zwischen der Markensuche in den Segmenten Bezug genommen.

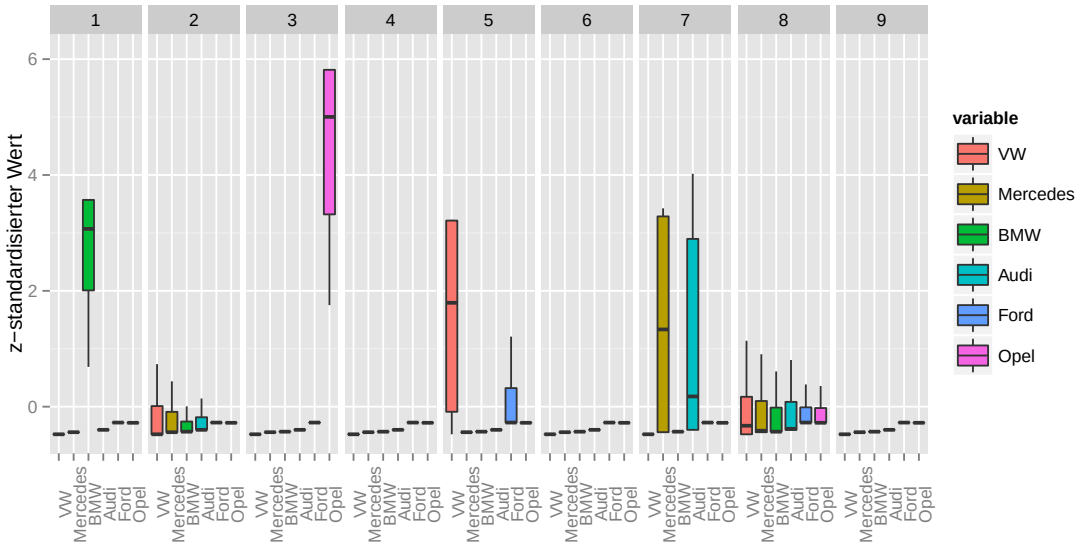


Abbildung 23: Boxplots der Markenunterschiede / K-Means

Das zweite und achte Segment filtert im Vergleich zu den anderen Gruppen vermehrt nach verschiedenen Marken. BMW wird sehr häufig von Nutzern des ersten und achten Clusters gefiltert. Cluster drei interessiert sich ausschließlich für Opel. Die Marke Volkswagen erscheint oft in den Filterkriterien der Cluster zwei, fünf und acht. Mercedes und Audi wird stark von Gruppe sieben gefiltert. Gruppe fünf filtert neben VW zusätzlich noch nach der Marke Ford. Cluster vier, sechs und neun filtern nicht auffällig oft nach den Top 5 Marken.

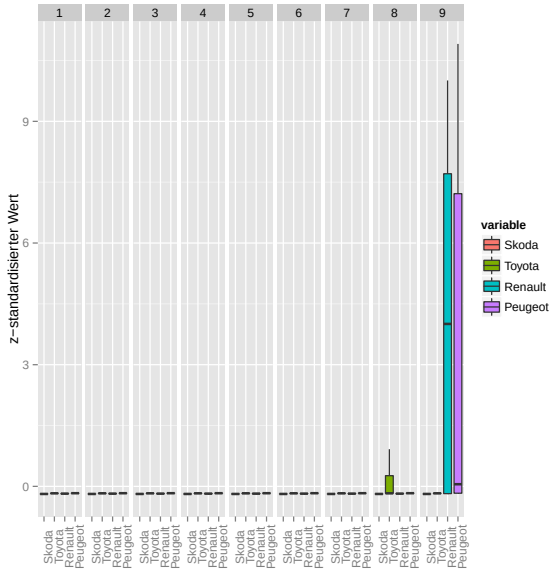


Abbildung 24: Boxplots der Top 6 bis 10 Marken / K-Means

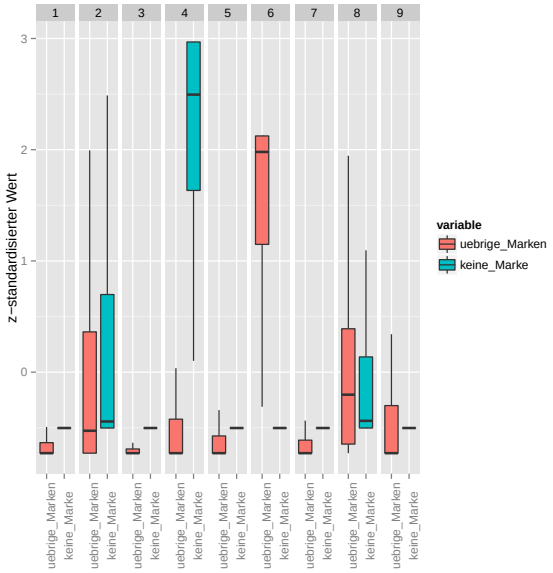


Abbildung 25: Boxplots der übrigen Marken & keine Marken / K-Means

Wie man Abbildung 24 entnehmen kann, ist die Varianz zwischen der Markensuche innerhalb der Cluster im Vergleich zu den Top 5 Marken nicht mehr so groß. Cluster neun, welches nicht nach den Top 5 Marken gesucht hat, filtert primär nach den französischen Marken Renault und Peugeot. Nutzer der achten Gruppe filtern zusätzlich neben allen Top 5 Marken nach der Marke Toyota.
Abbildung 25 macht deutlich, dass Personen aus Gruppe zwei und acht verstärkt nach den übrigen Marken filtern, jedoch auch nach keinen Marken. Gruppe vier wird charakterisiert dadurch, dass generell nach keiner Marke gefiltert wird, was bedeutet, dass die Marke für diese Nutzer von geringer Bedeutung ist. Cluster sechs sucht vorwiegend nach den übrigen Marken und interessiert sich auch nicht für die Top 10 Marken.
Insgesamt lässt sich sagen, dass die Cluster sich in der Markenpräferenz und -treue stark voneinander unterscheiden.

4.3 Die Funktionsweise des Two-Step-Cluster Algorithmus

Der Zwei-Stufen Clusteralgorithmus wurde von der Firma SPSS aufbauend auf dem BIRCH-Algorithmus [Zhang et al. (1999)] entwickelt und stellt ihn im Fachbericht „The SPSS Two-Step-Cluster-Component“ [SPSS (2006)] als „angepasste Komponente zur effizienteren Kundensegmentierung“ vor.
Die zweistufige Clusteranalyse ist nach SPSS (2006) für große Datenmengen konzipiert und kann im Vergleich zu BIRCH sowohl mit metrischen als auch mit katego-

riellen Variablen umgehen. Wie der Name bereits verrät, wird der Algorithmus in zwei Schritten durchgeführt.
Als erstes werden unter Verwendung sequentieller Verfahren vorläufige Unterclus-ter gebildet, anschließend verdichtet das hierarchische Clusterverfahren die Unter-cluster in die gewünschte Anzahl an Segmenten. Die nachfolgenden Erklärungen entstammen sowohl dem White-Paper „The SPSS Two-Step-Cluster-Component“ [SPSS (2006)], als auch „SPSS 21“ von Brosius (2013) und „A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment“ von Chiu et al. (1999).
Im ersten Schritt werden alle Objekte sukzessive bis zu den Subclustern in einen „Cluster-Feature“ - Baum (CF-Baum) eingeordnet.
SPSS verwendet für die CF-Bäume voreingestellt eine Tiefe von drei und die Kno-tenanzahl acht.
Wie Abbildung 26 zu entnehmen ist, ermöglicht diese Kombination insgesamt ma-ximal 8 x 8 = 64 Blätter und jeweils 8 x 64 = 512 Sub-Cluster.
Dieses grobe Vorgehen ist notwendig, um große Datenmengen bearbeiten zu kön-nen. „Mögliche Fehler in der Clusterzuordnung scheinen dabei vertretbar, weil das Ergebnis der ersten Stufe noch nicht abschließend ist und Fehler der ersten Stufe in der zweiten durchaus wieder korrigiert werden können“ [Brosius (2013)].

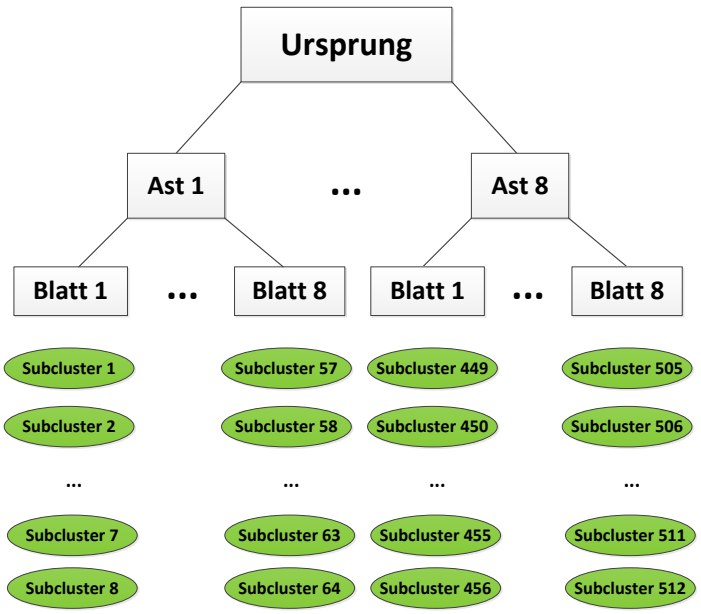


Abbildung 26: Cluster-Feature Baum der ersten Stufe [Brosius (2013)]

In diesen Subcluster sind die „Cluster Features“ (CF) beziehungsweise Kennzah-

len „Anzahl der Dateneinträge zu jedem Cluster j“: N_j, „absolute Häufigkeiten der stetigen Merkmale jeder Kennzahl N_j“: s_j, „absolute quadrierte Häufigkeiten der stetigen Merkmale jeder Kennzahl N_j“: ${s_j}^2$ und „Anzahl der Dateneinträge, bei denen das k-te kategorielle Merkmal der l-ten Kategorie zu jedem Cluster j angehört“: N_{jkl} enthalten. Der Vektor der Cluster Features hat folgende Form:

$$CF_j = (N_j, s_j, {s_j}^2, N_{jkl}) \tag{6}$$

Diese Informationen dienen als Grundlage zur Berechnung der Distanzen, welche man zur Einteilung in die Subcluster benötigt.
Wie bereits erwähnt, kann dieser Algorithmus stetige und kategorielle Variablen behandeln.
Liegen ausschließlich stetige Variablen vor, wird die Distanz zweier Objekte X und Y mit der euklidische Distanz berechnet:

$$d_E(X,Y) = \sqrt{\sum_{i=1}^{v} (X_i - Y_i)^2} \tag{7}$$

wobei
v: Anzahl der in die Analyse mit einbezogenen Variablen
und
X_i beziehungsweise Y_i: die Ausprägungen der Variablen i der beiden Objekte X und Y

Die euklidische Distanz entspricht der Wurzel aus der Summe der quadrierten Abweichungen zwischen zwei Clustern X und Y.
Falls jedoch zusätzlich kategorielle Variablen geclustert werden sollen, wird das Log-Likelihood Distanzkriterium, welches auf der BIRCH-Methode [Zhang et al. (1999)] aufbaut, verwendet.
„Steht man nun vor der Frage, ob ein Fall einem Cluster A oder Cluster B zugeordnet werden sollte, lässt sich für beide Varianten die Log-Likelihood-Distanz berechnen und es wird die Zuordnung vorgenommen, die mit der höchsten Wahrscheinlichkeit (Likelihood) verbunden ist“ [Brosius (2013)]. BIRCH kann ausschließlich stetige Variablen behandeln und verwendet die „Anzahl der Dateneinträge“, die „absoluten Häufigkeiten der stetigen Merkmale“ und die „absoluten quadrierten Häufigkeiten der stetigen Merkmale“, was analog die ersten drei CF für die Two-Step-Clusteranalyse sind.
Der SPSS Algorithmus erweitert BIRCH's CF, indem die „Anzahl der Dateneinträge, bei denen das k-te kategorielle Merkmal der l-ten Kategorie zu jedem Cluster j

angehört“ hinzugefügt wird.
Die genaue Berechnung der Distanz zwischen zwei Beobachtungen, welche kategorielle Ausprägungen enthalten, ist an dieser Stelle nicht näher erläutert und unter Zhang et al. (1999) genauer ausgeführt.
Im weiteren Verlauf wird genauer darauf eingegangen wie der CF-Baum aufgebaut wird. Jede zu klassifizierende Beobachtung wird chronologisch in den Baum eingeordnet und dem Knoten zugewiesen, zu dem die Distanz am kleinsten ist. Der Abstand wird durch das zuvor beschriebene Distanzmaß berechnet.
Um Heterogenität zwischen den Unterclustern zu sichern, wird ein Schwellenwert T, welcher den Distanzwert berücksichtigt, vorgegeben. Die einzuordnenden Beobachtungen dürfen diesen Wert nicht überschreiten.
Ist dies jedoch der Fall, wird der Baum abhängig von den unterschiedlichen Beobachtungen und deren zugehörigen Schwellenwerten T reorganisiert. Grundsätzlich gilt: Je größer T, desto kleiner ist der Entscheidungsbaum, da die Regeln der Clusterzuweisung umso flexibler sind.
Des Weiteren hängt der Aufbau des CF-Baums stark von der Reihenfolge der eingeordneten Punkte ab. Um diesen Effekt zu umgehen, bietet sich eine randomisierte Einordnung an.
Das Ablaufschema zur Aufnahme eines Falles/Objekts in den CF-Baum wird in Abbildung 27 dargelegt.

Nachdem nun der CF-Baum erstellt wurde, werden die Untercluster im zweiten Schritt unter Verwendung der hierarchischen Clusteranalyse zu der gewünschten Anzahl an Gruppen zusammengefasst.
Die Fusion erfolgt durch die schrittweise Zusammenfassung zweier Untercluster, welche gemäß dem Distanzmaß als ähnlich erachtet werden. Der Algorithmus kann zu einer vorgegebenen Clusterzahl zusammenfassen oder eine Anzahl vorschlagen.

