**Projektbericht zum Modul Information Retrival und Visualisierung**

Sommersemester 2021

**Visualisierung von verschiedenen Weindaten**

Marie-Luise Lindner

Matrikelnummer: 211231747

GitHub Repository: <https://github.com/RicBre/Elm-Projekt-WineInformation>

# Inhaltsverzeichnis

[Inhaltsverzeichnis II](#_Toc82161120)

[1. Einleitung 1](#_Toc82161121)

[1.1 Anwendungshintergrund 1](#_Toc82161122)

[1.2 Zielgruppen 3](#_Toc82161123)

[1.3 Überblick und Beiträge 4](#_Toc82161124)

[2. Daten 5](#_Toc82161125)

[2.1 Technische Breitstellung der Daten 6](#_Toc82161126)

[2.2 Datenvorverarbeitung 6](#_Toc82161127)

[3. Visualisierung 7](#_Toc82161128)

[3.1 Analyse der Anwendungsaufgaben 7](#_Toc82161129)

[3.2 Anforderungen an die Visualisierungen 8](#_Toc82161130)

[3.3 Präsentation der Visualisierung 9](#_Toc82161131)

[3.3.1 Visualisierung Eins 9](#_Toc82161132)

[3.3.2 Visualisierung Zwei 10](#_Toc82161133)

[3.3.3 Visualisierung Drei 12](#_Toc82161134)

[3.4 Interaktion 12](#_Toc82161135)

[4. Implementierung 13](#_Toc82161136)

[5. Anwendungsfälle 14](#_Toc82161137)

[5.1 Anwendung Visualisierung Eins 14](#_Toc82161138)

[5.2 Anwendung Visualisierung Zwei 14](#_Toc82161139)

[5.3 Anwendung Visualisierung Drei 16](#_Toc82161140)

[6. Verwandte Arbeiten 17](#_Toc82161141)

[7. Zusammenfassung und Ausblick 18](#_Toc82161142)

[8. Literatur i](#_Toc82161143)

[Anhang ii](#_Toc82161144)

[Git Historie ii](#_Toc82161145)

# Einleitung

Bereits zur Zeit des antiken Griechenlands herrschte der Wunsch nach einer Maschine mit der Fähigkeit menschlichen Denkens. Mehr als einhundert Jahre bevor programmierbare Computer wissenschaftliche Wirksamkeit erzielten, stellte sich die Frage nach künstlicher Intelligenz (KI). [11] Heute ist diese fester Bestandteil des menschlichen Alltags. Es ist ein blühendes Feld mit vielen praktischen Anwendungen und aktiven Forschungsthemen. Im bestehenden gesellschaftlichen Kontext bildet maschinelles Lernen die Basis für automatisierte Routinearbeiten, das Verständnis von Sprache oder Bildern, Diagnosen in der Medizin, sowie der Unterstützung wissenschaftlicher Grundlagenforschung. [7] Sie sind geeignet zur Vorhersage von Ereignissen und Modellierung komplexer zeitabhängiger Systeme. Insbesondere finden sie Anwendung, wenn die Einflussfaktoren für einen bestimmten Ausgang nicht oder nur unvollständig bekannt und die Zusammenhänge komplex und nichtlinear sind. [14]

Automatisierte Routinearbeiten sind unter anderem im Sektor des Automobilvertriebes beobachtbar. Berichten prognostizieren ein Wachstum bei Neuwagen in den kommenden 5 Jahren durchschnittlich um 3,5 %. Ein Anstieg im Gebrauchtwagensektor sei von 5% zu registrieren. Kausal der gesteigerten Nachfrage des Gebrauchtwagenmarktes sind veränderte Bedürfnisstrukturen der Verbraucher zu verzeichnen. Primäres Ziel ist die Optimierung des Wiederverkaufspreises das Automobiles. Die Validierung dessen ist verbunden mit Werkstattbesuchen, sowie Kostenvoranschlägen und wird als zeit -und ressourcenaufwändiger Prozess betrachtet. Ein indisches Unternehmen namens „avl“ fokussiert diese Thematik in einem Projekt. Ziel ist die Entwicklung eines Modells für Drittunternehmen, zur Schätzung des Preises des Autos des Kunden im Online-Portal. Folglich erfolge eine Rationalisierung des Verkaufsprozesses von Gebrauchtwagen. Methodisch soll die Umsetzung dieses Projektes mit Hilfe von maschinellem Lernen auf Basis eines Trainingsdatensatzes von rund 7000 Daten über Gebrauchtwagen erfolgen. [9]

Der vorliegende Bericht nimmt Rückschlüsse auf diesen Sachverhalt und die vorliegenden Daten. Ziel ist die Aufarbeitung des Datensatzes, sowie dessen anschließende Visualisierung in drei differenzierten Visualisierungstechniken.

## Anwendungshintergrund

Um ein Verständnis für die Wichtigkeit einer signifikanten Datengrundlage für künstliche Intelligenzen zu schaffen, steht im Folgenden der Anwendungshintergrund im Fokus.

Die Anfänge der Entwicklung von Künstlicher Intelligenz fokussierten die Umsetzung von Problemen, die sich durch eine Liste formaler, mathematischer Regeln beschreiben lassen. Als dessen größte Herausforderung ergaben sich Fragen, welche der Mensch intuitiv und automatisiert löst, wie beispielhaft das Erkennen von Gesichtern in Bildern. Ansätze zur Lösung sind künstliche neuronale Netze oder Entscheidungsbäume. Diese simulieren Prozesse des Zentralnervensystems höherer Lebewesen. Im Detail bilden sie ein abstrahiertes Modell miteinander verbundener Neuronen ab, welche auf Grundlage spezieller Anordnung und Verbindung eine Klassifizierung unstrukturierter Daten ermöglichen. [1]

Der Aufbau eines künstlichen neuronalen Netzes definiert sich über ein Schichtsystem. In erster Instanz steht die Eingabeschicht aus Neuronen (Input Layer), welche mittels Zwischenneuronen über mehrere Schichten (Hidden Layer) verknüpft ist. Das Konstrukt endet mit einer Ausgabeschicht (Output Layer), welche Ergebnisse zurückliefert. [1] Abbildung 1 verweist auf diese Struktur.

