**Projektbericht zum Modul Information Retrieval und Visualisierung Sommersemester 2022**

# Eine Analyse von Anzeigen auf einer deutschen Gebrauchtwagenplattform aus dem Jahr 2023

Johannes Göbel Matr.Nr.:

18. Dezember 2023

---

letzter commit:
Link zum Github Repository:
(https://github.com/johannesgoebel/germany-used-cars-dataset)
Link zur Live Demo:
(xxx)

# Inhaltsverzeichnis

# 1 Einleitung

Internetplattformen sind Digitalfirmen, die zwei oder mehr distinkte Nutzergruppen zusammenbringen. Sie werden zwischen B2B, bei denen zwei Geschäftskunden miteinander interagieren, und B2C, bei denen eine Seite durch einen Endkunden gebildet wird. In jedem Fall profitiert jede Seite durch eine Partizipation der gegenüberliegenden Seite am Netzwerk [**MUZELLEC2015139**]. Dieser Effekt konnte bereits im späten 19.Jahundert bei der schleppend laufenden Adoption des Telefons beobachtet werden. Ein Telefon war nutzlos, wenn es niemanden gab, den man anrufen konnte. Das Netzwerk war erst wertvoll, als es viele Teilnehmer hatte – ein positiver direkter Netzwerkeffekt. [**evans2016matchmakers**] Internetmarktplattformen gewannen zuerst in den 1990er mit Firmen wie e-bay und Craiglist Bedeutung und fanden Anfang des Jahrtausends mit AirBnB, Booking.com und heutigen Entwicklungen wie Maschinensucher, wo ganze Industriemaschinen gehandelt werden können, einen steilen Aufstieg. Heute, 30 Jahre später, interagieren Menschen wie selbstverständlich mit solchen Plattformen. Selbst große E-Commerce Händler, wie Amazon, haben schon vor Jahren Marktplatzmechaniken in ihre Webseiten integriert. Diese Integration ist nahtlos, dass viele Nutzer gar nicht wissen, dass auf Amazon ein großer Preiskampf zwischen Drittanbietern ausgebrochen ist. (?????) Gerade deshalb lohnt es sich einen Blick auf die Mechanismen von diesen Plattformen zu werfen und nachzuvollziehen, wie Verkäufer ihre Angebote bewerben. Dies soll in dieser Arbeit am Beispiel von einer Gebrauchtwagenplattform geschehen. // Dabei sollen in dieser Arbeit aber auch Markteffekte von PKWs mit unterschiedlichen Eigenschaften nicht außer Acht gelassen werden. Zu diesem Zweck werden die folgenden Forschungsfragen formulieren:

- Wie hängt der Angebotspreis mit den angegebenen Merkmalen eines Gebrauchtwagens zusammen?
- Was sind die Werte für ein bestimmtes Fahrzeug? Wo befindet es sich im Vergleich zu den anderen Fahrzeugen?
- Sind Preis- und Merkmalsunterschiede zwischen unterschiedlichen Automarken festzustellen, und wenn ja, wie?

## 1.1 Anwendungshintergrund

Zu diesem Zweck soll eine Webseite mit verschiedenen Visualisierungen erstellt werden, die es Nutzern ermöglicht schnell einen Überblick über die auf der Plattform vorhandenen Angebote zu erlangen und diese einzuordnen.
Die Daten hierzu sollen von einer echten Marktplattform programmatisch abgerufen werden um ein möglichst reales Bild einer solchen Plattform zu geben. Dies führt allerdings auch zu einem hohen Datenaufbereitungsaufwand.

## 1.2 Zielgruppen

Die Zielgruppe dieser Arbeit sollen Gebrauchtwagenverkäufer sein.
Die erste Zielgruppe zeichnet sich durch eine hohe Fachkenntnisse aus. Im Markt profitiert diese Gruppe von bestehenden Informationsasymmetrien, die sich durch Marktintransparenzen bilden. Durch Internetmarktplattformen sind diese Marktintransparenzen seltener zu finden. Beispielsweise können Käufer Preise vergleichen und können damit besser abschätzen, wie hoch der reale Marktpreis ist.
Daher interessieren professionelle Marktteilnehmer, wie auch trotz dieser technologischen Entwicklungen, eine relevante Marge erzielen können. Dieser Informationsbedürfnisse möchte diese Arbeit mithilfe von Visualisierungen zu Fahrzeugmerkmalen und Bewerbungsstilen decken.

