Projektbericht zum Modul Information Retrieval und Visualisierung Sommersemester 2021

Bike Buyers 1000

Floyd Spuhler

10.09.2021

Inhaltsverzeichnis

1 Einleitung

Durch das wachsende Umweltbewusstsein in der Bevölkerung rückt besonders das Fahrrad als klimaneutrales Verkehrsmittel in den Fokus [1]. Bereits 2019 betrug der Gesamtbestand an Fahrrädern in Deutschland fast 76 Millionen [2, 3]. Die anhaltende Covid-19 Krise verstärkt den anhaltenden Fahrrad-Boom für Freizeit und Pendeln [4]. Durch die Krise bedingte Lieferverzögerungen und Knappheiten führen zu einem Nachfrageüberhang und steigenden Preisen. Auch wärmere Winter begünstigen eine längere Fahrradsaison und sorgen dadurch zusätzlich für Lieferengpässe [5]. Neuentwicklungen wie E-Bikes werden immer beliebter, was sich an der Gesamtabsatzbeteiligung im Jahr 2020 von 38,7% bermerkbar macht [5]. In der Schweiz ist das Fahrrad nicht zuletzt durch den staatlich geförderten Ausbau von Fahrradspuren zum beliebtesten Verkehrsmittel geworden [4].
Fragestellungen: Welche Eigenschaften lassen sich aus den Bike-Buyers-Daten ableiten? Darstellungen: Scatterplot, Parallele Koordinaten

1.1 Anwendungshintergrund

Diese Forschungsarbeit bereitet Informationen auf, die Interessante Einblicke in das Kaufverhalten von Fahrrad-Kunden geben. So lassen sich anhand von Einkommensdaten, dem Alter und dem Bildungshintergrund Kundengruppen ableiten, auf welche in allen Wertschöpfungsebenen eingegangen werden kann. Hersteller müssen beispielsweise die Rahmengröße auf das Alter abstimmen. Einkommensstarke Kunden setzten den Fokus beispielsweise auf hochwertige Materialien (z.B. Carbon). Personen die das Fahrrad zum Pendeln nutzen sind eher auf ein Stadt- als ein Mountainbike angewiesen. Anhand der Distanz zur Arbeit können Bauunternehmen in Innenstädten gezielter Fahrradwege bauen. Diese Anwendungsfelder werden über die drei verschiedenen Visualisierungsanwendungen aufgegriffen.

Mit einem Scatterplot, bestehend aus X und Y Achse, können Zusammenhänge von Y in Abhängigkeit von X festgestellt werden. Diese Zusammenhänge werden in Form von Punkten, die zwischen beiden Achsen liegen und einen Achsenwert darstellen, visualisiert. Dabei fallen die Beziehungen zwischen den beiden Punkten je nach Verwendungszweck unterschiedlich stark aus [6, 7, 8]. Der Scatterplot wird für die Daten zu Fahrradkunden als erste Darstellung von Zusammenhängen verwendet.

Mit der zweiten Visualisierung, Parallelen Koordinaten, können Zusammenhänge multi-dimensionaler Daten besser als über Punkte dargestellt werden. Punkte werden dabei zu Achsenbeschriftungen und Linien [9]. (Inselberg) Jede betrachtete Variable wird nebeneinander angeordnet. Alle dazugehörigen Datenwerte werden über eine Linie minenander verbunden [10].

Die vertikalen Achsen sind über Linienn miteinander verbunden. Die Auf- und Abbewegungen der Linien zeigen Werteveränderungen auf [11]. Dabei sind die Achsen parallel zueinader angeordnet. Parallele Koordinaten bieten einen giten Datenüberblick. Eine Gefahr stellt allerdings die Überlappung von Linen dar, wenn zu viele Daten verwendet werden [12].

Über die Buamhierarchie lassen sich Daten und deren Beziehungen untereinander anordnen, wodurch eine übersichtliche Datenstruktur entsteht und sich Daten schnell wiederfinden lassen. Die Baumstruktur besteht aus Knoten und Kanten. Zwei Knoten sind jeweils über eine Kante miteinander verbunden. In der Baumstruktur muss ein Knoten vorhanden sein, der keinen Vorgänger hat. DIeser Knoiten wird Wurzel genannt. Dessen Folgeknoten werden Nachfolger genannt. Über die Wurzel führen nur azyklische Pdade und zu jedem Knoten nur ein Pfad. Durch die unterschiedlichen Pfade und Verzweigungen der unterschiedlichen Daten entsteht die Baumstruktur [13]. (S. 289) Diese Visualisierungstechnik eignet sich für die übersichtliche Darstellung bestimmter Daten aus den BikeBuyern, da diese Daten Strukturen aufweisen, die sich als Hierarchien darstellen lassen. Der hintergrund für die Auswahl dieser Form ist, dass sich die DAten in DFahrradkäufer und nicht-Käufergruppen unterteilen lassen. Des Weiteren sollen globale Unterschiede dargestellt werden, um auf territoriale Umfelder eingehen zu können. EIn weiter Faktoir ist die Pendlerdistanz vom Wohnsiotz der Befragten zur Arbeit, welche verschiedene Antwortmöglichkeiten zulässt. All diese Faktoren können Kombiniert in die Baumhierarchie übertragen werden.

1.2 Zielgruppen

Dieser Forschungsbericht richtet sich vor allem an die Anbieterseite auf den B2C Fahrradmarkt. Auf der Anbieterseite sind alle in der Lieferkette vorhandenen Unternehmensbrachen betroffen. Die Hersteller haben mit dem Materialmangel zu kämpfen.Den Fahrradverkäufern macht der Onlinehandel Konkurrenz und auch Bauunternehmen, die Fahrradspuren bauen haben mit Rohstoffmangel Probleme. Hierzu lassen sich drei Hauptzielgruppen, neben Fahrradinteressierten, herausfiltern, an welche sich dieser Visualisierungsbericht richtet.

- **Fahrradhersteller**:
  Fahrradhersteller benötigen besonders wegen der Materialknappheit spezifische Informationen zu den personenbezogenen Merkmalen potenzieller Kunden, wie z.B. Größe, Alter, Einkommen um einen Fahrradrahmen mit entsprechend wertigen / nicht wertigen Materialien für eine Verwendungszweck (z.B. Mountainbike, Stadtrad)herzustellen Für Fahrradhersteller sind Informationen zum Alter der Kundengruppe für Rahmengröße, Fahrradart, sowie zum Einkommen in Hinblick auf die Auswahl der Materialien und deren Qualität wichtig.