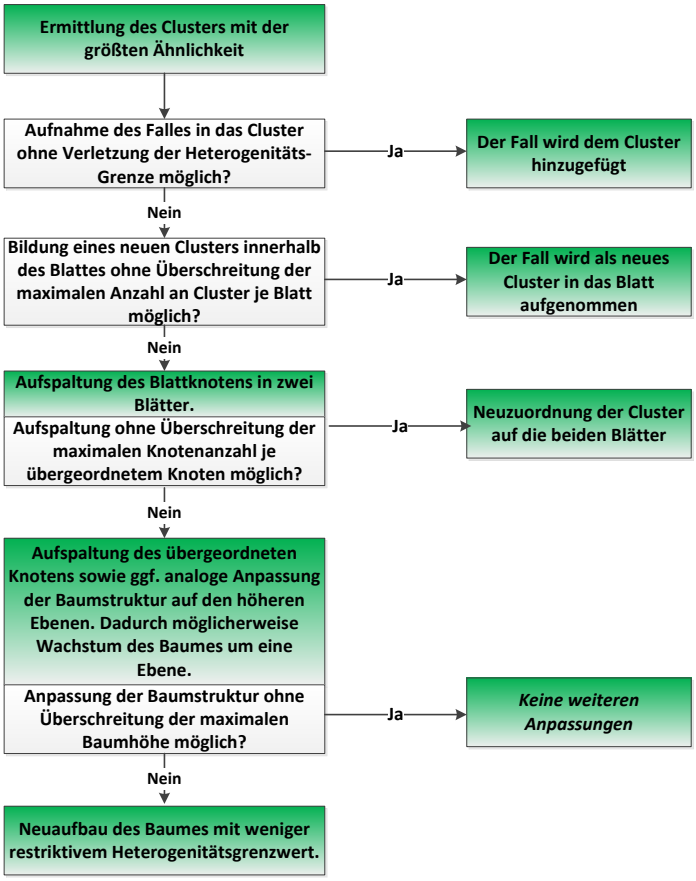


Abbildung 27: Ablaufschema der Aufnahme eines neuen Falles in den CF-Baum [Brosius (2013)]

4.4 Two-Step-Clustering in der Anwendung

Der Algorithmus wird mit einer vorgegebenen Clusteranzahl von neun in SPSS durchgeführt. Die Anzahl wurde an K-Means angelehnt, um eine bessere Vergleichbarkeit der beiden Clusteralgorithmen zu gewährleisten. Die Verteilung der Beobachtungen in den einzelnen Clustern ist folgender Tabelle zu entnehmen.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
7,62%	5,41%	7,04%	15,76%	9,35%	8,79%

Cluster 7	Cluster 8	Cluster 9
10,41%	15,49%	20,13%

Tabelle 6: Resultat Clustering: Verteilung der Nutzer auf alle neun Cluster / Two-Step-Cluster-Component

Man kann, wie bei K-Means auch, drei größere Segmente (Cluster vier, acht und neun), drei mittelgroße (Cluster fünf, sechs und sieben), und drei kleine Gruppen (Cluster eins, zwei und drei), erkennen.
Um die beide Algorithmen gegenüberstellen zu können, werden analog zu K-Means erneut die Boxplots für die Variablen der unterschiedlichen Cluster verglichen.
Alle dazu produzierten Grafiken finden sich im Anhang B.
Zunächst werden die Variablen „alle_Aktionen“, „aktiv_Score“ und „Filter_Initialsuche“ betrachtet. Bei den ersten beiden Variablen fallen Cluster zwei und vier auf. Nutzer dieser Gruppe werden im Vergleich zu den anderen Segmenten durch eine große Anzahl täglicher Aktionen sowie viele aktive Tage gekennzeichnet.
Betrachtet man die Anzahl der Initialfilter, ist der Wert für diese beiden Cluster jedoch niedrig. Auffällig ist Gruppe neun bei den Initialfiltern. Dieses Segment setzt zu Beginn jeder Suche vergleichsweise viele Filter. Analysiert man die Aktivität, bewegen sich diese Personen eher im unteren Bereich. Des Weiteren fällt Cluster zwei wieder bezüglich des prozentualen Anteils an verschiedenen Seitentypen auf. Personen dieser Gruppe haben einen geringen relativen Anteil an Suchseiten, betrachten aber viele Inserate im Detail. Außerdem versenden diese Nutzer viele Emails und legen vergleichsweise viele Lesezeichen an. Alle übrigen Segmente werden für die Variablen „Such- und Detailseiten“ durch annähernd gleiche Mediane und Interquartilsabstände beschrieben. Das zweite Cluster ausgeschlossen fällt keine andere Gruppe durch einen starken Emailversand oder eine Vielzahl an Lesezeichen aus.
Außerdem ist der prozentuale Anteil an Suchseiten bei allen Segmenten außer dem zweiten etwas höher als an Detailseiten.
Im nächsten Schritt werden analog zur Analyse der K-Means Ergebnisse aufs Neue

die Variablen „Durchschnittspreis“, „Minimum-“ und „Maximumpreis“ betrachtet. Cluster sechs, acht und neun grenzen sich von den übrigen durch hohe Durchschnittspreise ab. Ergänzend stellen diese drei Gruppen vergleichsweise auch sehr hohe Maximumpreise ein.
Segment vier setzt einen hohen Minimumpreisfilter und einen niedrigen Maximumpreis. Dies gibt Auskunft darüber, dass Nutzer dieser Gruppe in einem kleinen Preisintervall suchen. Der durchschnittliche Preis aller im Detail angesehenen Fahrzeuge ist für dieses Cluster im Vergleich zu den anderen eher im mittleren Bereich. Gruppe sieben fällt durch im Schnitt niedrige Fahrzeugpreise auf. Außerdem wird ein niedriger Minimum- und Maximumpreis gesetzt.
Beim Vergleich der Markenpräferenzen werden, wie bei K-Means auch, deutliche Unterschiede zwischen den Gruppen klar. Cluster zwei und vier interessiert sich primär für die Top Vier Marken VW, Mercedes, BMW und Audi, sucht jedoch auch nach Fahrzeugen ohne Markeneinschränkung.
Das vierte Segment filtert außerdem noch nach Ford und Opel.
Cluster fünf betrachtet ausschließlich VW-Fahrzeuge und Nutzer der Gruppe sechs sind Mercedesliebhaber. Personen aus Segment acht interessieren sich nur für die Marken BMW und Audi während Nutzer der Gruppe eins sich für Ford, Opel und die übrigen Top elf bis 353 Marken begeistern. Cluster drei geht auf die Suche nach Skoda, Toyota, Renault und Peugeot. Personen der Gruppe neun schenkt den Top 10 Marken keine Beachtung und suchen lediglich nach den 353 übrigen Marken. Für Cluster sieben sind Marken unwichtig, da der relative Anteil an Suchen ohne spezifische Marke für diese Nutzergruppe am höchsten ist.
Zur besseren Übersicht kann man neben Anhang B auch Tabelle 7 entnehmen, wie sich verschiedene Variablenwerte innerhalb der Cluster unterscheiden.

	Aktivität	Initialfilter	Anteil Detailseiten	Anteil Suchseiten	Anteil Emails	Anteil Merkzettel	Durchschnittspreis	Marken
Cluster 1	-	-	o	o	-	-	o	Ford, Opel, übrige und restliche Marken
Cluster 2	+	-	+	-	+	+	o	VW, BMW, Mercedes, Audi, übrige und restliche Marken
Cluster 3	-	-	o	o	o	o	o	Skoda, Toyota, Renault, Peugeot, übrige Marken
Cluster 4	+	-	o	o	o	o	o	VW, BMW, Mercedes, Audi, Ford, Opel, übrige und restliche Marken
Cluster 5	-	-	o	o	o	o	o	VW
Cluster 6	-	o	o	o	o	o	+	Mercedes
Cluster 7	o	o	o	o	o	o	-	keine Marken
Cluster 8	-	-	o	o	o	o	+	BMW, Audi
Cluster 9	-	+	o	o	o	o	+	übrige Marken

Tabelle 7: Ausprägungen ausgesuchter Variablenwerte innerhalb der Cluster, +: hoher Wert, o: mittel, -: niedrig

5 Ex-Post Analyse

5.1 K-Means

Im Folgenden werden die neun Cluster auf vier in der Praxis relevante Nutzergruppen reduziert, da sich diese in den wichtigsten Merkmalen, wie der relative Anteil an Seitentypen oder dem Aktivitätsmaß nicht wesentlich voneinander unterscheiden. Ergänzend wird die Interpretation von neun Nutzergruppen, welche sich nur durch marginale Feinheiten von den übrigen distanzieren, als zu komplex erachtet und aus diesem Grund erscheint die Einschränkung auf vier Segmente, welche sich im Wesentlichen unterscheiden, sinnvoller.
Hierzu wird das arithmetische Mittel jeder Variable pro Cluster berechnet und heuristisch in vier Koordinatenplots (Abbildung 28 bis 31) eingeordnet.
Drei der vier Nutzergruppen spiegeln die „consumer journey“[1] wider. Diese beschreibt die einzelnen Phasen, welche jeder Nutzer vor dem Fahrzeugerwerb durchläuft: Orientierungsphase, Detailsuche und Entscheidungsphase.
Die vierte Gruppe wird charakterisiert durch „Fahrzeugliebhaber“, was eine andauernde Phase ist, welche jedoch nicht auf den Erwerb eines Fahrzeugs abzielt.

Der Anteil der Nutzer in den Gruppen ist in folgender Tabelle dargestellt.

Orientierungsphase	aktive Suche	Entscheidungsphase	Fahrzeugliebhaber
35,15%	12,74%	4,30%	47,81%

Tabelle 8: Clusterzuweisung des K-Means Algorithmus

Die Cluster sind farblich getrennt, wie es in der jeweiligen Legende dargestellt ist.

[1] unternehmensinterne Auswertung (basierend auf qualitativen Interviews)

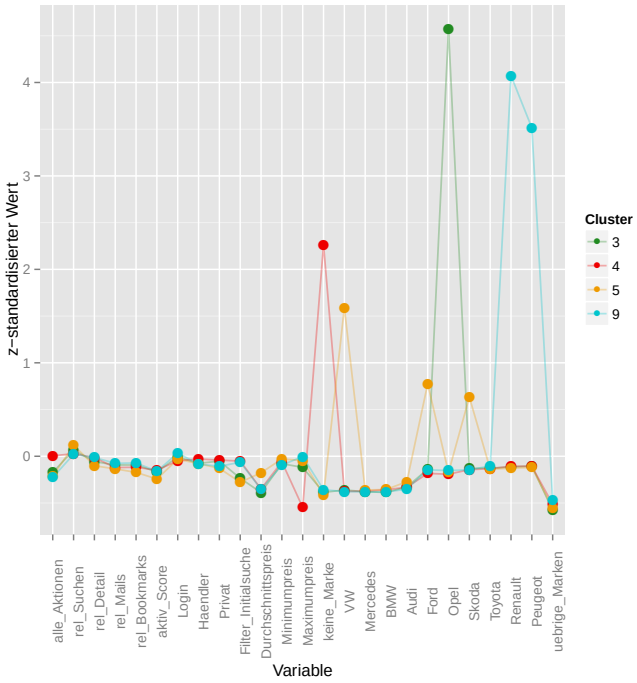


Abbildung 28: Koordinatenplot der „Orientierungsphase“/ K-Means

In Abbildung 28 werden Cluster drei, vier, fünf und neun charakterisiert durch Nutzer, die in der Suchphase noch am Anfang zu stehen und sich orientieren.
Der relative Anteil an Suchseiten ist im Vergleich zu Detailseiten, Mailversand und dem Anlegen von Merkzetteln etwas höher als der Durchschnitt, was vermuten lässt, dass sich diese Nutzer zunächst einen groben Überblick verschaffen möchten.
Des Weiteren sind diese Personen im Vergleich zum Erwartungswert weniger aktiv, als es in einer frühen Phase üblich ist. Befragungen[1] von Nutzern, welche sich noch in der Orientierung befinden, ergeben, dass sich diese Personen zunächst bezüglich anderer Faktoren, wie Unfallstatistiken des ADAC, informieren oder zunächst Meinungen aus dem Bekanntenkreis einholen. Das Thema des Fahrzeugvergleichs ist bei den meisten laut der Umfrage noch nicht so relevant, weswegen die Plattform „autoscout24.de“ auch noch nicht so häufig besucht wird.
Es werden außerdem wenige Initialfilter festgelegt, was darauf schließen lässt, dass diese Personengruppe noch keine konkrete Vorstellung vom gewünschten Fahrzeug hat, sondern sich in dieser Phase noch „durchklickt“. Diese Personen filtern auch noch nicht unter- oder überdurchschnittlich oft nach Händler- oder Privatfahrzeugen. Des Weiteren wird mit Ausnahme einzelner Cluster und Marken unterdurchschnittlich selten nach den Top 10 Marken, aber auch nach keinen und übrigen Marken gesucht.

[1] Wird aus rechtlichen Gründen nicht veröffentlicht und kann beim Autor angefragt werden

Abbildung 29 veranschaulicht Cluster acht. Dieses Segment spiegelt Nutzer wider, welche aktiv auf der Fahrzeugsuche sind.

