Deutsche Hochschule

für angewandte Wissenschaften

Datenanalyse und lineare Regression des Auto-Datensatzes aus ISLR mit Python

Modul bzw. Unit: Big Data Programming

Name: Benjamin Schneider

Datum: 30.07.2025

Inhalt

Abkürzungsverzeichnis

EU	Europäische Union
kg	Kilogramm
kW	Kilowatt
ISLR	An Introduction to Statistical Learning
MAE	Mean Absolut Error
mpg	Miles per gallon (Meilen pro Gallone)
mph	Miles per Hour (Meilen pro Stunde)
NaN	Not a Number
PS	Pferdestärke (Leistungsangabe bei PKWs)
R2	Bestimmtheitsmaß
RMSE	Root Mean Squared Error
USA	United States of America

Abbildungsverzeichnis

Tabellenverzeichnis

Hinweis zur verwendeten Sprache

Zur besseren Lesbarkeit wird in dieser Studienarbeit das generische Maskulinum verwendet. Die in dieser Arbeit verwendeten Personenbezeichnungen beziehen sich – sofern nicht anders kenntlich gemacht – auf alle Geschlechter.

1 Einleitung

Die explosionsartige Zunahme verfügbarer digitaler Daten stellt sowohl Unternehmen als auch Forschungseinrichtungen vor neue Herausforderungen, eröffnet aber zugleich erhebliche Potenziale. Die zielgerichtete Analyse großer, strukturierter Datensätze ermöglicht es, Zusammenhänge sichtbar zu machen, Prognosemodelle zu erstellen und datenbasierte Entscheidungen zu treffen. Diese Entwicklungen markieren einen fundamentalen Wandel in der datengetriebenen Wissensgenerierung (Kitchin 2014; Chen et al. 2014) Insbesondere im Bereich des maschinellen Lernens und der prädiktiven Modellierung bietet die Kombination klassischer statistischer Methoden mit skalierbaren Technologien wie Dask oder Apache Spark eine praxisnahe Antwort auf diese Herausforderungen (Hashem et al. 2015)

Ziel dieser Studienarbeit ist es, auf Basis eigener Analysen mit dem Auto-Datensatzes aus dem Werk An Introduction to Statistical Learning (Huang 2014) ein vollständiges Analyseprojekt zu realisieren. Die Umsetzung erfolgt unter Verwendung der Programmiersprache Python und relevanter Bibliotheken wie pandas, scikit-learn und Dask. Darüber hinaus wird das lineare Regressionsmodell sowohl mit gängigen Bibliotheken als auch in einer eigenständigen Implementierung (Least Squares) realisiert, um ein tieferes Verständnis für die mathematischen Grundlagen zu fördern.

Die zentrale Fragestellung der Analyse besteht darin, den Zusammenhang zwischen verschiedenen Einflussgrößen, wie PS-Zahl, Beschleunigung, dem Baujahr – und dem Kraftstoffverbrauch (in Miles per gallon) zu modellieren und zu bewerten.

Die Arbeit verfolgt einen methodischen Ansatz, der typische Schritte eines Data-Science-Projekts umfasst und darüber hinaus die Herausforderungen im Kontext wachsender Datenmengen beleuchtet. Dabei steht nicht nur die Modellgüte im Fokus, sondern auch die Skalierbarkeit der eingesetzten Methoden und deren Performance bei der Verarbeitung großer Datenmengen.

2 Datengrundlage

2.1 Beschreibung des Datensatzes

Der in dieser Arbeit verwendete Datensatz trägt den Namen „Auto“ und entstammt dem Buch An Introduction to Statistical Learning (Huang 2014). Der Datensatz enthält Informationen zu technischen und leistungsbezogenen Eigenschaften von 392 Automobilmodellen, die in den Jahren 1970 bis 1982 in den USA verkauft wurden.

Die Tabelle umfasst acht erklärende Variablen sowie eine Zielvariable, ergänzt durch den Namen des Fahrzeugmodells. Eine Übersicht über die enthaltenen Merkmale ist in Tabelle 1 dargestellt.

