

Multivariate Datenanalyse

**Thema: Autopreise**



vorgelegt von

**Clement Mabou Fodjo**

Matrikelnummer: 5336614

SS 2025

Prof. Dr. Julia Meister

Gießen, 13 September 2025

**Inhaltverzeichnis**

[Einleitung 1](#_Toc208658951)

[A. Datengrundlage 2](#_Toc208658952)

[A.1 Quantitative Prädiktoren 2](#_Toc208658953)

[A.2 Qualitative Prädiktoren 2](#_Toc208658954)

[B. Datenbereinigung 3](#_Toc208658955)

[B.1 Duplikate 3](#_Toc208658956)

[B.2 Unlogische Werte 3](#_Toc208658957)

[B.3 Datentypen 4](#_Toc208658958)

[B.4 Fehlende Werte 4](#_Toc208658959)

[B.5 Außreiser 5](#_Toc208658960)

[C. Analyse der Daten und Ableitung von Aussagen 5](#_Toc208658961)

[C.1 Univariate Datenanalyse 5](#_Toc208658962)

[C.2 Multivariate Datenanalyse 7](#_Toc208658963)

[C.2.1 Scatterplots : Zielvariable vs quantitative Prädiktoren 7](#_Toc208658964)

[C.2.2 Korrelationsplot: lineare Zusammenhänge zwischen quantitativen Variablen 8](#_Toc208658965)

[C.2.3 Boxplots : Zielvariable vs qualitative Prädiktoren 9](#_Toc208658966)

[C.3 Multiple lineare Regression 11](#_Toc208658967)

[C.3.1 Datenaufteilung für Modellbildung und Validierung 11](#_Toc208658968)

[C.3.2 Feature Selektion und Modellbildung 11](#_Toc208658969)

[C.3.3 Überprüfung der Modellannahmen 12](#_Toc208658970)

[C.3.4 Verbesserung des Modells 13](#_Toc208658971)

[C.3.5 Performance des verbesserten Modells 14](#_Toc208658972)

[D. Fazit und Ausblick 16](#_Toc208658973)

[Literatur 17](#_Toc208658974)

**Abbildungsverzeichnis**

[Abbildung 1: quantitative Prädiktoren 2](#_Toc208658549)

[Abbildung 2: qualitative Prädiktoren 3](#_Toc208658550)

[Abbildung 3: Struktur des Datensatzes cars\_df 4](#_Toc208658551)

[Abbildung 4: fehlende Werte im Datensatz cars\_df 5](#_Toc208658552)

[Abbildung 5: Ausreißer im Datensatz cars\_df 5](#_Toc208658553)

[Abbildung 6: Univariate Analyse der Zielvariable Preis 6](#_Toc208658554)

[Abbildung 7: Univariate Analyse der qualitative unabhängige Variable Hersteller 6](#_Toc208658555)

[Abbildung 8: Preis vs quantitative Prädiktoren 7](#_Toc208658556)

[Abbildung 9: lineare Zusammenhänge zwischen quantitativen Variablen 8](#_Toc208658557)

[Abbildung 10: Boxplots: Preis vs qualitative Prädiktoren 9](#_Toc208658558)

[Abbildung 11: Boxplots: Preis vs qualitative Prädiktoren nach Binning 10](#_Toc208658559)

[Abbildung 12: Modell nach der Stepwise-Selektion 11](#_Toc208658560)

[Abbildung 13: Diagnoseplots zur Überprüfung der Modellannahmen 12](#_Toc208658561)

[Abbildung 14: Box-Cox Transformation 13](#_Toc208658562)

[Abbildung 15: Das verbessertes Modell 13](#_Toc208658563)

[Abbildung 16: Diagnoseplots zur Überprüfung der Modellannahmen des verbessertes Modells 14](#_Toc208658564)

[Abbildung 17: Modellgüte im Trainings- und Testdatensatz (ohene Außreiser und fehlende Werte) 14](#_Toc208658565)

[Abbildung 18: Tatsächliche Werte vs vorhergesagte Werte 14](#_Toc208658566)

[Abbildung 19: Modellgüte im Trainings- und Testdatensatz mit fehlenden daten 15](#_Toc208658567)

[Abbildung 20: Modellgüte im Trainings- und Testdatensatz mit Außreiser 15](#_Toc208658568)

[Abbildung 21: Modellgüte im Trainings- und Testdatensatz mit fehlenden Werte und Außreiser 15](#_Toc208658569)

**Ehrenwörtliche Erklärung**

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig und unter ausschließlicher Verwendung der angegebenen Literatur und Hilfsmittel erstellt habe. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde in gleicher oder ähnlicher Form vorgelegt und auch nicht veröffentlicht.

**Gießen, 13. 09. 2024**

**Clement Mabou Fodjo**

# Einleitung

Automobile sind nicht nur ein bedeutendes Konsumgut, sondern auch ein zentrales Element moderner Gesellschaften. Für viele private Haushalte zählt der Kauf eines Autos zu den größten finanziellen Entscheidungen im Alltag. Sowohl für Privatpersonen als auch für Unternehmen, insbesondere in der Automobilindustrie und im Handel, sind Kenntnisse über die Preisbildung von Fahrzeugen daher von zentraler Bedeutung.