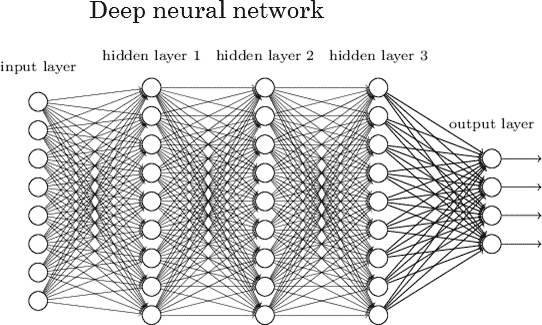


Abbildung 1: Struktur eines künstlichen neuronalen Netzes, Quelle: [3]

Der signifikante Unterschied zu klassischen statistischen Berechnungsverfahren definiert sich über die Lernfähigkeit eines künstlichen neuronalen Netzwerkes. Im Detail enthält ein solches System anfänglich keine Informationen.[14] Anhand von Trainingsdaten lernt das künstliche neuronale Netz. In diesem Prozess werden die Verknüpfungen, sowie die Verbindungsstärke innerhalb des Neuronenverbandes angepasst, sodass aus unbekannten Daten, Muster extrahiert und klassifiziert werden können. Die grundlegende Trainingsmethode erfordert große Datenmengen, welche den Analyseprozess stetig erneut durchlaufen. In jedem Analyseschritt gewonnene Erkenntnisse lassen sich mit weiteren Daten verknüpfen. Durch getroffene Entscheidungen erhalten die Verbindungen zwischen den Neuronen Gewichtungen. Positiv konjungierte Entscheidungen erhöhen das Gewicht der Verbindung. Ein Widerruf verringert das Gewicht. Folglich besitzt das künstliche neuronale Netzwerk die Fähigkeit Entscheidungen auf Basis von Verknüpfungen zu treffen.[1]

Auf Grundlage dessen wird die Bedeutung der Verfügbarkeit qualifizierter Daten- als Trainingsdatensatz- deutlich. Unzureichend verzerrte oder zu geringe Datenmengen führen zu verzerrten Ergebnissen, welche die Anwendung eines künstlichen neuronalen Netzes entkräften würden.[1] Entsprechend ist eine Analyse des Trainingsdatensatzes unabdinglich und von hoher Priorität. Aus diesem Grund wird im vorliegenden Bericht auf diese abgezielt.

Methodisch werden drei Visualisierungstechniken eingesetzt. Hauptziel ist es, anhand dieser den bereitgestellten Datensatz graphisch darzustellen und darauf basierend Rückschlüsse auf die Validität des Trainingsdatensatzes vornehmen zu können. Die graphische Umsetzung der Zusammenhänge gegebener Informationen ist vorteilhaft bei der Analyse von Ausreißern, der Betrachtung von Verhältnismäßigkeiten, sowie dem Erkennen signifikanter Werte eines Datensatzes. Auf Grundlage dessen wurden die Visualisierungstechniken der Baumhierarchie, des Scatterplots und der parallelen Koordinaten ausgewählt. Diese besitzen zu Teilen die Disposition der Interaktivität. Diese Eigenschaft ermöglicht den Nutzer den Vergleich von mehr als 2 Charakteristika der Informationen.

Die Visualisierungstechniken werden nachfolgend kurz vorgestellt.

Die erste Technik kennzeichnet eine Baumhierarchie. Die Baumhierarchie ist ein abstrakter Datentyp, welcher es ermöglicht hierarchische Strukturen abzubilden. Ausgehend von einem Wurzelelement können mehrere gleichartige Objekte miteinander verkettet werden, sodass lineare Strukturen aufgebrochen werden und folglich eine Verzweigung stattfindet. Diese Verzweigung kann beliebig oft wiederholt werden, welches das Addieren von Hierarchieebenen beinhaltet. Entsprechend ist Baumhierarchie geeignet Daten zu ordnen und hierarchische Zusammenhänge dazustellen. [8]

Bei der zweiten Visualisierungstechnik handelt es sich um einen Scatterplot. Mit wird die Gegenüberstellung zweier verschiedener numerischer Variablen ermöglicht. Die Werte werden in einem kartesischen Koordinatensystem eingetragen. Dies bietet die Möglichkeit die Beziehung zwischen diesen Variablen zu beobachten. Korrelationsbeziehungen, Muster oder Trends können erkannt und analysiert werden. Additional können mit Hilfe dieser Technik Daten gruppiert werden, welches Rückschlüsse auf mögliche Ausreißer oder Datenlücken vereinfacht. [15]

Die dritte Visualisierungstechnik repräsentiert den Ansatz der Parallelen Koordinaten. Auf Grund ihrer Beschaffenheit ergibt sich der Vorteil der Untersuchung mehrdimensionale Daten im zweidimensionalen Raum auf Trends, sowie Anomalien. Im Gegensatz zum Scatterplot, in welchem die zwei Koordinatenachsen rechtwinklig zueinander angeordnet sind, verlaufen die Achsen der Parallelen Koordinaten parallel und in gleichem Abstand. „Jede Linie von links nach rechts entspricht einem Datenpunkt und wird durch einen Polygonzug mit Ecken auf den parallelen Achsen dargestellt. Die Position der Ecke auf der i-ten Achse entspricht der i-ten Koordinate des Punktes.“ [2]