## 1.3 Überblick und Beiträge

Die Daten werden hierzu in zwei unterschiedlichen Arten und Weisen aufgearbeitet.
Für die ersten beiden Visualisierungen – einen Scatterplot und einen Parallelplot – werden die beschafften Daten lediglich bereinigt und auf Ebene der einzelnen anzeigen betrachtet. Der Scatterplot soll es dem Nutzer hierbei ermöglichen Attribute miteinander gegenüberzustellen und der Parallelplot alle relevanten Merkmale überblicksartig einzusehen.
Für die dritte Visualisierung wurden die Daten auf Ebene der einzelnen Automarken betrachtet. Mittels eines Starplots soll dem Nutzer Aufschluss über Trends auf dieser Ebene gegeben werden.

# 2 Daten

## 2.1 Datenverarbeitung

Der Datensatz auf der Webseite Kaggle unter dem Titel "Germany Used Cars Dataset 2023"Kaggle Datensatz bereitgestellt. Dieser umfasst über 200.000 Datensätze von der Gebrauchtwagenplattform Autoscout24 aus dem Jahr 2023, die durch Scraping gesammelt wurden.
In einem IPython Notebook wurden zunächst grundlegende Überlegungen angestellt, gefolgt von der Bereinigung des Datensatzes. Die Bewertung der Attribute erfolgte in Bezug darauf, ob sie für spätere Visualisierungen als (1) essenziell oder (2) sekundär betrachtet werden sollten. Fehlende Werte wurden entsprechend behandelt. Das Ziel der Bereinigung ist es in jedem Feld einen verarbeitbaren Wert vorfinden zu können.
Nach dem Löschen der unbenannten Indexspalte und der Zeilen ohne Werte in den Spalten öffer$_d$escriptionöder"mileage$_i$n$_k$m", wurdenfehlendeWerteinder"colorSpaltemitünknownäufgefüllt.Die
SpaltewurdeaufechteAntriebsartenüberprüft, undunsinnigeDateninanderenSpaltenwurdenentfernt.Da

Die Spalte "fuel$_c$onsumption$_{gk}$m"wurdeentfernt, dasieviele fehlendeWerteenthieltundkeinezusätzliche
Spalteundder"registration$_d$ateSpaltedurchgeführt.

Dabei war ein Ziel die Vergleichbarkeit von Fahrzeugen unterschiedlicher Antriebsarten zu Gewährleisten, was beispielsweise bei den Daten zum Verbrauch ein Problem aufwarf. Wie sind die Daten von Pkws mit Verbrennungsmotor in l/100km und der Reichweitedaten für Elekoautos vergleichbar? Da diese nicht gewehrleistet werden konnte, - Erweiterung um Daten zu Anzeigetiteln (Zeichenlänge des Titels, Sentiment Analysis?, Anzahl Großbuchstaben)

Hier wurde der Datensatz auch um erste Merkmale wie offer$_l$en, diedieAnzahlderBuchstabendesAnzeigete

Im Anschluss daran wurde der Datensatz im CSV-Format exportiert und in Elm geladen.
Dabei wurden drei Datensätze exportiert: (1) "data.csv": Der Standarddatensatz, der rein wie oben beschreiben erstellt wurde.
(2) ävg$_s$tar$_d$ata.csvünd(3)ßum$_s$tar$_d$ata.csv" : DieDatensätzefürdenStarplot, welchedurcheinenzusätzlich
EineletzteBemerkungsollzueinemgrö„erenProblembeimLadenderDateninElmgewidmetsein.Leidertrate
NotebooksProblemebeiderKodierungderDatenauf, sodassunsichtbareCharactersinderCSVvorzufinden

## 2.2 Eignung der Daten

In diesem Abschnitt wird diskutiert, inwieweit die vorliegenden Daten geeignet sind, die zuvor formulierten Forschungsfragen zu beantworten.

