Final Project Report

Comprehensive Vehicle Price Analysis Using Regression Analysis

Mandira Ghimire, Harini Sanamandra, Manohar Bathina, and Abdul Sajid Mohammad

MSDA 3055: Linear Regression & Time Series, Clark University

Professor Aghil Alaee Khangha

December 11, 2024

Table of Contents

Introduction

The automotive market is highly dynamic, with prices influenced by numerous factors such as mileage, age, fuel type, engine specifications, and consumption rates. Accurate price prediction is essential for dealerships, buyers, and sellers to make informed decisions. This project aims to analyze vehicle data and build a regression model to predict vehicle prices based on key characteristics.

Objective

This project aims to predict vehicle prices based on attributes such as mileage, age, fuel type, and horsepower and to build a robust regression model for price prediction and discuss findings. In this project, we can discuss model assumptions and model fit analysis. The goal is to provide actionable insights for pricing strategies and to discuss the limitations and next steps.

Research Questions

1. What are the significant factors affecting vehicle prices?
    - This question aims to predict vehicle prices based on attributes such as mileage, age, fuel type, and horsepower.
2. How accurately can these prices be predicted using a regression model?
    - This question explores how fuel consumption and horsepower affect pricing patterns.

Hypothesis

Our research is all about predicting vehicle prices using regression analysis involving simple, multiple, and polynomials, etc. To get accurate vehicle prices, firstly we cleaned the Vehicle Pricing Rides dataset from Kaggle. By analyzing the vehicle price, horsepower, age, and mileage, we can predict the vehicle prices. Below are the few hypotheses we are working on in this project

H1: Engine power significantly influences vehicle prices.

H2: Vehicle features, and inspection positively impact pricing.

H3: Mileage and age have a negative correlation with vehicle prices.

Description of Data

Data Collection**:** The dataset used in this project, the Vehicle Prices dataset, consists of 15,915 records and 23 variables. It provides detailed information about vehicle pricing, specifications, and performance metrics, offering range of variables of data for regression analysis. The dataset was collected to represent a diverse range of vehicle attributes, including technical specifications, aesthetic features, and performance indicators.

Variables Overview:

A. Target/Response Variable

- **Price: (Quantitative, Continuous)** This is the target variable representing the price of a vehicle. It serves as the dependent variable to be predicted based on the other attributes.

B. Predictor Variables

Quantitative Variables

1. **Kilometers (km):** Distance traveled by the vehicle, indicating its usage**.**
2. **Age:** The age of the vehicle in years.
3. **Horsepower (hp_kW):** Engine power in kilowatts.
4. **Displacement (Displacement_cc):** Engine displacement measured in cubic centimeters.
5. **Weight (Weight_kg):** Weight of the vehicle in kilograms.
6. **Gears:** Number of gears in the vehicle's transmission.
7. **Previous Owners:** Number of previous owners of the vehicle.
8. **Inspection New (Inspection_new): (**Binary Quantitative**)** A binary indicator (1 for a new inspection, 0 otherwise).
9. **Combined Fuel Consumption (cons_comb):** The combined fuel consumption of the vehicle in liters per 100 km.

Categorical Variables

1. **Make/Model:** The brand and model of the vehicle (e.g., "Audi A1").
2. **Body Type:** Vehicle type (e.g., "Sedans," "SUVs").
3. **VAT:** Tax applicability, such as "VAT deductible" or "Price negotiable."
4. **Type:** Condition of the vehicle (e.g., "Used," "New").
5. **Fuel:** Type of fuel used, such as "Diesel" or "Benzine."

6. **Upholstery Type:** Type of upholstery in the vehicle, such as "Cloth" or "Leather."

7. **Gearing Type:** Transmission type, e.g., "Automatic" or "Manual."

8. **Drive Chain:** The drivetrain type, e.g., "Front" or "All-wheel drive."

9. **Paint Type:** Type of paint finish, such as "Metallic" or "Non-metallic."

Vehicle Price(*A.1*) dataset provides a robust mix of quantitative and categorical variables, enabling a detailed exploration of the factors that influence vehicle pricing. These variables are further preprocessed and encoded to ensure compatibility for regression analysis.

Data Pre-processing:

In the data preprocessing(*A.2*) phase, several steps were taken to prepare the dataset for regression analysis**.**

1. Handling Missing Values:

The dataset was thoroughly examined for missing values, and no missing values were found. A detailed check was also performed for placeholder values like empty strings, "NA," or "Unknown" to ensure data consistency.

2. Removal of Duplicate Rows:

Duplicate entries were identified and removed to prevent redundancy and ensure the dataset's integrity. After removal, the dataset was rechecked to confirm no duplicates remained.

3. Encoding Categorical Variables:

Categorical variables were processed in two ways:

- One-Hot Encoding: Applied to single-valued categorical variables to convert them into binary columns, ensuring they are machine-readable for regression modeling.
- Multi-Valued Categorical Columns: Columns like "Comfort_Convenience" and "Safety_Security" containing multiple values were split into separate binary features, representing the presence or absence of each feature.

4. Boolean Conversion:

Boolean columns were converted into integers (0 or 1) to make them compatible with numerical analysis techniques.

5. Removal of Redundant Columns:

Columns that were multi-valued categorical variables were removed after their features were extracted as binary indicators. This reduced the dataset's complexity and dimensionality.

6. Dataset Validation

The dataset was validated after all preprocessing steps to ensure that the transformations were successful, and the data structure was suitable for analysis.

Regression Analysis:

1. **Exploratory Data Analysis (EDA):** Conduct a thorough examination of patterns, trends, and outliers
2. **Feature Engineering:** Develop relevant features for predicting vehicle prices
3. **Model Development:** Apply Multiple Linear Regression using pertinent variables
4. **Model Evaluation:** Assess performance through R-squared, MSE, and RMSE
5. **Model Deployment:** Create interpretable insights for stakeholders

Data Splitting:

To ensure reliable and unbiased model evaluation, the dataset was split into training and testing sets for regression analysis(*B.1*). Using a 70/30 ratio, the training set contains 9,958 rows (70% of the data) for model training, while the testing set consists of 4,268 rows (30%) for model validation and performance evaluation. A random seed (123) was set to ensure reproducibility of the split, so results can be consistently replicated.