## Zielgruppen

Die Zielgruppen der graphischen Umsetzung des Trainingsdatensatzes lassen sich anhand des Daten Lebenszyklus herauskristallisieren. Dieser beinhaltet fünf Schritt. Schritt ein wird als „Data Collection“ (deutsch: Daten Kollektion) definiert. Dieser erläutert das Zusammentragen von Daten. Nachdem die Daten aufgenommen wurden, müssen sie in Formaten und an Orten gespeichert werden, an denen sie für andere Dienste leicht zugänglich sind. Dieser Vorgang beschreibt das „Data Cleaning“ (deutsch: Daten Säuberung). Um aus den erfassten Daten Erkenntnisse und Informationen zu gewinnen, müssen diese im dritten Schritt „Exploratory Data Analysis“ (deutsch: Explorative Daten Analyse) verarbeitet und analysiert werden. Im Detail bedeutet dies die Normalisierung aus der Quelle, Bereinigung und Speicherung in Analysesystemen, welche Abfragen und Untersuchungen ermöglichen. Schritt vier beschreibt das „Model Building“ (deutsch: Model Bauen). In diesem Vorgang werden Rückschlüsse aus der Analyse gezogen. Im finalen Schritt dem „Model Deployment“ (deutsch: Modell-Einsatz) wird die Bereitstellung von Modellen mit einer Daten-Pipeline in einer Unternehmensumgebung für die endgültige Benutzerabnahme angestrebt. Die Profession eines Data Scientist übernimmt die Aufgaben des gesamten Lebenszyklus. Der Fokus eines Data Engineers liegt vermehrt im Aufgabensektor der „Data Collection“ und des „Data Cleaning“. Die Rolle des Data Analyst zielt auf die Anforderungen des „Data Cleaning“ und der „Exploratory Data Analysis“ ab. Machine Learning Engineers verantworten die „Model Building“ und „Model Deployment“ Prozesse.[13] Da in diesem Bericht die Visualisierung des Trainingsdatensatzes, sowie dessen Analyse bearbeitet wird, kann in Konklusion von der Zielgruppe der Berufsbilder des Data Scientist, Data Analyst und des Maschine Learning Engineers ausgegangen werden. Berufsbedingt ist bei dieser Zielgruppe von einem exzellenten Vorwissen, sowie von Sicherheit im Umgang und der Auswertung der drei Visualisierungstechniken auszugehen. In diesem Zusammenhang ist eine Anwendung der gewonnen Kenntnisse im Rahmen einer Implementierung einer Maschine-Learning Anwendung in der Unternehmensumgebung denkbar. Additional besteht die Möglichkeit der Fehlererkennung im Datensatz und folglich dessen Ergänzung, Umstrukturierung oder Weiterverarbeitung.

Als zweite mögliche Zielgruppe können Studenten der Studiengänge im informatischen Sektor, wie beispielsweise Data Science, Informatik oder Automatisierungstechnik, betrachtet werden. Das deutsche Bildungssystem im universitären Bereich mit informatischer Ausprägung legt eine hohe Gewichtung auf die Ausbildung in mathematisch / statistischen Zusammenhängen, sowie analytischen Denken. Basierend dieser Annahme kann bei der Zielgruppe der Studenten von sehr guten Vorkenntnissen ausgegangen werden. Die Techniken des Baumdiagrammes, der parallelen Daten und des Scatterplots, sowie deren Auswertung gehören zur Grundausbildung eines Studenten mit wissenschaftlicher Ausrichtung. Eine mögliche Verwendung der Visualisierungen in diesem Kontext ist die des Lernobjektes. Die Visualisierungen ermöglichen einen fachlichen Zugewinn im Erkennen von Fehlern, Trends, Zusammenhängen oder Anomalien innerhalb des Datensatzes. Additional ist ein Einsatz als Grundlage in Projekten zum maschinellen Lernen im studentischen Kontext möglich.

Die dritte Zielgruppe wird durch Interessierte repräsentiert. Diese beschreibt Menschen, welche ein Grundinteresse für Gebrauchtwagen in Indien aufweisen. Da es sich in diesem Kontext vermehrt um eine Wissensaneignung auf persönlicher Ebene und weniger im beruflichen oder universitären Spektrum handelt, kann kein fachliches Grundwissen vorausgesetzt werden. Entsprechend besteht die Möglichkeit von geringen bis mäßigen Vorkenntnissen. Insbesondere im Hinblick auf die Techniken ist bei dieser Zielgruppe nicht davon auszugehen, dass die Funktionsweise der Parallelen Koordinaten bekannt ist. Trotzdem wird dieser Gruppe ausreichend Interesse unterstellt, Zeit und Energie in das Verständnis der Visualisierungstechnik zu investieren. Schlussfolgernd kann auch diese Visualisierungstechnik als Anwendung für interessierte Personen verstanden werden. In diesem Zusammenhang ermöglicht das Diagramm der Parallelen Koordinate eine Entnahme des Preis-Leistungs-Verhältnisses der einzelnen Objekte im Vergleich. Eine weitere relevante Visualisierungstechnik für diese Zielgruppe ist die explizite Baumhierarchie. Mit Hilfe dessen besteht die Möglichkeit einen Überblick über die Verfügbarkeiten von Gebrauchtwagenmodellen zu erhalten. Dadurch können Interessierte bei der Entscheidung eines Modelles unterstützt werden. Auf Basis der Einfachheit des Baumdiagrammes kann ein Verständnis der Funktionsweise dessen angenommen werden.

