Analyse

Die Datenanalyse bildet den Kern der Untersuchung. Aufbauend auf den in den vorangegangenen Kapiteln beschriebenen detaillierten Datenerhebungen und standardisierten Testbedingungen, konzentriert sich dieser Abschnitt auf die systematische Verarbeitung und Analyse der gesammelten Daten. Der Prozess umfasst die Aggregation und Vorverarbeitung der Fahrdaten, die Analyse von Key Performance Indicators (KPIs), die Anwendung geeigneter statistischer Methoden und ML-Verfahren zur Bewertung der Hypothesen.

Datenaggregation und -vorverarbeitung

Bei den erhaltenen Fahrdaten handelt es sich pro Fahrt um csv-Dateien, welche durch das LCMM-System z. T. bereits anhand von ISO 23795-1:2022 voraggregiert wurden. Die Daten enthalten beispielsweise i. d. R. sekündliche Informationen zu Zeitstempeln, Geschwindigkeiten, Beschleunigungen, Distanzen und Energieverbräuchen.

Zunächst werden die Daten in die verwendete Analyseumgebung importiert und ein erster Überblick über die Struktur und den Inhalt der Datensätze gewonnen, z. B. durch `str(dataset)`. (Quelle)

Da für die meisten Analysen die Zellen zum Spritverbrauch bzw. den daraus resultierenden CO2-Emissionen nicht benötigt werden und bei den Elektrofahrzeugen ohnehin leer sind, werden diese Spalten entfernt. Ebenso sind in der ersten Zeile einer Fahrtaufzeichnung einige ‘NA’-Werte enthalten. Dies liegt daran, dass diese z. B. im Falle der Geschwindigkeit anhand der Differenz zum vorherigen Zeitpunkt berechnet werden, dieser jedoch nicht existiert. Solche Felder werden durch den Wert 0 ersetzt. Eine weitere Imputation ist nun nicht weiter notwendig, da die Daten dann vollständig und konsistent sind.

Als nächster Schritt folgt die Zusammenführung der einzelnen Fahrten zu einem Gesamtdatensatz. Die einzelnen Datenpunkte werden mithilfe einer weiteren Spalte zu dem jeweiligen Fahrzeugtyp zugeordnet.

Um vor den Analysen bereits einen ersten Überblick über die Daten zu erhalten, werden die Verteilung der Energieverbräuche nach Geschwindigkeit und Fahrzeugtyp visualisiert. Dies ermöglicht es, erste Unterschiede und Muster zu erkennen, die für die weiteren Analysen relevant sein könnten.

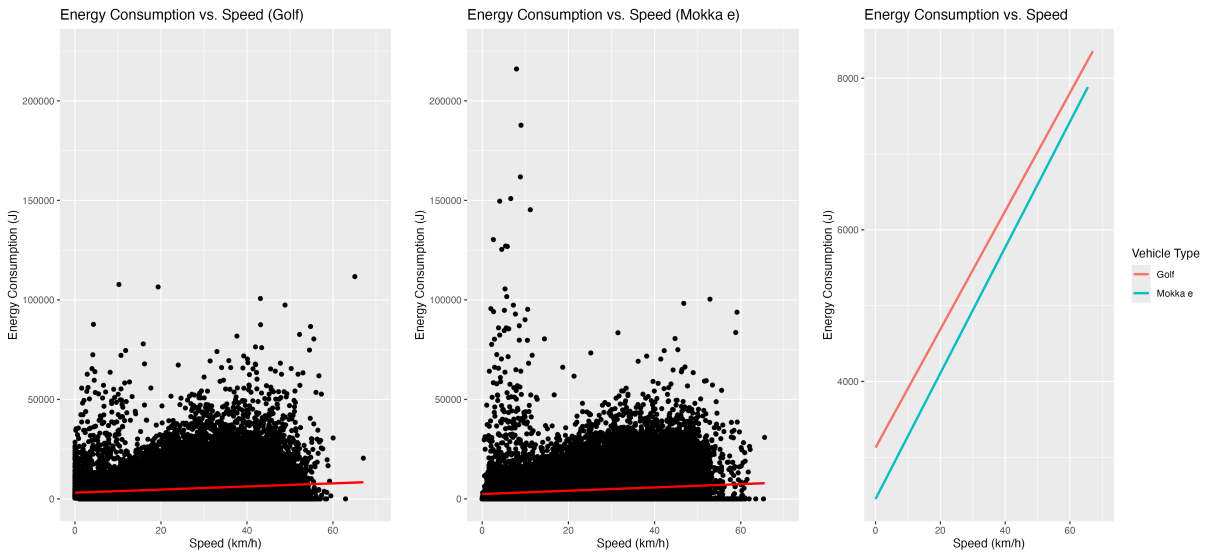

Figure 1: Verteilung des Energieverbrauchs nach Geschwindigkeit und Fahrzeugtyp.

