Einflussfaktoren auf die Fahrzeugflotten in deutschen Landkreisen

...

Data Analytics with Statistics
Projektpräsentation
16. Januar 2025

Übersicht

1. Einleitung
2. Daten
3. Explorative Datenanalyse (EDA)
4. **Modellierung**
5. **Fazit**

Einleitung

- Hintergrund
- Thesen
- Forschungsfrage
- Datenquellen
- Datenwörterbuch

Einleitung - Hintergrund

Die Reduktion von Emissionen und die Förderung umweltfreundlicher Fahrzeuge sind zentrale Ziele der Verkehrspolitik. Die Einführung neuer Emissionsvorschriften sowie die Verbreitung neuer Technologien, wie Elektrofahrzeuge und Plug-in-Hybride, haben das Potenzial, die Zusammensetzung der Fahrzeugflotten in den Landkreisen erheblich zu beeinflussen.

Die gesetzliche Einführung neuer Emissionsvorschriften wird voraussichtlich einen signifikanten Einfluss auf die Bestandsflotte haben. Es wird erwartet, dass der Anteil von Euro4-Fahrzeugen in Landkreisen mit einem hohen Anteil neuer Technologien (wie Plug-in-Hybriden und Elektrofahrzeugen) abnimmt. Gleichzeitig könnten sozioökonomische Faktoren wie das verfügbare Einkommen und die Unfallrate ebenfalls eine Rolle bei der Erneuerung der Fahrzeugflotten spielen. In Landkreisen mit einer älteren Fahrzeugflotte, die durch einen hohen Anteil von Fahrzeugen der Emissionsgruppen Euro2 und Euro3 gekennzeichnet ist, wird jedoch erwartet, dass der Anteil von Euro4-Fahrzeugen trotz neuer Emissionsvorschriften und

Einleitung - Thesen

Die gesetzliche Einführung neuer Emissionsvorschriften wird voraussichtlich einen signifikanten Einfluss auf die Bestandsflotte haben. Es wird erwartet, dass der **Anteil von Euro4-Fahrzeugen**:

- in Landkreisen mit einem hohen Anteil neuer Technologien (wie Plug-in-Hybriden und Elektrofahrzeugen) abnimmt.
- Gleichzeitig könnten sozioökonomische Faktoren wie das verfügbare Einkommen und die Unfallrate ebenfalls eine Rolle bei der Erneuerung der Fahrzeugflotten spielen.
- In Landkreisen mit einer älteren Fahrzeugflotte, die durch einen hohen Anteil von Fahrzeugen der Emissionsgruppen Euro2 und Euro3 gekennzeichnet ist, wird jedoch erwartet, dass der Anteil von Euro4-Fahrzeugen trotz neuer Emissionsvorschriften und Technologien robust bleibt.

Einleitung - Forschungsfrage

Welche Faktoren beeinflussen den Anteil von Euro4-Fahrzeugen in deutschen Landkreisen und wie stark ist dieser Einfluss?

Das multiple lineare Regressionsmodell soll die folgenden Fragen zu beantworten:

- **Identifikation relevanter Prädiktoren**: Welche Variablen haben einen signifikanten Einfluss auf den Anteil von Euro4-Fahrzeugen?
- **Quantifizierung des Einflusses**: Wie stark ist der Einfluss der identifizierten Prädiktoren auf den Anteil von Euro4-Fahrzeugen?
- **Modellgüte und Generalisierbarkeit**: Wie gut erklärt das Modell die Varianz im Anteil von Euro4-Fahrzeugen und wie gut generalisiert es auf neue Daten?

Einleitung - Datenquellen

Für die Analyse soll ein Gesamt-Datensatz aus den folgenden vier Datenquellen erstellt werden:

1. **Daten über Fahrzeugbestand** (nach Kraftstoffart und Emissionsgruppen), Quelle: [Statistik des Kraftfahrzeug- und Anhängerbestandes, Statistisches Bundesamt, Code: 46251-0021 (https://www-genesis.destatis.de/genesis/online/data)
2. **Bevölkerung am Hauptwohnort nach Altersgruppen und Geschlecht**, Quelle: Regionalstatistik, Code: 12211-Z-03 (https://www.regionalstatistik.de/genesis/online/)
3. **Verfügbares Einkommen je Einwohner**, Quelle: Regionalstatistik, Code: AI-S-01 (https://www.regionalstatistik.de/genesis/online/)
4. **Straßenverkehrsunfälle bezogen auf Kfz**, Quelle: Regionalstatistik, Code: AI013-3 (https://www.regionalstatistik.de/genesis/online/)