- **Fahrradhandel**:
  Für Fahrradverkäufer spielt vor allem der Verwendungszweck des potenziellen Kunden eine übergeordnete Rolle beim Fahrradkauf. Die Wahl des richtigen Modells unterscheidet sich für die Freizeit (Mountainbike) mit weiten Distanzen vom Gebrauch für die Stadt mit geringeren Distanzen (Stadtrad). Für weitere Distanzen eignen sich Mountainbikes besser als für die Fahrt in ebenerdigem Terrain, wie asphaltierten Straßen. für Stadträder. Optimalerweise betreibt diese Zielgruppe einen eigenen Onlineshop zum Fahrradvertrieb

und benötigt Informationen für die zielgerichtete Kundengruppenwerbung (ggf. Verweis einbauen?).

- **Bauunternehmen mit dem Fokus auf Fahrradinfrastruktur**:
  Auch für Unternehmen aus der Baubranche mit dem Fokus auf die Infrastruktur für Fahrradwege ist diese Arbeit eine geeignete Anlaufstelle für Informationen zum Einsatz des Fahrrads in Bezug auf den Arbeitsweg. Daten zu Pendlerwegen müssen für diese Zielgruppe besonders aufgegliedert vorliegen, da Bauunternehmen somit Informationen über die benötigten Distanzen neuer Fahrradwege erhalten und besonders in Städten nur begrenzt Raum zur Verfügung haben. Dadurch muss der EInsatz von Baumaschinen beonsders abgewogen werden.

Dieser Visualisierungsbericht ermöglicht es den oben genannten Unternehmen eine bessere Kundenmarktsegmentierung zu betreiben. Kurzfristig können durch die Ergebnisse dieses Berichtes Ressourcen sparsam eingesetzt werden (v.a. Hersteller, Baubranche). Langfristig können besonders Fahradhändler von diesem Bericht profitieren, da sie durch die personenbezogenen Daten optimale Kundenakquise / Kundenberatung garantieren können und anhand von Einkommensparametern Preise bestmöglich bilden können.

1.3 Überblick und Beiträge

Die durch Kaggle bereitgestellten Daten bestehen aus demographischen Kundeninformationen, wie Alter, Geschlecht, Familienstand etc. Diese Daten werden über die drei Visualisierungstechniken Scatterplot, Parallele Koordinaten und Baumhierarchie abgebildet, um den in 1.2 angesprochenen Zielgruppen einen Überblick in diese Kundendaten zu vermitteln.Über den Scatterplot können jeweils zwei Dateneigenschaften einander gegenübergestellt werden. Bei der Anwendung kann über Buttons selbst ausgewählt werden, welche beiden Eigenschaften angezeigt werden sollen. Mit den parallelen Koordinaten haben Interessenten die Möglichkeit über vertikale Achsen Werte miteinander zu vergleichen. Buttons ermöglichen hierbei die dynamische Achsenverschiebung. Die Baumhierarchie lässt die Daten anhand wesentlicher Eigenschaften in verschiedenen Ebenen darstellen.

2 Daten

Die diesem Projektbericht zugrundeliegenden Rohdaten entstammen einem Datensatz des "KaggleAccount von Heeral Dedhia [14], welche Antworten von 1.000 NutzerInnen zum Thema Fahrradkauf bereitstellt. Die Nutzerin hat diese Daten zuletzt im Jahresverlauf 2020 erweitert. Datum- und Erhebungsform sind hierbei unbekannt. In dieser aktuellsten Version liegen 13 verschiedene Attribute zu den 1000 befragten Personen vor. Die Nutzerin hat zwei verschiedene Datensätze bereitgestellt, die sich lediglich durch NA-Werte unterscheiden. Um bei einer Datenvorverarbeitung keine Daten zu vergessen und die Funktionsfähigkeit des Elm CSV-Decoders zu

gewährleisten, stellt die bereinigte Datei "bikebuyersclean.csv"die Grundlage für dieses Visualisierungsprojekt dar (ToDo: Kaggle Seite zitieren).

zu allen befragten personen wurde eine eindeugige ID vergeben, welche ein INT-Typ ist. Zur quantifizierung der Baumhierarchie wurde diese Tabellenspalte für die dritte Visualiserung übernommen. Die nächsten beiden Spalten "Marital Status", "Genderünd "Childrenünd Ägeäls Int geben als String-Datentyp Aufschluss über den sozialen Familienstand und Geschlecht der befragten Person. Die Spalten Ïncome", Ëducation", Öccupation"geben Auffschluss über die berufliche Karriere. In Verbindung mit den Spalten "HomeOwneründ "Cars"lässt sich der Status der Person interpretieren. Das Attribut "Commute Distance"gibt Aufschluss über die Distanz zur Arbeit"dient der Befragung zur Entfernung zwischen Wohnort und Arbeitsstätte, wodurch wertvolle Informationen zwischen Cars, Bikes und co gewonnen werden können. Durch purchasedBike und Region können diese Daten weiter voneinander unterschieden oder für einen globalen Einblcik als ganzes betrachtet werden.

Die Daten eignen sich besonders für Analysten von Fahrradunternehmen, die beispielsweise einen Online-Fahhrad-Shop betreiben wollen. Hierdurch erhalten sie eine Grundlage über mögliche Kundengruppen, wodurch wertvolle Informationen, wie die Arbeitsentferung vorhanden sind und sich insbesondere in der zukünftigen Infrastruktur von Großstädten bemerkbar machen werden. Auch in Hinblick auf die anhaltende covid-19 Krise und den sicheren Aspekt des Individualverkehrs bietet eiun Fahrad auf kurze bis mittlere Distanz eine umweltschonende und kostengünstige Alternative zum Auto. Diese Informationen in Verbindlung mit dem Alter, Einkommen und Beruf können individuelle Kundengruppen angesprochen werden.

Um eine geeignete Überblicksmöglichkeit über diese potenziellen Kundengruppen zu schaffen, musste der dafür notwendige Datensatz für die Baumhierarchie angepasst werden.