Die ersten beiden Abbildungen zeigen die Verteilung des Energieverbrauchs des VW Golf und des Opel Mokka E in Abhängigkeit von der Geschwindigkeit inklusive einer Trendlinie. Die dritte Abbildung zeigt die beiden Trendlinien in einem gezoomten Bereich, um Unterschiede besser sichtbar zu machen.

Die hohe Variabilität des Energieverbrauchs bei verschiedenen Geschwindigkeiten, die in den Streudiagrammen für beide Fahrzeuge sichtbar wird, könnte durch unterschiedliche Fahrstile, Straßenbedingungen oder externe Faktoren wie Wetter bedingt sein. Die Ausreißer im Diagramm des Mokka e könnten auf spezifische Situationen oder Fehlmessungen hindeuten, die einer genaueren Untersuchung bedürfen.

Insgesamt ist zu erkennen, dass der Energieverbrauch bei höheren Geschwindigkeiten tendenziell steigt, wobei der Mokka E im Vergleich zum Golf einen niedrigeren Energieverbrauch aufweist.Dies bietet eine erste Orientierung für die weiteren Analysen und Hypothesenprüfungen.

Key Performance Indicators

In diesem Abschnitt werden die bereits definierten KPIs, wie der Energy Performance Index (EPI) und der Acceleration Performance Index (API), berechnet und analysiert, um die Hypothese 1 (H1) zu überprüfen. Diese KPIs ermöglichen es, Energieverbräuche norminalisert pro 100 km und Tonne zu vergleichen und die Effizienz der Fahrzeuge zu bewerten.

Energy Performance Index (EPI)

Der EPI wird berechnet, indem die gesamte verbrauchte Energie (in kWh) durch die gesamte zurückgelegte Strecke (in 100 km) und das Gewicht des Fahrzeugs (in Tonnen) geteilt wird. Dadurch wird der durchschnittliche Energieverbrauch pro 100 km und Tonne Fahrzeuggewicht ermittelt.

$$\text{EPI} = \frac{\text{Gesamte aufgewendete Energie (kWh)}}{\left(\frac{\text{Gesamtstrecke (km)}}{100}\right) * \text{Fahrzeuggewicht (t)}}$$

Die spezifische Implementierung sind im Anhang X zu finden.