Alle Daten sind aus dem Jahr 2019

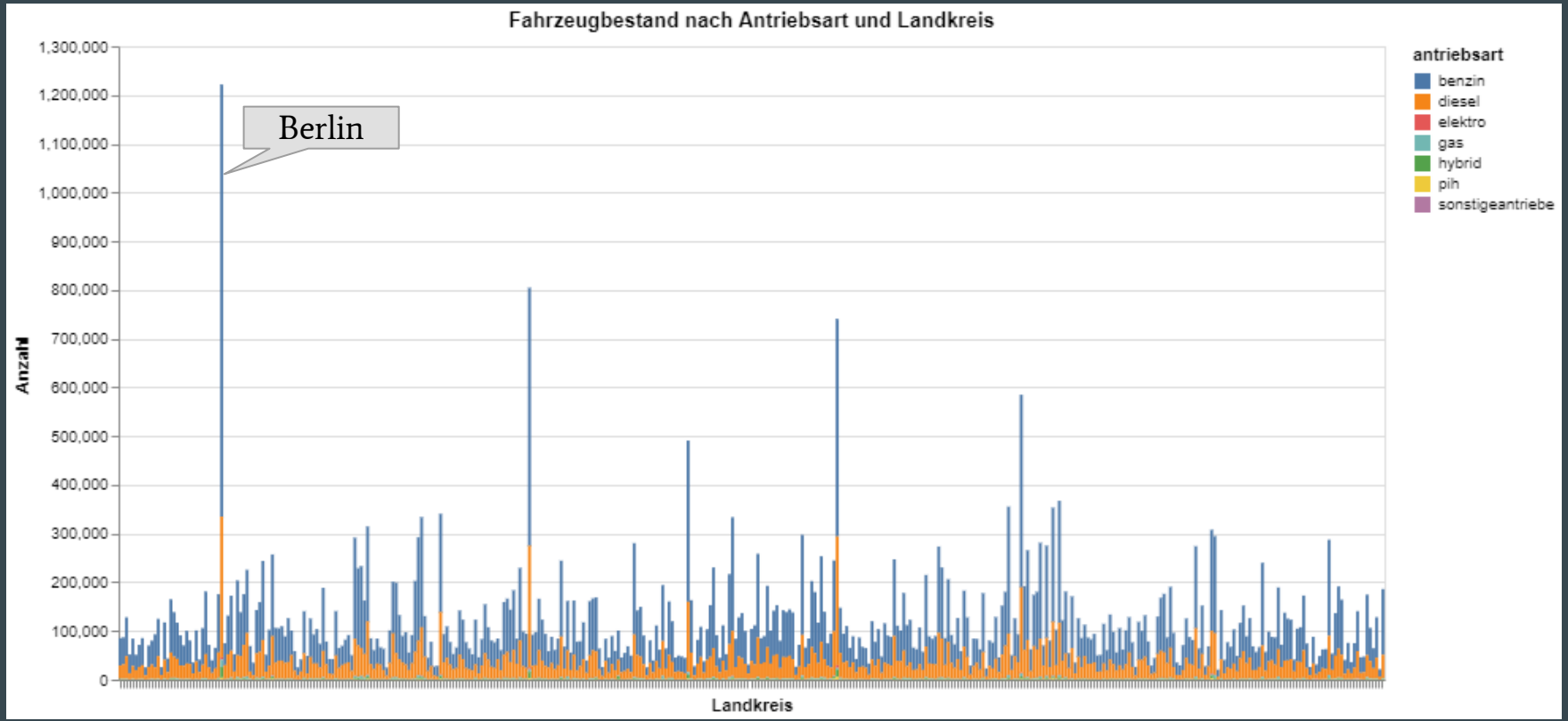
Einleitung - Datenwörterbuch

Name	Beschreibung	Rolle	Typ	Format
landkreis_id	Landkreis ID	ID		String
anzahl_personen_1000	Bevölkerungsanzahl in Tausend	Prädiktor	numerisch	Float
vee	Verfügbares Einkommen je Einwohner in EUR	Prädiktor	numerisch	Float
anzahl_kfz_je_person	Anzahl der Kraftfahrzeuge pro Person	Prädiktor	numerisch	Float
unfaelle_je_10k_kfz	Unfaelle je 10000 Kfz	Prädiktor	numerisch	Float
elektro	Anteil der Elektrofahrzeuge	Prädiktor	numerisch	Float
pih	Anteil der Plug-in-Hybridfahrzeuge	Prädiktor	numerisch	Float
euro2	Anteil der Fahrzeuge mit EURO 2 Norm	Prädiktor	numerisch	Float
euro3	Anteil der Fahrzeuge mit EURO 3 Norm	Prädiktor	numerisch	Float
euro4	Anteil der Fahrzeuge mit EURO 4 Norm	Antwort	numerisch	Float
euro6	Anteil der Fahrzeuge mit EURO 6 Norm	Prädiktor	numerisch	Float
euro6dt	Anteil der Fahrzeuge mit EURO 6d-TEMP Norm	Prädiktor	numerisch	Float