Es stellt sich die Frage: Welche Faktoren beeinflussen den Preis eines Autos? Sind es der Motor, der Kraftstoffverbrauch, die Anzahl der Türen oder die Marke? Und wie lassen sich aus diesen Faktoren die Autopreise vorhersagen? In dieser Arbeit wird untersucht, welche dieser und weiterer Variablen den Preis eines Fahrzeugs maßgeblich bestimmen und wie stark ihr jeweiliger Einfluss ist. Mithilfe statistischer Verfahren sollen die relevanten Zusammenhänge aufgezeigt werden. Im Mittelpunkt stehen dabei Methoden der explorativen und der multivariaten Datenanalyse, insbesondere die Regressionsanalyse.

Die Arbeit gliedert sich in vier Hauptabschnitte: Der erste Abschnitt stellt die Datengrundlage vor und beschreibt sie systematisch. Im zweiten Abschnitt werden die erforderlichen Schritte der Datenbereinigung detailliert erläutert, um eine saubere und belastbare Basis für die anschließende Analyse zu gewährleisten. Der dritte Abschnitt widmet sich der eigentlichen Datenanalyse. Abschließend fasst der vierte Abschnitt die Ergebnisse zusammen und gibt einen Ausblick auf mögliche Erweiterungen der Untersuchung.

# A. Datengrundlage

Für die vorliegende Untersuchung stehen Daten zu über 200 Automobilen zur Verfügung. Die Daten wurden in der Datei autos.csv bereitgestellt und enthalten insgesamt 25 Merkmale, die unterschiedliche Aspekte von Fahrzeugen beschreiben. In der Arbeit wird diese CSV-Datei unter dem Namen **"cars\_df"** in einem Dataframe gespeichert. Die Merkmale lassen sich in folgende Gruppen einteilen:

* **Zielvariable** : Preis des Autos .
* **Unabhängige Variablen (Prädiktoren)** : Die unabhängigen Variablen, auch Prädiktoren genannt, sind Merkmale, die den Zielwert ( preis) beeinflussen können. Diese Gruppe lässt sich in zwei Kategorien unterteilen: die quantitative Prädiktoren und die qualitative Prädiktoren

## A.1 Quantitative Prädiktoren

|  |  |
| --- | --- |
| **Merkmale** | **Beschreibung** |
| Risikofaktor | Preis-Risiko des Autos (-3 bis 3) |
| radstand | Abstand zwischen Vorder- und Hinterachse |
| laenge | Fahrzeuglänge |
| breite | Fahrzeugbreite |
| hoehe | Fahrzeughöhe |
| gewicht | Leergewicht des Fahrzeugs |
| anzahl\_zylinder | Beschreibt, wie viele Zylinder der Motor hat |
| motor\_groesse | Größe des Motors |
| kolben\_durchmesser | Durchmesser der Zylinder im Motor |
| Motortaktung | Beschreibt die Drehzahl oder Frequenz, mit der der Motor arbeitet |
| compressionsrate | Verhältnis des Zylindervolumens bei Kolben unten zu Kolben oben |
| PS | Pferdestärke |
| Drehzahl\_max | Maximale Motordrehzahl |
| Verbrauch\_stadt\_mpg | Kraftstoffverbrauch im Stadtverkehr |
| Verbrauch\_autobahn\_mpg | Kraftstoffverbrauch auf der Autobahn |
| tueren | Anzahl der Türen des Fahrzeugs |

**Abbildung 1:** quantitative Prädiktoren

## A.2 Qualitative Prädiktoren

|  |  |
| --- | --- |
| **Merkmale** | **Beschreibung** |
| hersteller | Hersteller oder die Marke des Autos |
| kraftstoff | Art des verwendeten Kraftstoffs (Diesel oder Gas) |
| motoraufladung | Zeigt, wie der Motor Luft bekommt. „Standard“ = normale Ansaugung, „Turbo“ = Luft wird mit Lader hineingepresst |
| karosserie\_form | Karosserieform: Äußere Bauweise des Fahrzeugs (Hatchback: Kompakt, Hecktür öffnet nach oben, Wagon: Längere Karosserie, großer Laderaum, Convertible: Offenes Fahrzeug, klappbares Dach, Sedan: Klassisch, vier Türen, separater Kofferraum, Hardtop: Sportlich-elegantes Design) |
| antrieb\_typ | Gibt an, welche Räder die Motorleistung auf die Straße übertragen( 4wd: Allradantrieb, fwd: Frontantrieb, rwd: Heckantrieb) |
| motor\_pos | Position des Motors im Fahrzeug (front = vorne, rear = hinten). |
| motorsteuerung | Art der Steuerung des Motors (dohc, dohcv, l, ohc, ohcf, ohcv, rotor) |
| kraftstoff\_system | beschreibt, wie der Kraftstoff dem Motor zugeführt und eingespritzt wird. (1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi) |

**Abbildung 2:** qualitative Prädiktoren

# B. Datenbereinigung

Die Datenbereinigung ist ein zentraler Schritt jeder empirischen Analyse. Nur wenn die verwendeten Daten korrekt, konsistent und weitgehend fehlerfrei sind, lassen sich verlässliche und aussagekräftige Ergebnisse erzielen. In Datensätzen treten häufig Unstimmigkeiten auf, wie z. B. falsche Datentypen, fehlende Werte, Duplikate, unplausible Angaben oder Ausreißer. Ziel der Datenbereinigung ist es, diese Probleme zu erkennen und angemessen zu behandeln.