Exploratory Data Analysis (EDA)

1. **Strong Correlation:**
    - hp_kW and price (Corr: 0.77)
    - Weight_kg and displacement (Corr: 0.566)
2. **Strong Negative Correlation:**
    - Age and price(-0.495)
    - Km and price (-0.418)
3. **Weak Relationship:**
    - Inspection and price

The exploratory data analysis (*A.3*) reveals that vehicle price is strongly positively correlated with horsepower (hp_kW) and engine displacement, suggesting more powerful and heavier vehicles tend to command higher prices. Conversely, price is negatively correlated with age and mileage, indicating that older and more used vehicles are priced lower, while inspection status shows a weak relationship with price, implying limited influence.

Final Regression Model:

A multiple linear regression model(*B.3*) incorporating interactive and polynomial terms was selected as the primary approach due to its balance of simplicity, interpretability, and ability to capture nonlinear relationships. Predictor variables were chosen based on correlation analysis, domain knowledge, and statistical significance. Key predictors identified include mileage (km), age, horsepower (hp_kW), engine displacement (Displacement_cc), fuel consumption (cons_comb), and vehicle weight (Weight_kg). Among these, mileage, age, and fuel consumption exhibited negative relationships with price, while horsepower and weight had positive effects. The number of previous owners (Previous_Owners) was not statistically significant and showed limited impact on price(B.2).

Interpretations and Findings from the Coefficients:

The regression analysis(*B.3*) provides several key insights into the factors influencing vehicle prices. Statistical significance of coefficients was assessed using **p-values** (<0.05 threshold), which supported our hypotheses as follows:

1. **Engine Power (hp_kW) and Price**:

   A highly significant positive coefficient for engine power (p<0.001) confirms that vehicles with higher horsepower command higher prices. This finding aligns with **H1 (Engine power significantly influences vehicle prices)**, as engine performance is a critical determinant of desirability and market value. The high t-value of 81.96 further emphasizes the robustness of this relationship.

2. **Vehicle Features and Inspection**:

   Some vehicle features, such as weight (p<0.001), showed a significant positive impact on pricing, supporting **H2 (Vehicle features positively impact pricing)**. However, inspection status (binary variable) and certain additional features were less significant, indicating a mixed relationship with pricing. The analysis suggests that features like weight contribute more strongly to vehicle valuation than inspection alone.

3. **Mileage (km) and Price**:

   The coefficient for mileage is negative and highly significant (p<0.001), indicating that higher mileage reduces vehicle prices due to depreciation. This supports **H3 (Mileage has a negative correlation with vehicle prices)**. For every additional kilometer traveled, the price decreases by approximately $0.05, emphasizing the importance of mileage in determining value.

4. **Age and Price**:

   A significant negative coefficient for vehicle age (p<0.001) corroborates **H3 (Age has a negative correlation with vehicle prices)**. Older vehicles lose value due to wear and tear and reduced consumer demand. Specifically, vehicle prices decrease by approximately $1,804 per additional year of age.

5. **Fuel Consumption (cons_comb) and Price**:

   Fuel consumption (combined) has a significant negative coefficient (p<0.001), indicating that less fuel-efficient vehicles are associated with lower prices. This highlights consumer preference for efficiency in modern markets.

6. **Previous Owners**:

   The number of previous owners showed no significant effect on price (p=0.327), suggesting that this variable may not heavily influence vehicle valuation in the dataset.

Model Assumptions & Model Fit Analysis:

Model Assumptions

Multiple linear regression relies on several key assumptions to ensure the validity and reliability of the results. Below, we evaluate whether these assumptions were met in our model:

1. **Linearity**:
   - Assumption: The relationship between predictor variables and the response variable (price) is linear.
   - Evaluation: The residuals vs. fitted plot showed some non-random patterns, particularly at higher fitted values, suggesting mild non-linearity. This may indicate that certain predictors have interactions or polynomial relationships that the linear model does not fully capture.
2. **Independence**:
   - Assumption: Observations are independent of one another.
   - Evaluation: Independence was assumed based on the nature of the data collection process, as no time-series or hierarchical data structure was present.
3. **Homoscedasticity (Constant Variance)**:
   - Assumption: The variance of residuals is constant across all levels of fitted values.
   - Evaluation: The Scale-Location plot indicated slight heteroscedasticity, as the variance increased at higher fitted values. This suggests that the model's error terms may not have a consistent spread, which could affect prediction accuracy at extreme price ranges.
4. **Normality of Residuals**:

   - Assumption: Residuals are normally distributed.
   - Evaluation: The Q-Q plot showed that residuals largely follow a normal distribution, except for deviations at the tails. Outliers such as observations 2922 and 32850 affected the normality of the residuals.

5. **No Multicollinearity**:
   - Assumption: Predictor variables are not highly correlated with each other.
   - Evaluation: Variance Inflation Factor (VIF) values were calculated, with most predictors showing acceptable multicollinearity levels (VIF<5). However, some variables such as km and cons_comb had higher interactions with others. Adjusted GVIF1/(2Df) values confirmed that multicollinearity was within manageable thresholds.

Model Fit Analysis

The model fit analysis(*B.8*) evaluates how well the regression model explains the variability in vehicle prices and predicts outcomes.

1. **Performance Metrics**:
   - **R²**: The model achieved an R2 value of **77.51%**, indicating that approximately 77.51% of the variability in vehicle prices is explained by the predictors.
   - **Adjusted R²**: The adjusted R2 value accounts for the number of predictors, ensuring that only significant variables contribute to the fit. It remained close to R2 confirming the model's robustness.
   - **Root Mean Squared Error (RMSE)**: An RMSE of **3,237** suggests that the model predicts vehicle prices with a reasonable level of accuracy.