## Überblick und Beiträge

Die Informationsgrundlage dieses Projektes bilden Daten der Webseite „Kaggle“. [9] Dieser sind zwei Datensätze zu entnehmen. Ein Testdatensatz und ein Trainingsdatensatz. Dementsprechend ist davon auszugehen, dass diese im Zusammenhang mit Maschinellen Lernen stehen. Die Datensätze fokussieren Informationen über Gebrauchtwagen in Indien. Im Detail zentriert sich der Inhalt dessen auf die Komponenten *Name* (deutsch: Name), *Location* (deutsch: Ort), *Year* (deutsch: Jahr), *Kilometers\_Driven* (deutsch: gefahrene Kilometer), *Fuel\_Type* (deutsch: Kraftstofftyp), *Transmission* (deutsch: Schaltgetriebe), *Owner Type* (deutsch: Vorbesitzer), *Mileage* (deutsch: Kilometerstand), *Engine* (deutsch: Hubraum), *Power* (deutsch: Pferdestärken), *Seats* (deutsch: Sitze), *New\_Price* (deutsch: neuer Preis) und *Price* (deutsch: Preis). Erkennbar geht die Thematik des Gebrauchtwagenhandels mit einer Vielzahl an Informationen einher. Insbesondere in einem Verkaufsszenario, stellt ein hoher Informationsgehalt eine unabdingbare Notwendigkeit dar. Anhand genannter Eigenschaften besteht die Möglichkeit der Validierung des Preises eines gebrauchten Fahrzeugs. Handelsüblich wird in diesem Kontext auf die Expertise einer Fachkraft aus dem Autosektor zurückgegriffen. Werkstätten und Gebrauchtwagenhändler dienen als Bezugspunkte. Hingegen gilt der Besuch in einer Werkstatt als ressourcenaufwendig. Neben der energetischen Ressource stellt die zeitliche Komponente ein Problem dar. Aus diesem Grund besteht der Bedarf der Optimierung des Validationsprozesses von Gebrauchtwagen. Ein Online-Portal soll dem Nutzer die Möglichkeit geben durch Eingabe spezieller Kennzahlen seines Gebrauchtwagens, dessen Preis ermitteln zu können. In diesem Szenario entfällt der ressourcenaufwendige Prozess beim Fachmann. Um eine Ausgabe gültiger Werte der Software des Online-Portals sicher stellen zu können, ist der in Kapitel 1.1 beschriebene Prozess des maschinellen Lernens notwendig. Basierend eingespeister Trainingsdaten lernt das System und ist folglich fähig, anhand gegebenen Wertes, den Preis eines Gebrauchtwagens zu prognostizieren. Basierend dieser Grundlage ist die Unverzerrtheit des Trainingsdatensatzes unabdingbar. Schlussfolgernd wird eine genaue Analyse dessen als essentiell betrachtet. Aus diesem Grund wird in diesem Projekt der Analyseprozess des Trainingsdatensatzes fokussiert. Der Testdatensatz bleibt unbeachtet. Die gewählten Visualisierungstechniken dienen der Unterstützung der Untersuchung des Trainingsdatensatzes auf gewählte Eigenschaften. Ziel ist es anhand der Visualisierungen den Trainingsdatensatz auf Anomalien, Ausreißer, Verteilungen und verzerrten Werte zu untersuchen, sowie zu bereinigen, um folglich eine Unverzerrtheit dessen zu gewährleisten. Entsprechend ist das Baumdiagramm aufgrund der Möglichkeit hierarchische Daten darstellen zu können, ausgewählt worden. Durch dieses kann eine übersichtliche Darbietung der Verteilung der Automarken erreicht werden.

Die Verwendung des Scatterplots wird durch dessen Eigenschaft der Simplizität begünstigt. Im Hinblick auf die Zielgruppen des Projektes, kann fachliche Expertise nicht vorausgesetzt werden. Entsprechend bietet der Scatterplot eine leicht verständliche Visualisierung von Daten, welche auf Grundlage seiner Beschaffenheit dennoch einen erheblichen Mehrwert gegenüber Daten in tabellarischer Form bietet.

Die Zusammenhänge im System der Parallelen Koordinaten sind komplexer und bieten mehr Raum für Vergleichbarkeit von Eigenschaften. Mittels dieser Visualisierungen können multidimensionale Daten im zweidimensionalen Raum dargestellt werden. Mithilfe der interaktiven Funktion besteht der Vorteil des Vergleiches von mehr als 2 Attributen. In Kapitel wird ein detaillierterer Bezug zur Umsetzung und Auswertung der Visualisierungstechniken genommen.

# Daten

Die Grundlage dieses Projekts bilden Datensätze der Webseite „Kaggle“. [9] Die Spezialisierung dieser Plattform liegt auf den Sektoren Datenanalyse und Maschinellen Lernen. Eine Community stellt auf diesem Kontext basierende Datensätze öffentlich für die Weiterverarbeitung zur Verfügung. *Used Cars Price Prediction* des Nutzers Avi Kasliwalist einer dieser und findet im vorliegenden Projekt Anwendung. Als Lokation Avi Kasliwals ist die indische Stadt Noida im Bundesstaat Uttar Pradesh angegeben. Genauere Beschreibungen zum Datensatz sind von der Organisation Co-Learning Lounge angegeben. Diese erläutern die Relevanz des Gebrauchtwagenhandels und erklären das daraus resultierende Bedürfnis der Optimierung des Validierungsprozesses von Gebrauchtwagen. Additional verweisen sie auf optionale Aufgabenstellungen diesbezüglich. Co-Learning Lounge ist eine Plattform indischen Ursprungs, welche einer Community Raum für gemeinsames Lernen, Zusammenarbeiten, Innovieren und Mentorieren bietet. Die Ausrichtung ist in diesem Zusammenhang auf die Schwerpunkte künstliche Intelligenz, Blockchain, Big Data, Automatisierung und Technik ausgerichtet. [4]

Das letzte Update der Datensätze wurde vor einem Jahr vorgenommen. Entsprechend kann von einer Aktualität der Daten ausgegangen werden. Die Originaldatensätze sind ein Trainingsdatensatz namens *train.csv* und ein Testdatensatz namens *test.csv.* Zusätzlich existiert eine *.xlsx* Datei namens *dictionary.xlsx*, welche Auskunft über Inhalte der verwendeten Variablen innerhalb der Datensätze gibt. Die Originaldatei des Trainingsdatensatzes verfügt über 13 Spalten mit insgesamt 6019 Datensätzen. Das original des Testdatensatzes enthält 12 Spalten mit 1234 Datensätzen. Innerhalb der Datensätze wird sich auf die Informationen *Name* (deutsch: Name), *Location* (deutsch: Ort), *Year* (deutsch: Jahr), *Kilometers\_Driven* (deutsch: gefahrene Kilometer), *Fuel\_Type* (deutsch: Kraftstofftyp), *Transmission* (deutsch: Schaltgetriebe), *Owner\_Type* (deutsch: Vorbesitzer), *Mileage* (deutsch: Kilometerstand), *Engine* (deutsch: Hubraum), *Power* (deutsch: Pferdestärken), *Seats* (deutsch: Sitze), *New\_Price* (deutsch: neuer Preis) und *Price* (deutsch: Preis) bezogen. Der *dictionary.xlsx* Datei sind charakterisierende Informationen der Spaltennamen zu entnehmen. Der *Name* gibt die Marke und das Model des Fahrzeugs aus. Die *Location* deklariert den Ort an welchem das Auto verkauft wurde oder zum Kauf angeboten wird. Die Auflage beziehungsweise das Baujahr des Modells wird durch die Variable *Year* wiedergegeben. *Kilometers\_Driven* geben Auskunft über die gefahrenen Kilometer des Fahrzeugs durch Vorbesitzer. *Fuel\_Type* deklarierte die Art des Kraftstoffes mit welchem der Gebrauchtwagen angetrieben wird. Die Variable *Transmission* definiert die Art des Schaltgetriebes. Mittels des *Owner\_Types* wird beschrieben wie viele Vorbesitzer das Fahrzeug in Benutzung hatten. *Mileage* gibt den Verbrauch an. Wie viel Kilometer können mit einem Liter Kraftstoff gefahren werden. Das Hubraumvolumen des Motors definiert sich über die Variable *Engine. Power* deklariert die Leistungsfähigkeit des Motors. Die Anzahl der Sitze wird durch *Seats* ausgegeben. *New\_Price* gibt den Preis des Neuwagens des Models aus. Die *Price* Variable enthält Informationen des Gebrauchtwagens.