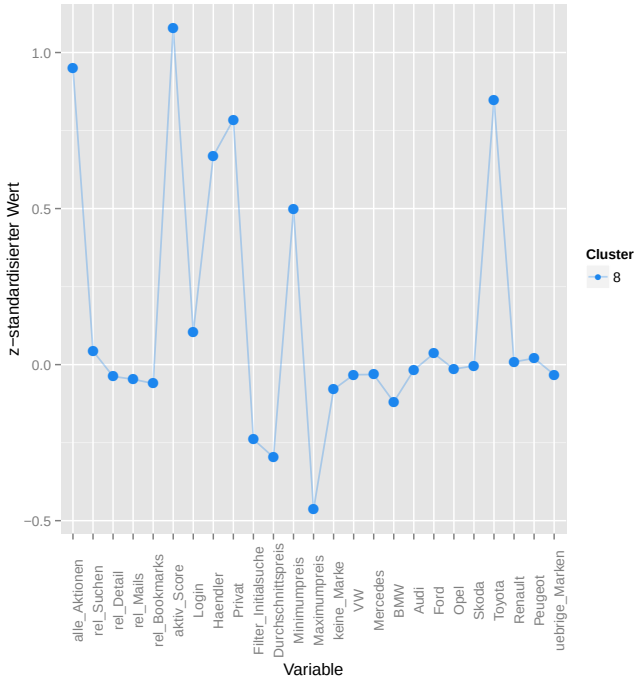


Abbildung 29: Koordinatenplot der „aktiven Suche“/ K-Means

Personen dieses Clusters differenzieren sich von der vorherigen Gruppe durch eine hohe Präsenz auf der Plattform, da sie sowohl überdurchschnittlich viele Aktionen pro Tag tätigen, als auch im Verhältnis zu ihrer Zeitspanne an vielen Tagen die Webseite besuchen.
Der prozentuale Anteil an Suchseiten ist wie in der Anfangsphase leicht überdurchschnittlich, während der Anteil an Detailseiten, Emails und Merkzetteln leicht unterdurchschnittlich ist.
Des Weiteren werden auch wenige Initialfilter gesetzt, was darauf schließen lässt, dass diese Nutzer immer noch keine klare Vorstellung vom gewünschten Automobil haben.
Außerdem loggen sich diese Nutzer im Vergleich zum Durchschnitt oft ein, was ein Hinweis darauf ist, dass sie Kontakt zum Verkäufer aufnehmen möchten. Dies ist nur möglich, sofern man eingeloggt ist.
Zusätzlich wird im Vergleich zu der vorherigen Phase ein hoher Minimumpreis und ein niedriger Maximumpreis festgelegt, was bedeutet, dass diese Nutzer in einem konkreteren Preisintervall suchen.
Überblickend zeichnen sich Nutzer dieser Phase dadurch aus, im Vergleich zu Personen, welche sich in der Anfangsphase befinden, öfter und intensiver die Plattform

zu nutzen, ein kleineres Preisintervall festzulegen und sich überdurchschnittlich oft einzuloggen.

Im weiteren Verlauf werden die Nutzer beschrieben, die in der Entscheidungsphase weit fortgeschritten sind und vermutlich kurz vor dem Fahrzeugkauf stehen.

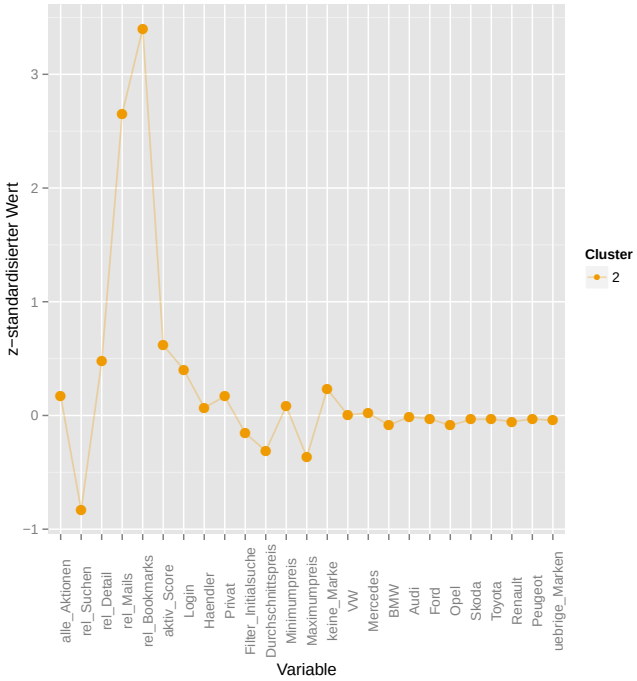


Abbildung 30: Koordinatenplot der „Entscheidungsphase“ / K-Means

Wie Abbildung 30 zu entnehmen ist, sind diese Nutzer im Vergleich zu denjenigen der aktiven Suche nicht mehr so stark, jedoch trotzdem noch überdurchschnittlich oft, auf der Webseite aufzufinden, da nur noch gezielt nach einer kleinen Grundgesamtheit an Fahrzeugen gesucht wird.
Außerdem fällt auf, dass der prozentuale Anteil des Anlegens von Merkzetteln und Versenden von Emails vergleichsweise hoch ist. Nutzer dieses Segments loggen sich, wie in der aktiven Suche, überdurchschnittlich oft ein und setzen in den wenigen Suchen, die sie tätigen, bewusst Filter um zwischen Privat- und Händlerfahrzeugen zu unterscheiden.
Es wird, wie in der aktiven Suche, ein überdurchschnittlich hoher Minimum- und unterdurchschnittlich tiefer Maximumpreis eingestellt, was auch wieder auf eine sehr klare Preisvorstellung schließen lässt.
Insgesamt wird diese Phase also, im Vergleich zu der aktiven Suche, durch den Versand überdurchschnittlich vieler Emails und Kontaktanfragen sowie durch das Anlegen vieler Merkzettel charakterisiert. Analog zur aktiven Suche besitzen diese

Nutzer einen hohen Aktiv-Score und tätigen vergleichsweise viele Aktionen.

Der vierte Nutzertyp besteht aus „Fahrzeugliebhabern“ beziehungsweise Automobilenthusiasten.

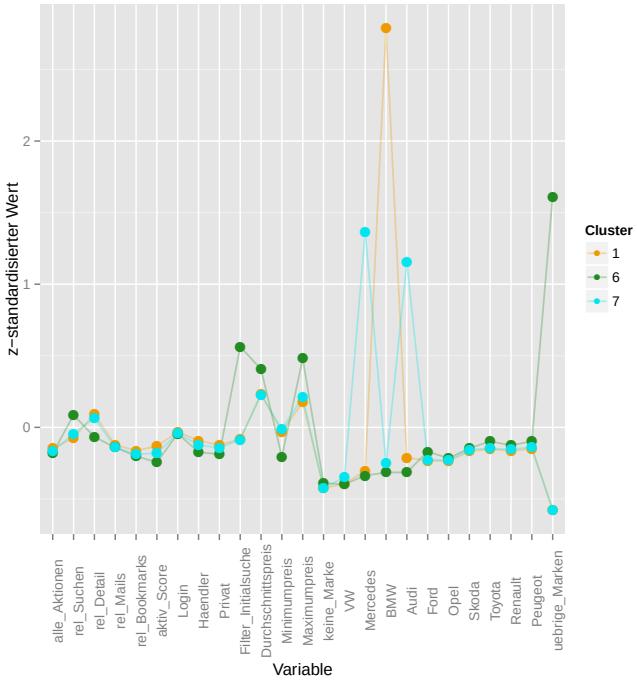


Abbildung 31: Koordinatenplot „Fahrzeugliebhaber“ / K-Means

Übergreifend kann man sagen, dass diese Nutzergruppe die Webseite nicht oft (unterdurchschnittlicher Aktiv-Score) und wenig intensiv (unterdurchschnittlich wenige Aktionen pro Tag) besucht. Zwei von drei Cluster zeichnen sich durch einen überdurchschnittlichen Anteil an Detailseiten- im Vergleich zu Suchseitenaufrufen aus. Außerdem werden gegenüber anderen Nutzertypen am seltensten Emails verschickt und Merkzettel angelegt. Personen dieser Cluster loggen sich nicht überdurchschnittlich häufig ein und ein großer Teil ist daher vermutlich auch nicht auf der Plattform registriert. Nach Händler- und Privatfahrzeugen wird im Vergleich zu anderen Nutzern nicht gefiltert. Cluster sechs filtert im Vergleich zu den anderen Clustern dieser dritten Nutzergruppe am Anfang jeder Suche sehr genau. Alle Cluster interessieren sich für überdurchschnittlich teure Fahrzeuge und es wird nach einem sehr hohen Maximumpreis gefiltert. Diese Nutzergruppe sieht sich häufig die Marken Mercedes, BMW, Audi und die übrigen Marken an.
Zusammenfassend kann man sagen, dass dieser Nutzertyp nicht daran interessiert ist, die angesehenen Fahrzeuge zu kaufen, sondern diese „Wunsch-/Traumfahrzeuge“ nur betrachten möchte.

Überblickend werden nun die arithmetischen Mittel jeder Variable in den verschiedenen Clustern durch eine Heatmap (Abbildung 32) veranschaulicht. Die Verschmelzung einzelner Cluster wird anhand der Werte nochmals verdeutlicht.
Da die Werte z-standardisiert sind, repräsentieren dunkelrote Flächen pro Variable überdurchschnittlich hohe Werte und umgekehrt weiße unterdurchschnittlich niedrige Werte.

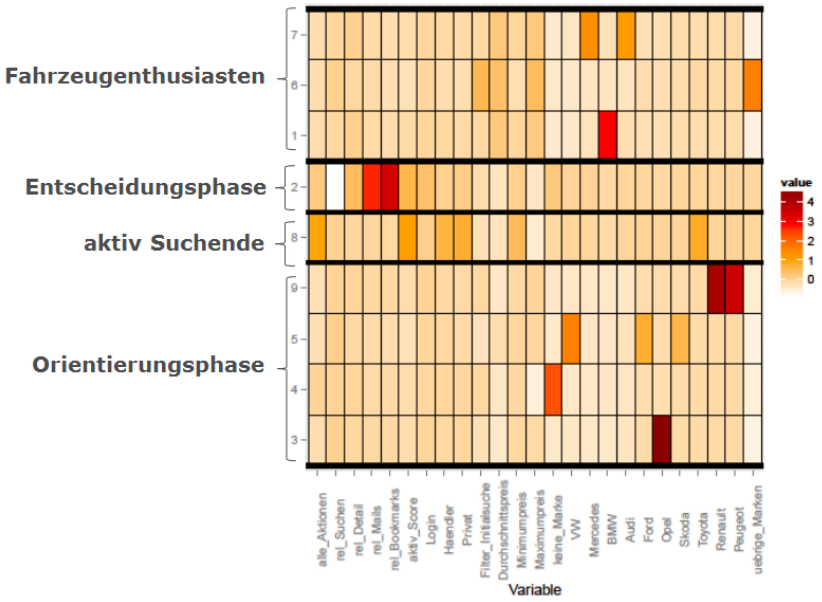


Abbildung 32: Heatmap / K-Means

Die Werte der Gruppe „Orientierungsphase“ (Cluster drei, vier, fünf und neun) unterscheiden sich nur im Interesse für die einzelnen Marken, wie „keine Marke“, VW, Ford, Opel, Skoda und die französischen Automobile Renault und Peugeot. Mit einer niedrigen Aktivität und ähnlichen prozentualen Anteilen der Seitentypen sind sie insgesamt sehr homogen untereinander. Diese Personen haben die gleichen Verhaltensweisen, suchen jedoch verstärkt nach verschiedenen Marken.
Segment acht, welches Nutzer der aktiven Suche enthält, und zwei, welches Personen in der Entscheidungsphase repräsentiert, bilden jeweils eigene Nutzergruppen. Diese heben sich durch die Messzahlen der Aktivität, der verschiedenen Seitentypen und der Filtereinstellungen von den übrigen Clustern ab.
Die vierte Gruppe, die Fahrzeugenthusiasten, besteht aus Cluster eins, sechs und

sieben. Personen dieses Segments fallen durch im Vergleich hohe Durchschnitts- und Maximumpreise auf.
Des Weiteren wird in der Heatmap noch einmal die Präferenz teurer Marken im Vergleich zu anderen Clustern deutlich.

5.2 Two-Step-Clustering

Wie bei K-Means sind die Variablenwerte einzelner Cluster wieder ähnlich und können daher zu insgesamt vier verschiedenen Gruppen zusammengefasst werden.
Somit wird die Gruppenanzahl neun auch auf vier reduziert.
Wie Tabelle 9 entnommen werden kann, sind die einzelnen Nutzergruppen und deren Verteilung sehr ähnlich zu den Ergebnissen des K-Means Algorithmus.

Orientierungsphase	aktive Suche	Entscheidungsphase	Fahrzeugliebhaber
34,30%	14,76%	6,40%	44,54%

Tabelle 9: Clusterzuweisung des Two-Step-Cluster Algorithmus

Die durchschnittlichen Variablenwerte für jede Phase unterscheiden sich nicht wesentlich zwischen den beiden zu vergleichenden Algorithmen. Aus diesem Grund wird an dieser Stelle auf Anhang D verwiesen, in dem die Koordinatenplots zu finden sind und lediglich überblickend im Folgenden auf die Heatmap eingegangen, welche die Ähnlichkeiten innerhalb einzelner Cluster und die Unterschiede zwischen den Gruppen verdeutlicht.

Wie man Abbildung 33, welche ähnlich zu der Heatmap von K-Means ist, entnehmen kann haben Nutzer der Orientierungsphase (Cluster eins, drei, fünf und sieben) alle Variablenwerte gemeinsam und unterscheiden sich nur in der Markenpräferenz durch verschiedene Vorlieben. Cluster vier repräsentiert Personen in der aktiven Suche, was durch überdurchschnittlich hohe Werte aller Aktionen und des Aktiv-Scores in der Grafik deutlich wird. Segment zwei enthält Personen in der Entscheidungsphase, wie man den dunkelroten Felder des relativen Anteils sowohl für Merkzettel, als auch für Emails entnemen kann. Cluster sechs, acht und neun stellen Fahrzeugliebhaber dar. Für diese Nutzergruppe sind insbesondere die Werte des hohen Durchschnitts- und Maximalpreises ähnlich und die starke Vorliebe für Premiummarken, wie Mercedes, BMW oder Audi, wird durch einen hohen Wert der entsprechenden Spalten deutlich.