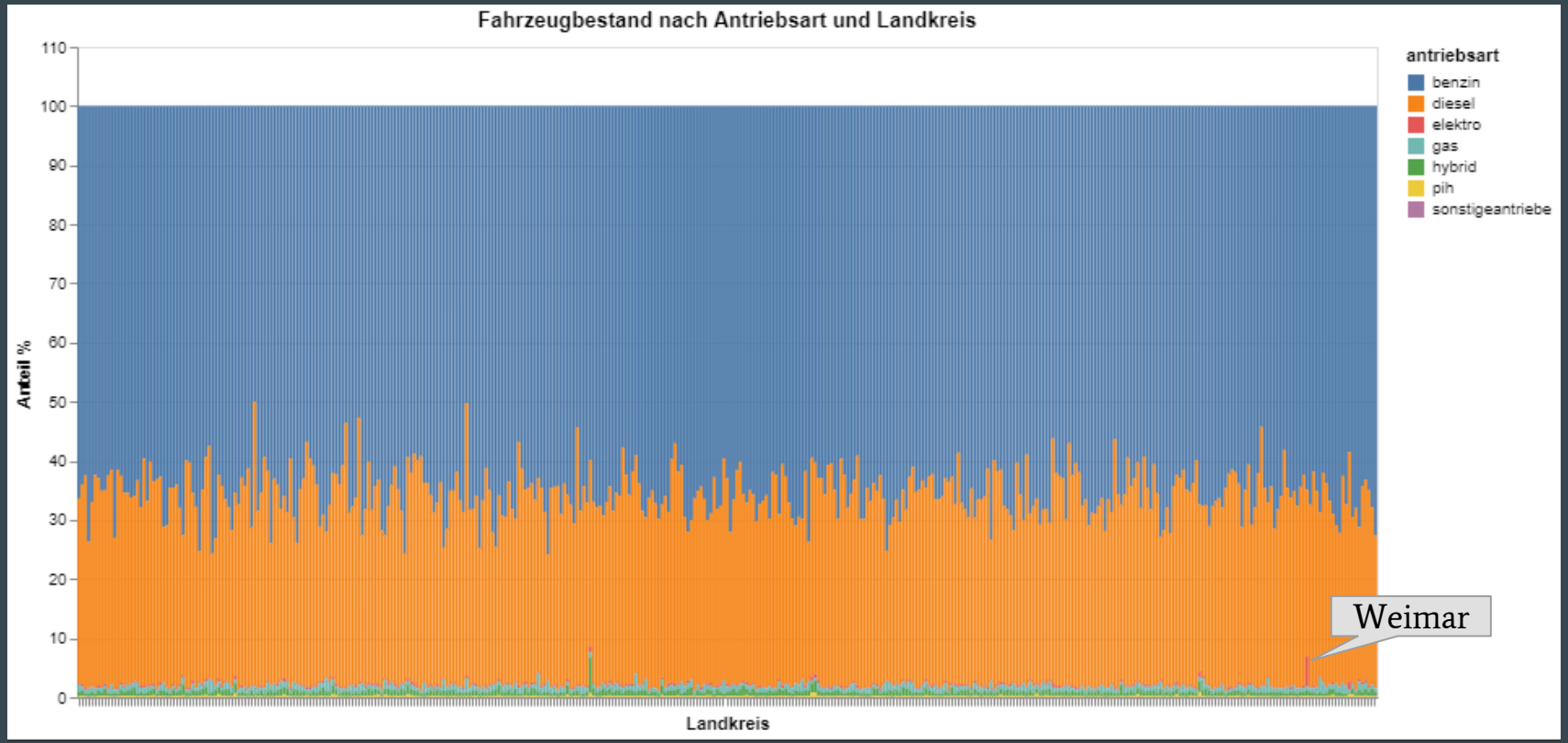
Explorative Datenanalyse

- gestapelte Balkendiagramme
- Histogramme
- Boxplots (und Dichteplots)
- Scatterplots (mit Residuen)