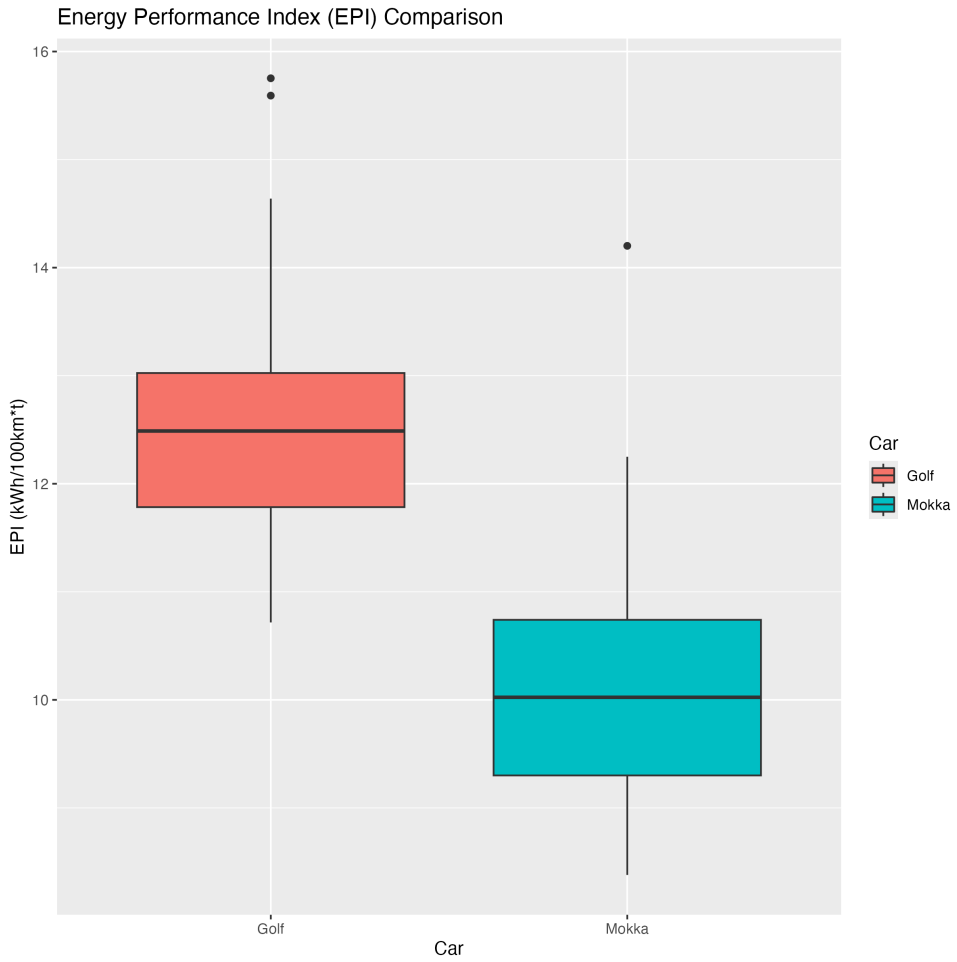

Figure 5: Energy Performance Index zwischen Golf und Mokka.

Die obige Abbildung zeigt einen Vergleich des Energy Performance Index (EPI) zwischen einem VW Golf (Verbrennungsmotor) und einem Opel Mokka (Elektrofahrzeug).

Die Boxplots verdeutlichen, dass der Opel Mokka E einen deutlich niedrigeren mittleren EPI-Wert (Median von etwa 11 kWh/100km*t) aufweist als der VW Golf (Median von etwa 12,5 kWh/100km*t). Der Interquartilsabstand (IQR) des Mokka erstreckt sich von etwa 10 bis 12 kWh/100km*t, während der IQR des Golfs von etwa 11,5 bis 13,5 kWh/100km*t reicht. Der Gesamtbereich der EPI-Werte des Golfs, einschließlich Ausreißern, liegt zwischen etwa 10 und 15 kWh/100km*t, wobei einige Ausreißer oberhalb von 15 kWh/100km*t zu erkennen sind. Im Gegensatz dazu zeigt der Mokka eine geringere Streuung der Werte, die von etwa 9 bis 13 kWh/100km*t reichen, mit einem Ausreißer oberhalb von 14 kWh/100km*t.

Die Analyse des EPI zeigt somit, dass der Opel Mokka E im Vergleich zum VW Golf eine höhere Energieeffizienz aufweist, was auf seinen niedrigeren Energieverbrauch pro 100 km und Tonne Fahrzeuggewicht zurückzuführen ist.

Acceleration Performance Index (API)

Der Acceleration Performance Index (API) ist ein weiterer wichtiger KPI, der den Energieverbrauch für Beschleunigungsmanöver normiert und zwischen Fahrzeugen vergleichbar macht.

Der API wird berechnet, indem die gesamte Energie für Beschleunigungsmanöver (in kWh) durch die gesamte zurückgelegte Strecke (in 100 km) und das Gewicht des Fahrzeugs (in Tonnen) geteilt wird.

$$\text{API} = \frac{\text{Gesamte Beschleunigungsenergie (kWh)}}{\left(\frac{\text{Gesamtstrecke (km)}}{100}\right) * \text{Fahrzeuggewicht (t)}}$$

Die spezifische Implementierung sind im Anhang X zu finden.