## B.1 Duplikate

Duplikate entstehen, wenn ein Datensatz mehrfach denselben Eintrag enthält. Diese wiederholten Daten können Analysen verzerren, z. B. Mittelwerte verschieben oder Häufigkeiten verfälschen. In "cars\_df" wurden keine doppelte Zeilen identifiziert.

## B.2 Unlogische Werte

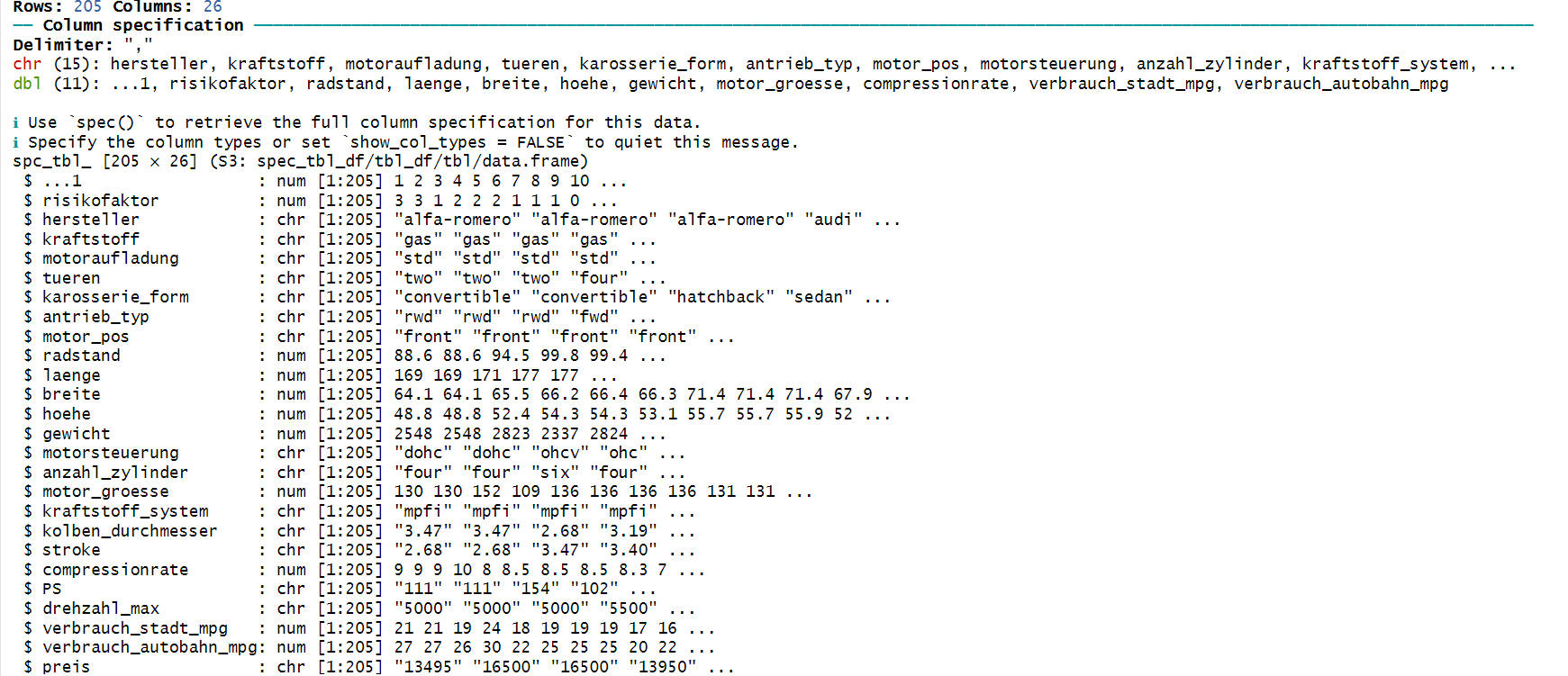
In Datensätzen können unlogische Werte auftreten. Aus der zusammenfassenden Statistik der numerischen Variablen sowie den Häufigkeitstabellen der kategorialen Variablen lassen sich solche Auffälligkeiten durch eine genaue Betrachtung von Minimum- und Maximumwerten sowie der Kategorienamen identifizieren. Für den Datensatz "cars\_df" ergeben sich folgende Unstimmigkeiten:

* **kolben\_durchmesser, stroke, PS, drehzahl\_max, preis, tueren**: In diesen Variablen treten Einträge wie „?“ auf. Diese Kodierung für fehlende Werte ist unlogisch und sollte durch NA ersetzt werden, um eine korrekte statistische Verarbeitung zu gewährleisten.
* **motoraufladung**: In dieser Spalte sind die Werte als *std* und *turbo* gespeichert, obwohl laut Datenbeschreibung die Bezeichnungen *standard* und *turbo* vorgesehen sind.