2. **Residual Diagnostics**:
    - **Residuals vs. Fitted Plot**: Some residual patterns at higher fitted values suggest potential specification issues, which may require higher-order or interaction terms.
    - **Scale-Location Plot**: Highlighted non-constant variance (heteroscedasticity), especially for extreme price values.
    - **Q-Q Plot**: Confirmed near-normal distribution of residuals, with some deviation at the tails caused by outliers.
3. **Influential Points**:
    - High-leverage observations(*B.4*) (e.g., 33256 and 32850) were identified in the Residuals vs. Leverage plot. A total of **859 high-leverage points** exceeded Cook's Distance thresholds, suggesting these points may disproportionately influence the model. Addressing these points could further stabilize the coefficients and improve predictions.
4. **Goodness-of-Fit**:
    - The model performed consistently across the training and test datasets, with no significant overfitting. This was evident from the similarity of RMSE and MAE values between the two datasets.

Discussion & Limitations

Our regression analysis of vehicle pricing provides valuable insights into the complex factors influencing vehicle market valuations. However, it is crucial to critically examine our methodology, findings, and potential limitations.

Predictions and Conclusions

The multiple linear regression model reveals several key predictive factors for vehicle pricing:

1. **Engine Performance**: Horsepower emerges as a critical determinant of vehicle value. The highly significant positive coefficient demonstrates that more powerful vehicles command higher prices, reflecting consumer preferences for performance.
2. **Depreciation Factors**: The model conclusively shows the negative impact of mileage and age on vehicle prices. Each additional kilometer traveled reduces the price by approximately $0.05, while vehicle age decreases value by around $1,804 per year. This aligns with expected depreciation patterns in the automotive market.
3. **Efficiency Considerations**: The negative correlation between fuel consumption and price suggests a growing market preference for fuel-efficient vehicles. This trend likely reflects increasing environmental consciousness and economic considerations among buyers.

Methodological Critiques

Data Limitations:

- For the detail evaluation , sampling the data for less than 500 observations would have helped to have better understanding.
- The data appears to be from a specific market or time, potentially limiting the generalizability of findings.
- The binary nature of some categorical variables (e.g., inspection status) may oversimplify complex market dynamics.

Model Constraints:

1. **Non-Linearity**: The residuals vs. fitted plot revealed non-random patterns, suggesting that some relationships might be more complex than a linear model can fully capture.
2. **Heteroscedasticity**: The Scale-Location plot indicated increasing variance at higher fitted values, implying that the model's predictive accuracy may vary across different price ranges.
3. **Outlier Influence**: 859 high-leverage points were identified, potentially skewing the model's coefficients and predictive power.

Reliability and Validity Considerations

Data Reliability:

- While no missing values were found, the preprocessing steps (such as one-hot encoding and feature extraction) may have introduced some information loss.
- The removal of multi-valued categorical columns might have eliminated nuanced information about vehicle characteristics.

Model Validity:

- The R^2 value of 77.51% indicates a strong explanatory power but also suggests that approximately 22.49% of price variation remains unexplained.
- The RMSE of 3,237 provides a reasonable measure of prediction accuracy but indicates potential for improvement.

Suggested Improvements

1. **Advanced Modeling Techniques**:
   - Explore non-linear regression models or machine learning approaches like Random Forest or Gradient Boosting.
   - Implement more sophisticated feature engineering techniques.
   - Conduct more in-depth analysis of interaction terms and polynomial relationships.
2. **Data Enrichment**:
   - Incorporate additional variables such as market trends, regional pricing differences, and economic indicators.
   - Collect a more diverse and larger dataset to improve model generalizability.
3. **Outlier Handling**:
   - Develop more robust methods for identifying and treating high-leverage points.
   - Consider robust regression techniques that are less sensitive to outliers.

Practical Implications and appropriateness of regression analysis

Despite its limitations, the model provides actionable insights for:

- Used car dealerships in pricing strategies
- Potential buyers in understanding vehicle valuation factors
- Automotive market researchers studying pricing dynamics

Conclusion:

The research project successfully developed a multiple linear regression model to predict vehicle prices by analyzing 15,915 vehicle records. By examining critical factors such as horsepower, mileage, age, and fuel consumption, the study revealed significant insights into automotive pricing dynamics. The model achieved an impressive 77.51% explanatory power (R^2 value), demonstrating that engine power, vehicle features, and depreciation substantially influence vehicle valuation. Specifically, the analysis confirmed our hypotheses that more powerful vehicles command higher prices, while factors like increasing mileage and age consistently decrease a vehicle's market value.

The findings provide actionable insights for stakeholders in the automotive market, including dealerships, buyers, and market researchers. By quantifying the impact of key variables—such as each kilometer reducing vehicle price by $0.05 and each year of age decreasing value by approximately $1,804—the research offers a data-driven approach to understanding pricing mechanisms. While the model presents a robust framework for price prediction, it also underscores the complexity of automotive valuations, suggesting that future research could benefit from more advanced modeling techniques, expanded datasets, and deeper exploration of market dynamics.

Additional Regression Analysis

Simple Linear Regression

Simple Linear Regression evaluated individual relationships between predictors engine horsepower and the target variable. While useful for initial insights, this model fails to account for the combined influence of multiple factors.

Assumptions and Model Evaluation:

1. Model Assumptions:
   - Linearity: The model assumes a linear relationship between horsepower (hp_kW) and vehicle price.
   - Independence: Observations are assumed to be independent of each other.
   - Homoscedasticity: The model assumes constant variance of residuals across all predictor values.
   - Normality of Residuals: The residuals are expected to be normally distributed.
2. Model Characteristics:
   - Baseline Model: Simple linear regression with price as the dependent variable and horsepower as the independent variable.
   - Initial Model Coefficients:
     - Intercept: $2,448.95 (base price when horsepower is zero)
     - Coefficient: $174.56 increase in price per unit increase in horsepower
3. Model Performance:
   - R-squared: 0.4576 (45.76% of price variation explained)
   - Residual Standard Error: 4,995

    - Statistically Significant: Both intercept and horsepower coefficient are highly significant (p-value < 2e-16)

4. Diagnostic Findings:
    - Breusch-Pagan Test: Extremely low p-value (5.44e-264) indicates significant heteroscedasticity
    - High Leverage Points: 246 identified points
    - Influential Points: 605 points identified through Cook's Distance

Reasons for Not Selecting the Model:

1. Limited Predictive Power:
    - The model explains only 45.76% of price variation, indicating substantial unexplained variance.
    - Suggests that horsepower alone is insufficient to accurately predict vehicle prices.
2. Diagnostic Issues:
    - Significant heteroscedasticity (unequal variance of residuals)
    - Large number of influential and high-leverage points
    - Non-linear relationship between horsepower and price.