Grundsätzlich stammen die Daten aus dem Scraping einer Marktplatz-Webseite. Dies könnte ein realistisches Abbild des Marktplatzes liefern, sofern die Datenerhebung repräsentativ und stichprobenartig erfolgt ist. Es ist jedoch zu beachten, dass die Übertragbarkeit der Erkenntnisse aus der Datenanalyse und -visualisierung auf andere Plattformen möglicherweise eingeschränkt ist.

Zusätzlich ist anzumerken, dass keine Informationen darüber vorliegen, ob die vorliegenden Daten authentisch sind. Es ist unklar, ob die Daten ohne Vorauswahl oder Verfälschung durch das Scraping-Prozess gesammelt wurden. Andererseits gibt es derzeit keine erkennbaren Gründe, an der Authentizität der Daten zu zweifeln.

Eine letzte allgemeine Anmerkung ist, dass es sich bei der Beurteilung der Preise immer um Angebotspreise handelt. Es ist nicht bekannt, ob diese Fahrzeuge zu den genannten Preisen verkauft wurden oder nicht. Ebenfalls gehen natürlich weitere als die hier aufgeführten Merkmale in die Preisbildung ein. Beispielsweise ist es nicht klar, ob ein Auto eine Klimaanlage hat oder nicht, was die Preisvorstellung des Anbieters stark beeinflussen würde.

Die Daten enthalten Informationen über den Angebotspreis sowie verschiedene Merkmale von Gebrauchtwagen. Allerdings ist eine detaillierte Analyse der Beziehung zwischen diesen Merkmalen und dem Preis abhängig von der Qualität und Vollständigkeit der Daten. Einschränkungen könnten in fehlenden oder ungenauen Merkmalen liegen, die die Analyse beeinträchtigen könnten.

Die Datengrundlage um die Werte für ein einzelnes Fahrzeug und diese mit anderen zu vergleichen ist gegeben. Allerdings sind hier auch die allgemeinen Einschränkungen zu beachten.

Die Daten könnten eine Analyse der Preis- und Merkmalsunterschiede zwischen verschiedenen Automarken ermöglichen. Hierbei sind jedoch mögliche Einschränkungen zu berücksichtigen, wie ungleichmäßige Verteilung der Daten, so sind Einträge der Marke Audi viel häufiger im Datensatz vorhanden als Daten der Marke Proton.

Insgesamt sind die vorliegenden Daten geeignet, diese Forschungsfragen zu beantworten, vorausgesetzt, dass bestimmte Einschränkungen und Unklarheiten in der Datenqualität sorgfältig berücksichtigt werden.

# 3 Visualisierungen

## 3.1 Analyse der Anwendungsaufgaben

Analysieren sie die konkreten Anwendungsaufgaben, die die Lösung des Zielproblems durch die Anwender:innen bearbeitet werden müssen. Welche sinnvollen mentale Modelle helfen den Personen bei der Bearbeitung. Sind diese mentalen Modelle für sie notwendig, um die Aufgaben lösen zu können? Gehen sie bei ihrer Argumentation von den Anwendungsaufgaben aus und kommen sie dann zu den mentalen Modellen, deren Aufbau durch Visualisierungen unterstützt wird.

## 3.2 Anforderungen an die Visualisierungen

Bei diesem Visualisierungsprojekt sind klare Anforderungen an die Darstellung von Daten von entscheidender Bedeutung. Um potenziellen Käufern eine hilfreiche Entscheidungsgrundlage zu bieten, sollten die Visualisierungen Aspekte wie Fahrzeugpreise, Kilometerstand, Baujahr, etc. einschließen. Eine intuitive Benutzeroberfläche ist unabdingbar, damit die Interessenten mühelos durch die Vielzahl von Angeboten navigieren können. Die Visualisierungen sollten daher die Möglichkeit bieten, Fahrzeuge anhand verschiedener Parameter zu filtern und zu vergleichen. Ein interaktiver Ansatz mit Filteroptionen und Hover-Funktionen könnte dabei helfen, detaillierte Informationen zu einzelnen Fahrzeugen abzurufen. Die visuelle Darstellung sollte sowohl informativ als auch ansprechend sein, um die Benutzererfahrung zu optimieren und eine effektive Entscheidungsfindung zu ermöglichen.
Basierend auf diesen Metaanforderungen werden konkrete User Stories formuliert, die im Rahmen des Projektes addressiert werden sollen.
Funktionale User Stories:

1. Als Nutzer möchte Fahrzeuge anhand verschiedener Parameter wie Preis, Kilometerstand und Baujahr zu vergleichen, um schnell einen Überblick über die verfügbaren Optionen zu erhalten.