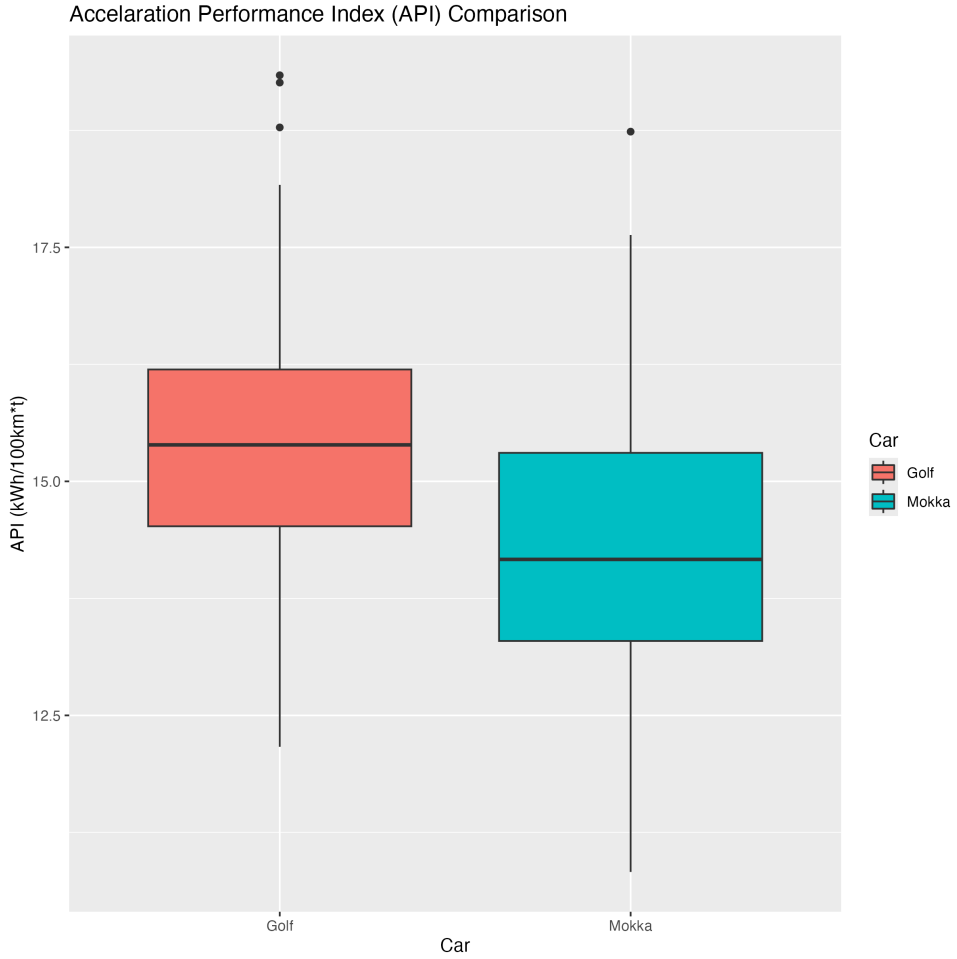

Figure 6: Acceleration Performance Index zwischen Golf und Mokka.

Die obige Abbildung zeigt einen Vergleich des Acceleration Performance Index (API) zwischen den beiden Fahrzeugen.

Die Boxplots verdeutlichen, dass der Opel Mokka einen deutlich niedrigeren mittleren API-Wert (Median von etwa 14 kWh/100km*t) aufweist als der VW Golf (Median von etwa 15,5 kWh/100km*t). Der Interquartilsabstand (IQR) des Mokka erstreckt sich von etwa 13 bis 15 kWh/100km*t, während der IQR des Golfs von etwa 14 bis 16 kWh/100km*t reicht. Der Gesamtbereich der API-Werte des Golfs, einschließlich Ausreißern, liegt zwischen etwa 13 und 17,5 kWh/100km*t, wobei einige Ausreißer oberhalb von 17,5 kWh/100km*t zu erkennen sind.

Im Gegensatz dazu zeigt der Mokka eine größere Streuung der Werte, die von etwa 12 bis 18 kWh/100km*t reichen, mit einem Ausreißer unterhalb von 12,5 kWh/100km*t und einem weiteren oberhalb von 18 kWh/100km*t.

Die Analyse des API zeigt somit, dass der Opel Mokka E im Vergleich zum VW Golf eine höhere Energieeffizienz bei Beschleunigungsmanövern aufweist, da er im Mittel weniger Energie für Beschleunigungen verbraucht.

Statistische Modellierung

Die statistische Analyse und Modellierung der Fahrdaten ist entscheidend, um die Hypothesen zu überprüfen und die Unterschiede zwischen Elektrofahrzeugen (EVs) und Fahrzeugen mit Verbrennungsmotor (ICEs) zu quantifizieren. In diesem Abschnitt werden verschiedene statistische Methoden und Machine-Learning-Verfahren angewendet, um die Beziehung zwischen Fahrzeugtyp und Energieverbrauch zu untersuchen.