2.1 Technische Bereitstellung der Daten

Die dem Kaggle Account (hier Ztat) entstammenden, bereinigten Rohdaten in der Datei "bike buyers clean.csv"wurden in das Github Repository des Autors hochgeladen. Diese ist im Ordner "Daten zum Ladenäbgelegt. Für die beiden Elm Dateien SScatterplot.elmünd "Parallele Koordinaten.elm"wurden die vollständigen Daten der durch Kaggle bereitgestellten Datei als String in die jeweiligen ELM Dateien geladen. Dieses Vorgehen ermöglicht die dauerhafte Visualisierungdsdarstellung und ist unabhängig von Linkveränderungen. Das für die Baumhierarchie notwenige JSON-Format wird im Ordner "JSON"durch die Datei "Datenvorverarbeitungohne-CarWolrdwide.json"bereitgestellt und über einen Link in die entsprechende Elm-Datei "Baumhierarchie.elm"geladen.

Der zugrundeliegende Datensatz wurde um keine zusätzlichen Daten erweitert. Die Daten bilden eine gute Verteilung in verschiedenen Regionen ab, sind ausgewogen verteilt und bieten eine vielzahl an Informationen, mit denen Fahrradhersteller / Verkäufer wie Onlineshiops gezielt Kundengruppen ansprechen können.

2.2 Datenvorverarbeitung

Um die CSV Daten in den jeweiligen ELM Programme zu verwenden war für die Dateien SScatterplot.elmünd "ParalleleKoordinaten.elm"keine Datenvorverarbeitung notwendig. Die Rohdaten wurden in der jeweiligen Datei als String hinzugefügt und entsprechend decodiert. Die Datei "Visualisierung3 Vorverarbeitung.xlsxßeigt das Ergebnis, dass aus dem CSV-String der Rohdatei einzelne Excel Spalten gemacht wurden. Für die Datrstellungsziele der Baumhierarchie sind die Spalten ID, Cars, Commute Distance, Region und Purchased Bike notwenig. Die übrigen Datenattribute wurden für die Vorerarbeitung für die JSON-Transformation entfernt. Die wichigen Attribute sind in der Datei "Visualsierung3 Vorverarbeitung.csvënthalten. Die Spalte ID wurde für die letzte Hierarchieebene in eine neue Spalte zusammen mit "data id "übertragen, damit die Daten vom ELm Json Decoder erkannt werden. Dies erfolgte mit dem Excelbefehl "Verketten(...)". Anschließend wurden die Daten ausgewählt, welche 0 Autos aufweisen, damit der Effekt zwischen Fahrradkäufern und Nicht-Käufern und deren Pendler-Distanz vergleichbar wird. Für die JSOn Atei wurde die Länderliste aus der Übung als Vorlage für den Syntax genommen. Dabei wurden alle Länder raus gelöscht. In den Syntax wurden die Daten übertragen und um die Ebene mit den IDs aus der CSV- Datei erweitert.

Vorverarbeitung: Die Vorverarbeitung erfolgt im filtern leerer Felder

3 Visualisierungen

3.1 Analyse der Anwendungsaufgaben

Über den Scatterplot erhält besonders die Zielgruppe der Fahrradhersteller wichtige Informationen auf einen Blick, die bei der Produktion unterstützen können. Mit den Paralleln Koordinaten können die Zusammenhaänge der Zahlenwerte aus dem Datensatz besser untersucht werden. Die Röntgendarstellung ermöglicht es, Überschneidungsmuster gut zu erkennen. Hierdurch haben vor allem neue Fahradhandelsgeschäfte (auch mit Onlineshop) die Möglichkeit Ihren Kundengruppen zielgrerichteter Fahrräder anzubieten, wodurch die Value Proporsition un d die mit dem Fahrrradkauf verbundenen, positiven Erinnerungen stärker hervorgerufen werden können.

3.2 Anforderungen an die Visualisierungen

Im ersten Kapitel wurde die eingehende Motivation beschrieben, den verschiedenen Zielgruppen bestmögliche Anhaltspunkte zu finden, um die Fahrradkungengruppe nachhaltig an die jeweilige Unternehmensbranche zu binden. Die Designs müssen die angesprochene Übersichtlichkeit einhalten. Des Weiteren liegt ihr Wertversrpechen für die Zielgruppe im Aufzeigen von Zusammenhängen und vereinfachen von Datenmengen. Dabei müssen die Visualisierungen besondere Merkmale gut hervorheben, was auf Grund der Datensatzgröße (n=1000) eine Herausforderung darstellt.

3.3 Präsentation der Visualisierungen

In diesem Abschnitt werden die drei dem Projektbericht zugrunde liegenden Visualisierungen, der Scatterplot, Parallele Koordinaten und die Baumhierarchie.

3.3.1 Visualisierung Eins

Für die erste Darstellung wurde der Scatterplot gewählt. Dessen einfache Grundfunktionen bieten eine gute Einleitung in die Visualiserung der Fahrradwerte. Die in der Einleitung beschriebenen Merkmale, wie die gegenüberstellung der X und Y Achse erfolgt über die Datenwerte Alter auf der X Achse und in dieser Abbildung über Einkommen. Anwendende haben darüber hinaus die Möglichkeit, über die oben gelegenen Buttons die Y-Achse des Scatterplots dynamisch anzupassen. Dabei können Einkommen, die Kinder- und Autoanzahl des Datensatzes in Abhängigkeit des Alters dargestellt werden und Die Punkte stellen jeweils einen koordinatenpunkt aus dem BikeBuyers Index dar. Um die Zielgruppe mit Zusatzinformationen zu versorgen können beim Hovern mit der Maus über die Punkte neben dem exakten Einkommen auch Informationen ob ein Fahrradkauf getätigt wurde und welcher Berufsgruppe die befragte Person angehört, eingeholt werden. Die Wahl den Beruf beim Hovern über die punkte anzuzeigen, lässt sich damit begründen, dass String-Werte nicht in den Achsen dargestellt werden können. Die in der Ein-