Zur Bereinigung werden alle „?“ konsistent durch NA ersetzt. Zusätzlich wird in der Spalte *motoraufladung* der Wert *std* mit *standard* ersetzt, um die Schreibweise zu vereinheitlichen.

## B.3 Datentypen

Eine erste Überprüfung der Datenstruktur (str(cars\_df)) zeigt, dass mehrere Spalten nicht im passenden Datentyp vorliegen. Dazu gehören:tueren**,** anzahl\_zylinder**,** PS**,** motor\_groesse**,** kolben\_durchmesser**,** stroke**,** drehzahl\_maxund preis.Diese Spalten sind als Zeichenketten (chr) gespeichert, obwohl sie numerische Werte repräsentieren sollten.Die Spalte tueren enthält Einträge wie "two" oder "four". Für die Analyse müssen diese Angaben in numerische Werte umgewandelt werden ("two" → 2, "four" → 4). Ähnliche Anpassungen werden für die übrigen Spalten vorgenommen, um eine konsistente und analysierbare Datenbasis zu schaffen.



**Abbildung 3:** Struktur des Datensatzes cars\_df

## B.4 Fehlende Werte

Fehlende Werte sind ein häufiges Problem in realen Datensätzen und können die Ergebnisse statistischer Analysen verfälschen. Im Datensatz "cars\_df" wurden fehlende Einträge in den Spalten Kolben\_durchmesser, stroke, preis, tueren und PS identifiziert.

Da der Anteil der fehlenden Werte in diesen Spalten weniger als 2 % der Gesamtdaten ausmacht, werden die betroffenen Zeilen aus dem Datensatz entfernt. Dieses Vorgehen ist gerechtfertigt, da der Verlust einer so geringen Datenmenge die Repräsentativität des Datensatzes nicht wesentlich beeinträchtigt und somit eine robuste und verlässliche Grundlage für die anschließende Analyse sichergestellt wird.

Ein Bild, das Text, Quittung, Screenshot, Schrift enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Abbildung 4:** fehlende Werte im Datensatz cars\_df

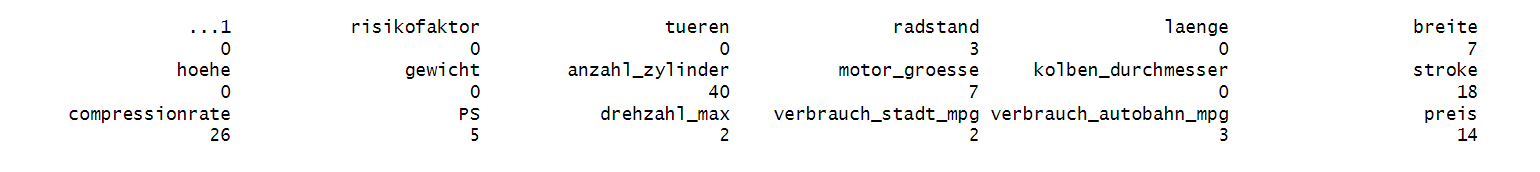
## B.5 Außreiser

Ausreißer sind Beobachtungen, die sich deutlich von der Mehrheit der Daten unterscheiden und dadurch die Ergebnisse statistischer Analysen verzerren können. Im cars\_df werden Ausreißer in numerischen Spalten mithilfe der **IQR-Methode (Interquartilsabstand)** identifiziert:

* Für jede numerische Spalte werden das 1. Quartil (Q1), das 3. Quartil (Q3) und der IQR berechnet:

IQR=Q3−Q1

* Die Grenzen für Ausreißer werden definiert als:
* untere Grenze=Q1−1,5⋅IQR
* obere Grenze=Q3+1,5⋅IQR
* Werte außerhalb dieser Grenzen gelten als Ausreißer.



**Abbildung 5:** Ausreißer im Datensatz cars\_df

Die Behandlung erfolgt durch **Winsorisierung**:

* Werte unterhalb der unteren Grenze werden auf die untere Grenze gesetzt.
* Werte oberhalb der oberen Grenze werden auf die obere Grenze gesetzt.

Dieses Vorgehen reduziert die Verzerrung durch extreme Werte, ohne Datenpunkte vollständig zu entfernen.

# C. Analyse der Daten und Ableitung von Aussagen

## C.1 Univariate Datenanalyse

Die univariate Datenanalyse dient dazu, jede Variable einzeln zu betrachten und ein grundlegendes Verständnis ihrer Verteilung, Skalierung und Wertebereiche zu gewinnen. Sie ermöglicht zudem die Überprüfung der Datenaufbereitung, etwa im Hinblick auf Ausreißer, fehlende Werte oder andere Unstimmigkeiten. Durch diese Vorarbeit können potenzielle Probleme frühzeitig identifiziert und geeignete Maßnahmen wie Transformationen oder Bereinigungen ergriffen werden. Im Folgenden wird als Beispiel die univariate Analyse einer quantitativen Variable („Preis“) und einer kategorialen Variable („Hersteller“) dargestellt und

Interpretiert.