Lineare Regression

Die lineare Regression ist eine grundlegende statistische Methode, die verwendet wird, um die Beziehung zwischen einer abhängigen Variable und einer oder mehreren unabhängigen Variablen zu modellieren. In diesem Kontext dient die lineare Regression dazu, den Energieverbrauch (TotalWork.J.) in Abhängigkeit von verschiedenen Einflussfaktoren wie Fahrzeugtyp, Distanz, Rollarbeit, Steigungsarbeit, Beschleunigungsarbeit und Beschleunigung zu untersuchen. Die Wahl der linearen Regression ermöglicht es, die Stärke und Richtung dieser Beziehungen quantitativ zu bestimmen und Vorhersagen über den Energieverbrauch basierend auf den unabhängigen Variablen zu treffen.

Auswahl der Variablen für das Regressionsmodell

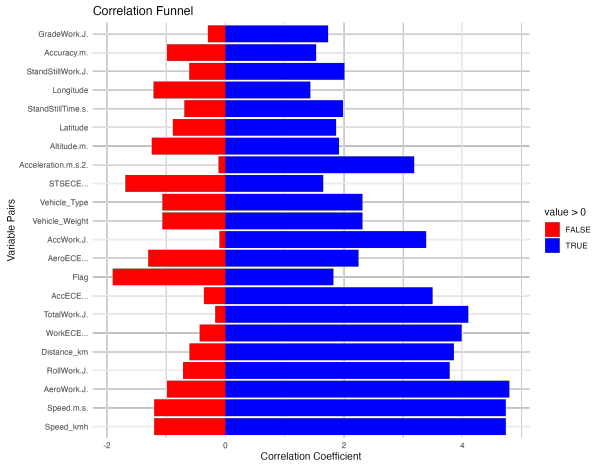

Figure 7: Korrelationsdiagramm der Variablen.

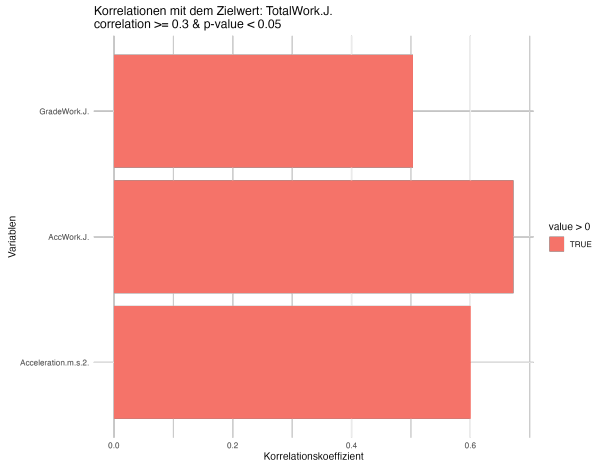

Figure 8: Korrelationsdiagramm der ausgewählten Variablen.

Die Auswahl der Variablen für das Regressionsmodell basiert auf ihrer theoretischen und empirischen Relevanz für den Energieverbrauch. Der Prozess der Variablenauswahl wurde durch die Analyse der Korrelationen zwischen den potenziellen Prädiktoren und dem Energieverbrauch unterstützt. Diese Korrelationen wurden in Form eines Correlation Funnels (Abbildung 8) und einer Korrelationsmatrix visualisiert, um die Stärke und Richtung der Zusammenhänge zu bewerten.

Der Correlation Funnel (Abbildung 8) zeigt die Korrelationen der unabhängigen Variablen mit der abhängigen Variable (TotalWork.J.). Die im Bild dargestellten Variablen „GradeWork.J.“, „AccWork.J.“ und „Acceleration.m.s.2.“ haben die höchsten positiven Korrelationen mit dem Energieverbrauch. Diese Variablen wurden aufgrund ihrer signifikanten Korrelationen und ihres theoretischen Zusammenhangs mit dem Energieverbrauch ausgewählt.

Abbildung 9 verdeutlicht zusätzlich die Korrelationen dieser ausgewählten Variablen untereinander. Es wird gezeigt, dass „GradeWork.J.“, „AccWork.J.“ und „Acceleration.m.s.2.“ starke Korrelationen aufweisen und somit wichtige Prädiktoren für den Energieverbrauch darstellen.

Neben diesen Variablen wurde auch „Vehicle_Type“ in das Modell aufgenommen, obwohl es nicht in der höchsten Korrelation mit dem Energieverbrauch stand. Dies ist notwendig, um die Hypothese

H1 zu überprüfen. „Vehicle_Type“ als Variable ermöglicht es, den Einfluss des Fahrzeugtyps direkt zu modellieren und die Unterschiede zwischen Elektrofahrzeugen und Verbrennerfahrzeugen zu quantifizieren.