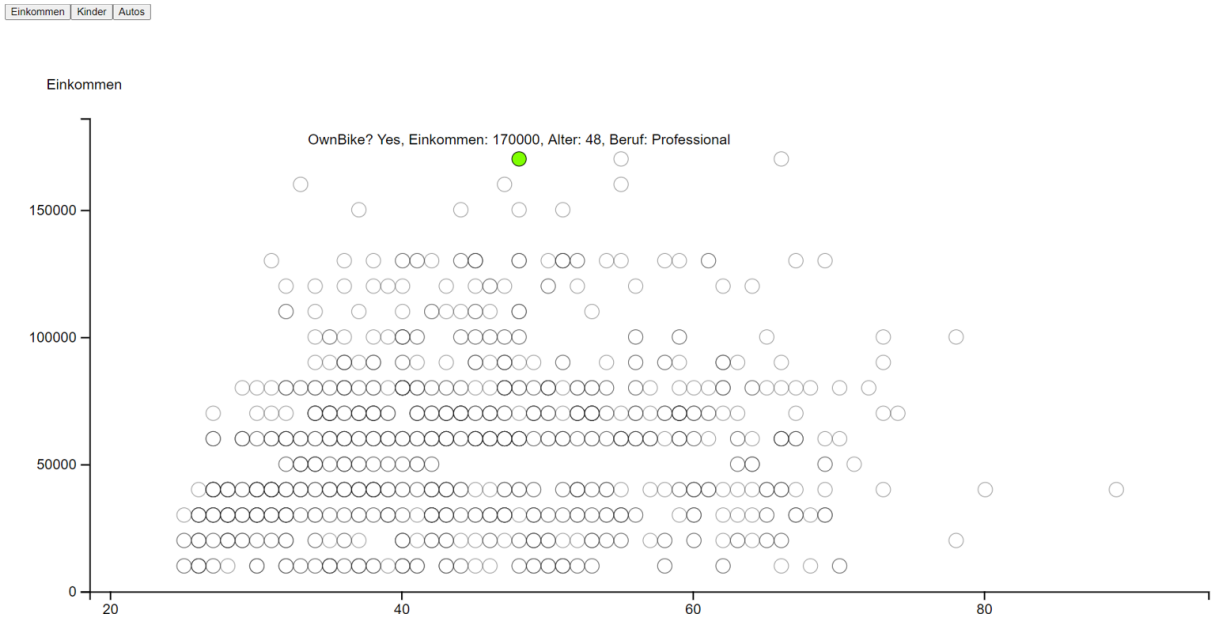

Abbildung 1: Scatterplot für Bike Buyers, Quelle: Eigene Darstellung

leitung herausgearbeiteten Scatterplot-Anforderungen können wie in Abbildung 1 entnehmbar, umgesetzt werden. Die beliebige Y-Achsenverändrung erweitert die Anforderungen darhingehend, das Anwendendende den Darstellungs-Inhalt und somit bestinmnte Schwerpunkte selbt setzten können. Stärker sichtbare Punkte bedeuten, dass Angaben der befragten Personen in die-

sen Punkten stärker übereinstimmen, als in helleren Punkten. Eine Alternative zum Scatterplot stellen die Zeitreihendiagramme dar, wodurch Zeitliche Verläufe anhand vorbestimmter Faktoren als Linien dargestellt werden. Da der vorliegende Datensatz keine zeitlichen Ausprägungen hat, welche chronologisch dargestellt werden könnten, sondern zeitlich unabhängige Werte beinhaltet wurde der Scatterplot als zweidimeinsionale Datenvisualiserung gewählt.

3.3.2 Visualisierung Zwei

Mit der zweiten Visualisierung, den Parallelen Koordinaten, können Im Gegensatz zum Scatterplot mehrdimensionale Zahlenwerte gleichzeitig, über vertikale Achsen ohne X Achse dargestellt werden. Für den Datensatz der bikeBuyers ergibt sich hierdurch den Vorteil, gleichzeitig alle relevanten Zahlenwerte in einer Visualisierung darzustellen, was sich in Abbildung 2 bemerkbar macht. Ebenso wie beim Scatterplot haben Anwendende die Möglichkeit die Darstellung über Buttons zu verändern. Durch die Buttons können die Achsen beliebig miteinander vertauscht werden und neue Zusammenhänge, je nach Betrachtungsziel,identifiziert werden.