Aus bereits beschriebenen Gründen stellt das Ziel dieser Arbeit die Visualisierung und Auswertung des Trainingsdatensatzes dar. Dies bezüglich wird im Folgenden ausschließlich Bezug auf diesen genommen. Generalisierend kann zunächst davon ausgegangen werden, dass die im Datensatz enthaltenen Informationen bezüglich der Validierung des Preises eines Gebrauchtwagens von Relevanz sind. Ob die Datenmenge und Aussagekraft der Informationen ausreichend für Maschinelles Lernen sind, ist dem Kapitel 7 dieser Arbeit zu entnehmen.

Um eine optimierte Datengrundlage für die Umsetzung der Visualisierungen bereitzustellen, mussten die Daten einer Analyse unterzogen werden. Auffällig in diesem Zusammenhang war es, dass die Spalten *Name, Location, Year, Kilometers Driven, Fuel Type, Transmission, Owner Type* und *Price* vollständig gedeckt sind. Die Informationen in den Spalten *Mileage, Engine, Power* und *Seats* sind lückenhaft. Additional ist liegt eine Diversität der Einheiten der Spalte Mileage vor. Generalisierend kann die Auswahl der Einheiten aller Spalten als semioptimal betrachtet werden. Die Kilometer per Liter wurden auf die Einheit *kmpl* angeglichen, der Hubraum des Fahrzeugs wird allgemein gültig in Liter angegeben, für die Stärke des Motors wird die Einheit Pferdestärken verwendet (PS). Additional hat eine Umrechnung des Preises des Gebrauchtwagens in Euro stattgefunden. Eine detailliertere Betrachtung des Modifizierungsprozesses der Daten kann Kapitel 2.2. entnommen werden.

## Technische Breitstellung der Daten

Die Datensätze werden als CSV-Datei bereitgestellt. [9] Eine CSV-Datei ist definiert als Comma-Seperated-Value. Entsprechend werden einzelne Datenwerte durch Komma getrennt aufgelistet. Vorteil dieses Dateityps ist die Kompatibilität der Integration mit einer Vielzahl von Programmen. Im Kontext des vorliegenden Projektes wurde der Trainingsdatensatz zur Bearbeitung in Microsoft Excel eingefügt und als modifizierter Datensatz exportiert.

Die technische Bereitstellung der Daten erfolgt über ein GitHub Repository. [10] In diesem wurde das Visualisierungsprojekt gehostet. In diesem sind die Daten der Visulisierungen Scatterplot und Parallele Koordinaten als *.csv* Datei im *data* Ortner vorzufinden. Die Baumhierarchie arbeitet mit dem Dateiformat *.json* ***(J****ava****S****cript* ***O****bject* ***N****otation).* Die Quelldaten liegen im Ortner *Quelldaten.* Alle durch Selektion oder Modifizierung veränderten Daten sind dem Ortner *AufbereiteteDaten* zu entnehmen.

## Datenvorverarbeitung

Bereits in Kapitel 2 wurde Bezug zur Notwendigkeit der Modifizierung des Trainingsdatensatzes genommen. Diese Anpassung basiert in drei Schritten. In Abfolge werden die Daten gesichtet, bearbeitet und überführt. Eine detailliertere Prozessbeschreibt folgt in diesem Kapitel.

Das Ziel des Sichtungsprozesses der Daten war es die Werte in ein gut zu bearbeitendes Format zu übertragen. Entsprechend fand die Konvertierung der Trainingsdatensatz *train.csv* in eine Exceldatei Anwendung. Die Umsetzung dessen zielt auf eine übersichtliche tabellarische Darstellung der Datenwertpaare, sowie den Zugriff auf den Umfang der Excelfunktionen ab.

Im zweiten Schritt wurden die Daten aufbereitet, sodass deren reibungslose Verarbeitung durch den Einsatz in Code gewährleistet werden konnte. Entsprechend wurden in erster Instanz, aufgrund fehlender Aussagekraft sowie Vergleichbarkeit, Datenwertpaare entfernt, welche Nullwerte oder Lücken in den Spalten *Mileage, Engine, Power* und *Seats* enthalten. Zusätzlich wurde die Spalte *New\_Price* auf Basis lückenhafter Informationen, sowie fehlender Irrelevanz zum vorliegenden Kontext, entfernt. In zweiter Instanz wurden alle englischen Bezeichnungen im Spaltenkopf ins deutsche Äquivalent überführt. Folgend fand eine Generalisierung der Einheiten innerhalb einer Spalte und die damit verbundene Umrechnungen Anwendung. Da der vorliegende Bericht in der deutschen Sprache verfasst ist, wurden zusätzlich alle englischen Informationen zur besseren Verständlichkeit ins deutsche übersetzt.