Regressionsformel

Die Regressionsformel für das Modell lautet wie folgt:

$$\text{TotalWork.J.} = \beta_0 + \beta_1 * \text{Vehicle\_Type} + \beta_2 * \text{GradeWork.J.} + \beta_3 * \text{AccWork.J.} + \beta_4 * \text{Acceleration.m.s.2.} + \varepsilon$$

Bewertung des Regressionsmodells

Das Regressionsmodell wurde mit den oben genannten Variablen geschätzt und die Ergebnisse wurden anhand verschiedener statistischer Metriken wie dem Bestimmtheitsmaß (R^2), dem Adjusted R^2, dem F-Test und den p-Werten der Koeffizienten bewertet.

```
Residuals:
   Min     1Q Median     3Q    Max
 -5607  -2873  -1021   1653 127202

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.501e+03  2.406e+01 228.594  < 2e-16 ***
Vehicle_TypeMokka e -6.474e+02  2.990e+01 -21.653  < 2e-16 ***
GradeWork.J.         5.626e-01  2.095e-03 268.533  < 2e-16 ***
AccWork.J.           5.653e-01  3.497e-03 161.673  < 2e-16 ***
Acceleration.m.s.2. -1.267e+02  3.837e+01  -3.302 0.000961 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4365 on 93384 degrees of freedom
Multiple R-squared:  0.6916,  Adjusted R-squared:  0.6916
F-statistic: 5.236e+04 on 4 and 93384 DF,  p-value: < 2.2e-16
```

Code-Ausschnitt 1: Zusammenfassung des Regressionsmodells.

Die Ergebnisse zeigen, dass das Modell etwa 69,16 % der Varianz im Energieverbrauch erklären kann (R^2 = 0.6916). Alle unabhängigen Variablen haben signifikante Koeffizienten mit sehr niedrigen p-Werten ($p < 0.001$), was ihre Bedeutung im Modell bestätigt. Der F-Statistik-Wert und der damit verbundene p-Wert (< 2.2e-16) weisen darauf hin, dass das Modell insgesamt signifikant ist.

Die signifikanten Koeffizienten für die Variablen „Vehicle_Type“, „GradeWork.J.“, „AccWork.J.“ und „Acceleration.m.s.2.“ zeigen, dass diese Faktoren wichtige Determinanten des Energieverbrauchs sind. Insbesondere zeigt die negative Schätzung für „Vehicle_TypeMokka e“, dass der Mokka E im Vergleich zum Referenzfahrzeug (Golf) einen geringeren Energieverbrauch hat. Dies unterstützt die Annahme, dass Elektrofahrzeuge effizienter sind als Verbrennerfahrzeuge. Somit wird die Hypothese H1 unterstützt.

ANOVA

Die ANOVA (Analysis of Variance) ist eine statistische Methode, die verwendet wird, um festzustellen, ob es signifikante Unterschiede zwischen den Mittelwerten von zwei oder mehr Gruppen gibt. In diesem Fall wurde eine einfaktorielle ANOVA durchgeführt, um den Unterschied im Energieverbrauch (TotalWork.J.) zwischen verschiedenen Fahrzeugtypen zu analysieren.

Durchführung der ANOVA

```
> anova_result <- aov(TotalWork.J. ~ Vehicle_Type, data = combined_data)
> summary(anova_result)
                Df    Sum Sq   Mean Sq F value Pr(>F)
Vehicle_Type     1 5.925e+09 5.925e+09   96.01 <2e-16 ***
Residuals    93387 5.763e+12 6.171e+07
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Code-Ausschnitt 2: Zusammenfassung der ANOVA-Analyse.

Die Ergebnisse der ANOVA zeigen, dass der Fahrzeugtyp einen signifikanten Einfluss auf den Energieverbrauch hat (F-Wert = 96.01, p < 2e-16). Dies bedeutet, dass es signifikante Unterschiede im Energieverbrauch zwischen den verschiedenen Fahrzeugtypen gibt.

Tukey-HSD-Test

Um genauere Informationen über die Unterschiede zwischen den Fahrzeugtypen zu erhalten, wurde ein Tukey-HSD-Test (Tukey's Honest Significant Difference Test) durchgeführt. Der Tukey-HSD-Test ist ein post-hoc-Test, der nach der ANOVA durchgeführt wird, um festzustellen, welche spezifischen Gruppen sich signifikant voneinander unterscheiden. Der Vorteil des Tukey-HSD-Tests liegt darin, dass er die Fehlerwahrscheinlichkeit kontrolliert und somit zuverlässigere Ergebnisse liefert, wenn mehrere paarweise Vergleiche durchgeführt werden.