Tabelle 1: Variablen

Variable	**Bedeutung**	**Datentyp**
mpg	Miles per gallon (Verbrauch)	float (Ziel)
cylinders	Anzahl Zylinder	integer
displacement	Hubruam in Kubikzoll	float
horsepower	Leistung in PS	float
weight	Gewicht in Pfund	float
acceleration	Beschleunigung (Zeit von 0 auf 60 mph)	float
year	Baujahr des Fahrzeugs	integer
origin	Herkunftsland	integer
name	Modellbezeichnung des Fahrzeugs	string

Die Spalte name dient rein deskriptiven Zwecken und wurde für die Modellierung nicht berücksichtigt.

2.2 Zielvariable und Features

Im Mittelpunkt der Untersuchung steht die Zielvariable mpg, die den Kraftstoffverbrauch beschreibt. Diese Kennzahl ist sowohl aus ökonomischer als auch aus ökologischer Perspektive von hoher Relevanz, da sie zentrale Auswirkungen auf

Betriebskosten, Energieeffizienz und Treibhausgasemissionen hat (Gerst et al. 2010). Ziel dieser Arbeit ist es, diesen Verbrauch mithilfe geeigneter Einflussgrößen möglichst genau vorherzusagen.

Als potenzielle Features (Prädiktoren) wurden insbesondere folgende Merkmale identifiziert:

- horsepower: Die Motorleistung wirkt sich typischerweise negativ auf den Verbrauch aus.
- weight: Ein höheres Fahrzeuggewicht führt meist zu einem erhöhten Kraftstoffverbrauch.
- displacement: Der Hubraum korreliert mit Leistung und Masse des Fahrzeugs.
- acceleration: Kann Rückschlüsse auf das Fahrzeugkonzept und dessen Effizienz zulassen.
- year: Modelljahr, möglicherweise im Zusammenhang mit technologischem Fortschritt.
- origin: Kodierte geografische Herkunft, die Unterschiede im Fahrzeugdesign abbilden könnte.

Die Auswahl dieser Features erfolgte sowohl theoriegeleitet (technisches Verständnis von Fahrzeugen) als auch empirisch gestützt durch die im nächsten Kapitel dargestellte explorative Analyse.

2.3 Relevanz und Eignung des Datensatzes

Der Auto-Datensatz eignet sich in mehrfacher Hinsicht für das Ziel dieser Studienarbeit:

- Er ist ausreichend umfangreich, um Modellbildung und Evaluation sinnvoll durchzuführen, zugleich aber kompakt genug, um eine manuelle Implementierung zu ermöglichen.

- Die enthaltenen Merkmale sind realitätsnah, interpretierbar und spiegeln typische Herausforderungen bei der Datenanalyse wider – etwa fehlende Werte, Ausreißer oder gemischte Datentypen.

- Der Datensatz lässt sich leicht künstlich skalieren, was ihn ideal für den Vergleich klassischer Methoden mit skalierbaren Tools wie Dask macht.

- Schließlich ist mpg eine anschauliche Zielgröße, deren Interpretation auch für ein nicht-technisches Publikum nachvollziehbar ist.

Insgesamt bietet der Datensatz damit eine geeignete Grundlage, um sowohl klassische Regressionsmethoden als auch Big-Data-Technologien im Rahmen eines vollständigen Analyseprojekts zu demonstrieren.

3 Explorative Datenanalyse

Die explorative Datenanalyse dient dazu, den Auto-Datensatz hinsichtlich seiner Struktur, Verteilungen, Korrelationen und potenziellen Besonderheiten zu untersuchen. Ziel ist es, erste Hypothesen über Zusammenhänge zwischen den technischen Merkmalen und dem Kraftstoffverbrauch (mpg) zu entwickeln und geeignete Features für die spätere Modellierung zu identifizieren.

3.1 Datenstruktur und erste Beobachtungen

Beim Einlesen des Datensatzes fällt auf, dass die Variable horsepower als object erkannt wird. Eine manuelle Prüfung ergibt, dass einzelne Werte nicht-numerisch kodiert sind (z. B. mit "?"). Diese Werte verhindern die Berechnung von Korrelationen und müssen daher im Preprocessing bereinigt werden.