Um die Werte innerhalb der Zellen als Zahl identifizieren zu können, bestand zunächst die Notwendigkeit alle Einheitsbezeichnungen innerhalb der Zellen zu entfernen. Nachfolgend wurden alle Werte der Einheit *km/kg* in der Spalte *Mileage* in *kmpl* umgerechnet. Der Umrechnungsfaktor beträgt 1.4. Alle Ergebnisse wurden auf zwei Dezimalstellen gerundet. Zusätzlich fanden Modifikationen in der Spalte *Engine* Anwendung. Alle Werte waren in *CC* angegeben. Es ist davon auszugehen, dass *CC* die Einheit Kubikzentimeter wiedergibt. Da diese gebräuchliche Anwendung im Kontext zu Angaben des Hubraums ist. Die Einheitsangaben wurden entfernt um den Inhalt der Zelle als zahl definieren zu können. Alle Informationen über die *Power* des Fahrzeugs sind im Originaldatensatz mit der Einheit *bhp* beschrieben. *Bhp* steht für Brake horsepower. 1 bhp entspricht 1.0139 Pferdestärken (PS). Da PS im europäischen Raum als handelsübliche Einheit für die Leistung des Motors eines Fahrzeugs betrachtet wird, fanden Umrechnungen in PS statt. Zusätzlich musste die Ziffer innerhalb der Zelle als Zahl definiert werden. Die Spalte *Price* war in der Währung *Lakh* angegeben. *Lakh* gilt als südasiatischen Zahlenwort für einhundertausend. [5] Insbesondere im Zusammenhang mit der Angabe der indischen Währung *Indische Rupien (INR)* findet dieser Wortlaut Anwendung. Entsprechend bezeichnet 1 *Lakh, einhunderttausend* Indischen Rupien*.* Zusätzlich liegt der Wechselkurs von *Indischen Rupien* in *Euro* bei 83,5096 INR zu 1€. [6] Entsprechend dieser Annahmen fand die Umrechnung der Werte innerhalb der Spalte *Price* statt. Wiederholt hat eine Entnahme der Einheitsangaben innerhalb der Zellen stattgefunden, um den Wert in dieser als Zahl definieren zu können. Die Ergebnisse wurden auf ganze Zahlen ohne Dezimalstellen gerundet. Erklärend für diesen Sachverhalt ist der Vergleich der Preisangaben zu erfolgreichen Gebrauchtwagenplattformen, wie *mobile.de* oder *autoscout24.de*. [3, 12]Diesen Webseiten zu entnehmen sind Preisangaben in ganzen Zahlen, ohne Dezimalstellen. Dementsprechend wurde diese Richtlinien übernommen und bezüglich der Preisangaben in den Visualisierungen angewandt.

Für die Darstellung der Baumhierarchie war zusätzlich eine Umwandlung des Dateiformats von *.csv* in *.json* notwendig. In diesem Zusammenhang besteht die Möglichkeit die Daten aus Microsoft Excel in eine *.json* zu exportieren. Mittels dieser Anwendung liegt allerdings keine Baumstruktur der Daten vor. Diese gilt als unabdingbar für Anordnung eines Baumes. Folglich bestand die Notwendigkeit der manuellen Bearbeitung der exportierten Exceldatei mit Visual Studio Code. Demgemäß konnte eine Baumhierarchie manuell erzeugt werden. Die Daten wurden dafür nach Gebrauchtwagen, Marken und final nach Modellen geordnet. Im direkten Vergleich der Vorverarbeitung der Daten der 3 Visualisierungstechniken, kann die Vorverarbeitung der Baumhierarchie als aufwendigste betrachtet werden. Die entsprechende .*json* Datei ist im Ordner *AufbereiteteDaten* unter dem Namen *LocationCar.json* vorzufinden.

Zuletzt wurde eine Spalte ID für jeden Gebrauchtwagen hinzugefügt. Diese verhilft der eindeutigen Identifizierung des Datensatzes. Jeder ID-Wert ist eindeutig innerhalb der Tabelle. Auf diese Weise besteht bei alleiniger Angabe der ID, die Möglichkeit für den Nutzer bei Bedarf ergänzende Informationen über das Objekt anhand der Datentabelle zu erfahren.

Im dritten Schritt der „Überführung der Daten“, bestand das Ziel die Werte für den Programmcode lesbar zu machen. Entsprechend wurden alle innerhalb von Microsoft Excel modifizierten Daten, in ein *.csv* Format zurückgeführt. In diesem Prozess lag der Fokus darauf, dass die Daten durch Kommas getrennt, sowie als Dezimaltrennzeichen ein Punkt Verwendung fand. Im folgenden Kapitel wird detaillierter auf die Entwicklung der Visualisierungen eingegangen.

# Visualisierungen

Wie bereits verdeutlicht, erfolgt die Darstellung des Trainingsdatensatzes von Gebrauchtwagen anhand von 3 Visualisierungstechniken. In erster Instanz wird mit Hilfe eines Baumdiagrammes die quantitative Verteilung der vorhandenen Datensätze analysiert. Anschließend wird ein Scatterplot erzeugt, um den Vergleich zweier Attribute des Gebrauchtwagens zu ermöglichen. Ziel ist es, dass der Nutzer diese Attribute interaktiv wählen kann. Mitunter besteht allerdings die Notwendigkeit mehr als zwei Attribute miteinander zu vergleichend. Entsprechend wurde die dritte Visualisierung der Parallelen Koordinaten gewählt, welch einen solchen Vergleich gewährleistet. Folglich können die Zusammenhänge von mehr als zwei Eigenschaften betrachtet werden. Die Reihenfolge dessen soll wiederholt für den Nutzer wählbar sein.

Im folgenden Kapitel wird detaillierter auf die Anforderungen an die Visualisierungen, sowie deren technische Umsetzung eingegangen. Des Weiteren werden im Verlauf

## Analyse der Anwendungsaufgaben

Die Visualisierungen sollen eine Unterstützung für in Kapitel 1.2 erläuterte Zielgruppen darstellen. Vordergründing sollen sie verhelfen die Informationen der Gebrauchtwagen darzustellen, sodass bei dem Nutzer ein Zugewinn an Erkenntnissen besteht, welcher ohne die Visualisierung nicht stattgefunden hätte.