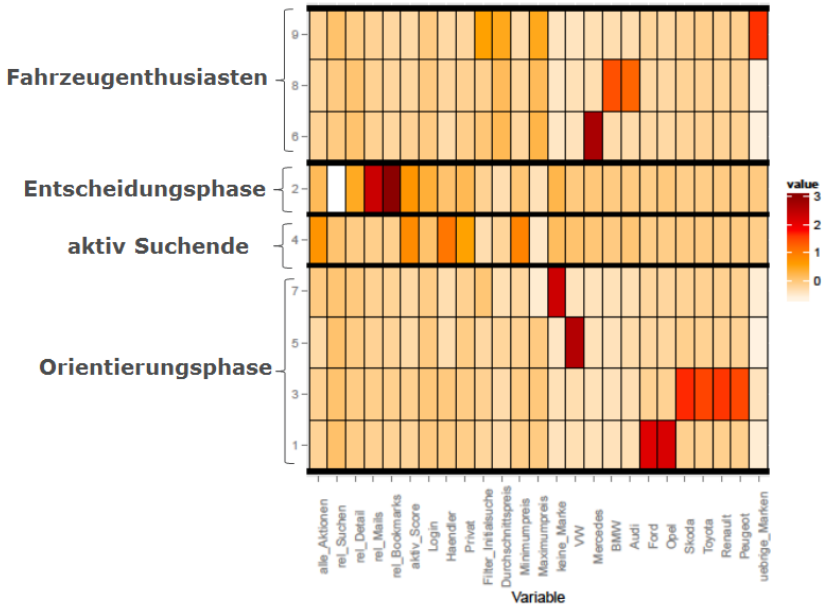


Abbildung 33: Heatmap / Two-Step-Cluster Algorithmus

5.3 Vergleich K-Means und Two-Step-Clustering

Beide verwendeten Algorithmen resultieren in vier sehr ähnliche Nutzergruppen, welche unter anderem die Entscheidungsphasen im Laufe einer Fahrzeugsuche wiederspiegeln. Segment eins wird charakterisiert durch Personen in einer frühen Kaufphase, Cluster zwei beinhaltet Nutzer in einer aktiven Suche des Fahrzeugkaufs und Gruppe drei enthält entschlossene Käufer, welche sich in der Entscheidungsphase befinden und vermutlich bald ein Fahrzeug erwerben möchten. Die vierte Kategorie stellt Fahrzeugenthusiasten dar, welche nicht am Kauf interessiert sind und sich Fahrzeuge ansehen, die nicht zu den beliebtesten zehn gehören und teuer sind. Der prozentuale Anteil der Beobachtungen für alle vier Cluster wurde bereits in Tabelle 8 und 9 gezeigt. In untenstehender Tabelle werden die Clusterzuweisungen beider Algorithmen verglichen.

k/t	1	2	3	4	$\sum$
1	87,52%	9,40%	1,19%	1,89%	100%
2	17,99%	73,77%	2,91%	5,33%	100%
3	2,23%	0,35%	97,23%	0,19%	100%
4	2,40%	6,30%	0,93%	90,37%	100%

Tabelle 10: Klassifikationssvergleich beider Clusteralgorithmen, k:K-Means, t:Two-Step-Cluster Algorithmus

Die grüne Diagonale stellt eine übereinstimmende Clusterzuweisung beider Algorithmen dar. 87,52% aller Nutzer aus Cluster Eins des K-Means Algorithmus werden auch durch den zweistufigen Clusteralgorithmus als Käufer in der frühen Phase klassifiziert. 8059 Personen, welche 73,77% der durch K-Means dem zweiten Cluster zugeordneten Personen entsprechen, werden auch vom Two-Step-Cluster-Verfahren als Nutzer in der aktiven Suche befunden. 97,23% der durch K-Means als Nutzer kurz vor dem Fahrzeugkauf beurteilten Personen werden auch durch die Two-Step-Clustermethode als solche eingestuft. 90,37% aller sich durch K-Means in Gruppe vier befindlichen Personen, welche 41.000 entsprechen, klassifiziert der Two-Step-Cluster Algorithmus auch als Fahrzeugliebhaber.
Betrachtet man die Klassifikationen, welche nicht übereinstimmen, fällt auf, dass die Algorithmen primär bezüglich der Einteilung in Cluster eins, Orientierungsphase, und zwei, aktive Suche, unterschiedlich zuteilen. Von allen Personen, die K-Means als Nutzer in der aktiven Suche definiert, werden circa 18% vom Two-Step-Cluster Algorithmus als Nutzer in der Anfangsphase klassifiziert. Umgekehrt werden 9,4% der 30.137 Personen, die durch K-Means als Nutzer in der frühen Phase beurteilt werden, durch das Two-Step-Cluster-Verfahren als Nutzer der aktiven Suche erachtet.
Des Weiteren ist noch eine mangelnde Trennschärfe zwischen Cluster zwei und vier festzustellen. 6,3% der Nutzer, die von K-Means als Fahrzeugliebhaber befunden werden, weist Two-Step-Clustering dem zweiten Segment, welches die aktive Suche darstellt, zu. Andererseits werden von denjenigen Personen, welche K-Means in die aktive Suche einteilt, circa 5% vom zu vergleichenden Algorithmus als Fahrzeugliebhaber eingestuft.
Insgesamt werden circa 87% aller Personen von beiden Algorithmen in die gleiche Klasse eingeordnet, während um die 13% nicht eindeutig von beiden Clusteralgorithmen segmentiert werden.
In Allgemeinen lässt sich zu den übereinstimmenden Ergebnisen sagen, dass sich anhand der beiden Clusteralgorithmen circa 30% aller Nutzer in der frühen Kaufphase, 9% in einer Zwischenentwicklung und um die 4% in einem fortgeschrittenen Entscheidungsabschnitt befinden. Bemerkenswert ist, dass fast die Hälfte aller Beobachtungen als Fahrzeugenthusiasten gesehen werden, welche nicht am Kauf interessiert sind.
Übergreifend kann man sagen, dass die Klassifikationen der beiden voneinander unabhängigen Algorithmen sehr ähnlich sind. Die vier Segmente unterscheiden sich im Wesentlichen deutlich und ein Großteil der Nutzer kann klar in einzelne Gruppen eingeteilt werden.

5.4 Random Forest

Nachdem mit Hilfe der zwei Clusteralgorithmen einzelne Nutzertypen des gegebenen Datensatzes gefunden wurden, stellt sich nun die Frage, wie man neue Beobachtungen einteilen kann.
Hierfür bietet sich ein Ensemble mehrerer Entscheidungsbäume, ein Random Forest, an. Random Forest ist ein Klassifikationsverfahren aus dem maschinellen Lernen und liefert gerade bei vielen Einflussvariablen, wie es in dieser Arbeit der Fall ist, gute Ergebnisse [Cleve & Lämmel (2014)].
Grundsätzlich besteht der Random Forest aus mehreren unkorrelierten Entscheidungsbäumen. Für jeden Entscheidungsbaum werden Beobachtungen aus einer Bootstrap - Lernstichprobe verwendet und für jeden Knotenpunkt stehen üblicherweise nur $m \approx \sqrt{p}$ Einflussvariablen zur Verfügung, wobei p die Summe aller Einflussvariablen ist [Hastie et al. (2013)]. Die Stichprobe des Bruchteils an Prediktoren wird an jedem Knoten neu bestimmt. Die zufällige Auswahl an Entscheidungsknoten stellt sicher, dass die einzelnen Bäume nicht wiederholt aus den gleichen und wichtigsten Startvariablen aufgebaut werden, wie es bei „bagged-trees“ der Fall ist [Hastie et al. (2013)].
Durchschnittlich werden nach Hastie et al. (2013) $\frac{(p-m)}{p}$ Knoten die wichtigsten Zielvariablen nicht enthalten.
Folglich schließt diese Vorgehensweise die Korrelation einzelner Bäume aus.
In Abbildung 34 wird noch einmal verdeutlich wie sich die Korrelation zwischen den Bäumen in Abhängigkeit der Split-Variablen $\boldsymbol{m}$ verändert.

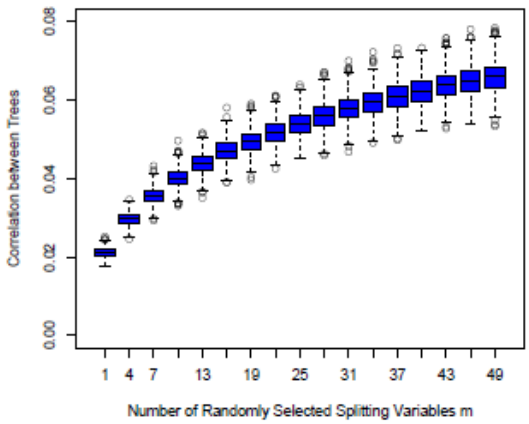


Abbildung 34: Korrelation zwischen Baumpaaren abhängig von den zufällig ausgewählten Split-Variablen $\boldsymbol{m}$ [Hastie et al. (2009)]

Wie man Abbildung 35 entnehmen kann besteht der Vorteil von Random Forest daraus, dass die Varianz mit weniger Variablen m, welche an jedem Knotenpunkt zur Verfügung stehen, reduziert wird.
Zu beachten ist bei dieser Grafik, dass die Werte der Varianz rechts abgetragen sind, während Werte des MSE und Bias links zu finden sind.
Zumal der Erwartungswert eines Baums gleich dem Erwartungswertes des Durchschnitts aller Bäume ist, verringert sich der Bias mit absteigender Split-Variablenanzahl nicht, was ebenfalls aus Abbildung 36 ersichtlich werden soll.
Da sich die mittlere quadratische Abweichung (MSE) aus Varianz plus $Bias^2$ berechnet, verläuft diese auch im oberen Bereich für wenige Split-Variablen.

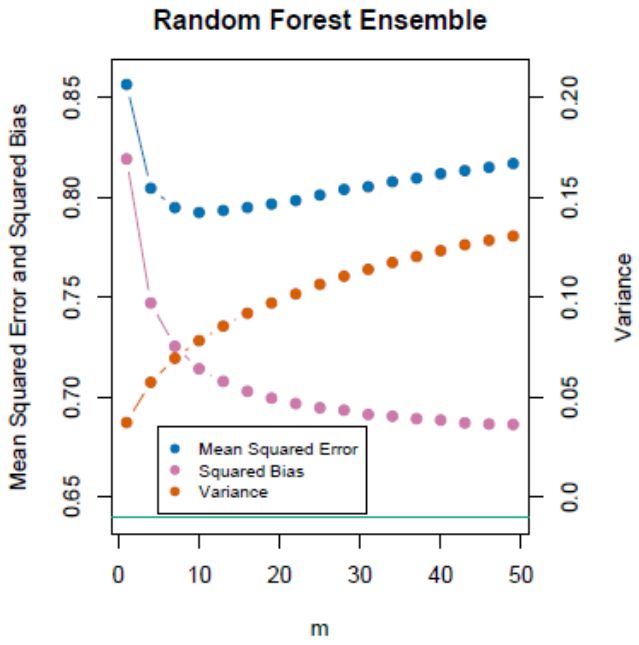


Abbildung 35: Varianz-Bias-MSE-Vergleich des Random-Forest Ensemble [Hastie et al. (2009)]

In Abbildung 36 wird exemplarisch ein Baum des Ensembles zur Klassifikation eines Nutzers anhand von drei Variablen verdeutlicht.
Die wichtigste Variable, welche den ersten Knotenpunkt darstellt, ist „relativer Anteil Bookmarks“. Falls mehr als 4% aller Aktionen für einen Nutzer das Anlegen von Merkzetteln ausmacht, wird dieser als Person in einer fortgeschrittenen Kaufphase klassifiziert.

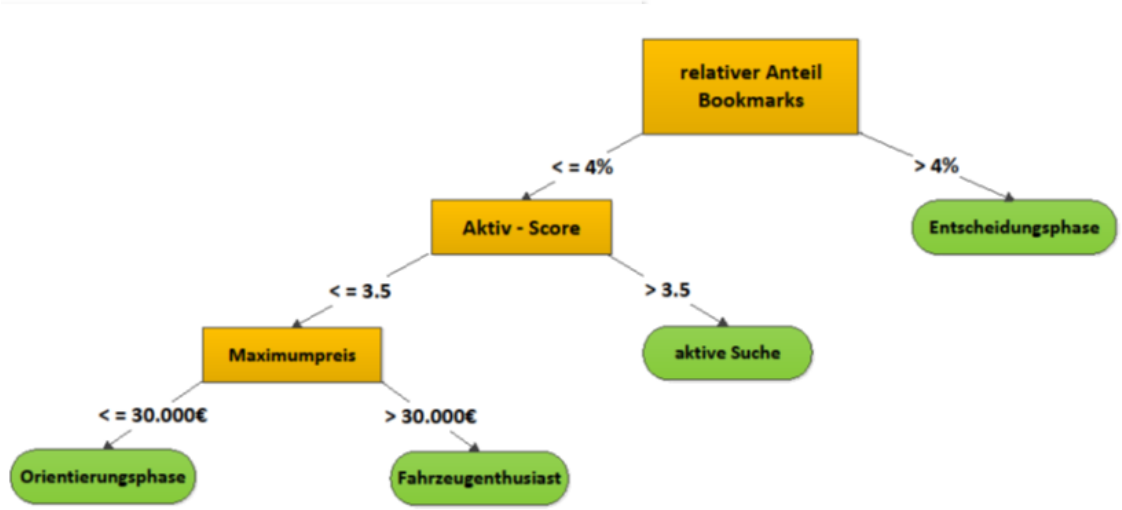


Abbildung 36: exemplarischer Entscheidungsbaum zur Nutzereinteilung

Ist dies nicht der Fall wird zusätzlich die zweite Variable „Aktiv-Score“ betrachtet. Handelt es sich um einen Nutzer, der einen hohen Score besitzt, welcher den Wert 3,5 übersteigt, wird dieser in die aktive Suche eingeteilt.
Hat diese Beobachtung jedoch einen prozentualen Anteil an Merkzetteln, welcher 4% nicht übersteigt und einen Aktiv-Score unter oder gleich 3,5, bezieht man eine dritte Variable, das Filterkriterium Maximumpreis, mit ein. Hat die zu klassifizierende Beobachtung im Laufe ihrer Suche eine minimale Obergrenze von 30.000€ überstiegen, wird sie den „Fahrzeugenthusiasten“ zugeordnet.
Falls dies nicht zutrifft, wird der Nutzer in die frühe Kaufphase eingeordnet.

Um nun einen Nutzertyp zu klassifizieren, werden alle Entscheidungsbäume mit den Eigenschaften dieses Nutzers einzeln analysiert. Eingeordnet wird nach dem Mehrheitsprinzip, folglich wird die Klasse, welche am häufigsten aus den einzelnen Bäumen resuliert, derjenigen Person zugewiesen.
In den nachfolgenden Abschnitten 5.4.1 und 5.4.2 wird versucht, die Klassifikation des K-Means beziehungsweise Two-Step-Cluster Algorithmus durch ein Random Forest Modell vorherzusagen.