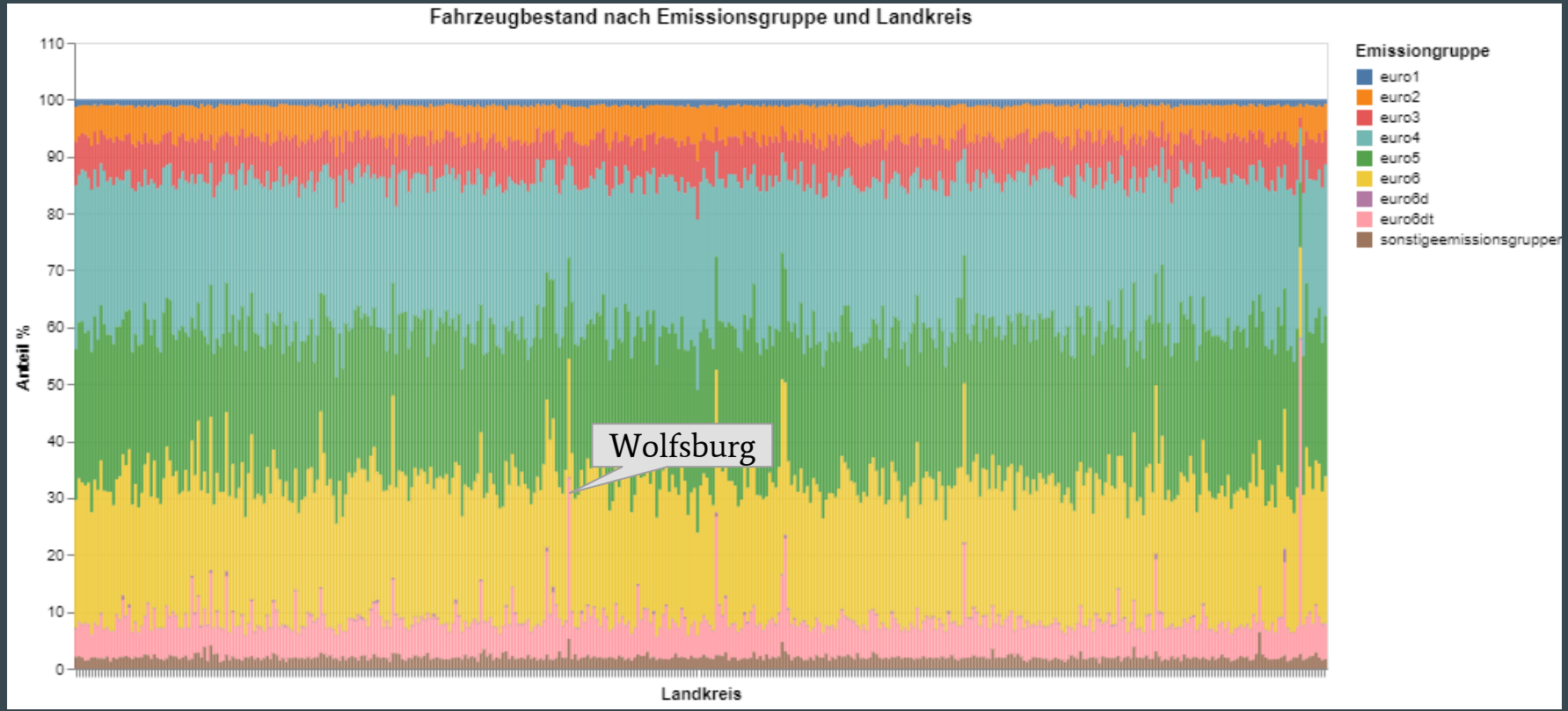
Explorative Datenanalyse



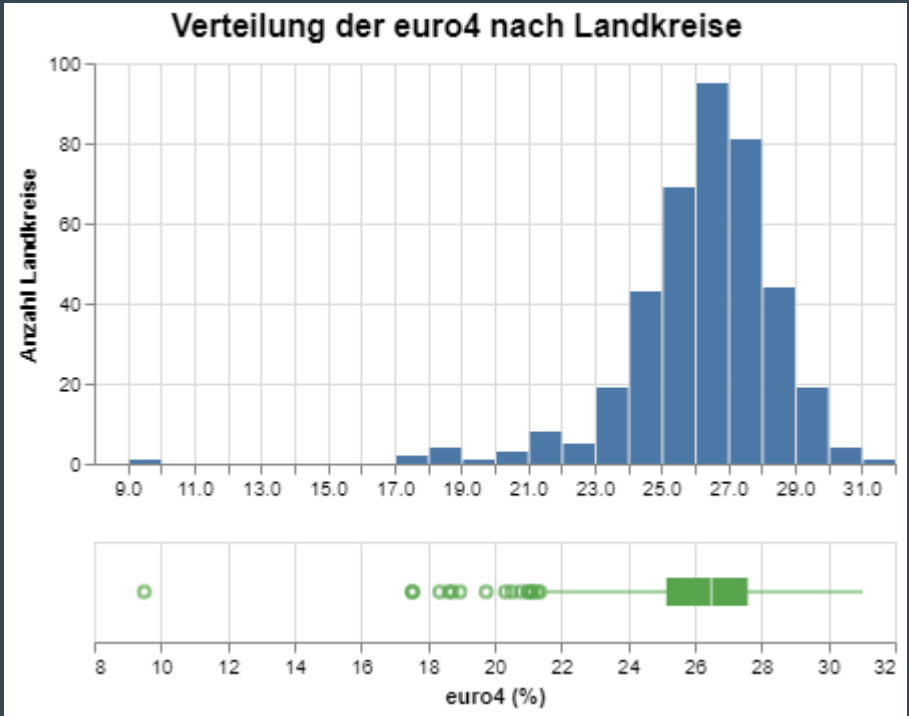
Explorative Datenanalyse



Explorative Datenanalyse



Explorative Datenanalyse



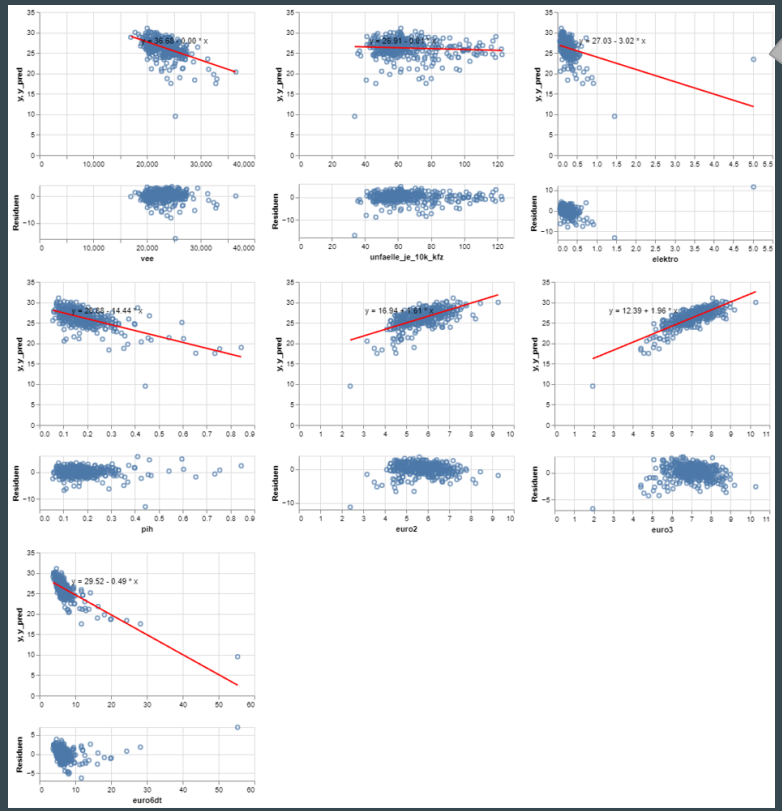
- normalverteilte Form mit leichter Linksschiefe
- Der zentrale Bereich (Box) liegt zwischen 25% und 28%
- Der Median bei 26,5% teilt die Box nahezu symmetrisch
- Ausreißer treten hauptsächlich am unteren Ende der Verteilung auf (unter 20%)
- Die Hauptmasse der Werte konzentriert sich im Bereich von 24% bis 28%