```
> tukey_result <- TukeyHSD(anova_result)
> print(tukey_result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = TotalWork.J. ~ Vehicle_Type, data = combined_data)

$Vehicle_Type
                   diff       lwr       upr p adj
Mokka e-Golf -527.2025 -632.6571 -421.7479     0
```

Code-Ausschnitt 3: Zusammenfassung des Tukey-HSD-Tests.

Der Tukey-HSD-Test bestätigt, dass der Mokka E im Vergleich zum Golf einen signifikant niedrigeren Energieverbrauch aufweist (Differenz: −527.2025 J, p-Wert: 0).

Bewertung der ANOVA und des Tukey-HSD-Tests

Die Ergebnisse der ANOVA und des Tukey-HSD-Tests bestätigen somit ebenso die Hypothese H1, dass Elektrofahrzeuge im Vergleich zu Verbrennerfahrzeugen einen geringeren Energieverbrauch haben.

Random Forest

Die Random Forest Regression ist ein leistungsstarkes maschinelles Lernmodell, das für die Vorhersage kontinuierlicher Werte verwendet wird. Es basiert auf der Aggregation von Vorhersagen mehrerer Entscheidungsbäume, um die Genauigkeit und Robustheit der Vorhersagen zu erhöhen. In diesem Abschnitt wird das Random Forest Modell zur Vorhersage des Energieverbrauchs (TotalWork.J.) verwendet.

Dazu werden die Daten zunächst in Trainings- und Testdaten aufgeteilt und norminalisert, um das Modell zu trainieren und zu validieren. Anschließend wird das Random Forest Modell mit den Trainingsdaten trainiert und auf den Testdaten getestet, um die Vorhersagegenauigkeit zu bewerten.

Durchführung des Random Forest Modells

```r
library(randomForest)
normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}

train_data_norm <- train_data %>%
  select(-TotalWork.J.) %>%
  mutate_if(is.numeric, normalize)
train_data_norm$TotalWork.J. <- train_data$TotalWork.J.

set.seed(123)
train_index <- createDataPartition(train_data_norm$TotalWork.J., p = 0.8, list =
FALSE)
train_set <- train_data_norm[train_index, ]
test_set <- train_data_norm[-train_index, ]

rf_model <- randomForest(TotalWork.J. ~ ., data = train_set, ntree = 100)
rf_pred <- predict(rf_model, test_set)
```

Code-Ausschnitt 4: Durchführung des Random Forest Modells.

Bewertung des Random Forest Modells

Die Ergebnisse des Random Forest Modells zeigen eine hohe Vorhersagegenauigkeit mit einem RMSE (Root Mean Squared Error) von 765.93, einem MAE (Mean Absolute Error) von 205.62 und einem Bestimmtheitsmaß (R^2) von 0.9909. Der RMSE in Prozent des durchschnittlichen Energieverbrauchs beträgt etwa 15.38 %, während der MAE in Prozent des gesamten Wertebereichs nur 0.095 % beträgt. Dies deutet darauf hin, dass das Modell die Energieverbräuche mit hoher Genauigkeit vorhersagen kann.

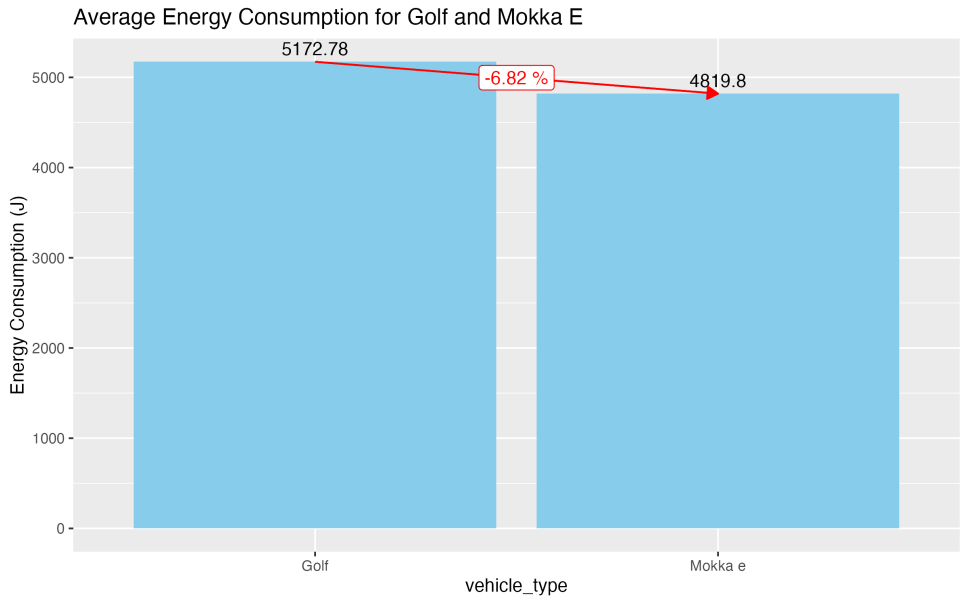

Figure 9: Ergebnisse des Random Forest Modells zu den Testdaten.

Die Vorhersageergebnisse zeigen, dass der durchschnittliche Energieverbrauch des Mokka E 6,82 % niedriger ist als der des Golfs. Dies unterstützt erneut die Hypothese H1, dass Elektrofahrzeuge im Vergleich zu Verbrennerfahrzeugen einen geringeren Energieverbrauch aufweisen.

Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) ist ein leistungsstarkes Ensemble-Lernverfahren, das schwache Lernalgorithmen, in der Regel Entscheidungsbäume, kombiniert, um ein starkes Vorhersagemodell

zu erstellen. Es arbeitet iterativ, indem es jedes neue Modell auf die Fehler der vorherigen Modelle trainiert.