Works Cited

1. (2024). Retrieved from Kaggle: https://www.kaggle.com/code/yaaryiitturan/predicting-vehicle-prices-using-regression-models/input

2. Allison, P. D. (n.d.). Testing for Interaction in Multiple Regression . *American Journal of Sociology*.

3. David Greene, A. H. (2018). Consumer willingness to pay for vehicle attributes: What do we Know? *Transportation Research Part A: Ploicy and Practice*, 258-279.

4. Kutner, M. H. (2004). Applied linear regression models. Ingram.

5. Sumeyra Muti, K. Y. (2023). Using Linear Regression For Used Car Price Prediction. *International Journal of Computational and Experimental Science and Engineering(IJCESEN)*.

Dataset

https://www.kaggle.com/code/yaaryiitturan/predicting-vehicle-prices-using-regression-models/input

Vehicle Price DATA.csv

Code

Vehicle Price analysis.R

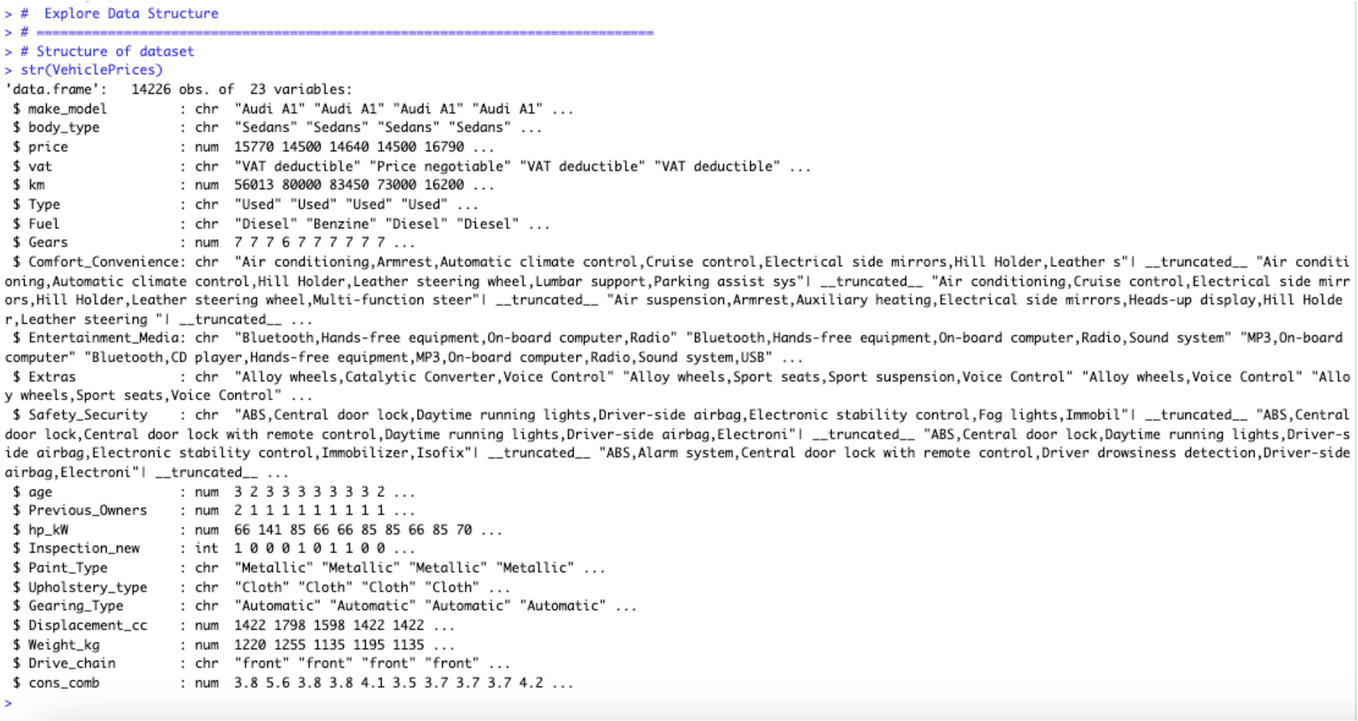
Appendix:

Appendix A

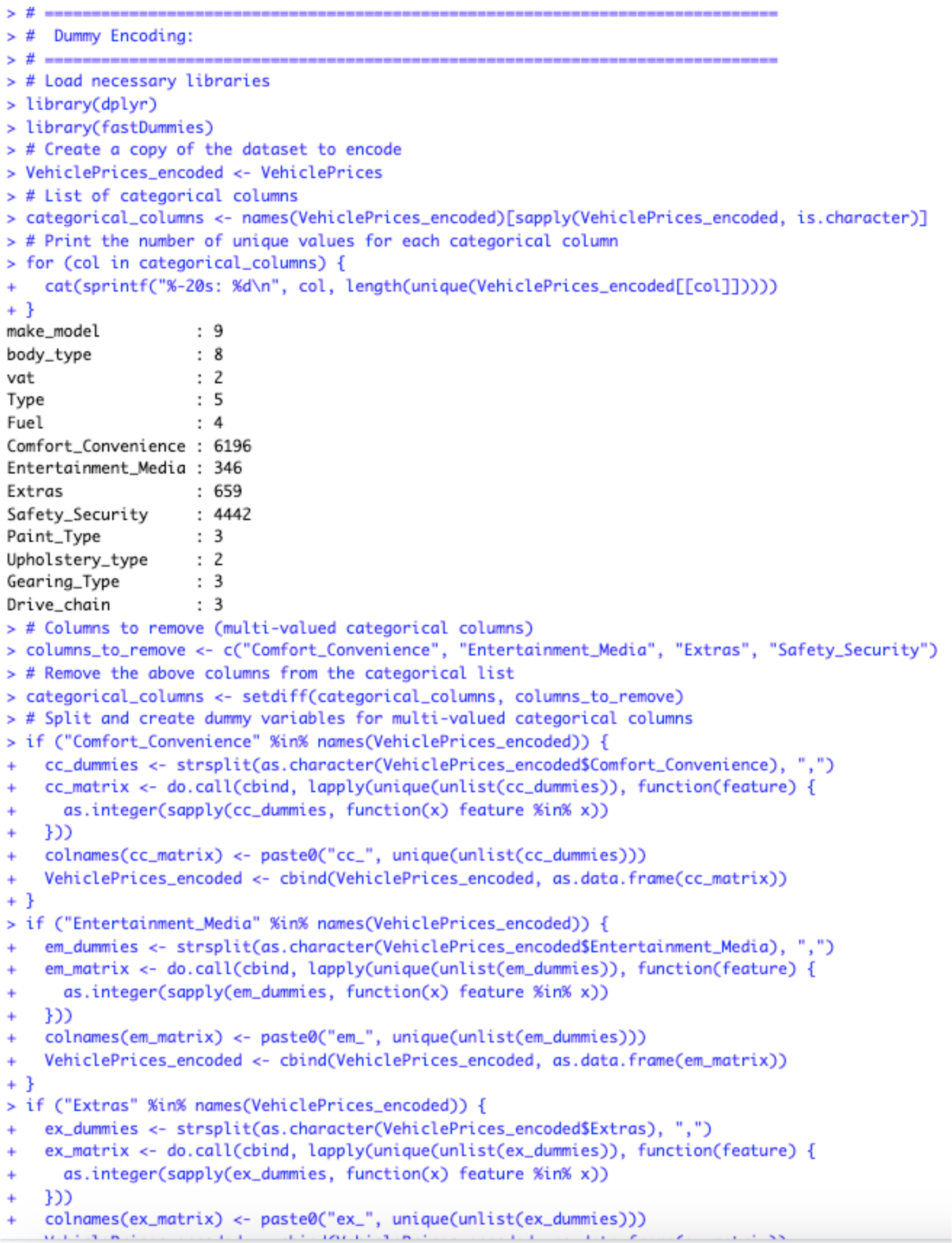
Appendix B

Appendix C

A.1

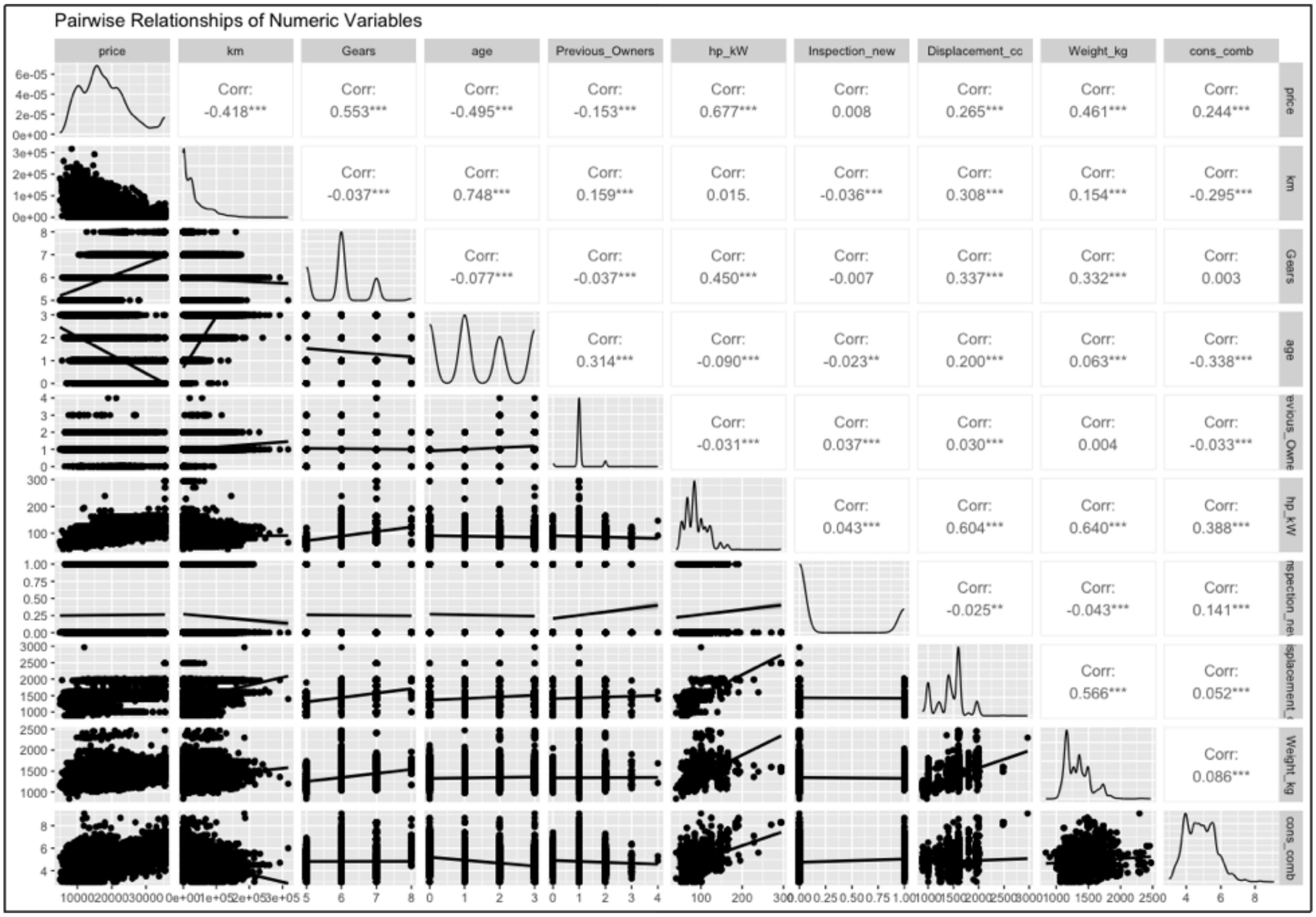


```
> #  Explore Data Structure
> # ==============================================================================
> # Structure of dataset
> str(VehiclePrices)
'data.frame':	14226 obs. of  23 variables:
 $ make_model         : chr  "Audi A1" "Audi A1" "Audi A1" "Audi A1" ...
 $ body_type          : chr  "Sedans" "Sedans" "Sedans" "Sedans" ...
 $ price              : num  15770 14500 14640 14500 16790 ...
 $ vat                : chr  "VAT deductible" "Price negotiable" "VAT deductible" "VAT deductible" ...
 $ km                 : num  56013 80000 83450 73000 16200 ...
 $ Type               : chr  "Used" "Used" "Used" "Used" ...
 $ Fuel               : chr  "Diesel" "Benzine" "Diesel" "Diesel" ...
 $ Gears              : num  7 7 7 6 7 7 7 7 7 7 ...
 $ Comfort_Convenience: chr  "Air conditioning,Armrest,Automatic climate control,Cruise control,Electrical side mirrors,Hill Holder,Leather s"| __truncated__ "Air conditi
oning,Automatic climate control,Hill Holder,Leather steering wheel,Lumbar support,Parking assist sys"| __truncated__ "Air conditioning,Cruise control,Electrical side mirr
ors,Hill Holder,Leather steering wheel,Multi-function steer"| __truncated__ "Air suspension,Armrest,Auxiliary heating,Electrical side mirrors,Heads-up display,Hill Holde
r,Leather steering "| __truncated__ ...
 $ Entertainment_Media: chr  "Bluetooth,Hands-free equipment,On-board computer,Radio" "Bluetooth,Hands-free equipment,On-board computer,Radio,Sound system" "MP3,On-board