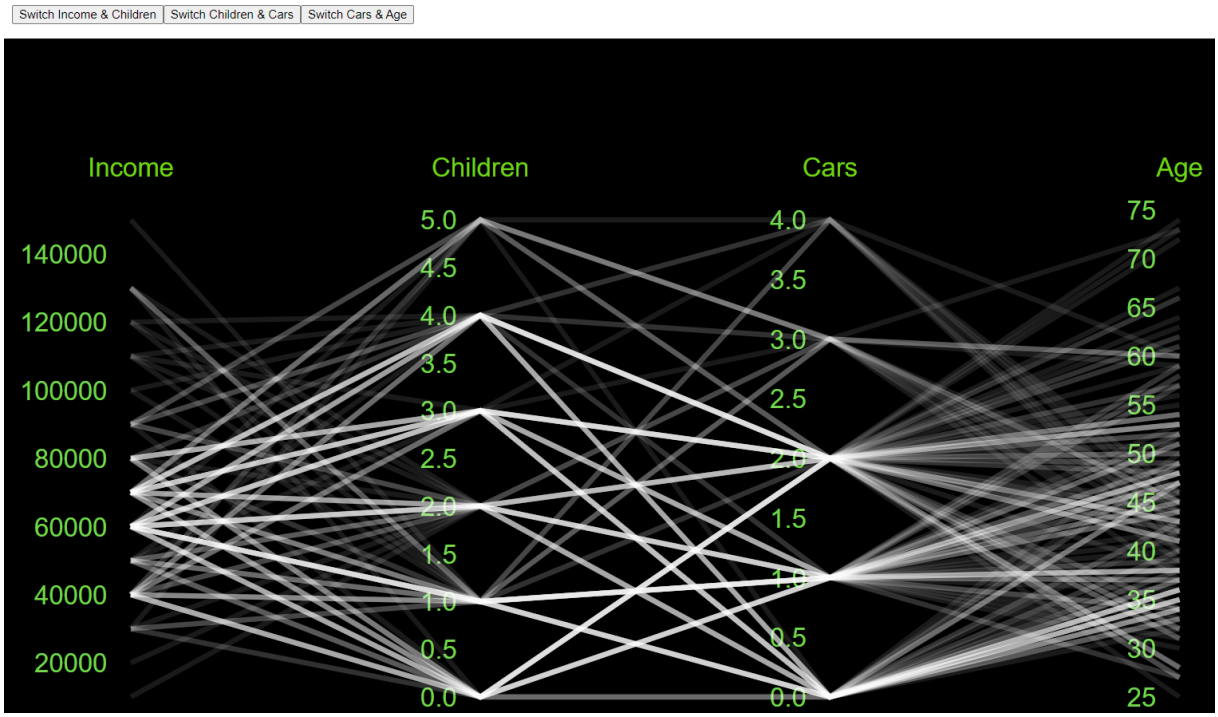

Abbildung 2: Parallele Koordinaten für Bike Buyers, Quelle: Eigene Darstellung

Die Abbildung 2 zeigt von links nach rechts den Verlauf eines mehrdimensionalen Punktes beginnend mit dem Einkommen, der Kinder- und Autoanzahl und dem Alter der befragten Person. Der Vorteil im Vergleich zum Scatterplot ist, dass mehrere Zusammenhänge und Auffälligkeiten gleichzeitig erkennbar sind. Im Direktvergleich mit der Darstellung 2 gibt es keine Hoverfunktion, wodurch die Linien individuell nach verfolgbar wären. Durch die Wahl des dunklen Hintergrunds

und der hellen Linienfarbe ergibt sich allerdings eine Röntgendarstellung. Durch die Linienüberlappung werden Identifikationsmuster sichtbar, was sich durch die kräftigeren Weißtöne erkennen lässt. Darüber hinaus bietet die durchgängige Achsenbeschriftung gute Anhaltspunkte für den Linienverlauf. Das Vorfiltern nach Region und Fahrradkauf im ELM-Programm verhindert eine "Überdarstellung". Somit wurden nicht nur alle Anforderungen gemäß der Theorie richtig abgeleitet, sondern auch die Gefahr der "Übervisualisierungërkannt und verhindert. Neben Parallelen Koordinaten gibt es weitere Visualisierungsmöglichkeiten mehrdimensionaler Daten. Da Scatterplots, am besten für zweidimensionale Daten geeignet, bereits verwendet wurden, stellen Projektion und Sektion, Sternkoordinaten, K-Means und Datentinte eine Alternative zu Parallelen Koordinaten dar. Mit Projektionen und Sektionen soll der mehrdimensionale Raum abgebildet werden. EIne hierfür anfallende hohe Anzahl an benötigten Darstellungen kommt in Bezug auf die Übersichtlichkeit und Anschaulichkeit für die Zielgruppen nicht in Frage. Mit Sternenkoordinaten können mehrdimensionale Daten in 2D oder in 3D abgebilfet werden. Der Name ist charakteristisch für die Achsenanordung. Auf Grund der ungewöhnlichen Erscheinung und der besseren intuitiven Nachvollziehbarkeit der Fahrradkäuferdaten wurde auf die Parallelen Koordinaten zurückgegriffen. Die Visualisierungstechniken K-Means und Datentinte fokussieren die Visualiserung durch Vorabberechnungen auf wenige Kerngedanken und verhindern somit einen Überblick über eine breitere Sichtweise, welche für alle Zielgruppen anvisiert wurde. Aus diesen Gründen wurde die Parallele Korrdinaten Visualiserung gewählt.

3.3.3 Visualisierung Drei

Als dritte Visualiserungstechnik wurde die Baumhierarchie ausgewählt. Diese stellt ein klassisches Baumdiagramm, wie in der Theorie beschrieben dar, durch welches die Fahrradkäuferdaten in eine Hierarchie gebracht werden. Die Knoten stellen hierbei Entscheidungsmöglichkeiten dar, welche durch die Linien miteinander verbunden werden. Die erste Unterscheidung besteht beim Fahrradkauf. Hierbei werden die Angaben in Ja oder Nein unterteilt. Anschließend erfolgt die Unterteilung nach den Regionen der befragten Personen. Die letzte Verweigung stellt die Pendlerdistan vom Wohnort zur Arbeit dar.

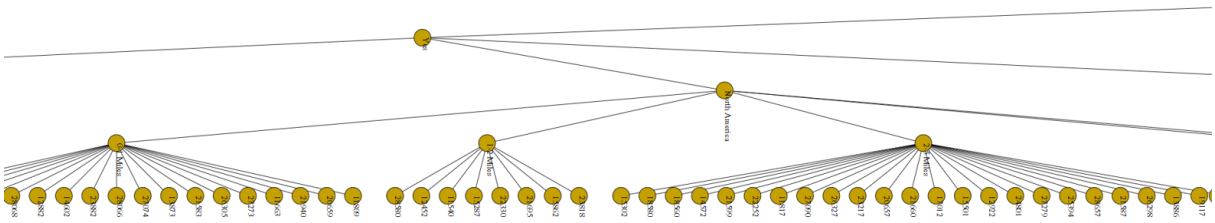

Abbildung 3: Ausschnitt aus der Baumhierarchie für Bike Buyers, Quelle: Eigene Darstellung

Die Abbildung 3 zeigt einen Ausschnitt aus der umfangreichen Baumdarstellung. Hierbei lassen

sich die eingefärbten Knoten gut unterscheiden. Im Vergleich zu den beiden vorangegangenen Visualiserungen wurde auf eine Interaktionsmöglichkeit verzichtet. Die angeleiteten Baumhierarchie-Bedingungen, nämlich die übersichtliche Darstellung und Einteilung in Gruppen konnte umgesetzt werden. Trotz Vorfilterung der Parameter, dass die befragten Personen keine Autos besitze, ist die Baumstruktur sehr umfangreich, was zu Überlappungen in der HTML-Darstellungen in der letzten Verzweigung geführt hat. Aus diesem Grund wurde auf weitere Parameter (wie Beruf, oder Bildungsgrad) verzichtet, da diese teilweise über den Scatterplot gezeigt werden. Neben der Möglichkeit die dritte Visualisierung als Baum darzustellen, gibt es weitere Kategorien, die die Darstellung komplexer machen. Hierunter fallen verschiedene Graphenanwendungen, die Mehrfachzuweisungen auf Knoten ermöglichen. Auf Grund der Zielsetzung, abgestuft aufzeigen, welche Pendlerdistanzen auftreten, wurde auf eine komplexere Darstellungsmethode verzichtet.