Im Kontext eines Baumdiagrammes lassen sich Datensätze in Hierarchie Ordnen und anzeigen. Dies ermöglicht eine übersichtliche Darstellung der Daten nach bestimmten Gesichtspunkten. Im vorliegenden Beispiel soll die Eigenschaft der hierarchischen Darstellung genutzt werden, um eine Verteilung der Datensätze analysieren zu können. Entsprechend ist das Wurzelelement als Gebrauchtwagen definiert und gliedert sich in die Marke und das Model des Automobils. Anhand dessen lassen sich quantitative Rückschlüsse über die Repräsentativität konkreter Fahrzeugtypen treffen.

Hingegen liegen die Vorteile des Scatterplots sowie der parallelen Koordinaten in der Möglichkeit des Erfassens von Trends und Anomalien innerhalb der Daten. Das Ziel des Scatterplots ist es, Zielgruppen eine Gegenüberstellung zweier Attribute zu ermöglichen. Dieser Ansatz erbringt den Vorteil, dass Zusammenhänge dieser Eigenschaften analysiert werden können. In diesem Kontext können beispielsweise Fragen beantwortet werden, wie: *Wie verhält sich der gefahrene Kilometerstand eines Gebrauchtwagens zu dessen Preis? Sind Gebrauchtwagen mit mehr PS grundsätzlich preisintensiver?* Insbesondere bei der Analyse statistischer Zusammenhänge gilt die Darstellung zweier Faktoren in zweidimensionalen Koordinatensystemen als effektives Hilfsmittel und findet in der Praxis Anwendung. Aufgrund der Simplizität des Systems kann ein leichtes Verständnis des Scatterplots beim Nutzer vorausgesetzt werden. Das Wählen der Attribute, welche miteinander verglichen werden sollen, kann über zwei Buttons selbstdefiniert gesteuert werden. Entsprechend werden die gewählten Eigenschaften auf der X-Achse, sowie Y-Achse dargestellt. Kausal folgt auf einen Wechsel der Attribute, eine Veränderung der Achsen und des Plots. Die Darstellung eines Scatterplots basiert vorerst ausschließlich auf einer Punktwolke. In diesem Fall kann nicht nachvollzogen werden, welcher Punkt welches Fahrzeug darstellt. Aus diesem Grund wurde eine „Hover“-Funktion der Namensanzeige und Farbänderung programmiert. Insbesondere bei der Analyse großer Datenmengen erweist sich diese Funktion als vorteilhaft. Basierend der Überschneidungen einer Vielzahl diskreter Variablen werden, tendiert die Darstellung unübersichtlich zu wirken. Mit Hilfe einer solchen „Hover“-Funktion kann diese Problematik umgangen werden.

Basierend der Grundlagen des Scatterplots kann das System der Parallelen Koordinaten als Erweiterung dessen betrachtet werden. Anstelle zweier Attribute tritt der Vergleich mehrerer Eigenschaften, welche gleichermaßen interaktiv ausgewählt werden können. Additional zu den bereits erwähnten Vorteilen können zusätzlich Nutzdefinierte Anforderungen gefunden und befriedigt werden. Beispielsweise können Antworten auf Fragen wie folgend gefunden werden. *Welcher Gebrauchtwagen ist Baujahr 2010, hat einen geringen Kilometerstand, 100 PS und kostet weniger als 10.000€?* Schlussfolgernd ist abzuleiten, dass diese Darstellung über die zwei Dimensionen des Scatterplots hinausgeht. Die gewählten Attribute werden miteinander durch Linien verbunden. Entsprechend sind der Darstellung Zusammenhänge zu erkennen. Wiederholt kristallisiert sich das Problem der Unübersichtlichkeit durch Überlagerung diskreter Werte heraus. Die Lösung dessen besteht, ähnlich wie beim Scatterplot, in der Implementierung der „Hover“-Funktion, welche mittels Namensänderung und Farbanzeige zu mehr Verständnis beiträgt. Anhand dessen kann der Nutzer die Ausgabe von Informationen eines konkreten Fahrzeugs durch die Kennzeichnung einer Linie entlang der Parallelen Koordinaten angezeigt bekommen. Grundlegend ist das System der Parallelen Koordinaten deutlich komplexer als das des Scatterplots. Zusätzlich verfügt diese Visualisierungstechnik nicht über einen gleichwertigen Bekanntheitsgrad. Entsprechend kann das Wissen der Zielgruppen über dessen Interpretation nicht vorausgesetzt werden. Aus diesem Grund bietet sich eine Beilage von Erklärung über die Interpretation der Parallelen Koordinaten an.

## Anforderungen an die Visualisierungen

In Kapitel 3.1 wurde bereits Bezug genommen, welche Ziele die gewählten Visualisierungen adressieren. Aus diesen Zielstellungen kann eine Reihe von Anforderungen abgeleitet werden.

Ziel der zweiten Visualisierung ist es zwei Attribute der Gebrauchtwagen miteinander ins Verhältnis zu setzen. Stilistisches Mittel in diesem Zusammenhang ist der Scatterplot. Um die Werte in einem Scatterplot zuordnen, sowie ablesen zu können ist eine fachgerechte Beschriftung der X-Achse und Y-Achse notwendig. Dies impliziert einerseits die ordnungsgemäße Angabe der Namen an den Achsen. Andererseits existieren Zahlenwerte zu den Attributen. Diese müssen an den Achsen mit entsprechenden Einheiten aufgeführt werden und folglich die Achse im Verhältnis dieser Zahlenwerte einteilen. Zusätzlich wird im Kontext des Scatterplots auf individuelle Interessenvertretung abgezielt. Schlussfolgend soll es den Nutzer möglich sein, selbstständig eine Auswahl der Attribute entsprechend seinem Interessengebiet auszuwählen. Die Umsetzung erfolgt über Buttons, welche beim Anklicken die gewünschte Eigenschaft der geforderten Achse zuordnen. Zusätzlich wird an die Visualisierungstechnik die Anforderung der Übersichtlichkeit an die Visualisierungstechnik gestellt. Die Ausgabe in Form einer Punktewolke kann diese Anforderung nicht erfüllen. Entsprechend wurde eine „Hover“-Funktion programmiert, welche dem Nutzer ermöglicht beim Anklicken eines Punktes dessen informatorischen Bezug zu dem konkreten Fahrzeug auszugeben. Neben diesen Informationen färbt sich der Punkt rot, für eine bessere Gesamtübersicht im Scatterplot. Schlussfolgernd konnte die Anforderung der Übersichtlichkeit mit Hilfe der integrierten Funktion gelöst werden.