5.4.1 Vorhersagekraft auf Basis der K-Means Ergebnisse

Um die Güte des Random Forest testen zu können, werden die Daten zunächst in Trainings- und Validierungsdaten zu einem Verhältnis von 60% zu 40% aufgeteilt. Das Random Forest Modell wird auf Basis der Trainingsdaten erstellt und unter Verwendung der predict-Funktion wird es auf die Validierungsdaten angewandt. Die Clusterzuweisungen des Random Forest Modells der Validierungsdaten werden in Tabelle 11 mit den Klassifikationen des K-Means Algorithmus verglichen.

k/rf	1	2	3	4	$\sum$
1	97,30%	0,96%	0,30%	1,44%	100%
2	8,65%	80,88%	0,43%	10,04%	100%
3	3,15%	2,55%	90,74%	3,56%	100%
4	0,94%	0,73%	0,15%	98,18%	100%

Tabelle 11: Klassifikationsvergleich K-Means - Random Forest, k:K-Means, rf: Random Forest

Der Tabelle ist zu entnehmen, dass Random Forest die Beobachtungen im Großteil richtig klassifiziert. Wie beim Vergleich zwischen den beiden Clusteralgorithmen K-Means und Two-Step-Clustering, existieren die meisten Fehlklassifikationen sowohl zwischen Cluster eins und zwei als auch zwischen Segment zwei und vier.
Durch Random Forest werden insgesamt 95,33% aller Beobachtungen richtig klassifiziert, wohingegen 1.602 Personen, welche circa 4% ausmachen, falsch segmentiert werden.
Die Parameter der randomForest-Funktion werden voreingestellt beibehalten, welche maximal 500 Bäume und 5 Split-Variablen an jedem Knoten zulassen.
Wie Anhang E zu entnehmen ist, ändert sich der prozentuale Anteil falsch eingeordneter Beobachtungen mit veränderten Parametern der Split-Variablen und der Anzahl der Bäume nicht in beachtlicher Form.
Nach der Klassifikation ist schlussfolgernd noch von Interesse, welche Variablen weit oben im Entscheidungsbaum stehen und daher ausschlaggebend für die Einteilung sind.
Um diese Frage zu beantworten wird nun Abbildung 37 betrachtet, welche die Wichtigkeit der Variablen im Random Forest Ensemble darstellt.

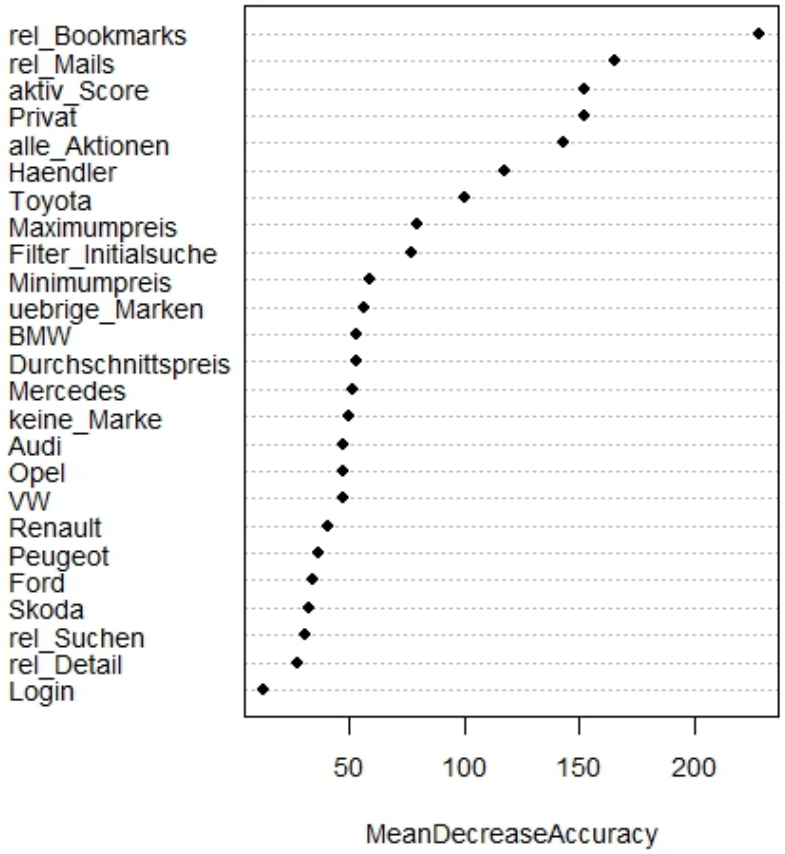


Abbildung 37: Variablenwichtigkeit / K-Means

Die Grafik zeigt jede Variable auf der y-Achse und die „mean decrease accuracy“ auf der x-Achse. Je höher der Wert der x-Achse für die jeweiligen Variablen, desto wichtiger sind diese für die Genauigkeit des Modells und desto mehr Einfluss hat diese darauf, die Fehlerrate zu erhöhen, falls die Variable nicht mit eingeht.
Die wichtigsten Variablen, welche weit oben im Baum vorkommen, sind die relative Häufigkeit an Merkzetteln und Mails, ob nach Privat- oder Händlerfahrzeugen gefiltert wird, sowie die Anzahl der Aktionen pro Tag, als auch der Aktiv-Score.

5.4.2 Vorhersagekraft auf Basis der Two-Step-Clustering Ergebnisse

Analog zu K-Means wird die Vorhersagekraft des Random Forest auch mit den Ergebnissen des Two-Step-Cluster Algorithmus verglichen. Oben stehende Funktionen werden ebenfalls auf Test- und Validierungsdatensätze der Two-Step-Clusterergebnisse angewandt. Die Matrix der falschen Klassifikationen liefert ähnlich niedrige Fehlklassifikationen wie in 5.4.1 (siehe Anhang F).
Wie Anhang G zu entnehmen ist, sind die wichtigsten Variablen für die Zuordnung, wie bei K-Means, ob ausdrücklich nach Händler- oder Privatfahrzeugen gefiltert wird, wie der prozentuale Anteil an Merkzetteln und Mails ausfällt und wie hoch der Aktiv-Score ist. Im Unterschied zu K-Means wird der Minimumpreis als drittwichtigste Variable erachtet, während dieser bei K-Means erst an zehnter Stelle steht. In Tabelle 12 werden jeweils die fünf wichtigsten Variablen der beiden Modelle verglichen.

K-Means	Two-Step-Clustering
Relativer Anteil der Merkzettel	Filter: Händlerfahrzeuge
Relativer Anteil der Mails	Relativer Anteil der Merkzettel
Aktiv-Score	Minimumpreis
Filter: Privatfahrzeuge	Aktiv-Score
Alle Aktionen	Filter: Privatfahrzeuge

Tabelle 12: Auflistung der fünf wichtigsten Variablen basierend auf Ergebnissen des K-Means- und Two-Step-Cluster Algorithmus

6 Fazit und Handlungsempfehlungen

Zusammenfassend lässt sich die Stichprobe der 85.743 beobachteten Nutzer durch vier Segmente beschreiben. Beide Algorithmen resultieren in annähernd identische Nutzergruppen, wovon drei die „consumer journey“, die jede Person vor dem Fahrzeugkauf durchläuft, und das vierte Segment die Fahrzeugenthusiasten beschreibt.
Das erste Segment, welches circa 35% der Grundgesamtheit ausmacht, wird charakterisiert durch Personen, die sich noch ganz am Anfang ihrer Suche befinden und sich einen groben Überblick über die Auswahl verschaffen möchten. Zweitere Gruppe hat die erste Phase schon durchlaufen und differenziert sich von den Übrigen durch genauere Vorstellungen vom gewünschten Fahrzeug. Diese Nutzer befinden sich bereits in einer aktiven Suche. Insgesamt weist diese Gruppe einen Anteil um die 15% aller

untersuchten Nutzer auf. Das dritte Cluster enthält mit einem prozentualen Anteil von 5% am wenigsten Nutzer, welche jedoch als am Wichtigsten erachtet werden: Diese Personen haben die ersten beiden Phasen bereits überwunden und befinden sich in einer Entscheidungsphase kurz vor dem Kauf eines Fahrzeugs. Theoretisch muss jeder kaufinteressierte Nutzer im Laufe seiner Suche die ersten drei Phasen mindestens einmal durchlaufen, kann aber jederzeit wieder in eine der ersten beiden Phasen zurückfallen, falls sich seine Vorstellungen ändern.
Personen des vierten und größten Segments, anteilig 45%, können als Fahrzeugliebhaber beschrieben werden, welche ausschließlich daran interessiert sind, Automobile anzusehen und (noch) nicht vorhaben, diese auch zu erwerben. Die Ergebnisse der Segmentierung werden in Abbildung 38 verdeutlicht.

Abbildung 38: Grafische Darstellung der vier Nutzergruppen

In Abschnitt 5.4.1 und 5.4.2 wurden die wichtigsten Variablen zur Bestimmung neuer Nutzer ermittelt, woraus resultierte, dass die Variablen „relativer Anteil an Merkzetteln“, „Filter: Privatfahrzeuge“ und „Aktiv-Score“ bei beiden zu vergleichenden Algorithmen unter den fünf wichtigsten Variablen vertreten sind und somit einen großen Einfluss besitzen. Aufbauend darauf könnten beispielsweise für die zehn wich-

tigsten Variablen Regeln der vier Nutzergruppen definiert werden und die einzelnen Personen anschließend durch Methoden des „supervised learning“ den vorgegebenen Klassen zugeteilt werden. Eine Idee zur Verbesserung der Werbestrategie wäre, den verschiedenen Nutzersegmenten gezielt angepasste Werbung auf der Webseite auszuspielen. Dies ist nicht nur ein Nutzen für das Unternehmen, sondern auch für die Besucher, da diesen gezielter Informationen zur Verfügung gestellt werden.
So könnten Nutzern der Orientierungsphase beispielsweise passend zu den Filtern, welche sie gesetzt haben, alternative (für ihn nicht in Erwägung gezogene) Fahrzeugmodelle vorgeschlagen werden, um sie in der Suche zu unterstützen. Außerdem hat der Automobilhersteller hier noch die Chance, die Aufmerksamkeit auf seine Marke zu lenken.
Nutzer, welche sich in der aktiven Suche befinden, könnten bei der Bewertung der einzelnen Fahrzeuge in Form von einer Preisanalyse oder Wertbeständigkeit unterstützt werden. Eine weitere Idee wäre, den Fokus auf neue Fahrzeuge zu lenken, da diese Nutzer sehr aktiv sind und die älteren Fahrzeuge bereits kennen. Dieser Aspekt könnte durch die Abänderung der Standardsortierung auf die Sortierung „Neueste Fahrzeuge zuerst“ erreicht werden. Übergreifend über die ersten beiden Phasen könnten diesen Personen in Kooperation mit externen Unternehmen auch Finanzierungsmöglichkeiten vorgeschlagen werden.
Nutzer, welche kurz vor dem Kauf stehen und sich in der Entscheidungsphase befinden, lassen sich möglicherweise nicht mehr von vorgeschlagenen Fahrzeugen überzeugen und sind vermutlich schon an einer Versicherung des Automobils interessiert. Dieser Zielgruppe könnte Werbung bezüglich verschiedener Versicherungsmöglichkeiten, in Absprache mit externen Firmen, vorgeschlagen werden. Des Weiteren könnte dieses Segment durch Verhandlungshilfen in Form von einer Liste mit wertbeständigen Equipments, sowie preistreibende oder -mindernde Eigenschaften des favorisierten Fahrzeugs unterstützt werden.
Fahrzeugliebhabern, welche vorwiegend teure Fahrzeuge ansehen, könnten abhängig von den Filtersetzungen neue Modelle der angesehenen Marken oder neue Testberichte oder Magazinartikel zu den angesehenen Fahrzeugen angezeigt werden. Außerdem könnten diese Nutzer zum Fahrzeugkauf angeregt werden, indem sie günstige Kaufzeitpunkte mitgeteilt bekommen.
Insgesamt gibt es zahlreiche Möglichkeiten die verschiedenen Nutzertypen anzusprechen, um somit zum einen den Suchenden auf seiner „Journey“ unterstützend zu begleiten und zum anderen dem Verkäufer die Möglichkeit zu bieten, die Nutzer individuell zu erreichen.

Literatur

AutoScout24 (2015), 'Unternehmensportrait autoscout24'. **URL:** `http://ww2.autoscout24.de/au-company/au-company-portrait.aspx` Zuletzt aufgerufen am: 10. August 2015.

Brosius, F. (2013), *SPSS 21*, Hüthig Jehle Rehm GmbH, Heidelberg. ISBN:978-3-8266-9454-7.

Chiu, T., Fang, D., Chen, J., Wang, Y. & Jeris, C. (1999), *A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment.* **URL:** `http://delivery.acm.org.emedien.ub.uni-muenchen.de/10.1145/510000/502549\\/p263-chiu.pdf?ip=129.187.254.47&id=502549&acc=ACTIVE\%20SERVICE&key=2BA2C\\432AB83DA15\%2EAF6136B168136FFE\%2E4D4702B0C3E38B35\%2E4D4702B0C3E\\38B35&CFID=542443909&CFTOKEN=82309108&__acm__=1441180282_feb98d7918\\dad32828ccdc1e9e3e88be` Zuletzt aufgerufen am: 2. September 2015.

Cleve, J. & Lämmel, U. (2014), *Data Mining*, Oldenbourg Wissenschaftsverlag GmbH. ISBN: 978 - 3 - 486 - 71391 - 6.

Enders, C. (2010), *Applied Missing Data Analysis*, Guilford Press, New York. ISBN: 978-1-60623-639-0.

Everitt, B. S. (1993), *Cluster Analysis, Third Edition*, John Wiley and Sons. ISBN:0-470-22043-0.

Fahrmeir, L., Künstler, R., Pigeot, I. & Tutz, G. (2004), *Statistik - Der Weg zur Datenanalyse*, Springer. ISBN:3-540-21232-9.

Fahrmeir, L., Tutz, G. & Hamerle, A. (1996), *Multivariate statistische Verfahren, 2. erweiterte Auflage*, Walter de Gruyter, Berlin, New York. ISBN:3110138069.

Flanagan, D. (2007), *JavaScript: das umfassende Referenzwerk*, O'Reilly.

Handl, A. (2010), *Multivariate Analysemethoden, 2. Auflage*, Springer. ISBN:9783642149863.

Hastie, T. J., Tibshirani, James, G. & Witten, D. (2013), *An Introduction to Statistical Learning*, Springer series in statistics, Springer, New York. ISBN:978-1-4614-7138-7.

Hastie, T. J., Tibshirani, R. J. & Friedman, J. H. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer series in statistics, Springer, New York. **URL:** http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf Zuletzt aufgerufen am: 20. August 2015.

Kopp, G. (2014), *Behavioral Targeting, Identifizierung verhaltensorientierter Zielgruppen im Rahmen der Online-Werbung*, disserta Verlag, Hamburg.

Liaw, A. & Wiener, M. (2002), 'Classification and regression by randomforest', *R News* **2**(3), 18–22. **URL:** http://CRAN.R-project.org/doc/Rnews/ Zuletzt aufgerufen am: 20. August 2015.