3.4 Interaktion

Für die Visualisierungen Scatterplot und Parallele Koordinaten wurden zwei verschiedene Interaktionsmöglichkeiten in den jeweils zugrunde liegenden Code integriert.

Bei der Visualiserung des Scatterplots haben Anwendende die Option, die Gegenüberstellung des Alters auf der X Achse mit einem beliebig verfügbaren Zahlenwert des Bike-Buyers-Datensatzes auf der Y-Achse über das Klicken auf die verschiedenen Buttons anzeigen zu lassen. Dadurch kann der Fokus individuell auf bestimmte Sachverhalte gelegt werden, was insbesondere für Fahrradhersteller und Fahrradhändler interessant ist. Um die eingeschränkte Funktionalität der zweidimensionalen Zahlengegenüberstellung zu erweitern, wurde die Hover-Funktion integriert. Durch das Hovern über den Punkten des aufgespannten Koordinatensystems werden auch String-Werte als Informationen, wie der Beruf ausgegeben. Daneben lassen sich die exakten Datenwerte beim Einkommen anzeigen, welche als Achseneinteilung zu ungenau wären.

Die zweite Visualsierung, Parallele Koordinaten, bietet Interessenten die Möglichkeit über Buttons die Achsen nach belieben zu verschieben. Je nach Einstellung lassen sich Zusammenhänge des mehrdoímensionalen Punktes hervorheben, was durch eine eindeutigere Überlappung der hellen Linien sichtbar wird. Anwendende werden dadurch ermutigt, selbst die Daten in eine für sie interessante Sichtweise zu bringen.

Für das Baumdiagramm wurde auf eine Interaktionsmöglichkeit verzichtet, da das Visualsierungsziel über eine Version erfüllt ist.

Auf eine Interaktion der Visualiserungen untereinander wurde verzichtet, da je nach Visualiserungszweck Datenmerkmale hervorgerufen werden sollen, die durch eine gemeinsame Verknüpfung nicht funktionieren würden.

4 Implementierung

Die Ausgangsbasis für diesen Projektbericht stellen die Programmcodes aus der Übung dar. Für die Scatterplot Visualiserung wurden die im Rahmen der Übungen eins, drei entwickelten

Codegerüste als Grundlage verwendet. Für die Parallelen Koordinaten ist die Übung 7 der Ausgangspunkt. Bei der Baumhierarchie galt die Übung 10 als Orientierung.

Während der frühen Entwicklungsphase zu Beginn des Projektes wurde versucht, mit Hilfe des CSV-Decoders nach dem Vorbild aus Übug 8 die Daten zu laden. Allerdings ist der unübersichtliche Aufbau des Decoders nicht für die Projektarbeit mit den Fahrraddaten geeignet. Hierzu sind weitere Decodr Funktionen notenig, Daten können nicht direkt nach ihren natürlichen Datentypen decodet werden, etc. Bereits nach der ersten Sichtung des Bike-Buyers-1000 Datensatzes wurde überlegt, die Daten als String in das Codeprogramm zu laden. damit verbundene Vorteile wurden im Kapitel 2 aufgeführt. Nach intensiver Alternativenrecherche wurden die Programmcoides für den Scatterplot und die Parallelen Koordinaten nach dem Vorbild des CSV-Decoders von Brian Hicks aufgebaut. Dieser besteht aus einer Funktion, welche problemlos Strings, Ints und Floats decodet. Drüber hinaus ist der Decoder explizit für das Decoden eines CSV-Strings geeignet.

Eine weitere Herausforderung stellt das Übertragen der Codefundamente aus der Übung auf den neuen Sachverhalt mit den Fahrraddaten dar. In der Übung stellt der Datensatz einen für ELM optimal geeignet Datensatz dar, über welchen eigene Datentypen problemlos über "type"definiert werden können und anschließend damit filterungen durchgeführt wrden können.

Der Code wurde in mehren Phasen erstellt. Nach erfolgreichen Decoder Tests wurden die Daten mit den zugrundeliegenden Funktionen für den Scatterplot verbunden. Komplikationen mit dem Einbinden des Decoders in die main Funktionen konnten gelöst werden. Nach Erweiterung und individueller Anpassung des Codes konnte ein Scatterplot komiliert werden. Hierbei war die Hoverfunktion bereits integriert. Die Überlegte weiterentwicklung mit Buttons konnte über die init model view update Elm-Grundgleiderung umgesetzt werden. Wie sieht die Elm-Datenstruktur für das Model aus, in dem die verschiedenen Zustände der Interaktion gespeichert werden können. Das Vorgehen wurde bei den Parallen Koordinaten ebenfalls angewendet. Nach erfolgreicher Implementierung und Verknüpfung des Decoders, Anpassen / Erweitern der Daten, inklusive neuer Vorfilterung konnte die Darstellung umgesetzt werden. Nach Abwägung der vOR- UND Nachteile der Röntgendarstellung im vergleich zu "normalem"Hintergrund, wurde die Röntegendatrstellung ausgewählt. Die Weiterentwicklung wurde mit dem Ziel umgesetzt, dass sich die Achsen beliebig verschieben lassen. Diese Umsetzung konnte anhand der Übung 7 erreicht werden. Bei der Visualisierung der Baumhierarchie gab es bis auf die Datenvorverarbeitung keine Probleme mit dem JSON-Decoder und dem verfügbaren Programmcode aus der Übung 10.

5 Anwendungsfälle

Im folgenden werden für die drei Darstellungen in Hinblick auf die Zielgruppe praktische Anwendungsfälle aufgezeigt.

5.1 Anwendung Visualisierung Eins

Grundsätzlich werden über die Hoverfunktion beim Scatterplot wichtige Zusatzinformationen geliefert, die genauer über die Buttons angezeigt werden können. DIe erste Möglichkeit vergleicht das Alter mit dem Einkommen.