Ein Bild, das Text, Diagramm, Screenshot, Design enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Abbildung 6:** Univariate Analyse der Zielvariable Preis

Die Zielvariable **„preis“** zeigt deutlich, dass die Preisverteilung der Autos rechtsschief ist. Dies erkennt man daran, dass der Mittelwert (12.829) deutlich höher liegt als der Median (10.245). Das bedeutet, dass die meisten Fahrzeuge im Datensatz im unteren Preissegment liegen. Das Histogramm hat seinen Höhepunkt unter 10.000, und die Dichtekurve zeigt eine starke Konzentration der Werte im Bereich unter 15.000. Einige wenige sehr teure Autos sind ebenfalls vorhanden, die den Mittelwert nach oben ziehen und somit die Schiefe der Verteilung verursachen. Wenn die Zielvariable eine rechtsschief aufweist wie hier , kann es manchmal sinnvoll sein, diese mithilfe einer logarithmischen oder Quadratwurzel-Transformation zu reduzieren, um ein besseres Regressionsmodell zu bekommen. Die Standardabweichung von 6.792,89 zeigt zudem eine relativ große Streuung der Preise um den Mittelwert.

Ein Bild, das Text, Screenshot, Diagramm, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Abbildung 7:** Univariate Analyse der qualitative unabhängige Variable Hersteller

Die Variable **„Hersteller“** ist eine kategoriale Variable mit insgesamt 21 Ausprägungen. Toyota stellt mit 16,6 % den größten Anteil der Fahrzeuge, gefolgt von Nissan. Die Mehrheit der übrigen Hersteller ist jedoch nur mit sehr wenigen Autos im Datensatz vertreten. Die Vielzahl an Ausprägungen kann zu einer unnötig hohen Modellkomplexität führen. Sollte sich zeigen, dass der Hersteller einen Einfluss auf den Preis hat, wäre es daher sinnvoll, die Ausprägungen durch **Binning** in Gruppen zusammenzufassen. Auf diese Weise ließe sich die Anzahl der Dummy-Variablen verringern und ein kompakteres sowie besser interpretierbares Regressionsmodell erstellen.

## C.2 Multivariate Datenanalyse

Die **multivariate Datenanalyse** untersucht Zusammenhänge zwischen zwei oder mehr Variablen gleichzeitig. Sie ermöglicht, Muster, Korrelationen oder Abhängigkeiten zwischen Variablen zu erkennen, die in der univariaten Analyse nicht sichtbar sind.

### C.2.1 Scatterplots : Zielvariable vs quantitative Prädiktoren

Ein Bild, das Text, Schrift, Screenshot, Zahl enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Abbildung 8:** Preis vs quantitative Prädiktoren

Die Abbildung 6 zeigt eine Reihe von Scatterplots (Streudiagrammen), in denen der Preis von Autos in Beziehung zu verschiedenen quantitativen Merkmalen gesetzt wird. Jeder Punkt im Diagramm repräsentiert ein einzelnes Fahrzeug, wobei seine Position auf der x- und y-Achse die Werte der beiden betrachteten Variablen angibt. Diese Visualisierung erleichtert die Identifikation potenzieller Zusammenhänge zwischen den Variablen.

Die Diagramme lassen sich anhand der erkennbaren Trends in drei Gruppen einteilen:

* **Starke positive lineare Zusammenhänge:** Bei diesen Variablen steigt der Preis deutlich mit zunehmendem Wert des Merkmals. Beispiele hierfür sind die Zusammenhänge zwischen Preis und Gewicht, Preis und PS sowie Preis und Motorgröße.
* **Starke negative lineare Zusammenhänge:** Hier sinkt der Preis, je höher der Wert des Merkmals ist. Dies ist insbesondere bei den Diagrammen Preis vs. Verbrauch auf der Autobahn und Preis vs. Verbrauch in der Stadt zu beobachten.
* **Schwache Zusammenhänge:** Bei diesen Variablen ist kein klarer linearer Trend erkennbar. Die Punkte sind breit gestreut, was auf einen geringen oder nichtlinearen Zusammenhang hinweist. Beispiele hierfür sind Preis vs. maximale Drehzahl oder Preis vs. Kompressionsrate.

### C.2.2 Korrelationsplot: lineare Zusammenhänge zwischen quantitativen Variablen

Ein Bild, das Text, Screenshot, Zahl, Schrift enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Abbildung 9:** lineare Zusammenhänge zwischen quantitativen Variablen

Der Korrelationsplot visualisiert die Stärke und Richtung der linearen Beziehungen zwischen den numerischen Variablen. Positive Korrelationen (blaue Kreise) bedeuten, dass ein Anstieg einer Variable mit einem Anstieg der anderen einhergeht, während negative Korrelationen (rote Kreise) das Gegenteil anzeigen.

Die Zielvariable **„preis“** zeigt eine **starke positive Korrelation** mit folgenden Prädiktoren: **motor\_groesse** (0,88), **gewicht** (0,88), **PS** (0,82), **anzahl\_zylinder** (0,75), **laenge** (0,73), **radstand** (0,61) und **kolben\_durchmesser** (0,57).