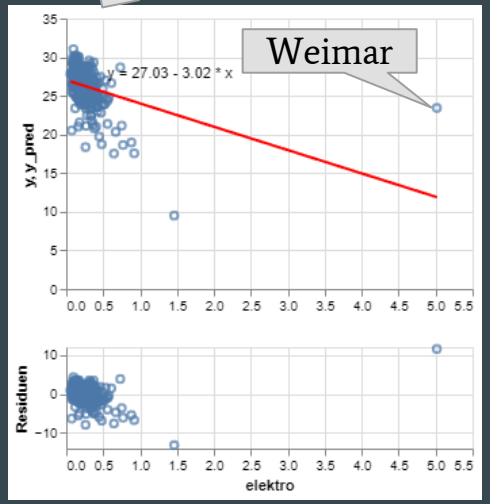
Explorative Datenanalyse

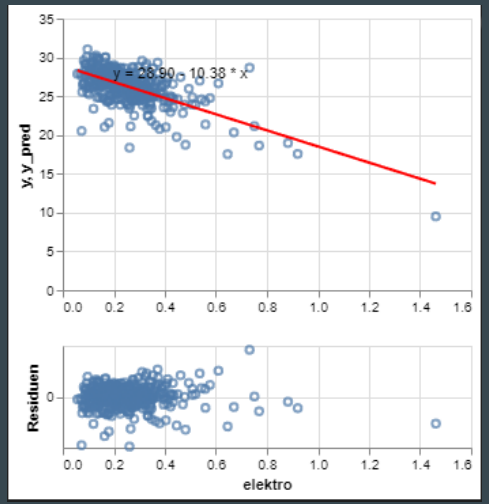
	anzahl_personen_1000	vee	anzahl_kfz_je_person	unfaelle_je_10k_kfz	elektro	pih	euro2	euro3	euro4	euro6	euro6dt
anzahl_personen_1000	1.000000	0.244382	-0.066238	0.117337	0.292294	0.339558	-0.054767	-0.058390	-0.109362	0.141360	0.246755
vee	0.244382	1.000000	0.402912	-0.163579	0.431506	0.328980	-0.110777	-0.171033	-0.304116	0.154861	0.112238
anzahl_kfz_je_person	-0.066238	0.402912	1.000000	-0.366200	-0.007752	-0.116796	0.097761	0.094401	0.022136	-0.171921	-0.222050
unfaelle_je_10k_kfz	0.117337	-0.163579	-0.366200	1.000000	0.115733	0.211471	0.032627	0.076399	-0.054337	0.065506	0.222919
elektro	0.292294	0.431506	-0.007752	0.115733	1.000000	0.598821	-0.075150	-0.146484	-0.367791	0.210161	0.327162
pih	0.339558	0.328980	-0.116796	0.211471	0.598821	1.000000	-0.089734	-0.165300	-0.395826	0.293661	0.483457
euro2	-0.054767	-0.110777	0.097761	0.032627	-0.075150	-0.089734	1.000000	0.662322	0.436581	-0.528772	-0.280286
euro3	-0.058390	-0.171033	0.094401	0.076399	-0.146484	-0.165300	0.662322	1.000000	0.563179	-0.657082	-0.360033
euro4	-0.109362	-0.304116	0.022136	-0.054337	-0.367791	-0.395826	0.436581	0.563179	1.000000	-0.673405	-0.556504
euro6	0.141360	0.154861	-0.171921	0.065506	0.210161	0.293661	-0.528772	-0.657082	-0.673405	1.000000	0.499326
euro6dt	0.246755	0.112238	-0.222050	0.222919	0.327162	0.483457	-0.280286	-0.360033	-0.556504	0.499326	1.000000

```
euro4                 1.000000
euro3                 0.563179
euro2                 0.436581
anzahl_kfz_je_person  0.022136
unfaelle_je_10k_kfz  -0.054337
anzahl_personen_1000 -0.109362
vee                  -0.304116
elektro              -0.367791
pih                  -0.395826
euro6dt              -0.556504
euro6                -0.673405
```

Explorative Datenanalyse







Modellierung

1. Datenteilung
    - Erstellung von zwei Datensätzen: Set A: Mit Ausreißer; Set B: Ohne Ausreißer
    - Aufteilung Set A und Set B jeweils in Trainings (20%) und Testdaten (80%)
2. Training mit Ausreißer
3. Training ohne Ausreißer & Performance Vergleich
4. Rückwärtselimination
5. Finales Modelltraining: Training mit den besten Prädiktoren aus der Rückwärtselimination

Modell mit Ausreißer validieren

```
# Bestimmtheitsmaß R² für Trainings- und
Test Daten mit Ausreißer berechnen
r2_train = regr.score(X_train, y_train)
r2_test = regr.score(X_test, y_test)

print(f'R² Training: {r2_train:.4f}')
print(f'R² Test: {r2_test:.4f}')
```

```
…
R² Training: 0.8823
R² Test: 0.8002
```

Bewertung

Diese Ergebnisse zeigen, dass das Modell eine hohe Erklärungskraft für die Trainingsdaten aufweist und auch auf den Testdaten eine gute Generalisierbarkeit besitzt. Das Modell kann somit als robust und zuverlässig angesehen werden, obwohl ein Ausreißer in den Daten vorhanden ist.

Leistungsabfall von ca. 8.2 Prozentpunkten

- typisches Phänomen.
- liegt im üblichen Rahmen
- kein Overfitting

Modell ohne Ausreißer validieren

```
# Bestimmtheitsmaß R² für Trainings- und Test Daten ohne Ausreißer berechnen
r2_train_filterd = regr.score(X_train_filtered, y_train_filtered)
r2_test_filterd = regr.score(X_test_filtered, y_test_filtered)

print(f'R² Training: {r2_train_filterd:.4f}')
print(f'R² Test: {r2_test_filterd:.4f}')
```

```
…
R² Training: 0.8806
R² Test: 0.8244
```

Metrik	Mit Ausreißer	Ohne Ausreißer	Differenz
R² Training	0,8823	0,8806	-0,0017
R² Test	0,8002	0,8244	+0,0242

Bewertung

Das Modell ohne Ausreißer zeigt eine konsistente Leistung auf Trainings- und Testdaten, was auf eine bessere Generalisierbarkeit hinweist. Daher sollte dieses Modell bevorzugt werden.