computer" "Bluetooth,CD player,Hands-free equipment,MP3,On-board computer,Radio,Sound system,USB" ...
 $ Extras             : chr  "Alloy wheels,Catalytic Converter,Voice Control" "Alloy wheels,Sport seats,Sport suspension,Voice Control" "Alloy wheels,Voice Control" "Allo
y wheels,Sport seats,Voice Control" ...
 $ Safety_Security    : chr  "ABS,Central door lock,Daytime running lights,Driver-side airbag,Electronic stability control,Fog lights,Immobil"| __truncated__ "ABS,Central
door lock,Central door lock with remote control,Daytime running lights,Driver-side airbag,Electroni"| __truncated__ "ABS,Central door lock,Daytime running lights,Driver-s
ide airbag,Electronic stability control,Immobilizer,Isofix"| __truncated__ "ABS,Alarm system,Central door lock with remote control,Driver drowsiness detection,Driver-side
airbag,Electroni"| __truncated__ ...
 $ age                : num  3 2 3 3 3 3 3 3 3 2 ...
 $ Previous_Owners    : num  2 1 1 1 1 1 1 1 1 1 ...
 $ hp_kW              : num  66 141 85 66 66 85 85 66 85 70 ...
 $ Inspection_new     : int  1 0 0 0 1 0 1 1 0 0 ...
 $ Paint_Type         : chr  "Metallic" "Metallic" "Metallic" "Metallic" ...
 $ Upholstery_type    : chr  "Cloth" "Cloth" "Cloth" "Cloth" ...
 $ Gearing_Type       : chr  "Automatic" "Automatic" "Automatic" "Automatic" ...
 $ Displacement_cc    : num  1422 1798 1598 1422 1422 ...
 $ Weight_kg          : num  1220 1255 1135 1195 1135 ...
 $ Drive_chain        : chr  "front" "front" "front" "front" ...
 $ cons_comb          : num  3.8 5.6 3.8 3.8 4.1 3.5 3.7 3.7 3.7 4.2 ...
>
```