2. Als Nutzer wünsche ich mir einen benutzerfreundlichen Parallelplot, der alle relevanten Informationen zu einzelnen Fahrzeugen auf einen Blick darstellt. Damit kann ich die Zusammenhänge zwischen den verschiedenen Parametern besser verstehen und gezielt nach meinen Präferenzen filtern.

3. Als Nutzer interessiere ich mich für einen Star Plot, der mir eine visuelle Zusammenfassung der wichtigsten Merkmale verschiedener Fahrzeugmarken bietet. Dadurch kann ich schnell erkennen, wie sich die Marken in Bezug auf verschiedene Parameter unterscheiden.

Nichtfunktionale User Stories:

[scale=0,5]img/uc$_d$*iagram.png*

1. Als Anwender möchte ich eine ansprechende und intuitive Oberfläche erleben, die meine Navigation durch die Gebrauchtwagenanzeigen erleichtert. Die Visualisierungen sollten informativ und leicht verständlich sein, um mir eine effiziente Entscheidungsfindung zu ermöglichen.

2. Als Nutzer erwarte ich, dass die Visualisierungen eine hohe Benutzerfreundlichkeit bieten, einschließlich interaktiver Elemente wie Filteroptionen und Hover-Funktionen. Dadurch kann ich gezielt nach meinen Anforderungen suchen und zusätzliche Details zu einzelnen Fahrzeugen abrufen.

3. Als Anwender möchte ich die Möglichkeit haben, die Visualisierungen individuell anzupassen, indem ich verschiedene Parameter auswähle oder abwähle. Dadurch kann ich meine Suche und Analyse personalisieren und mich auf die für mich relevanten Informationen konzentrieren.

Aus diesen User Stories wurden dann funktionale und nicht funktionale Anforderungen in einem Use-Case Diagramm (Abbildung **??**) erstellt.

## 3.3 Präsentation der Visualisierungen

Präsentieren sie die visuelle Abbildungen und Kodierungen der Daten und Interaktionsmöglichkeiten. Sie müssen begründen, warum und wie gut ihre Designentscheidungen die erstellten Anforderungen erfüllen. Weiterhin müssen sie begründen, warum die gewählte visuelle Kodierung der Daten für das zulösenden Problem passend ist. Typische Argumente würden hier auf Wahrnehmungsprinzipien und Theorie über Informationsvisualisierung verweisen. Die besten Begründungen diskutieren explizit die konkrete Auswahl der Visualisierungen im Kontext von mehreren verschiedenen Alternativen. Machen sie hier nicht den Fehler, einfach nur Visualisierung aus den vorgegebenen Bereichen zu diskutieren, weil das in der Regel nicht sinnvoll ist. Wenn sie sich für einen Scatterplot entschieden haben, ist ein Zeitreihendiagramm in der Regel keine Alternative. Diskutieren sie also nicht einfach Zeitreihendiagramme, weil sie in den Anforderungenen an das Projekt neben Scatterplots stehen, sondern suchen sie nach echten alternativen Visualisierungen, die zum Aufbau eines vergleichbaren mentalen Modells führen. Diskutieren sie die Expressivität und die Effektivität der einzelnen Visualisierungen.

Die eben beschriebenen Präsentationen und Begründungen sollen für jede der drei folgenden Visualisierungen durchgeführt werden.

### 3.3.1 Visualisierung Eins

### 3.3.2 Visualisierung Zwei

### 3.3.3 Visualisierung Drei

## 3.4 Interaktion

Die präsentierten Visualisierungstechniken müssen interaktiv zu einer Anwendung verknüpft werden. Die Interaktion mit einer Visualisierung soll in den anderen Visualisierungen zu einer Änderung führen. 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. Wenn sie keine der geforderten Interaktionen umsetzen, erhalten Sie im gesamten Projekt deutlichen Punktabzug.

# 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