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.

- **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. Diese 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

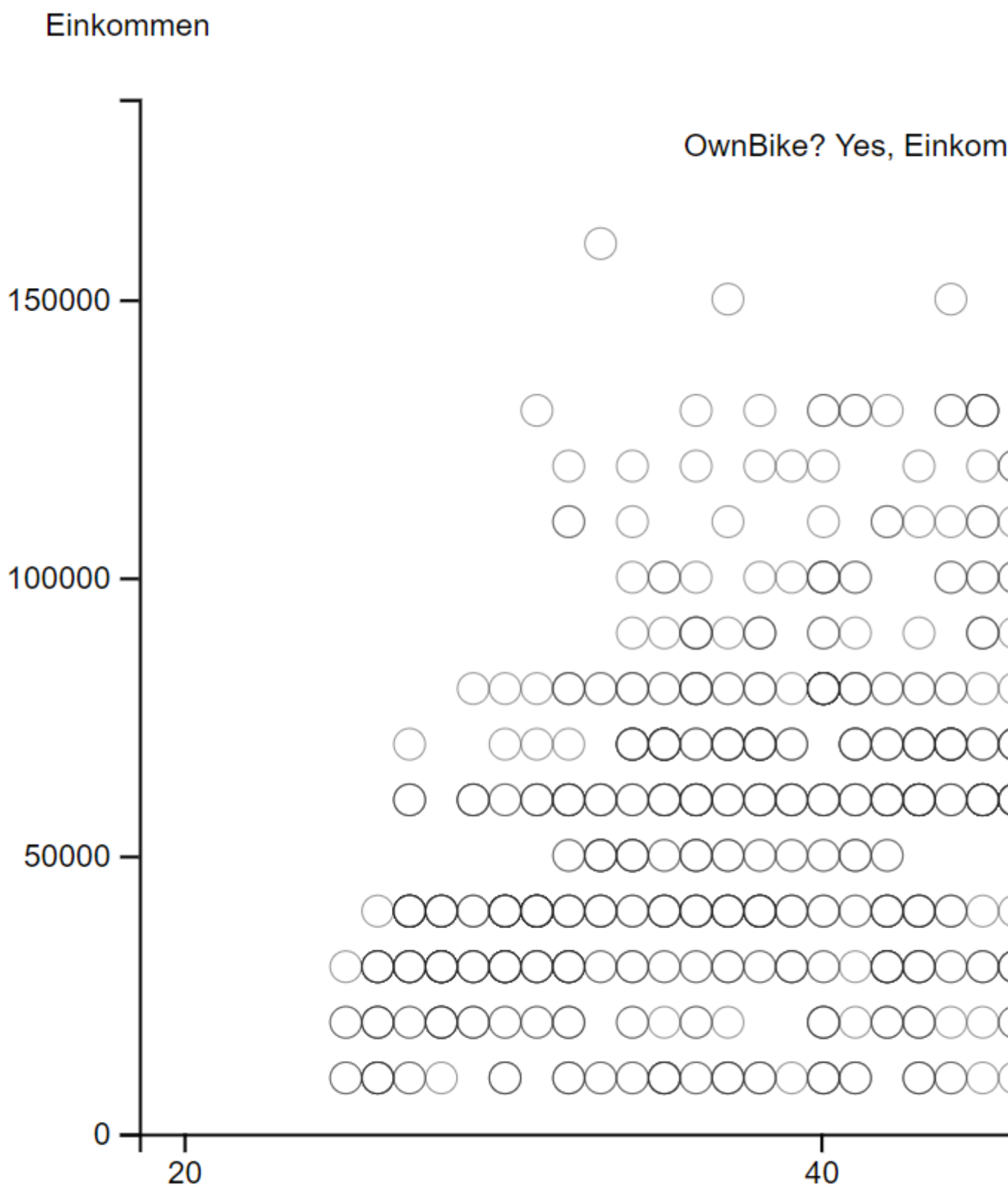
In diesem Kapitel werden die Visualisierungen und Ihre Aussagekraft beschrieben.

3.2 Anforderungen an die Visualisierungen

Leiten sie Anforderungen an das Design der Visualisierungen ab, die sich durch ihre Analyse des Zielproblems ergeben. 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.