```
library(gbm)

# GBM-Modell
gbm_model <- gbm(Energy_kWh ~ Vehicle_Type + Speed_kmh + Acceleration.m.s.2. +
Distance_km,
                 data = combined_data,
                 distribution = "gaussian",
                 n.trees = 5000,
                 interaction.depth = 4,
                 shrinkage = 0.01,
                 cv.folds = 5)
```

XGBoost

XGBoost (Extreme Gradient Boosting) ist ein optimierter verteilbarer Gradient Boosting Library, der speziell für Geschwindigkeit und Leistung entwickelt wurde. Es ist eine erweiterte Implementierung des Gradient Boosting Algorithmus, die Regularisierungsparameter verwendet, um Überanpassung zu vermeiden und die Leistung zu verbessern.

```
library(xgboost)

# Datenaufbereitung
data_matrix <- model.matrix(Energy_kWh ~ Vehicle_Type + Speed_kmh +
Acceleration.m.s.2. + Distance_km, data = combined_data)
labels <- combined_data$Energy_kWh

# XGBoost-Modell
xgb_model <- xgboost(data = data_matrix, label = labels, nrounds = 100, objective =
"reg:squarederror")

# XGBoost-Zusammenfassung
summary(xgb_model)
```

Vergleich mit WLTP-Werten

Dieses Kapitel untersucht den Energieverbrauch von Fahrzeugen unter realen Fahrbedingungen im Vergleich zu den WLTP-Normwerten (Worldwide Harmonized Light Vehicles Test Procedure), um die Hypothese 2 (H2) zu überprüfen.

Berechnung des WLTP-Verbrauchs

Zunächst wurden die Fahrten nach den WLTP-Grenzgeschwindigkeiten eingeteilt, um den Energieverbrauch in verschiedenen Geschwindigkeitsbereichen zu analysieren. Anschließend wurde der durchschnittliche Energieverbrauch in jedem Geschwindigkeitsbereich berechnet und mit den entsprechenden WLTP-Normwerten verglichen. Die Abweichung zwischen den realen Verbrauchswerten und den WLTP-Normwerten wurde in Prozent berechnet, um festzustellen, ob die Fahrzeuge die Normwerte einhalten oder überschreiten.

Diese Ergebnisse wurden in Diagrammen visualisiert, um eine klare Darstellung der Abweichungen zu ermöglichen und die Unterschiede zwischen den Fahrzeugtypen zu verdeutlichen.

Außerdem lag der Fokus auf der Verteilung der Verbrauchswerte in den verschiedenen WLTP-Kategorien, um zu verstehen, wie sich die Verbräuche über verschiedene Fahrbedingungen und -umgebungen verteilen. Dies ermöglichte es, festzustellen, ob bestimmte Bedingungen zu höheren oder niedrigeren Abweichungen vom WLTP-Normwert führen.

Der zugehörige Code zur Berechnung und Analyse der WLTP-Werte ist im Anhang X zu finden.