Rückwärtselimination mit und ohne Ausreißer

Modell mit Ausreißer:

- Startwert des adjustierten R^2 bei 0.8797
- Keine Prädiktoren wurden eliminiert
- Alle sieben Features tragen signifikant zur Modellgüte bei

Modell ohne Ausreißer:

- Startwert des adjustierten R^2 bei 0.8779
- Feature 'elektro' wurde als einziges eliminiert
- Die Elimination führt zu keiner Verschlechterung des adjustierten R^2

```
…
Bestes adjustiertes R²: 0.8779

Selektierte Features:
- vee
- unfaelle_je_10k_kfz
- pih
- euro2
- euro3
- euro6dt

Eliminations-Historie:

Entferntes Feature: elektro
Adjustiertes R²: 0.8779
```

Der Ausreißer befindet sich der Variable`elektro`. Nach dessen Entfernung verliert diese Variable ihre Bedeutung für das Modell, während die übrigen Prädiktoren weiterhin relevant bleiben. Die nahezu identischen R^2-Werte vor und nach der Elimination zeigen, dass`elektro` keinen substanziellen Beitrag zur Modellgüte leistet.

Kreuzvalidierung

```
# Kreuzvalidierung durchführen (z.B. 5-Fold)
cv_scores = cross_val_score(LinearRegression(),
X_train_filtered, y_train_filtered, cv=5, scoring='r2')

# Ergebnisse anzeigen
print(f'Kreuzvalidierungs-R² Scores: {cv_scores}')
print(f'Mittelwert der Scores: {cv_scores.mean():.4f}')
print(f'Standardabweichung der Scores: {cv_scores.std():.4f}')
```

```
…
Kreuzvalidierungs-R² Scores: [0.74759209 0.89525384 0.83436319
0.88378105 0.90148455]
Mittelwert der Scores: 0.8525
Standardabweichung der Scores: 0.0575
```

Diese Ergebnisse bestätigen die Robustheit und Zuverlässigkeit des Modells über verschiedene Teilmengen der Daten hinweg.

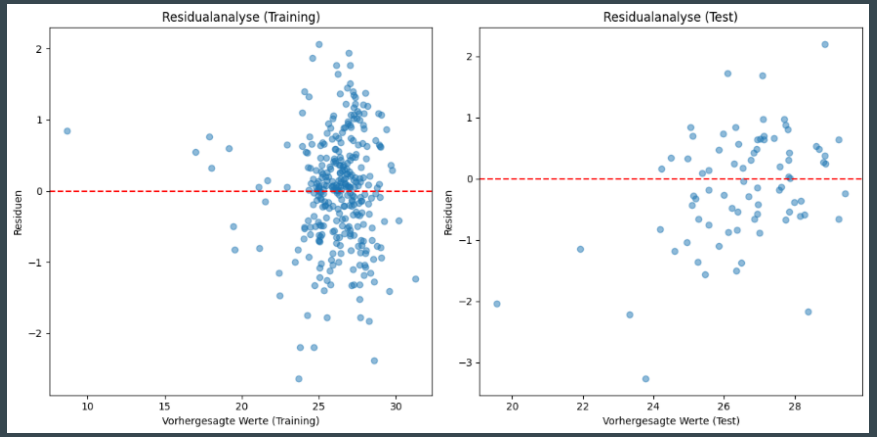
Bewertung

- mittl. Performance von 0.85 bestätigt die Güte des Modells und Generalisierbarkeit
- moderate Standardabweichung von 0.06 zeigt eine stabile Modellperformance
- schlechtester Fold (0.75) liefert noch gute Ergebnisse
- bester Fold (0.90) zeigt das Potenzial des Modells

Residualanalyse

Kriterien

1. Zufällige Verteilung:
   - Training: Residuen sollten zufällig um die Linie y=0 verteilt sein.
   - Test: Ähnliches Muster wie im Training. ✓
2. Homoskedastizität:
   - Training: Konstante Streuung der Residuen.
   - Test: Keine systematischen Muster. ✓



Residualanalyse

Kriterien

3. Normalverteilung:
   - Training: Residuen sollten normalverteilt sein.
   - Test: Ähnliches Muster wie im Training.
4. Ausreißer:
   - Training: Identifikation und Untersuchung von Ausreißern.
   - Test: Ähnliches Muster wie im Training.