Die Funktion describe() liefert zentrale statistische Kennzahlen. Die Zielvariable mpg weist einen Mittelwert von ca. 23,5 auf, mit Werten zwischen 9 und 46,6. Merkmale wie weight, displacement oder horsepower zeigen eine hohe Varianz, was auf stark unterschiedliche Fahrzeugtypen hinweist.

3.2 Verteilung der Ausreißer

Die Verteilungen wichtiger Merkmale wie mpg, horsepower, weight und acceleration werden mittels Histogramme und Dichtekurven visualisiert. mpg zeigt eine leicht linksschiefe Verteilung. Einzelne Fahrzeuge verbrauchen deutlich mehr oder weniger als

der Durchschnitt. Bei horsepower und mpg lassen sich potenzielle Ausreißer identifizieren, die jedoch nicht entfernt werden, da sie technisch plausibel sind.

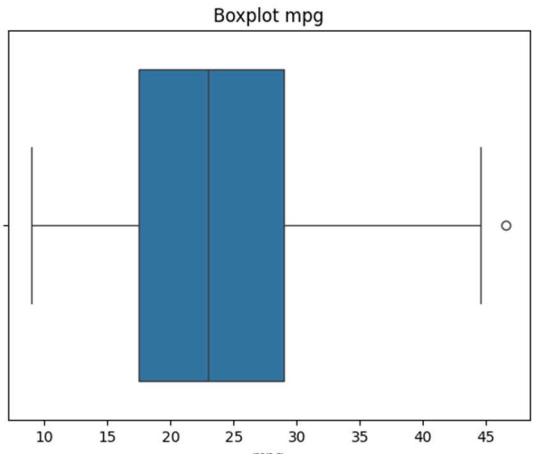

Abbildung 1 - Boxplot zu mpg

Auch die kategorialen Merkmale cylinders, origin und year werden analysiert. Fahrzeuge aus den USA sowie Modelle mit 4 Zylindern sind deutlich überrepräsentiert. Zur besseren Lesbarkeit wird origin später in Klartext (USA, Europa, Asien) überführt.

3.3 Korrelationen und Beziehungen

Die bereinigten numerischen Variablen werden in einer Korrelationsmatrix gegenübergestellt. Besonders auffällig ist die starke negative Korrelation zwischen mpg und weight ($r \approx -0{,}83$) sowie zwischen mpg und horsepower ($r \approx -0{,}78$). Gleichzeitig zeigen weight und displacement eine sehr hohe Korrelation ($r \approx 0{,}93$), was darauf hindeutet, dass sie ähnliche Informationen enthalten und später nicht gemeinsam in das Regressionsmodell aufgenommen werden sollten.

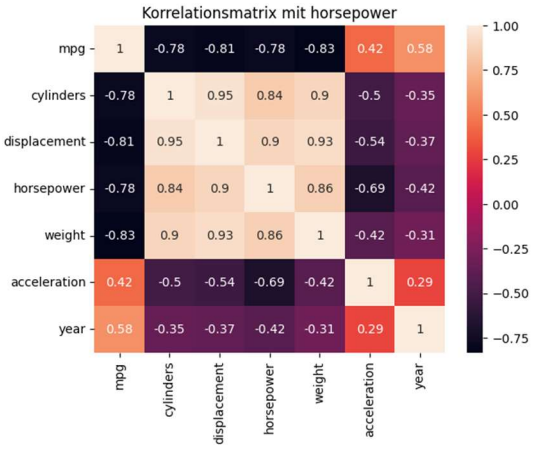

Abbildung 2 - Korrelations-Heatmap

Eine Heatmap visualisiert die Stärke dieser Zusammenhänge und unterstützt die Auswahl geeigneter, möglichst unabhängiger Features für die Regression.

3.4 Scatterplots und bivariate Muster

Zur detaillierten Betrachtung möglicher Zusammenhänge zwischen Zielgröße und Features werden Scatterplots zwischen mpg und Variablen wie horsepower, weight, displacement und acceleration erstellt. Die Visualisierungen zeigen:

- einen deutlich negativen linearen Zusammenhang zwischen mpg und weight,
- eine sehr ähnliche negative Beziehung zwischen mpg und horsepower,

- sowie eine ebenfalls negative Korrelation zwischen mpg und displacement (Hubraum), was darauf hindeutet, dass größere Motoren im Schnitt mehr Kraftstoff verbrauchen.