Müller, M. (2010), *Basiswissen Statistik*, W3L. ISBN: 3937137831.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. **URL:** http://www.R-project.org/ Zuletzt aufgerufen am: 10. August 2015.

Robinson, D. (2015), *broom: Convert Statistical Analysis Objects into Tidy Data Frames*. R package version 0.3.7, **URL:** http://CRAN.R-project.org/package=broom Zuletzt aufgerufen am: 15. August 2015.

Rutherglen, J., Wampler, D. & Edward, C. (2012), *Programming Hive*, O'Reilly Media, Inc. ISBN:978-1-449-31933-5.

SPSS (2006), The spss twostep cluster component, Technical Report TSCWP-0101. **URL:** http://www.spss.ch/upload/1122644952_The\%20SPSS\%20TwoStep\%20Cluster\\\%20Component.pdf Zuletzt aufgerufen am: 1. September 2015.

Steinhausen, D. & Langer, K. (1977), *Clusteranalyse: Einführung in Methoden und Verfahren der automatischen Klassifikation*, Walter de Gruyter Co. ISBN: 3 - 11 - 007054 -5.

van Groningen, M. (2009), 'Introduction to hadoop'. **URL:** http://blog.trifork.com/2009/08/04/introduction-to-hadoop/ Zuletzt aufgerufen am: 5. August 2015.

White, T. (2010), *Hadoop: The Definitive Guide, Second Edition*, O'Reilly Media, Inc., Canada. ISBN:978-1-4493-8973-4.

Wickham, H. (2009), *ggplot2: elegant graphics for data analysis*, Springer New York. **URL:** http://had.co.nz/ggplot2/book Zuletzt aufgerufen am: 15. August 2015.

Wickham, H. & Francois, R. (2015), *dplyr: A Grammar of Data Manipulation.* R package version 0.4.1, **URL:** `http://CRAN.R-project.org/package=dplyr` Zuletzt aufgerufen am: 15. August 2015.

Zhang, T., Ramakrishnan, R. & Livny, M. (1999), Birch: An efficient data clustering method for very large databases, Technical report, University of Wisconsin-Madison. **URL:** `http://www.cs.sfu.ca/CourseCentral/459/han/papers/zhang96.pdf` Zuletzt aufgerufen am: 1. September 2015.

Tabellenverzeichnis

Abbildungsverzeichnis

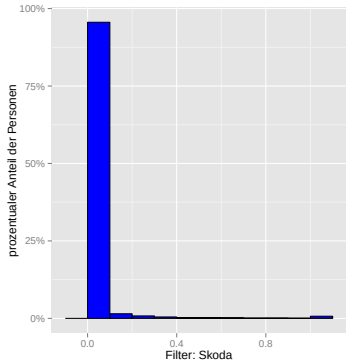
Appendix

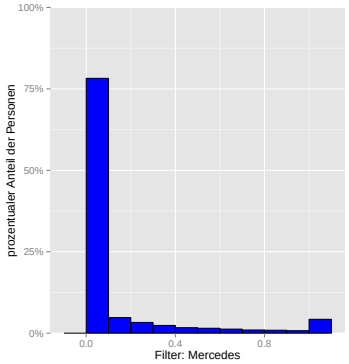
A Hive-Befehl zum Abfragen der Basisdaten aus dem Hadoop Cluster

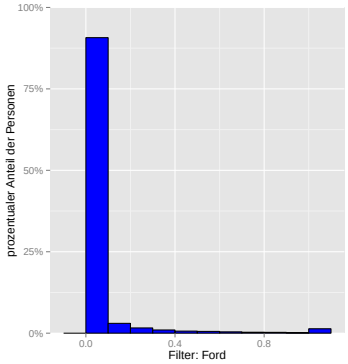
```
username <- 'ybarth'
system(paste('kinit ', username, '@AS24.LOCAL -k -t ~/', username,
             '.keytab && echo "kinit done" ||
              echo "kinit failed"', sep = ''))
library(RODBC)
library(sqldf)
# ODBC-Verbindung herstellen
Hive <- odbcConnect("Hive", readOnly = TRUE)
sqlQuery(Hive, "use temp")
sqlQuery(Hive, "show tables")
# Daten in eine Tabelle schreiben
userData <- sqlQuery(Hive,"
                     CREATE TABLE
                     hadoop_user_data_yvonne
                     STORED AS PARQUET
                     AS
                     SELECT
                     sdate, stime, csuristem, csuriquery, csreferer,
                     cshost, as24visitorguid
                     FROM weblogs.iislog_archive
                     WHERE time >= 20150503
                     AND time <= 20150517
                     AND as24visitorguid IN (
                     SELECT as24visitorguid
                     FROM weblogs.iislog_archive
                     WHERE time = 20150510
                     AND cshost LIKE '%autoscout24.de'
                     AND as24visitorguid != ''
                     LIMIT 300000)
                     ")
```

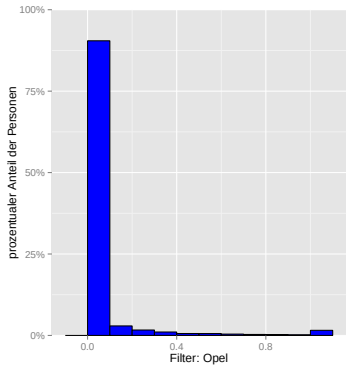
```
# relevante URL's aus der Tabelle extrahieren
userData <- sqlQuery(Hive,"
                     SELECT sdate, stime, as24visitorguid, csuristem,
                     csuriquery, csreferer, cshost
                     FROM temp.hadoop_user_data_yvonne
                     WHERE (csuristem = '/GN/CountV1.ashx' OR
                     csuristem LIKE '/classified/%' OR
                     csuristem = '/sendcontactmail/contact' OR
                     csuristem LIKE '/Parkdeck/Add/%' OR
                     csuristem = '/ArticleList/GetCounters' OR
                     csreferer LIKE '%login%' OR
                     csuristem = '/offerb2c/data/NewDecision/
                     Taxonomy/GetVehicleIdentificationData' OR
                     csuriquery LIKE '%tabg=guidedfull')
                     ORDER BY as24visitorguid
                     ")
# Testen, wieviele unterschiedliche Nutzer in Datensatz vorhanden
test <- sqldf("SELECT COUNT(DISTINCT as24visitorguid) FROM userData")

# Dataframe im csv-Format speichern
write.csv2(userData, file = "/data/R/export/yvonne_BA.csv")
```