Zudem besteht eine **starke negative Korrelation** mit dem Verbrauch, insbesondere mit **verbrauch\_autobahn\_mpg** (-0,74) und **verbrauch\_stadt\_mpg** (-0,76). Dies bedeutet, dass teurere Autos tendenziell weniger kraftstoffeffizient sind.

Auch zwischen den Prädiktoren selbst gibt es starke Beziehungen. **Gewicht** korreliert stark mit **motor\_groesse** (0,88) und **PS** (0,87). Ebenso sind **verbrauch\_autobahn\_mpg** und **verbrauch\_stadt\_mpg** mit 0,97 nahezu perfekt korreliert. Diese hohen Korrelationen deuten auf **Multikollinearität** hin, was bei der Erstellung von linearen Modellen unbedingt berücksichtigt werden muss, da stark korrelierte Prädiktoren die Stabilität und Interpretierbarkeit der Modellschätzungen beeinträchtigen können.

### C.2.3 Boxplots : Zielvariable vs qualitative Prädiktoren

Die **Boxplots** ermöglichen einen direkten visuellen Vergleich der Verteilung der Zielvariable **„preis“** über verschiedene kategoriale Prädiktoren hinweg.

Ein Bild, das Text, Diagramm, Screenshot, Plan enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Abbildung 10:** Boxplots: Preis vs qualitative Prädiktoren

Die rote Linie innerhalb der Boxen repräsentiert den **Median**. Der Medianpreis variiert je nach Hersteller stark: Jaguar liegt am höchsten, Toyota und Chevrolet niedriger, bei den meisten Marken ist die Preisspanne relativ gering. Bei der Karosserieform haben Hardtops und convertibles die höchsten Preise, Hatchbacks liegen am unteren Ende. Fahrzeuge mit Heckantrieb sind am teuersten, gefolgt von Allrad- und Vorderradantrieb, während Dieselfahrzeuge teurer als Benziner sind. Turbomotoren haben höhere Medianpreise als Standardmotoren.

Um die Vielzahl an Ausprägungen der kategorialen Variablen zu reduzieren und damit die Interpretierbarkeit sowie die Verwendbarkeit in späteren Analysen zu verbessern, wurde ein **Binning** durchgeführt. Dabei werden ähnliche Kategorien zusammengefasst und in übergeordneten Gruppen gebündelt.

Im cars\_df wurden mehrere Variablen basierend auf der **Erklärungsstärke ihrer Ausprägungen für den Preis** ( siehe Boxplots) entsprechend neu codiert.

* **Hersteller**: Die einzelnen Automarken wurden in drei Gruppen zusammengefasst:
  + **LS (Luxussegment)**: Jaguar, Mercedes-Benz, Porsche, BMW
  + **MS (mittleres Segment)**: Volvo, Audi, Peugeot, Mercury, Alfa-Romeo, Saab, Volkswagen
  + **NS (niedriges Segment)**: Toyota, Isuzu, Mazda, Mitsubishi, Nissan, Subaru, Plymouth, Honda, Dodge, Chevrolet
* **Kraftstoffsystem**: Unterschiedliche Einspritz- und Vergasersysteme wurden in drei Sammelkategorien gebündelt: „mpfi\_idi“, „mfi\_spfi\_spdi“ sowie „2bbl\_1bbl“.
* **Karosserieform**: Die Fahrzeuge wurden in die Kategorien „hardtop\_convertible“, „wagon\_sedan“ und „hatchback“ zusammengefasst.
* **Motorsteuerung**: Verschiedene Ventil- und Steuerungssysteme wurden in die Gruppen „ohcv“, „dohc\_l“ und „ohcf\_ohc“ eingeordnet.
* **Antriebstyp**: Zur Vereinfachung wurden Fahrzeuge mit Hinterradantrieb als „rwd“ bezeichnet, während Front- und Allradantriebe in der Kategorie „4wd\_fwd“ zusammengefasst wurden.

Durch dieses Vorgehen wird die **Anzahl der Kategorien deutlich reduziert**, wodurch die **Erklärungsstärke der neuen Ausprägungen qualitativer Prädiktoren** verbessert und die Datenstruktur für nachfolgende Analysen übersichtlicher wird.

Ein Bild, das Diagramm, Text, Plan, Screenshot enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Abbildung 11:** Boxplots: Preis vs qualitative Prädiktoren nach Binning

## C.3 Multiple lineare Regression

### C.3.1 Datenaufteilung für Modellbildung und Validierung

Zur Modellbildung wurden die verfügbaren Daten in Trainings- und Testdaten aufgeteilt. Dabei wurden 80 % der Beobachtungen per Zufallsauswahl als Trainingsdaten verwendet, während die verbleibenden 20 % als Testdaten dienten. Das Trainingsset (*train\_df*) wurde zur Schätzung der Modellparameter genutzt, während das Testset (*test\_df*) zur anschließenden Validierung der Modellgüte diente.