Bei acceleration zeigt sich dagegen kein klarer linearer Trend.

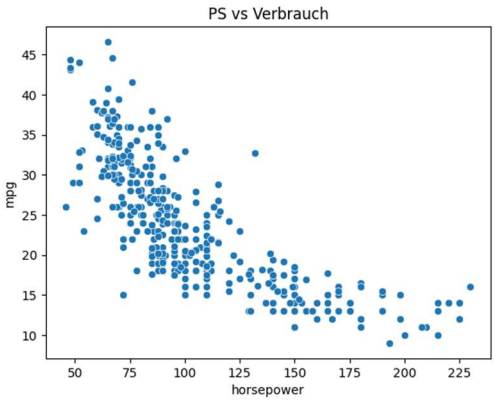

Abbildung 3 - Scatterplot PS, mpg

3.5 Zwischenfazit

Die explorative Analyse ergibt:

- teilweise schiefe Verteilungen und plausible Ausreißer,
- deutliche Korrelationen technischer Merkmale untereinander,
- sowie starke lineare Zusammenhänge zwischen mpg und Variablen wie weight und horsepower.

Diese Erkenntnisse bilden eine fundierte Grundlage für die nachfolgende Datenaufbereitung und Modellbildung.

4 Feature Engineering und Preprocessing

Bevor ein Regressionsmodell sinnvoll aufgebaut werden kann, müssen die Rohdaten in eine bereinigte und modellkompatible Form überführt werden. Dazu werden in diesem Abschnitt alle relevanten Vorverarbeitungsschritte sowie ergänzende Feature-Erweiterungen erläutert.

4.1 Behandlung fehlerhafter und fehlender Werte

Die Spalte horsepower wird beim Einlesen fälschlich als object erkannt, da einige Einträge nicht-numerisch vorliegen. Diese Einträge werden zunächst durch NaN ersetzt und anschließend aus dem Datensatz entfernt. Um den Informationsverlust gering zu halten, erfolgt das Entfernen nur für tatsächlich betroffene Zeilen, nicht pauschal für alle fehlenden Felder.

Auch die Spalten mpg und origin werden auf dieselbe Weise behandelt.

4.2 Transformation technischer Merkmale

Zur besseren Vergleichbarkeit der Daten, insbesondere für internationale Vergleiche, werden zusätzlich zu den Originalvariablen auch folgende transformierte Merkmale berechnet:

- weight_kg: Gewicht in Kilogramm statt Pfund
- displacement_liter: Hubraum in Litern statt Kubikzoll
- power_kw: Motorleistung in Kilowatt statt PS

Diese Umrechnungen haben keinen Einfluss auf die mathematische Form des Modells, verbessern aber die Interpretierbarkeit der Koeffizienten im internationalen Kontext.

4.3 Auswahl möglichst unabhängiger Features

Ein zentrales Ziel beim Feature Engineering besteht darin, redundante Informationen zu vermeiden. Da viele technische Variablen stark untereinander korrelieren (z. B. weight und displacement, $r \approx 0{,}93$), wird aus jeder Gruppe nur ein repräsentatives Merkmal verwendet. So bleibt horsepower im Modell, während stark verwandte Merkmale wie displacement ausgeschlossen oder ersetzt werden.

year wird trotz mäßiger Korrelation aufgenommen, da es möglicherweise technologische Entwicklungsschritte abbildet. Die Variable acceleration bleibt optional enthalten.

4.4 Interaktionsterm zu Verstärkung kombinierter Effekte

Da davon auszugehen ist, dass schwere Fahrzeuge mit starker Motorleistung besonders viel verbrauchen, wird ein Interaktionsterm eingeführt:

power_kw * weight_kg

Dieser Term berücksichtigt, dass beide Einflussgrößen gemeinsam einen verstärkenden Effekt auf den Kraftstoffverbrauch haben können. Die Idee basiert auf einem physikalisch motivierten Zusammenhang und ergänzt das lineare Modell um eine realitätsnahe Komponente.