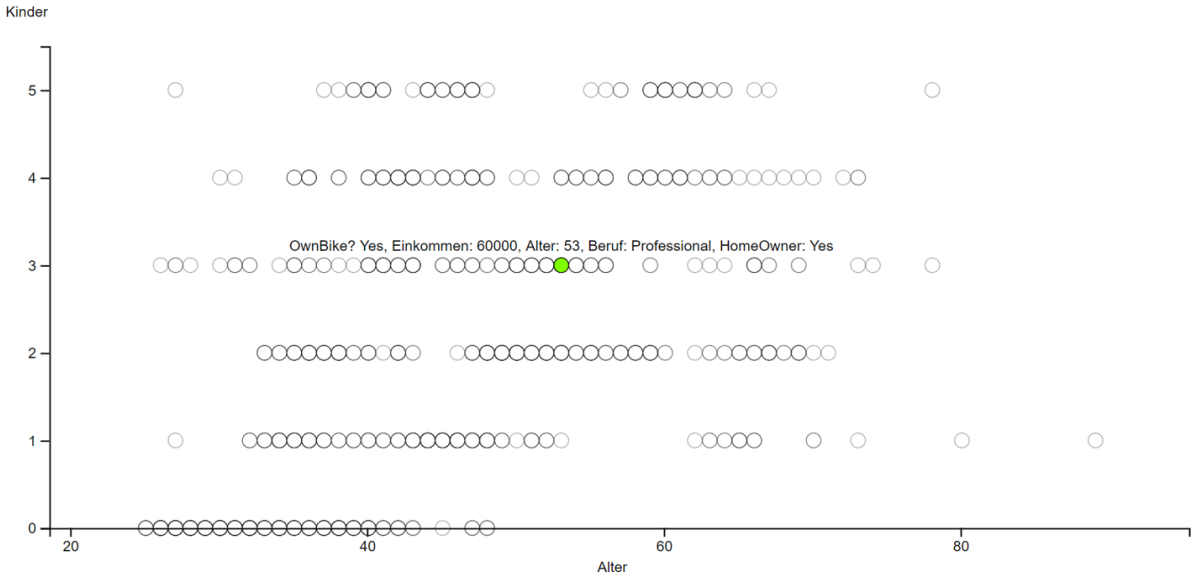

Abbildung 4: Anwendung Scatterplot, Quelle: Eigene Darstellung

In der Abbildujg 4 ist ein Scatterplot abgebilkdet, welcher sich besonders für die Zielgruppe des Fahrradhandels richtet. Durch Informationen, ob Kaufinteressierte eines Fahrrads ein Eigenheim besitzen kann auf eine Familie geschlossen werden, welche potenziell Platz für mehrere Fahrräder hat. Durch eine intensivierung der Kundenbeziehung, wie z.B. Rabtte beim Kauf eines zusätzlichen Kiunderfahrrads kann generationenübergreifend Fahrräder verkauft werden. Des weiteren kann aus dem Scatterplot entnommen werden, dass ab einem Alter von 30 Jahren die Kinderanzahl zunimmt und davor überwiegend zwischen 0 und 1 liegt. Mit der Zusatzinformation, ob die kaufinteressierte Person ein Eigenheim besitzt kann aber auch eine potenziell einkommensstarke Kundengruppe eingeordet werden, wodurch höherwertigere fahrräder für verschiedene Verwendungszwecke beworben werden können. Durch die Hover-Anzeige des Einkommens wird ersichtlich, dass Personen mit 4-5 Kindern ein hohes Einkommen aufweisen.

5.2 Anwendung Visualisierung Zwei

Die zweite Anwendung stellt die in Abbildung 5 gezeigte umstellung der Achsen dar. Diese ist besonders für Fahrradhersteller geeignet. Im Elm programmcode wurden Daten herausgefiltert, die "neinßum Fahrradkauf und als region nicht Europe angeben haben. Durch die erste Filterung lässt sich für die Fahrradhersteller besser erkennen, was gemacht werden muss. Die zweite Vorfilterung wurde vorgenaommen, damit der Datensatz auf Grund zu vieler Daten nicht zu unübersichtlich wird. Die meisten Datenwerte sind für die Region North America verfügbar. Danach folgen in Abstufender Reihenfolge Ëuropeünd "Pacific". Europa wurde auf Grubd des regionalen bezugs und der damit einhergehenden Assoziation für Regionale Bauunternehmen gewählt.

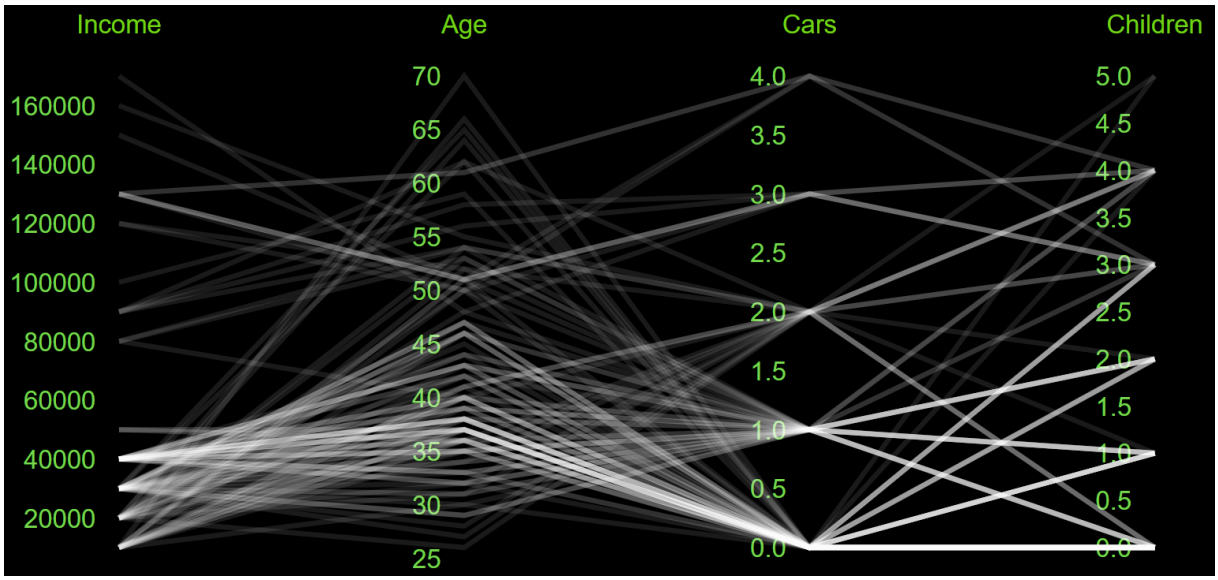

Abbildung 5: Anwendung Parallele Koordinaten, Quelle: Eigene Darstellung

In dieser Abbildung lässt sich erkennen, dass die höchsten EInkommen in den Altersgruppen zwischen 35 und 55 Jahren generiert werden. In Hinblick auf den Karrieweg vom Berufbeginn und Renteneintritt ist dieser Effekt logisch u erklären. In Hinblick auf das gestell des Fahrrads müssen somit keine Merkmale auf Aufstiegsfreundlichkeit gelegt werden.