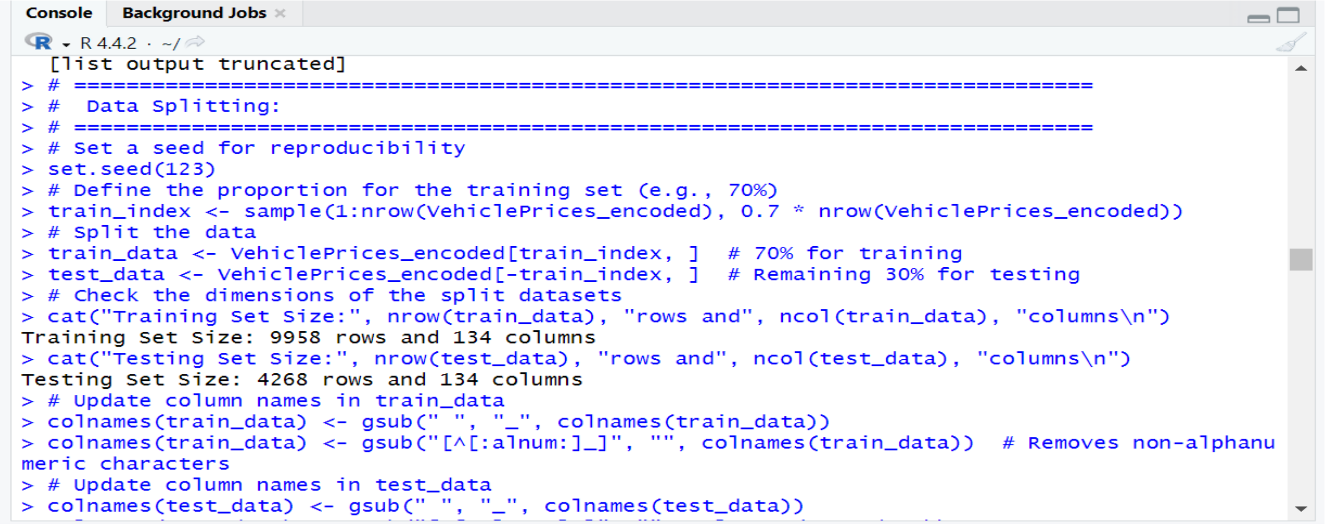


```
> # ==============================================================================
> #  Dummy Encoding:
> # ==============================================================================
> # Load necessary libraries
> library(dplyr)
> library(fastDummies)
> # Create a copy of the dataset to encode
> VehiclePrices_encoded <- VehiclePrices
> # List of categorical columns
> categorical_columns <- names(VehiclePrices_encoded)[sapply(VehiclePrices_encoded, is.character)]
> # Print the number of unique values for each categorical column
> for (col in categorical_columns) {
+   cat(sprintf("%-20s: %d\n", col, length(unique(VehiclePrices_encoded[[col]]))))
+ }
make_model          : 9
body_type           : 8
vat                 : 2
Type                : 5
Fuel                : 4
Comfort_Convenience : 6196
Entertainment_Media : 346
Extras              : 659
Safety_Security     : 4442
Paint_Type          : 3
Upholstery_type     : 2
Gearing_Type        : 3
Drive_chain         : 3
> # Columns to remove (multi-valued categorical columns)
> columns_to_remove <- c("Comfort_Convenience", "Entertainment_Media", "Extras", "Safety_Security")
> # Remove the above columns from the categorical list
> categorical_columns <- setdiff(categorical_columns, columns_to_remove)
> # Split and create dummy variables for multi-valued categorical columns
> if ("Comfort_Convenience" %in% names(VehiclePrices_encoded)) {
+   cc_dummies <- strsplit(as.character(VehiclePrices_encoded$Comfort_Convenience), ",")
+   cc_matrix <- do.call(cbind, lapply(unique(unlist(cc_dummies)), function(feature) {
+     as.integer(sapply(cc_dummies, function(x) feature %in% x))
+   }))
+   colnames(cc_matrix) <- paste0("cc_", unique(unlist(cc_dummies)))
+   VehiclePrices_encoded <- cbind(VehiclePrices_encoded, as.data.frame(cc_matrix))
+ }
> if ("Entertainment_Media" %in% names(VehiclePrices_encoded)) {
+   em_dummies <- strsplit(as.character(VehiclePrices_encoded$Entertainment_Media), ",")
+   em_matrix <- do.call(cbind, lapply(unique(unlist(em_dummies)), function(feature) {
+     as.integer(sapply(em_dummies, function(x) feature %in% x))
+   }))
+   colnames(em_matrix) <- paste0("em_", unique(unlist(em_dummies)))
+   VehiclePrices_encoded <- cbind(VehiclePrices_encoded, as.data.frame(em_matrix))
+ }
> if ("Extras" %in% names(VehiclePrices_encoded)) {
+   ex_dummies <- strsplit(as.character(VehiclePrices_encoded$Extras), ",")
+   ex_matrix <- do.call(cbind, lapply(unique(unlist(ex_dummies)), function(feature) {
+     as.integer(sapply(ex_dummies, function(x) feature %in% x))
+   }))
+   colnames(ex_matrix) <- paste0("ex_", unique(unlist(ex_dummies)))
```