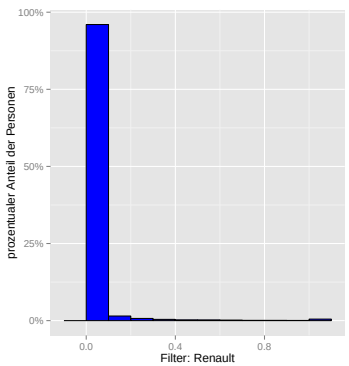
B Filterhäufigkeit einzelner Marken

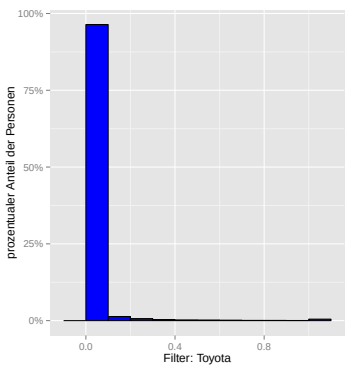


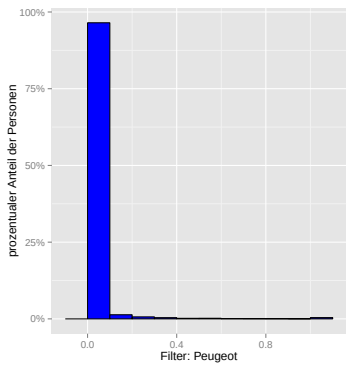


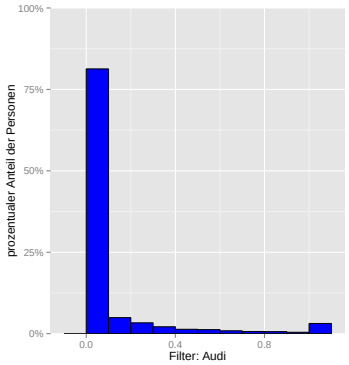


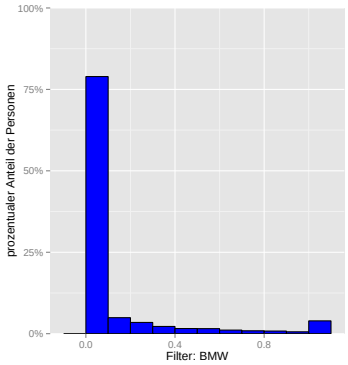


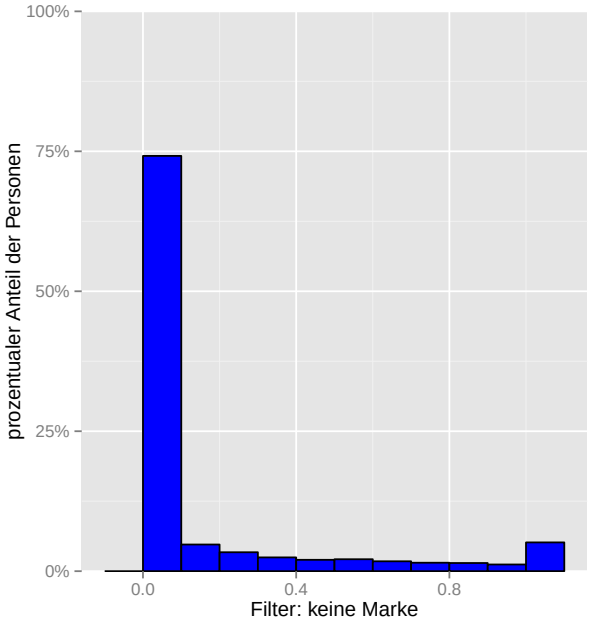


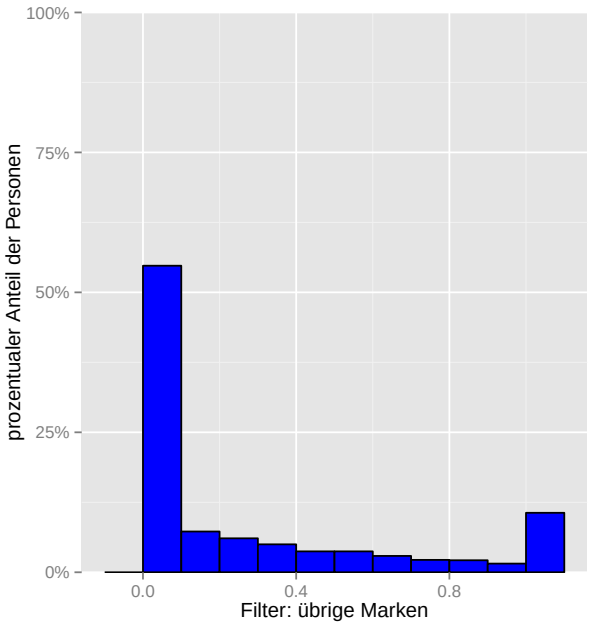




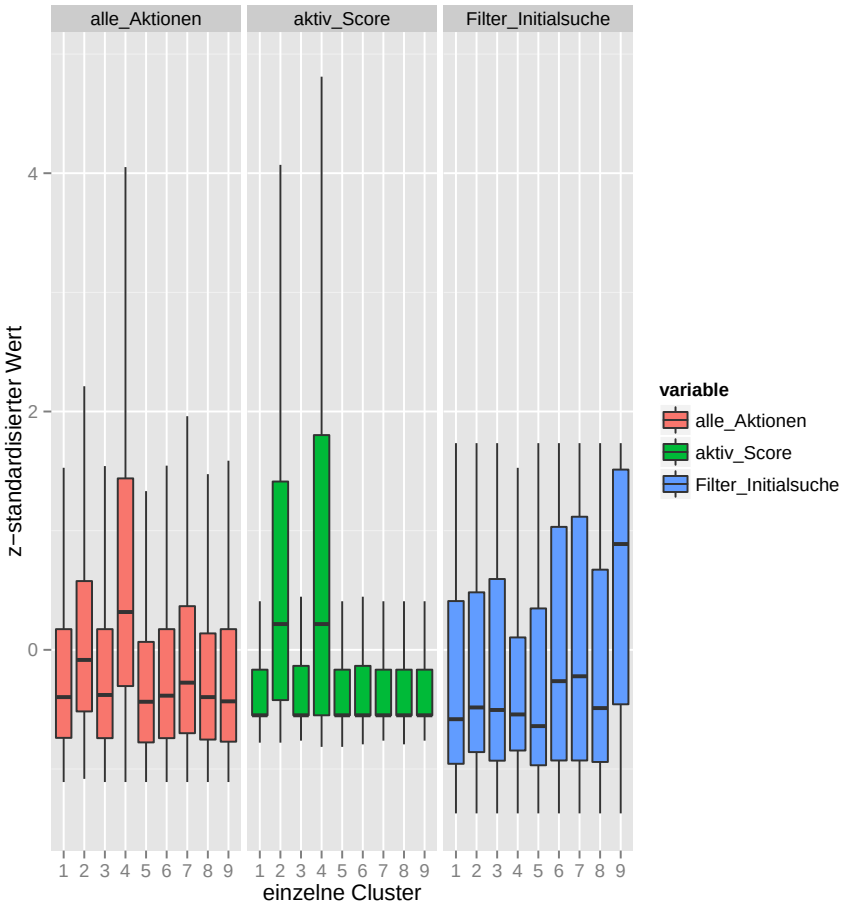


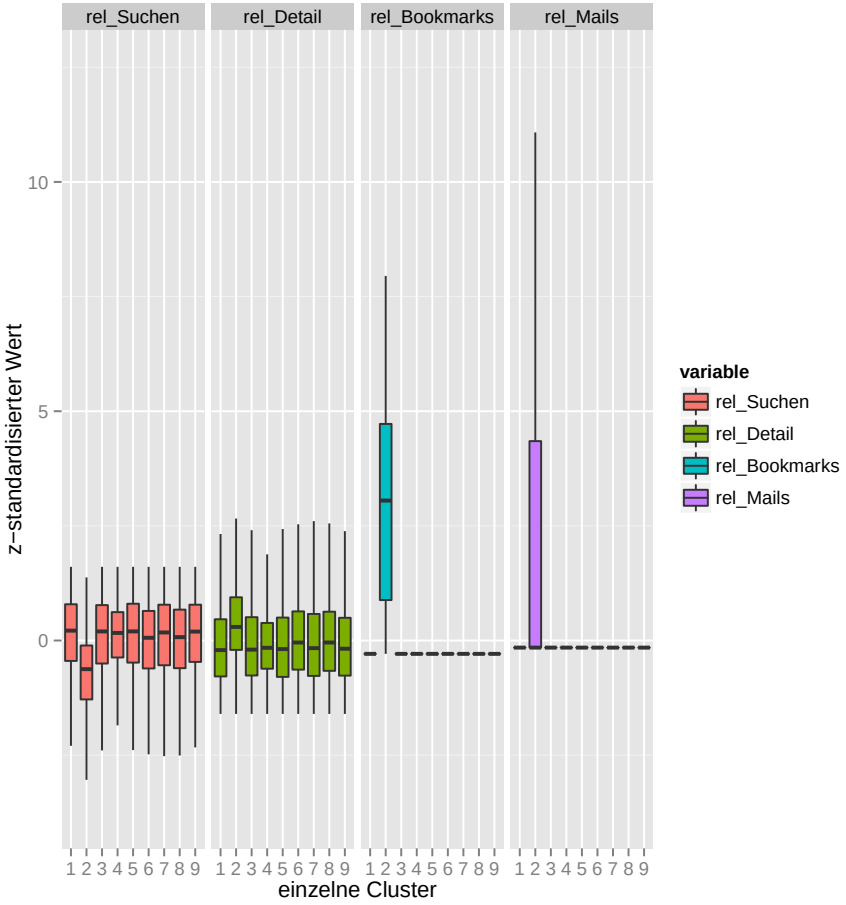


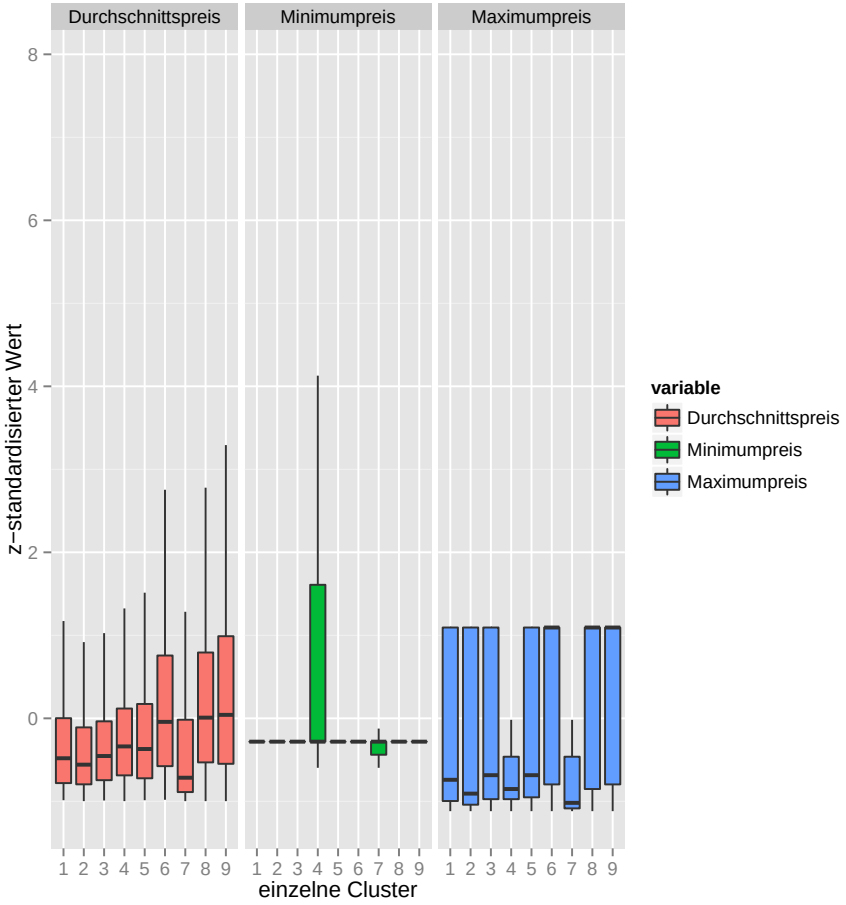

100%
75%
50%
25%
0%
prozentualer Anteil der Personen
0.0
0.4
0.8
Filter: keine Marke

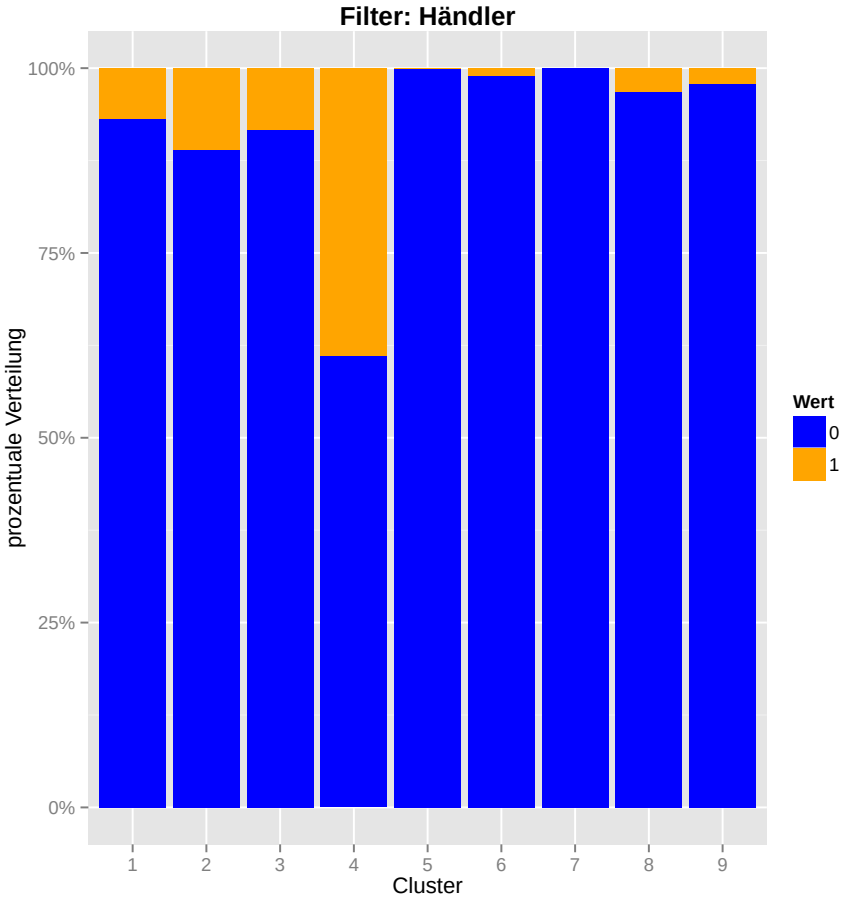

100%
75%
50%
25%
0%
prozentualer Anteil der Personen
0.0
0.4
0.8
Filter: übrige Marken

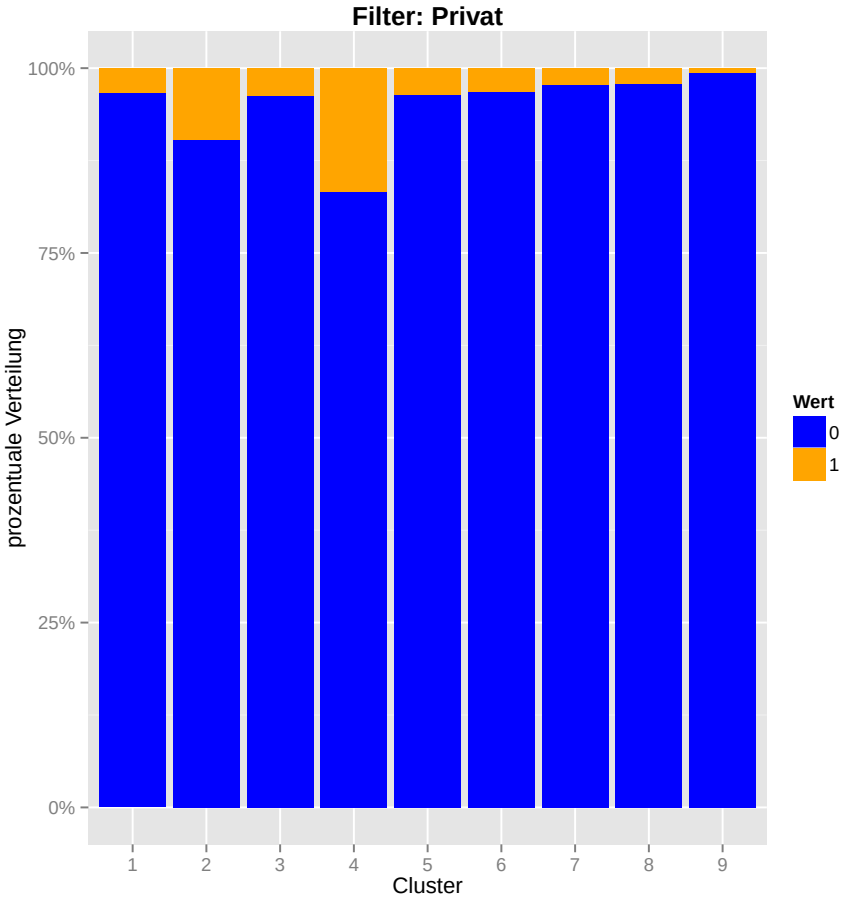
C Variablenunterschiede einzelner Cluster des Two-Step-Cluster Algorithmus

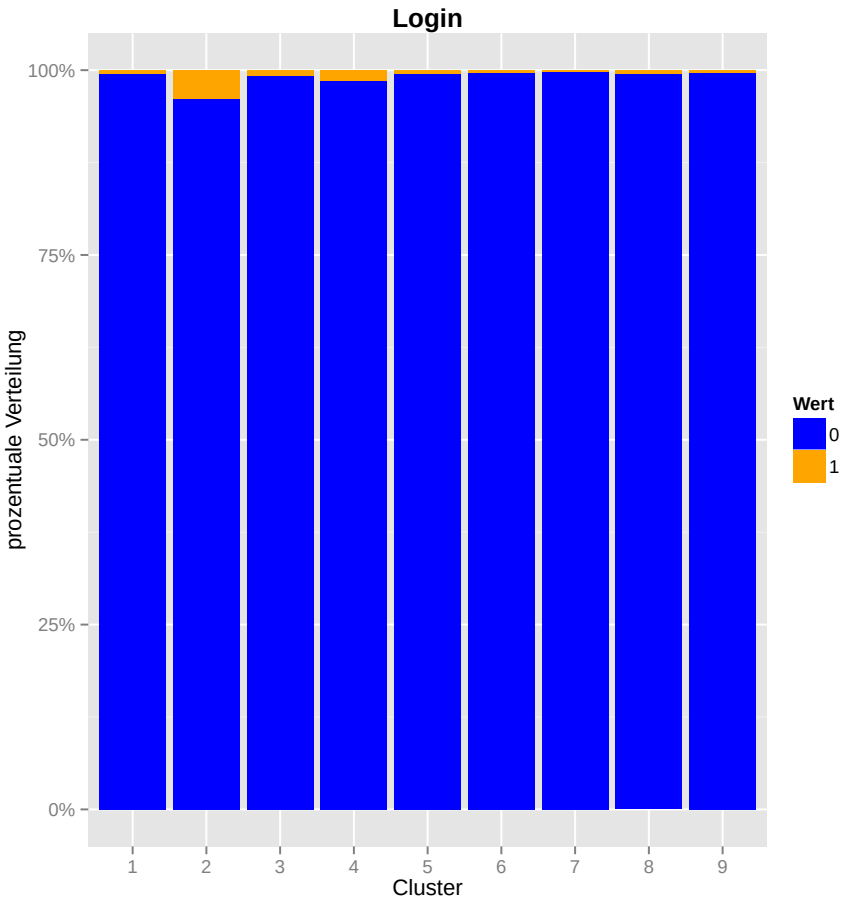


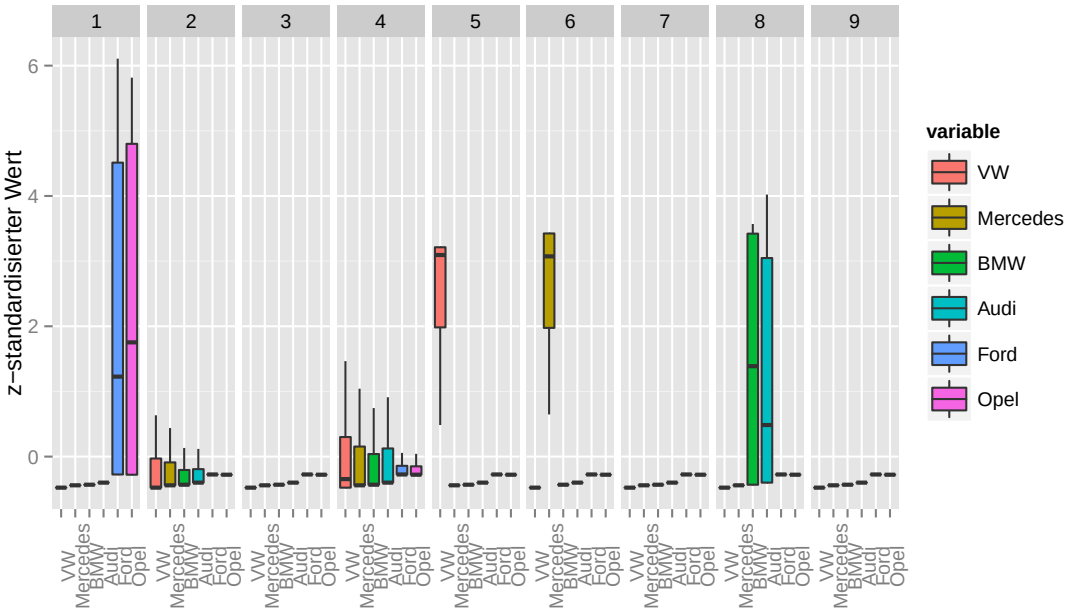

rel_Suchen
rel_Detail
rel_Bookmarks
rel_Mails
10
5
0
z−standardisierter Wert
1 2 3 4 5 6 7 8 9
einzelne Cluster
variable
rel_Suchen
rel_Detail
rel_Bookmarks
rel_Mails