Bei der Visualisierung der Parallelen Daten liegt der Fokus darauf, Zusammenhänge zwischen mehr als zwei Daten zu verdeutlichen.

## Präsentation der Visualisierung

### Visualisierung Eins

Abbildung 2: Scatterplot (Quelle: eigene Darstellung)

### Visualisierung Zwei

Abbildung 3: Parallele Koordinaten (Quelle: eigene Darstellung)

### Visualisierung Drei

Abbildung 4: Auszug aus der Baumhierarchie

## Interaktion

# Implementierung

# Anwendungsfälle

## Anwendung Visualisierung Eins

Abbildung 5: Scatterplot welcher die Eigenschaften Preis und Körper gegenüberstellt (Quelle: eigene Darstellung)

## Anwendung Visualisierung Zwei

Abbildung 6: Gegenüberstellung der Eigenschaften Körper, Gerbstoffe, Süße und Säuregehalt mithilfe der Parallelen Koordinaten (Quelle: eigene Darstellung)

## Anwendung Visualisierung Drei

Abbildung 7: Ausschnitt mit Asiatischen Weinen aus der Baumhierarchie (Quelle: eigene Darstellung)

# Verwandte Arbeiten

# Zusammenfassung und Ausblick

# Literaturverzeichnis

1. **Lovelace, A.** *Notes upon L. F. Menabrea’s “Sketch of the Analytical Engineinvented by Charles Babbage”.* 1842.

2. **Goodfellow, Ian, Bengio, Yoshua und Courville, Aaron.** *Deep Learning, Das umfassende Handbuch. Grundlagen, aktuelle Verfahren und Algorithmen, neue Forschungsansätze.* s.l. : MITP-Verlag, 2018. S. 912.

3. **https://www.kaggle.com/avikasliwal/used-cars-price-prediction?select=train-data.csv.** https://www.kaggle.com/avikasliwal/used-cars-price-prediction?select=train-data.csv. [Online] 26.11.2021. [Zitat vom: 26. 11 2021.] https://www.kaggle.com/avikasliwal/used-cars-price-prediction?select=train-data.csv.

4. **Von deskriptiven Analysen bis Machine-Learning-Algorithmen . [Online] [Zitat vom: 27. 11 2021.] https://www.industry-of-things.de/von-deskriptiven-analysen-bis-machine-learning-algorithmen-a-898215/.**

**5. Yi, Mike. A Complete Guide to Scatter Plots. *Chartio.* [Online] 27. 11 2021. https://chartio.com/learn/charts/what-is-a-scatter-plot/.**

# Anhang

## Git Historie

Literatur

1. (2020) Kurz gefasst: Künstliche neuronale Netze in 3 Minuten. https://blog.fintechcube.com/kuenstliche-neuronale-netze-in-3-minuten/. Zugegriffen: 29. November 2021

2. (2021) Parallele Koordinaten. https://de.wikipedia.org/wiki/Parallele\_Koordinaten. Zugegriffen: 29. November 2021

3. Auto Scout24. https://www.autoscout24.de/auto/gebrauchtwagen/. Zugegriffen: 05. Dezember 2021

4. Co-Learning Lounge. https://colearninglounge.com/. Zugegriffen: 01. Dezember 2021

5. dictionary.com. https://www.dictionary.com/browse/lakh. Zugegriffen: 03. Dezember 2021

6. finanzen.net. https://www.finanzen.net/devisen/euro-indische\_rupie-kurs. Zugegriffen: 24. November 2021

7. Goodfellow I, Bengio Y, Courville A (Hrsg) (2018) Deep Learning. Das umfassende Handbuch. Grundlagen, aktuelle Verfahren und Algorithmen, neue Forschungsansätze. MITP-Verlag

8. Gumm,H.,Peter Bäume. Bäume, Binärbäume, Traversierungen, abstrakte Klassen, Binäre Suchbäume, Balancierte Bäume, AVL-Bäume, Heaps, Heapsort, Priority queues. https://www.mathematik.uni-marburg.de/~gumm/Lehre/SS07/PraktischeInformatikII/Folien/08%20Baeume.pdf. Zugegriffen: 20. November 2021

9. Kasliwal A Used Cars price Prediction. Predict the price of an unknown car. Build your own Algo for cars 24!! https://www.kaggle.com/colearninglounge/used-cars-price-prediction. Zugegriffen: 29. November 2021

10. Lindner M-L (2021) Elm-Projekt-Used-Cars---main. https://github.com/milu1992/Elm-Projekt-Used-Cars---main. Zugegriffen: 19. Dezember 2021

11. Lovelace A (1842) Notes upon L. F. Menabrea’s “Sketch of the Analytical Engineinvented by Charles Babbage”

12. mobile.de. https://www.mobile.de/. Zugegriffen: 05. Dezember 2021

13. Nantasenamat C (2020) The Data Science Process. A Visual Guide to Standard Procedures in Data Science. https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b. Zugegriffen: 30. November 2021

14. Traeger M, Eberhart A, Geldner G, Morin AM, Putzke C, Wulf H, Eberahrdt,L.,H.,J. (2003) Künstliche neuronale Netze. Theorie und Anwendungen in der Anästhesie, Intensiv- und Notfallmedizin. Der Angesthesist:1055–1061

15. Yi M A Complete Guide to Scatter Plots. Chartio. https://chartio.com/learn/charts/what-is-a-scatter-plot/. Zugegriffen: 27. November 2021