### C.3.2 Feature Selektion und Modellbildung

Zunächst wurden alle Prädiktoren ausgeschlossen, deren Korrelation mit der Zielvariablen kleiner als 0,2 war, da ihr Erklärungsbeitrag als vernachlässigbar einzustufen ist. Betroffen waren die Variablen *risikofaktor*, *türen*, *höhe*, *stroke*, *compressionrate* und *drehzahl\_max*.

Zur Reduktion multikollinearer Variablen wurde bei den quantitativen Prädiktoren die Korrelationsmatrix verwendet. Wiesen zwei Variablen eine Korrelation von mehr als 0,8 auf, wurde jeweils die mit dem geringeren Einfluss auf die Zielgröße (Preis) entfernt. Auf diese Weise wurden *verbrauch\_stadt\_mpg*, *gewicht*, *radstand*, *PS* und *länge* eliminiert.

Die verbleibenden Variablen dienten als Grundlage für die Modellbildung. Hierzu wurde eine Stepwise-Regression eingesetzt, die Forward- und Backward-Selection kombiniert. Ausgehend von einem vollständigen Modell mit allen Variablen sowie einem leeren Modell entschied das AIC-Kriterium iterativ über die Aufnahme oder den Ausschluss von Prädiktoren. Das finale Modell umfasst somit nur die aussagekräftigsten Variablen bei gleichzeitig minimaler Redundanz.

Ein Bild, das Text, Screenshot, Schrift, Zahl enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Abbildung 12:** Modell nach der Stepwise-Selektion

### C.3.3 Überprüfung der Modellannahmen

Zur Beurteilung der Gültigkeit der Annahmen des linearen Regressionsmodells wurden vier Standard-Diagnoseplots herangezogen: Residuen vs. Fitted Values, Scale-Location, Q-Q Plot der Residuen sowie Residuen vs. Leverage. Diese Plots ermöglichen die Überprüfung von Linearität, Varianzkonstanz, Normalverteilung der Residuen und dem Einfluss potenzieller Ausreißer.

Ein Bild, das Text, Diagramm, Reihe, Karte enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Abbildung 13:** Diagnoseplots zur Überprüfung der Modellannahmen

Die Annahmen der Linearität und Homoskedastizität können anhand der Plots „Residuals vs. Fitted“ und „Scale-Location“ überprüft werden. Die leichte Krümmung der roten Linie im „Residuals vs. Fitted“-Plot weist lediglich auf geringe Abweichungen von der Linearität hin. Im „Scale-Location“-Plot deuten die Schwankung der roten Linie sowie die ungleichmäßige Streuung der Residuen um die vorhergesagten Werte auf eine Tendenz zur Heteroskedastizität hin.

Die Normalverteilungsannahme der Residuen wird durch den Q-Q-Plot überprüft. Die meisten Punkte liegen eng an der Linie, nur einige Extremwerte weichen leicht ab. Damit ist die Normalität der Residuen weitgehend erfüllt.

Potenzielle Ausreißer und einflussreiche Beobachtungen lassen sich aus dem Plot „Residuals vs. Leverage“ erkennen. Einzelne Punkte mit höheren Residuen oder Hebeleffekten überschreiten jedoch nicht die Schwellenwerte der Cook’s Distance und haben nur einen geringen Einfluss auf das Modell.

Zusammenfassend zeigt die Modellprüfung eine leichte Tendenz zur Heteroskedastizität, die weiterhin überprüft werden sollte. Gegebenenfalls könnten Transformationen der Variablen eingesetzt werden, um die Modellannahmen weiter zu verbessern. Eine mögliche Lösung zur Reduzierung des Heteroskedastizitätsproblems besteht darin, eine Logarithmus- oder Quadratwurzeltransformation der Zielgröße anzuwenden.

### C.3.4 Verbesserung des Modells

Die Box-Cox-Transformation unterstützt bei der Bestimmung einer geeigneten Transformation der Zielvariable. In unserem Fall liegt das Maximum des Schätzers für λ nahe bei 0,5, was darauf hindeutet, dass eine Quadratwurzeltransformation sinnvoll ist.

Ein Bild, das Text, Reihe, Diagramm enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Abbildung 14:** Box-Cox Transformation

Die Ergebnisse des verbesserten Modells lassen sich wie folgt darstellen:

**Ein Bild, das Text, Screenshot, Schrift, Zahl enthält.

KI-generierte Inhalte können fehlerhaft sein.**

**Abbildung 15:** Das verbessertes Modell

Das neue Modell erfüllt die Annahmen der linearen Regression deutlich besser: die Beziehung zwischen den Variablen ist linearer, die Residuen zeigen eine höhere Homoskedastizität und nähern sich stärker der Normalverteilung an.

Ein Bild, das Text, Diagramm, Reihe, parallel enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Abbildung 16:** Diagnoseplots zur Überprüfung der Modellannahmen des verbessertes Modells

### C.3.5 Performance des verbesserten Modells

Das Modell zeigt eine sehr gute Vorhersagekraft: Es erklärt 92,7 % der Varianz in den Trainingsdaten und 94,4 % in den Testdaten. Die ähnliche oder sogar leicht bessere Leistung im Testset deutet darauf hin, dass das Modell robust ist und gut generalisiert, ohne Anzeichen von Overfitting. Insgesamt liefert es konsistente und zuverlässige Preisvorhersagen.