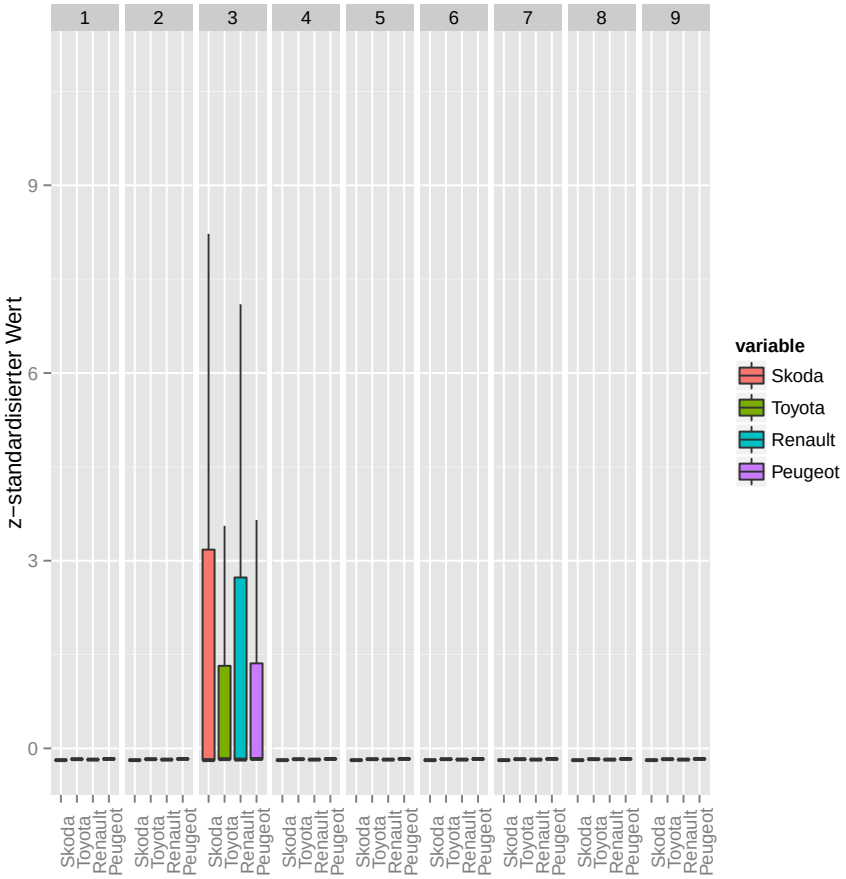

Durchschnittspreis
Minimumpreis
Maximumpreis
z−standardisierter Wert
8
6
4
2
0
1 2 3 4 5 6 7 8 9
einzelne Cluster
variable
Durchschnittspreis
Minimumpreis
Maximumpreis

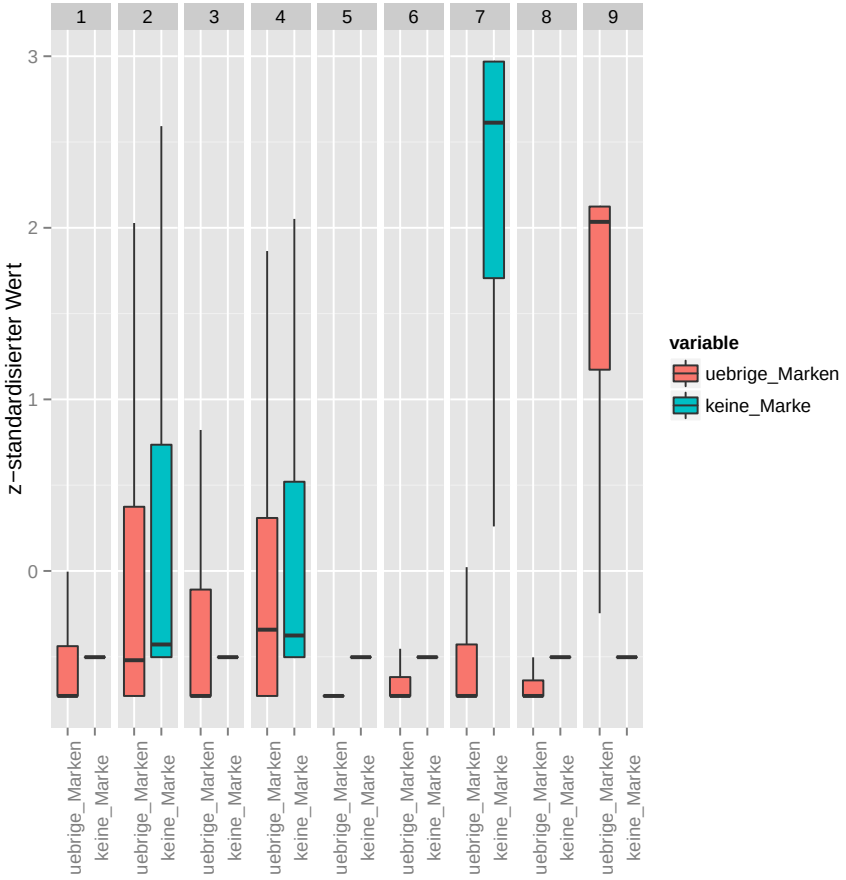

Filter: Händler
100%
75%
50%
25%
0%
prozentuale Verteilung
1
2
3
4
5
6
7
8
9
Cluster
Wert
0
1


Filter: Privat
100%
75%
50%
25%
0%
prozentuale Verteilung
1
2
3
4
5
6
7
8
9
Cluster
Wert
0
1

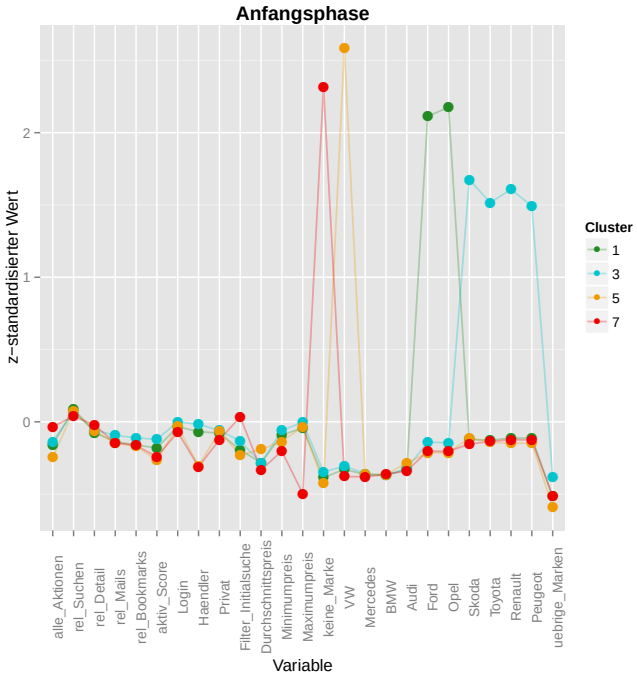

Login
100%
75%
50%
25%
0%
prozentuale Verteilung
1
2
3
4
5
6
7
8
9
Cluster
Wert
0
1

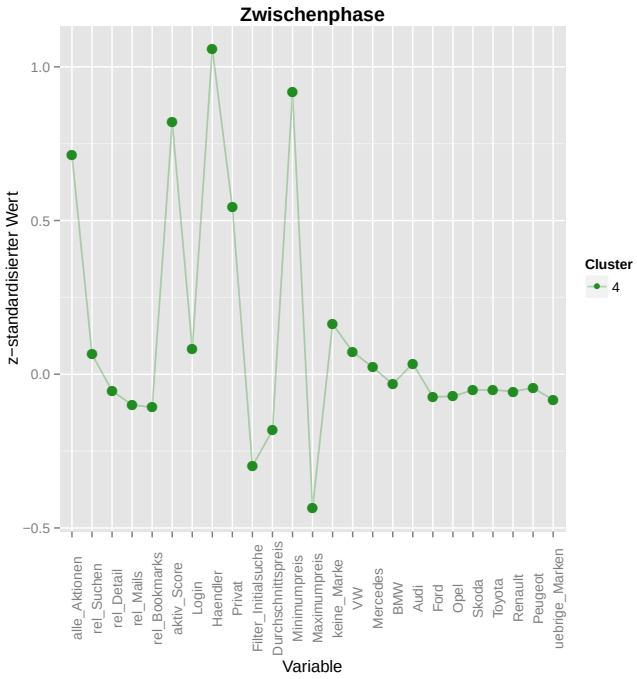

1
2
3
4
5
6
7
8
9
z−standardisierter Wert
6
4
2
0
VW
Mercedes
BMW
Audi
Ford
Opel
variable
VW
Mercedes
BMW
Audi
Ford
Opel

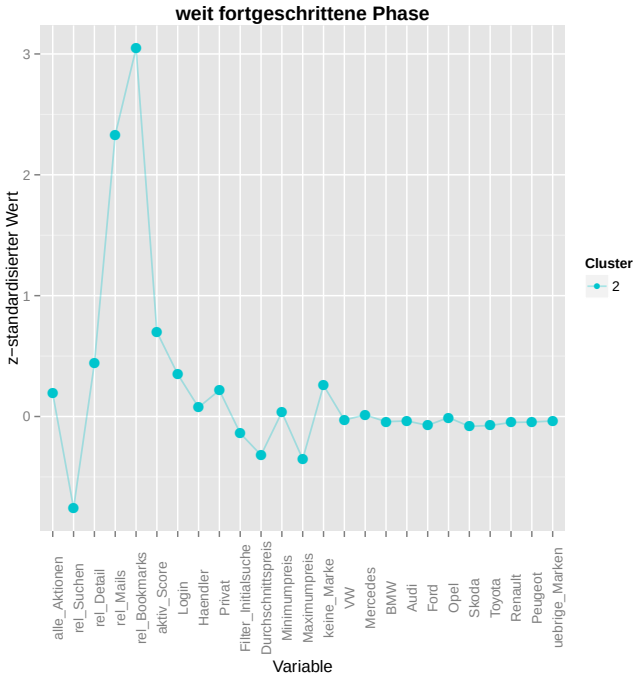

1
2
3
4
5
6
7
8
9
z−standardisierter Wert
9
6
3
0
Skoda
Toyota
Renault
Peugeot
variable
Skoda
Toyota
Renault
Peugeot

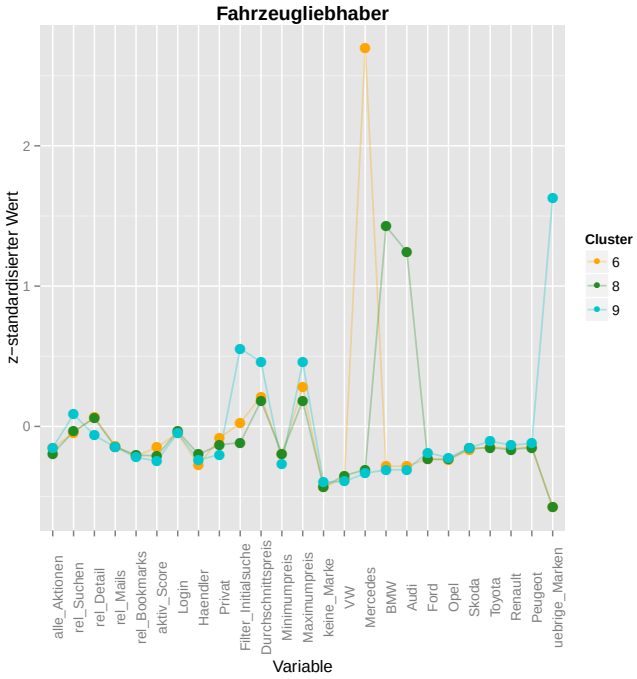

1
2
3
4
5
6
7
8
9
z−standardisierter Wert
3
2
1
0
uebrige_Marken
keine_Marke
variable
uebrige_Marken
keine_Marke

D Koordinatenplots der vier Nutzergruppen des Two-Step-Cluster Algorithmus






weit fortgeschrittene Phase
z−standardisierter Wert
3
2
1
0
alle_Aktionen
rel_Suchen
rel_Detail
rel_Mails
rel_Bookmarks
aktiv_Score
Login
Haendler
Privat
Filter_Initialsuche
Durchschnittspreis
Minimumpreis
Maximumpreis
keine_Marke
VW
Mercedes
BMW
Audi
Ford
Opel
Skoda
Toyota
Renault
Peugeot
uebrige_Marken
Variable
Cluster
2


Fahrzeugliebhaber
z−standardisierter Wert
2
1
0
alle_Aktionen
rel_Suchen
rel_Detail
rel_Mails
rel_Bookmarks
aktiv_Score
Login
Haendler
Privat
Filter_Initialsuche
Durchschnittspreis
Minimumpreis
Maximumpreis
keine_Marke
VW
Mercedes
BMW
Audi
Ford
Opel
Skoda
Toyota
Renault
Peugeot
uebrige_Marken
Variable
Cluster
6
8
9

E Klassifikationsvergleich K-Means - Random Forest mit veränderten Parametern

E.1 Anzahl Bäume: 500, Split-Variablen: 8

k/rf	1	2	3	4	$\sum$
1	97,12%	1,18%	0,32%	1,38%	100%
2	7,38%	83,72%	0,48%	8,42%	100%
3	3,23%	2,14%	91,41%	3,22%	100%
4	0,93%	0,87%	0,15%	98,05%	100%

E.2 Anzahl Bäume: 800, Split-Variablen: 5

k/rf	1	2	3	4	$\sum$
1	97,28%	0,97%	0,30%	1,45%	100%
2	8,4%	81,20%	0,43%	9,97%	100%
3	3,15%	2,55%	90,81%	3,49%	100%
4	0,93%	0,73%	0,14%	98,20%	100%

F Klassifikationsvergleich Two-Step-Clustering - Random Forest

F.1 Anzahl Bäume: 500, Split-Variablen: 5

t/rf	1	2	3	4	$\sum$
1	97,49%	1,04%	0,42%	1,05%	100%
2	4,23%	90,05%	0,65%	5,07%	100%
3	2,73%	3,90%	90,43%	2,94%	100%
4	0,80%	1,03%	0,30%	97,87%	100%

G Variablenwichtigkeit / Two-Step-Clustering