5.3 Anwendung Visualisierung Drei

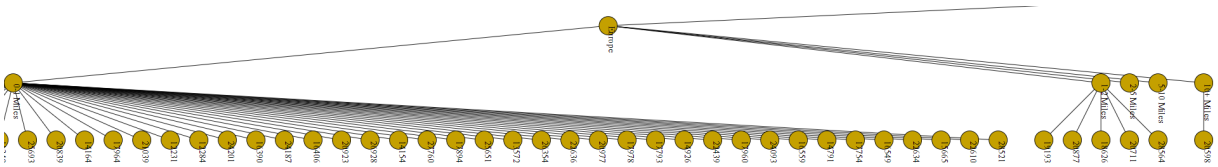

Abbildung 6: Anwendung Baumhierarchie, Quelle: Eigene Darstellung

Aus der Abbildung 6 geht hervor, dass in Europa von den Fahrradkäufern ohne Auto die überwiegende mehrheit eine geringe Distanz von Zuhause zur Arbeit aufweist. Da diese Distanz für den öffentlichen Nahverkehr zu gering wäre, gleichzeitig aber Distanzen von 1,2 km nicht mehr uzeitsparend zu Fuß zu bewältigen sind, kann davon augegangen werden, dass viele dieser Fahrradkäufer das Fahrrad zur Arbeit verwenden. Auf Grund der kurzen Entfernung ist davon auszugehen, dass die Daten in Großstädten gesammelt wurden. Die Infrastrukut ist hiervon von mehrspurigen Straßen mit vielen kruezungen und Abbiegungen gekennzeichnet. Infrastrukturprogeomme sehehn hierzu vor, dass besonders Fahrradwege besser ausgebaut werden sollen, damit weniger Verkehrsunfälle mit Fahrrädern passieren. Gerade deshalb ist dieser Ausschnitt aus der baumghierarchie für in Frage kommende Bauunternehmen spannend. Daran lässt sich erkennen, dass vermeintlich kurze Distanzen zwischen bereits vorhandener Infrastruktur gebaut werden müssen, um zu einer umweltneutalen, sicheren und verkehrsberuhigten Innenstadt beizutragen. Andere Darstellungsmethoden, die für eine Vergleichbare Datstellung in Frage kommen würden, wie die Graphendarstellung würden diesen auf grund der grpßen DAtenmenge bereits nicht optimal übersichtlichen Ausschnitt nicht verbessert dartellen können, sondern noch komplere Geflechte aufzeigen. Deshalb ist die BAumhierarchiestruktur für diesen Sachverhalt optimal gewählt.

6 Verwandte Arbeiten

Führen sie eine kurze Literatursuche in der wissenschaftlichen Literatur zu Informationsvisualisierung und Visual Analytics nach ähnlichen Anwendungen durch. Diskutieren sie mindestens zwei Artikel. Stellen sie Gemeinsamkeiten und Unterschiede dar.

7 Zusammenfassung und Ausblick

Fassen sie die Beiträge ihre Visualisierungsanwendung zusammen. Wo bietet sie für die Personen der Zielgruppe einen echten Mehrwert.

Was wären mögliche sinnvolle Erweiterungen, entweder auf der Ebene der Visualisierungen und/oder auf der Datenebene?

Anhang: Git-Historie

Literatur

[1] Heike Marquart, Julia Schuppan, Benjamin Heldt, Lisa Buchmann, Julia Jarass, Sarah Berg, Till Steinmeier, Philipp Masius, Meret Nathalie Batke, Arthur Zschäbitz, Jakob Bastian, Charlotte Blechner, David Brunner, Julian Maurer, Pascal Kraft, Leon Govinda Stephan, Tuan Anh Rieck, Konstantin Arndt, Lennart Goettsche, Robert Radloff, Nadja Martin, Lara Ann Steinert, and Fabian Drews. *Mobilität in Stadtquartieren*. Humboldt-Universität zu Berlin, 2021.

[2] Martin Kords. Statistiken zum thema fahrradfahrer, 2020.

[3] Statista. Corona-krise sorgt für fahrrad-boom, 25.08.2021.

[4] Martin Platter. Das virus bewegt aufs velo, 2020.

[5] Andreas Jöhrens. Boomendes geschäft, steigende preise: Lieferprobleme im fahrrad-handel, 2021.

[6] Mike Yi. A complete guide to scatter plots, 2019.

[7] F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17, 1973.

[8] William S. Cleveland and Robert McGill. The many faces of a scatterplot. *Journal of the American Statistical Association*, 79(388):807, 1984.

[9] A. Inselberg and B. Dimsdale. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In IEEE, editor, *Proceedings of the First IEEE Conference on Visualization: Visualization '90*, pages 361–378. IEEE Comput. Soc. Press, 1990.

[10] Rida Moustafa and Ed Wegman. Multivariate continuous data — parallel coordinates. In *Graphics of Large Datasets*, Statistics and Computing, pages 143–155. Springer New York, New York, NY, 2006.

[11] Stephen Few. Line graphs and irregular intervals. *Visual Business Intelligence Newsletter*, (11):1–11, 2008.

[12] Julian Heinrich and Daniel Weiskopf. Continuous parallel coordinates. *IEEE transactions on visualization and computer graphics*, 15(6):1531–1538, 2009.

[13] Heinz-Peter Gumm and Manfred Sommer. *Programmierung, Algorithmen und Datenstrukturen*, volume / Heinz-Peter Gumm, Manfred Sommer ; Band 1 of *De Gruyter Studium*. De Gruyter Oldenbourg, Berlin and Boston, 2016.

[14] Heeral Dedhia. Bike buyers 1000, 22.09.2020.