4.5 Kodierung kategorialer Variablen

Die Variable origin, welche die geografische Herkunft des Fahrzeugs angibt, 1 für USA - 2 für Europa - 3 für Asien (Auto: Auto Data Set in ISLR: Data for an Introduction to Statistical Learning with Applications in R 2025), wird mittels Dummy-Encoding in zwei binäre Spalten (origin_USA, origin_Japan) umgewandelt. Dadurch kann die Variable direkt im Regressionsmodell verarbeitet werden, ohne dass eine künstliche Reihenfolge angenommen wird.

4.6 Skalierung aller numerischen Variablen

Da die numerischen Merkmale unterschiedliche Wertebereiche aufweisen (z. B. Gewicht in kg, Leistung in kW, Jahr als Zahl), werden alle numerischen Features mit dem StandardScaler standardisiert. Dies führt zu Mittelwert 0 und Standardabweichung 1 und verhindert, dass einzelne Merkmale das Modell durch ihre Skala dominieren. Außerdem verbessert dies die Vergleichbarkeit der Koeffizienten im späteren Modell.

Tabelle 2 gibt einen Überblick über alle neu erzeugten Features, die im Rahmen der Datenaufbereitung und -transformation für die Modellbildung verwendet werden.

Tabelle 2: Neue Variablen

Neues Merkmal	Beschreibung	Berechnungsgrundlage
power	Motorleistung in Kilowatt	horsepower * 0.7355
weight_kg	Gewicht in Kilogramm	weight * 0.4536
displacement_liter	Hubraum in Liter	displacement * 0.0163871
power_weight	Interaktionsterm	power * weight_kg
origin_USA	Dummy für USA	aus origin == 1
origin_Japan	Dummy für Japan	Aus origin == 2

5 Modellbildung und Implementierung

Ziel dieses Abschnitts ist es, ein geeignetes Regressionsmodell zur Vorhersage des Kraftstoffverbrauchs (mpg) auf Basis technischer Fahrzeugdaten zu entwickeln und hinsichtlich seiner Prognosegüte zu bewerten.

5.1 Modellauswahl

Da der Zusammenhang zwischen mpg und den gewählten Prädiktoren in der explorativen Analyse weitgehend linear erscheint, wird ein lineares Regressionsmodell als Ausgangspunkt verwendet. Dieses Modell ist sowohl gut interpretierbar als auch mathematisch transparent, was eine manuelle Umsetzung zusätzlich ermöglicht.

5.2 Modell mit scikit-learn

Zunächst wird das Modell mithilfe der Bibliothek scikit-learn realisiert. Dafür werden die Daten in Trainings- und Testdaten aufgeteilt. Anschließend wird das Modell mit den zuvor standardisierten und bereinigten Daten trainiert.

Zur Beurteilung der Modellgüte werden die folgenden Metriken berechnet:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Bestimmtheitsmaß (R^2)

Im Ausgangslauf (Train/Test-Split) erzielt das Modell folgende Werte:

- RMSE ≈ 14,98
- MAE ≈ 3,0 mpg
- R^2 ≈ 0,70

Diese Werte deuten auf eine grundsätzlich gute Passung hin. Das Modell kann den Verbrauch also im Schnitt mit einer Abweichung von etwa ±3 mpg vorhersagen.

5.3 Kreuzvalidierung

Um die Aussagekraft des Modells über einen einfachen Trainings-/Test-Split hinaus zu validieren, wird eine 5-fache Kreuzvalidierung durchgeführt. Dabei wird der Datensatz in fünf gleich große Teile aufgeteilt; jeweils vier dienen zum Training, einer zum Test.

Die mittleren Fehlermaße über alle fünf Durchläufe betragen:

- MAE ≈ 3,64 mpg
- R^2 ≈ 0,40

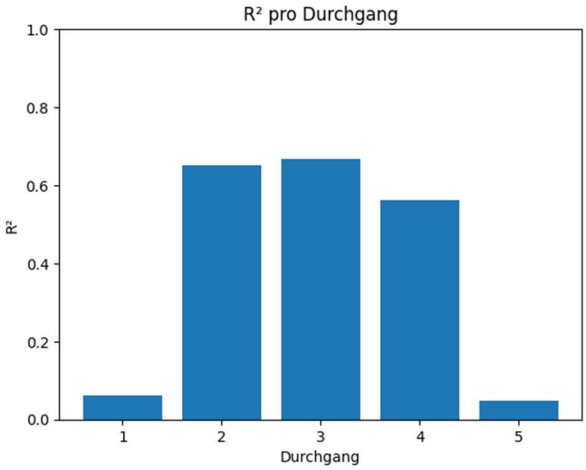

Abbildung 4 - R2 pro Durchgang

Die teils stark schwankenden Ergebnisse in den Durchgängen deuten darauf hin, dass das Modell in manchen Teilmengen relativ präzise, in anderen jedoch nahezu nutzlos vorhersagt. Dies spricht für eine eingeschränkte Robustheit und legt Optimierungspotenzial bei Feature-Auswahl oder Modellwahl nahe.

5.4 Eigene Implementierung mit Least Squares

Zur Vertiefung des Verständnisses wird das Modell zusätzlich manuell mittels der Normalengleichung implementiert.

Die Ergebnisse stimmen nahezu exakt mit jenen der scikit-learn-Lösung im ersten Versuch überein, was die Korrektheit der eigenen Umsetzung bestätigt. Die berechneten Regressionskoeffizienten sowie die Fehlermaße liegen im gleichen Bereich.

5.5 Fazit zur Modellierung

Das lineare Regressionsmodell kann einfache Zusammenhänge im Datensatz grundsätzlich erfassen. Dies zeigt sich in einem guten Ergebnis bei der klassischen Trainings-/Test-Aufteilung, wo ein MAE von rund 3,0 mpg und ein R^2-Wert von 0,70 erreicht werden.

Die Ergebnisse der 5-fachen Kreuzvalidierung relativieren diesen Eindruck jedoch deutlich. Die mittleren Fehlerwerte fallen spürbar schlechter aus, und insbesondere das Bestimmtheitsmaß (R^2 ≈ 0,40) zeigt, dass das Modell bei neuen Datenaufteilungen deutlich an Vorhersagekraft verliert.

Insgesamt lässt sich sagen, dass das lineare Modell erste Tendenzen abbildet, aber nicht stabil genug ist, um als verlässliches Prognosemodell zu gelten. Für robustere Ergebnisse wären eine gezieltere Auswahl von Merkmalen, alternative Modellansätze oder eine nichtlineare Erweiterung notwendig.

6 Big Data Skalierung

Ein zentraler Aspekt moderner Datenanalyse besteht darin, Methoden nicht nur auf kleinen Datensätzen zu testen, sondern sie auch auf größere Datenmengen skalierbar umzusetzen. Dazu wurde im Rahmen dieser Arbeit untersucht, wie sich gängige Analyseoperationen auf einen synthetisch vergrößerten Datensatz verhalten,

sowohl inhaltlich als auch hinsichtlich der Rechenzeit.

6.1 Big Data Skalierung

Da der ursprüngliche Auto-Datensatz lediglich 392 Zeilen umfasst, wurde er durch einfache Duplikation auf rund 40.000 Zeilen erweitert. Diese künstliche Skalierung verändert zwar nicht die inhaltlichen Strukturen oder Verteilungen, ermöglicht aber eine realitätsnahe Simulation größerer Datenmengen für Performancevergleiche.

6.2 Vergleich von pandas und Dask

Zur Durchführung der Mittelwert- und Korrelationsberechnung wurden zwei unterschiedliche Frameworks verwendet:

- pandas: Die Standardbibliothek für Datenanalyse in Python
- Dask: Eine skalierbare Alternative, die auch mit verteilten Daten arbeiten kann

Beide Bibliotheken liefern bei den betrachteten Operationen identische Ergebnisse, da Dask intern dieselben Algorithmen verwendet – jedoch für größere Datenmengen optimiert ist. Die Rechenzeiten unterscheiden sich jedoch: Für den skalierten Datensatz zeigt sich, dass pandas bei einfachen Berechnungen schneller arbeitet, da der Overhead von Dask sich erst bei größeren Datenmengen ab mehreren hunderttausend oder Millionen Zeilen auszahlt (Why Dask? — Dask documentation 2025).

Zum Beispiel dauert die Berechnung des Mittelwerts und der Korrelationsmatrix bei pandas nur wenige Millisekunden, bei Dask allerdings geringfügig länger, da die Ausführung erst explizit durch .compute() angestoßen wird.