Ein Bild, das Text, Screenshot, Schrift, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Abbildung 17:** Modellgüte im Trainings- und Testdatensatz (ohene Außreiser und fehlende Werte)

Ein Bild, das Text, Reihe, Diagramm, Screenshot enthält.

KI-generierte Inhalte können fehlerhaft sein.

**Abbildung 18:** Tatsächliche Werte vs vorhergesagte Werte

Wie wäre die Modellgüte ausgefallen, wenn zu Beginn der Analyse fehlende Werte und Ausreißer unberücksichtigt geblieben wären?

Ein Bild, das Text, Screenshot, Schrift, Reihe enthält.

KI-generierte Inhalte können fehlerhaft sein.

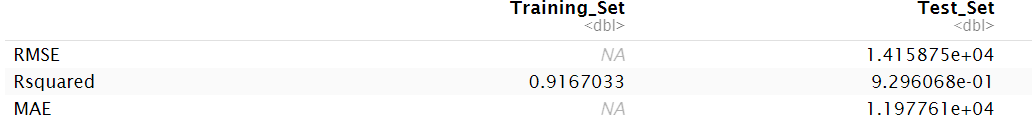
**Abbildung 19:** Modellgüte im Trainings- und Testdatensatz mit fehlenden daten

Abbildungen 17 und 19 zeigen ähnliche Ergebnisse, wobei bei den Datensätzen mit fehlenden Werten eine leichte Verbesserung erkennbar ist. Das Fehlen von Daten scheint die Modellgüte in diesem speziellen Fall nicht negativ zu beeinflussen. Dies könnte darauf liegen, dass die Anzahl der fehlenden Werte gering ist und die betroffenen Prädiktoren bei der Merkmalsselektion zur Modellbildung entfernt wurden.



**Abbildung 20:** Modellgüte im Trainings- und Testdatensatz mit Außreiser

Abbildung 20 im Vergleich zu Abbildung 17 zeigt, dass die Ausreißer einen negativen Einfluss auf die Modellgüte haben. Der R²-Wert sinkt, was darauf hinweist, dass das Modell einen geringeren Anteil der Varianz in den Daten erklären kann.



**Abbildung 21:** Modellgüte im Trainings- und Testdatensatz mit fehlenden Werte und Außreiser

Die Kombination aus fehlenden Werten und Ausreißern führt zu den schlechtesten Ergebnissen im Testdatensatz (Abbildung 21).

# D. Fazit und Ausblick

Die vorliegende Untersuchung zeigt, dass die Preisbildung von Automobilen durch eine Vielzahl quantitativer und qualitativer Merkmale beeinflusst wird. Quantitative Faktoren wie Motorgröße, PS, Gewicht und Anzahl der Zylinder haben einen starken positiven Einfluss auf den Fahrzeugpreis, während der Kraftstoffverbrauch negativ korreliert. Auch qualitative Merkmale, insbesondere der Hersteller, die Karosserieform und der Antriebstyp, zeigen deutliche Zusammenhänge mit den Preisen.

Die multivariate Analyse und die multiple lineare Regression ermöglichten die Erstellung eines robusten Vorhersagemodells, das über 92 % der Varianz in den Trainings- und Testdaten erklären kann. Durch Transformation der Zielvariable (Quadratwurzel) und eine gezielte Feature-Selektion konnte die Modellannahme der Linearität und Homoskedastizität deutlich verbessert werden.

Für zukünftige Untersuchungen bieten sich mehrere Erweiterungen an: Neben der Aufnahme weiterer relevanter Merkmale, etwa technischer Innovationen oder Markttrends, könnten auch nichtlineare Modelle wie Random Forests oder Gradient Boosting zur Verbesserung der Vorhersagekraft eingesetzt werden. Zudem könnte eine detaillierte Untersuchung von Interaktionen zwischen Prädiktoren zusätzliche Erkenntnisse liefern.

Insgesamt zeigt die Arbeit, dass eine sorgfältige Datenaufbereitung, gekoppelt mit systematischer Analyse und Modellierung, zuverlässige Vorhersagen der Fahrzeugpreise ermöglicht und wertvolle Einblicke in die Einflussfaktoren liefert.

# Literatur

Daniel P. McGibney. Applied Linear Regression for Business Analytics with R: A Practical Guide to Data Science with Case Studies, Springer, 2023.

Julia Meister, Vanessa Lange. MAM004 Multivariate Datenanalyse (SS2025). [online].[Zugriff im August 2025]. Verfügbar unter : <https://moodle.thm.de/course/view.php?id=9878>

Corporate Finance Institute . Data Prep for Machine Learning in Python. [online].[Zugriff im August 2025]. Verfügbar unter : <https://www.coursera.org/learn/data-prep-for-machine-learning-python>

Train in Data. Winsorization: handling outliers in machine learning. [online].[Zugriff im August 2025]. Verfügbar unter : <https://www.blog.trainindata.com/winsorization-handling-outliers-in-machine-learning/>