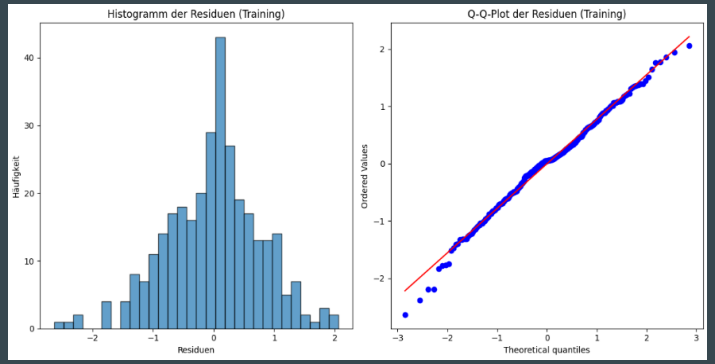
```
# Shapiro-Wilk-Test für Normalverteilung
shapiro_test = stats.shapiro(residuals_test)
print(f'Shapiro-Wilk-Test:
W={shapiro_test.statistic:.4f},
p={shapiro_test.pvalue:.4f}')
```

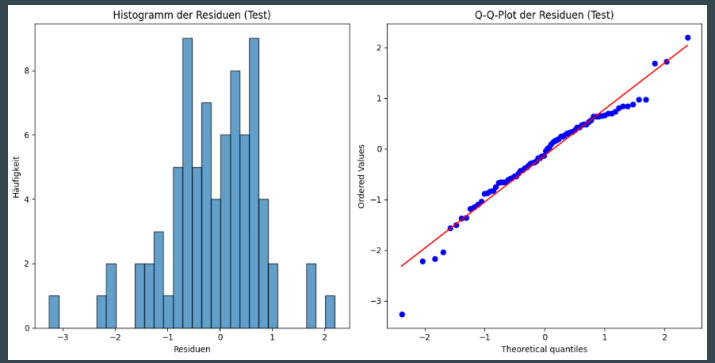
*1

*1: p-Wert > 0.05: Keine Ablehnung der Nullhypothese (Normalverteilung)

...

Shapiro-Wilk-Test: W=0.9923, p=0.0980





...

Shapiro-Wilk-Test: W=0.9710, p=0.0659

Achsenabschnitt und Steigungskoeffizienten

```
# Intercept (Achsenabschnitt)
intercept = pd.DataFrame({
    "Name": ["Intercept"],
    "Coefficient": [regr.intercept_]}
    )
# Steigungskoeffizienten
slope = pd.DataFrame({
    "Name": best_features,
    "Coefficient": regr.coef_}
)
# DataFrames kombinieren
table = pd.concat([intercept, slope], ignore_index=True, sort=False)

round(table, 3)
```

	Name	Coefficient
0	Intercept	25.857
1	vee	-0.000
2	unfaelle_je_10k_kfz	-0.014
3	pih	-2.099
4	euro2	-0.175
5	euro3	1.329
6	euro6dt	-0.224

Fazit

Das entwickelte multiple lineare Regressionsmodell konnte erfolgreich die relevanten **Einflussfaktoren auf den Anteil von Euro4-Fahrzeugen** identifizieren:

- Starke negative Korrelation mit neueren Technologien (Plug-In-Hybride)
- Positive Korrelation mit zeitlich naher Emissionsklasse (Euro 3)
- Moderate negative Korrelation mit neueren Emissionsklassen (Euro 6d-temp)
- Schwache Zusammenhänge mit sozioökonomischen Faktoren und Verkehrsunfällen

Bestätigte Annahmen:

- Signifikanter Einfluss neuer Emissionsvorschriften
- Deutlicher Rückgang bei hohem Anteil neuer Technologien
- Robustheit in Landkreisen mit älteren Fahrzeugen

	Name	Coefficient
0	Intercept	25.857
1	vee	-0.000
2	unfaelle_je_10k_kfz	-0.014
3	pih	-2.099
4	euro2	-0.175
5	euro3	1.329
6	euro6dt	-0.224

Fazit

Teilweise widerlegte Annahmen:

- Geringerer Einfluss sozioökonomischer Faktoren als erwartet
- Vernachlässigbarer Einfluss des verfügbaren Einkommens

Modellgüte und Generalisierbarkeit:

- Hohe Erklärungskraft (R^2)
- Stabile Kreuzvalidierungsergebnisse (mittleres R^2)
- Erfüllung aller statistischen Modellannahmen
- Robuste Performance auch nach Ausreißerbereinigung

Das Modell konnte die Forschungsfrage umfassend beantworten und die meisten Thesen bestätigen. Die identifizierten Zusammenhänge sind statistisch signifikant und inhaltlich plausibel. Die hohe Modellgüte und erfolgreiche Validierung unterstreichen die Zuverlässigkeit der Ergebnisse. Besonders hervorzuheben ist der starke Einfluss neuer Technologien auf die Verdrängung von Euro4-Fahrzeugen, während sozioökonomische Faktoren eine geringere Rolle spielen als ursprünglich angenommen.

Danke