6.3 Einschätzung zur Skalierbarkeit

Die durchgeführten Tests zeigen, dass pandas für kleine bis mittlere Datensätze vollkommen ausreichend und performant ist. Erst bei sehr großen Datenmengen – z. B.

bei speicherintensiven Prozessen, paralleler Verarbeitung oder Daten, die nicht mehr in den Hauptspeicher passen – bietet Dask einen echten Vorteil.

Die Implementierung unterscheidet sich nur minimal, was einen einfachen Umstieg ermöglicht. Dennoch sollte der Einsatz solcher Big-Data-Frameworks gezielt und abhängig von der Datenmenge erfolgen

6.4 Fazit zur Skalierung

Die Simulation großer Datenmengen durch Duplikation zeigt, dass einfache Analyseoperationen mit pandas auch bei zehntausenden von Zeilen problemlos möglich sind. Dask liefert identische Ergebnisse, eignet sich aber vor allem für Szenarien mit echten Big-Data-Herausforderungen, etwa bei Datenströmen, verteiltem Rechnen oder Arbeit im Gigabyte-/Terabytebereich.

Die Erkenntnis: Nicht jedes Datenanalyseprojekt erfordert sofort Big-Data-Tools – wohl aber ein grundsätzliches Verständnis für deren Einsatzgrenzen und Vorteile.

7 Diskussion der Ergebnisse

7.1 Aussagekraft des Regressionsmodells

Die Ergebnisse des linearen Regressionsmodells zeigen, dass sich der Kraftstoffverbrauch (mpg) grundsätzlich durch technische Merkmale wie horsepower, weight oder year vorhersagen lässt. Insbesondere die im klassischen Trainings-/Test-Split erreichten Werte (z. B. $R^2 \approx 0{,}70$, MAE ≈ 3.0 mpg) deuten auf eine gute Anpassung an die Trainingsdaten hin.

Allerdings zeigt die 5-fache Kreuzvalidierung ein deutlich differenzierteres Bild: Hier sinkt der mittlere R^2-Wert auf rund 0,40, mit teils starken Abweichungen zwischen den Folds. Dies weist auf eine begrenzte Generalisierungsfähigkeit des Modells hin. Besonders bei bestimmten Datenaufteilungen scheint das Modell kaum noch in der Lage zu sein, den Verbrauch zuverlässig vorherzusagen.

Diese Streuung könnte darauf hindeuten, dass die im Modell verwendeten Merkmale zwar relevante Informationen enthalten, jedoch nicht ausreichen, um alle Einflüsse auf den Kraftstoffverbrauch adäquat zu erfassen. Beispielsweise bleiben Aspekte wie Fahrzeugtyp, Aerodynamik oder Fahrverhalten unberücksichtigt.

7.2 Validität der Datengrundlage

Der verwendete Datensatz stellt ein praxisnahes Beispiel für typische Analyseaufgaben dar. Gleichzeitig bringt er aber auch Einschränkungen mit sich:

- Die Daten stammen aus den 1970er und frühen 1980er Jahren und bilden somit nur einen begrenzten technologischen Entwicklungsstand ab.
- Einige Variablen wie acceleration oder origin liefern nur schwache oder inkonsistente Beiträge zur Vorhersagequalität.
- Die künstliche Skalierung auf 40.000 Zeilen durch einfache Duplikation erlaubt zwar Performancevergleiche, ersetzt aber keine echte Big-Data-Situation mit vielfältigen, unabhängigen Beobachtungen.

7.3 Bewertung der Skalierung mit Dask

Die Skalierungstests zeigen, dass Dask bei einfachen Operationen (wie Mittelwert oder Korrelation) identische Ergebnisse liefert wie pandas, jedoch bei moderaten Datenmengen nicht schneller arbeitet. Der erwartete Performancegewinn tritt laut Dokumentation und Literatur erst bei sehr großen Datenmengen oder paralleler Verarbeitung ein (Why Dask? — Dask documentation 2025).

Die Umsetzung zeigt aber auch: Der Umstieg von pandas zu Dask ist konzeptionell einfach möglich. Wer später mit größeren Datenquellen arbeitet, kann viele Analysebausteine nahezu unverändert übernehmen.