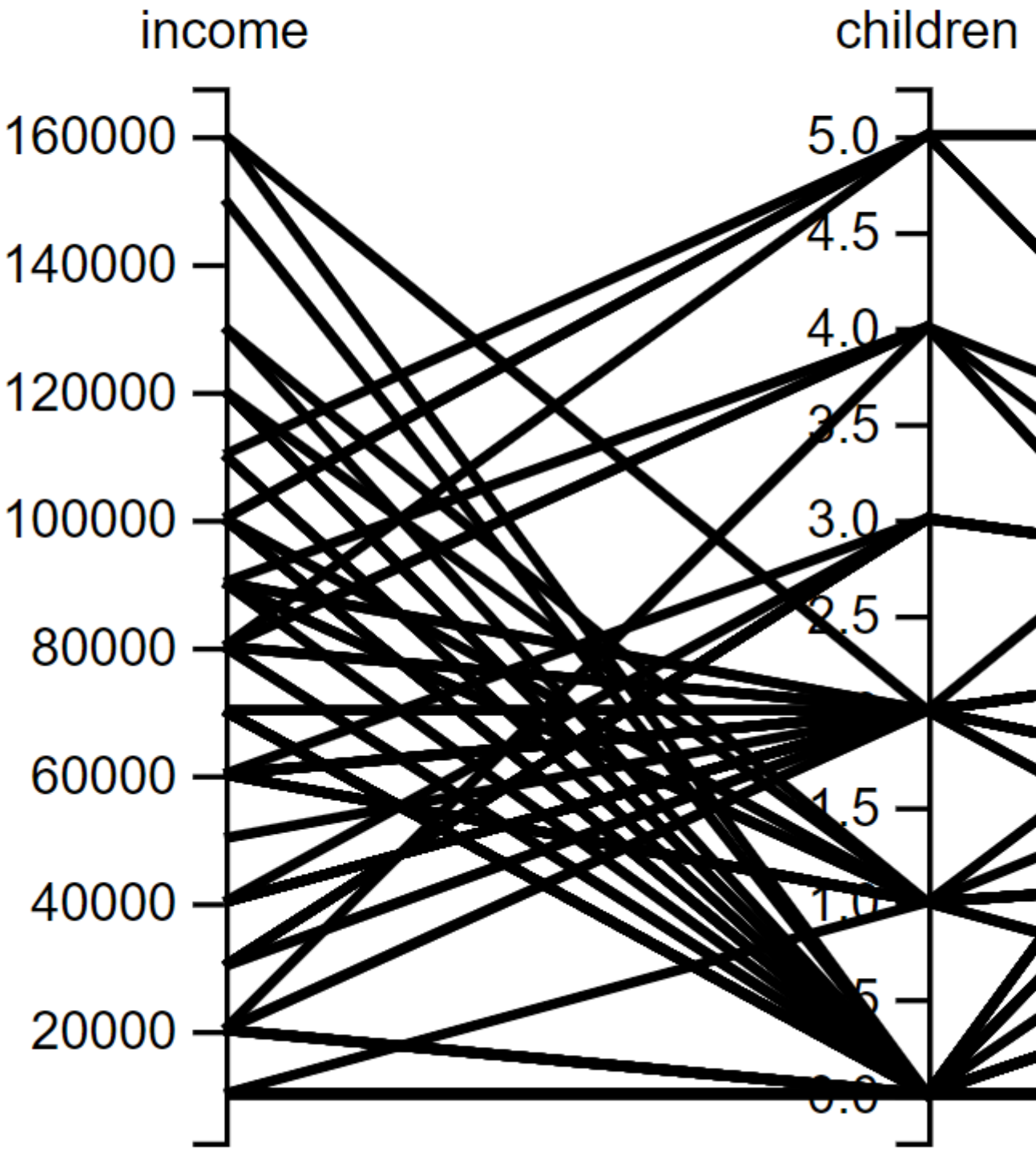
3.3 Präsentation der Visualisierungen

3.3.1 Visualisierung Eins



3.3.2 Visualisierung Zwei

TestText 123


income
children
160000
140000
120000
100000
80000
60000
40000
20000
5.0
4.5
4.0
3.5
3.0
2.5
1.5

TestText 123

3.3.3 Visualisierung Drei

3.4 Interaktion

Erklären sie die möglichen Interaktionen mit den einzelnen Visualisierungen und die möglichen Verknüpfungen zwischen ihnen. Begründen Sie warum die konkreten Interaktionen umgesetzt wurden und welche Zwecke für die Anwenderinnen mit ihnen unterstützt werden. Begründen sie ebenfalls warum sie andere Interaktionsmöglichkeiten nicht umgesetzt haben.

4 Implementierung

Beschreiben Sie die Implementierung ihrer Visualisierungsanwendung in Elm. Stellen die Gliederung ihres Quellcodes vor. Haben Sie verschiedene Elm-Module erstellt. Was war aufwändig umzusetzen, was ließ sich mit dem vorhanden Code aus den Übungen relativ einfach umsetzen?

Wie sieht die Elm-Datenstruktur für das Model aus, in dem die verschiedenen Zustände der Interaktion gespeichert werden können.

5 Anwendungsfälle

Präsentieren sie für jede der drei Visualisierungen einen sinnvollen Anwendungsfall in dem ein bestimmter Fakt, ein Muster oder die Abwesenheit eines Musters visuell festgestellt wird. Begründen sie warum dieser Anwendungsfall wichtig für die Zielgruppe der Anwenderinnen ist. Diskutieren sie weiterhin, ob die oben beschriebene Information auch mit anderen Visualisierungstechniken hätte gefunden werden können. Falls dies möglich wäre, vergleichen sie die den Aufwand und die Schwierigkeiten ihres Ansatzes und der Alternativen.

5.1 Anwendung Visualisierung Eins

5.2 Anwendung Visualisierung Zwei

5.3 Anwendung Visualisierung Drei

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.