7.4 Reflexion des Vorgehens

Die Durchführung der Analyse hat gezeigt, dass sich ein vollständiger Datenanalyseprozess – von der Bereinigung über die Modellbildung bis hin zur Skalierung – auch mit vergleichsweise einfachen Mitteln realisieren lässt. Besonders hilfreich war dabei die schrittweise Umsetzung in Jupyter Notebook, die sowohl das Testen einzelner Teilaufgaben als auch die Nachvollziehbarkeit aller Berechnungen unterstützt hat.

Die eigenständige Implementierung der Regressionsformel hat das theoretische Verständnis gestärkt, während der Vergleich mit scikit-learn die Richtigkeit der Ergebnisse bestätigt hat. Auch wenn Dask in dieser Arbeit keine spürbaren Geschwindigkeitsvorteile bot, konnte seine Anwendung dennoch sinnvoll erprobt werden.

Insgesamt hat sich der gewählte Ansatz als geeignet erwiesen, um die Zielsetzung der Arbeit systematisch zu bearbeiten und wichtige Erfahrungen im Umgang mit größeren Datenmengen und datengetriebenen Modellierungsverfahren zu sammeln.

8 Fazit und Ausblick

In dieser Arbeit wurde ein kompletter Analyseprozess zur Vorhersage des Kraftstoffverbrauchs mit Hilfe eines linearen Regressionsmodells durchgeführt – von der Datenaufbereitung über die Modellierung bis hin zum Test einer skalierbaren Lösung mit Dask.

Das Modell konnte grundlegende Zusammenhänge im Datensatz gut erfassen, zeigte aber auch Schwächen bei der Generalisierung. Besonders die Ergebnisse der Kreuzvalidierung machen deutlich, dass einfache lineare Ansätze schnell an ihre Grenzen stoßen, insbesondere dann, wenn die Daten intern stark streuen.

Auch Dask konnte sinnvoll getestet werden. Zwar brachte es bei der simulierten Datengröße noch keinen Vorteil, aber der Aufbau und die Anwendung sind einfach, sodass es sich bei größeren Datenmengen lohnen würde.

Als nächster Schritt wäre es interessant, andere Modellarten zu testen, zum Beispiel mit Entscheidungsbäumen oder regulären Regressionsverfahren. Auch ein echter, größerer Datensatz könnte zeigen, wie gut sich das Vorgehen übertragen lässt.

9 Literaturverzeichnis

Auto: Auto Data Set in ISLR: Data for an Introduction to Statistical Learning with Applications in R (2025). Online verfügbar unter https://rdrr.io/cran/ISLR/man/Auto.html, zuletzt aktualisiert am 24.07.2025, zuletzt geprüft am 24.07.2025.

Chen, Min; Mao, Shiwen; Liu, Yunhao (2014): Big Data: A Survey. In: *Mobile Netw Appl* 19 (2), S. 171–209. DOI: 10.1007/s11036-013-0489-0.

Gerst, Michael D.; Howarth, Richard B.; Borsuk, Mark E. (2010): Accounting for the risk of extreme outcomes in an integrated assessment of climate change. In: *Energy Policy* 38 (8), S. 4540–4548. DOI: 10.1016/j.enpol.2010.04.008.

Hashem, Ibrahim Abaker Targio; Yaqoob, Ibrar; Anuar, Nor Badrul; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015): The rise of “big data” on cloud computing: Review and open research issues. In: *Information Systems* 47, S. 98–115. DOI: 10.1016/j.is.2014.07.006.

Huang, Jianhua Z. (2014): An Introduction to Statistical Learning: With Applications in R By Gareth James, Trevor Hastie, Robert Tibshirani, Daniela Witten. In: *JABES* 19 (4), S. 556–557. DOI: 10.1007/s13253-014-0179-9.

Kitchin, Rob (2014): The data revolution. Big data, open data, data infrastructures and their consequences. Los Angeles: SAGE.

Why Dask? — Dask documentation (2025). Online verfügbar unter https://docs.dask.org/en/stable/why.html#performance, zuletzt aktualisiert am 14.05.2025, zuletzt geprüft am 24.07.2025.