A.3

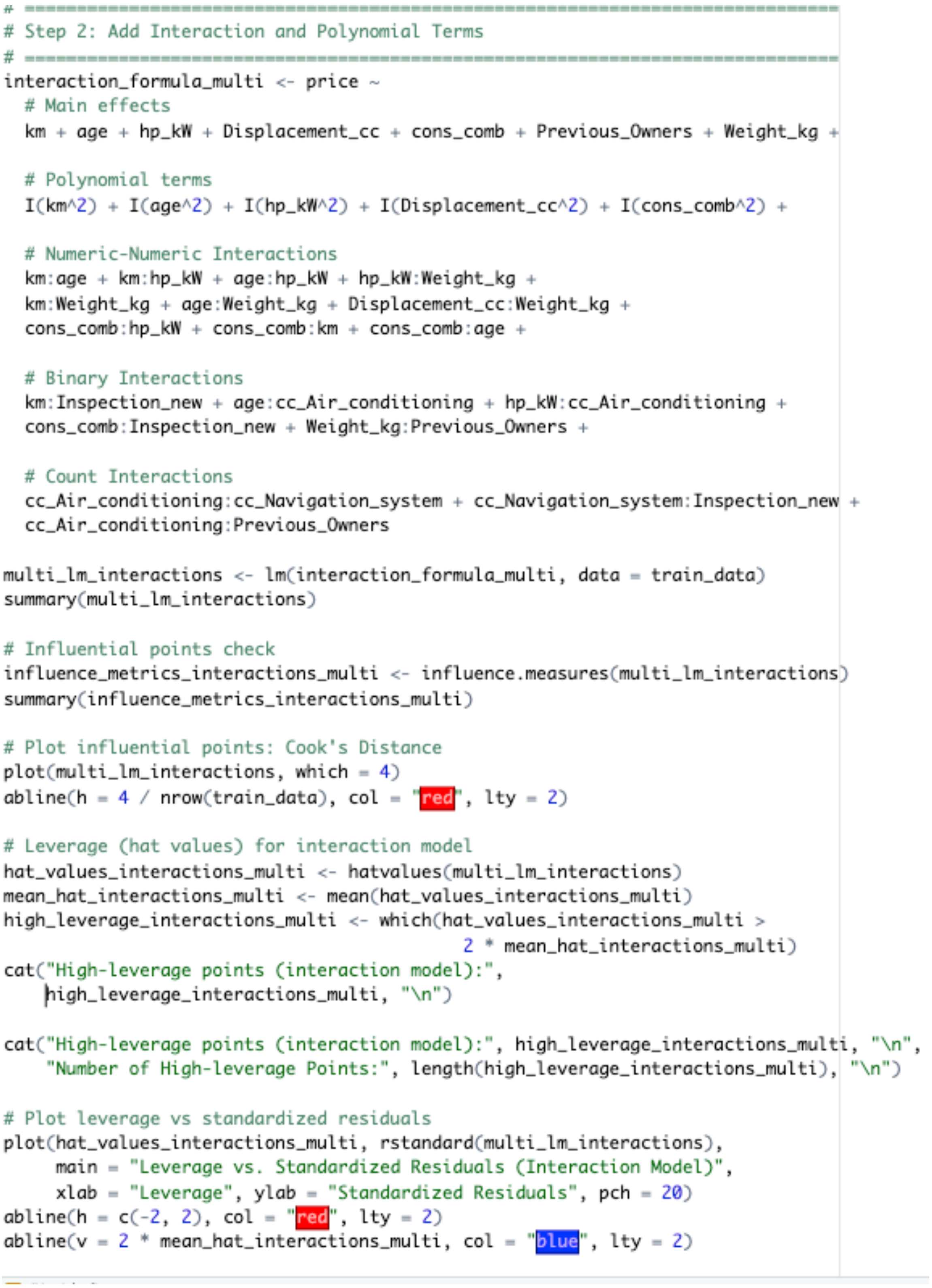


B.1



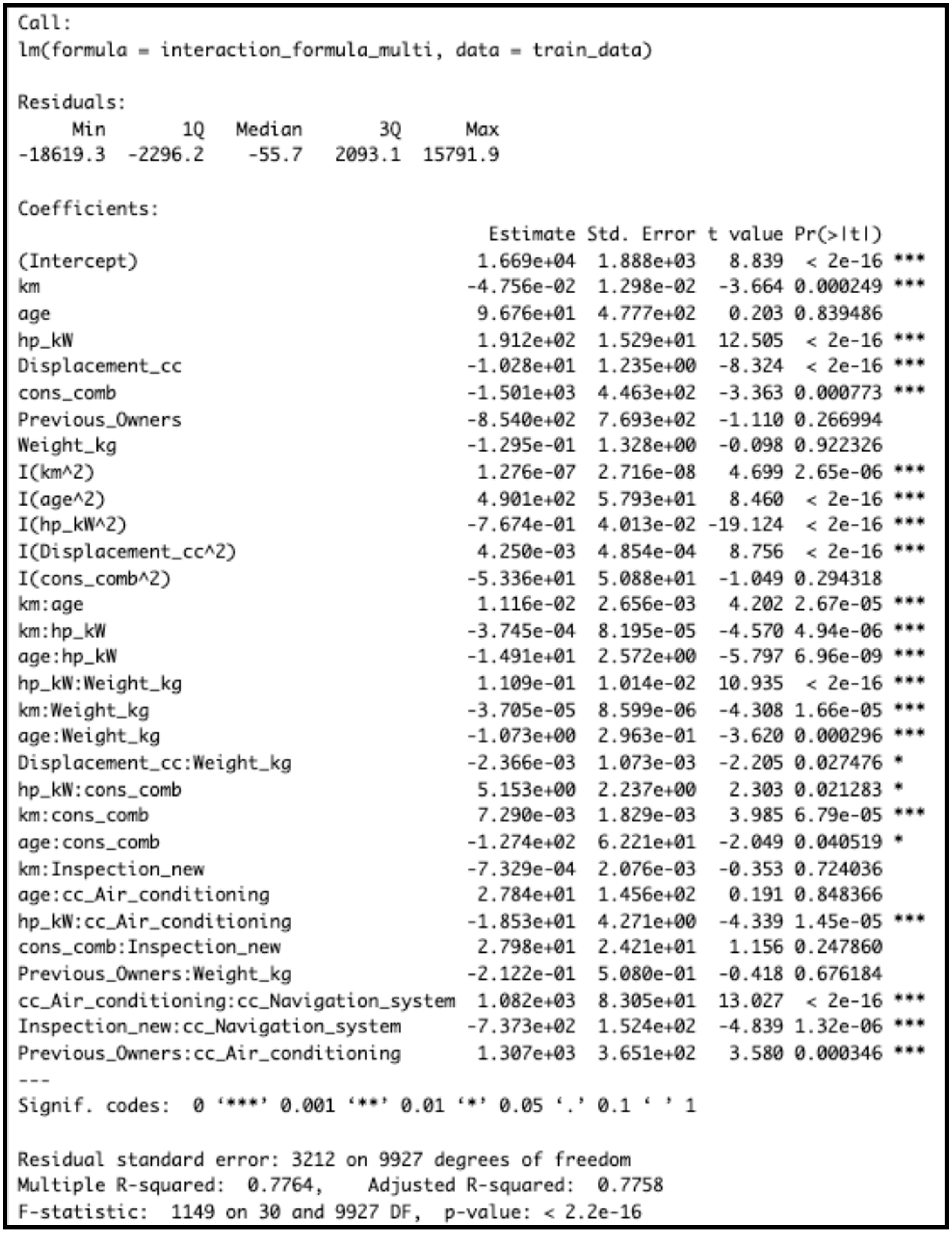
```
  [list output truncated]
> # ============================================================================
> #  Data Splitting:
> # ============================================================================
> # Set a seed for reproducibility
> set.seed(123)
> # Define the proportion for the training set (e.g., 70%)
> train_index <- sample(1:nrow(VehiclePrices_encoded), 0.7 * nrow(VehiclePrices_encoded))
> # Split the data
> train_data <- VehiclePrices_encoded[train_index, ]  # 70% for training
> test_data <- VehiclePrices_encoded[-train_index, ]  # Remaining 30% for testing
> # Check the dimensions of the split datasets
> cat("Training Set Size:", nrow(train_data), "rows and", ncol(train_data), "columns\n")
Training Set Size: 9958 rows and 134 columns
> cat("Testing Set Size:", nrow(test_data), "rows and", ncol(test_data), "columns\n")
Testing Set Size: 4268 rows and 134 columns
> # Update column names in train_data
> colnames(train_data) <- gsub(" ", "_", colnames(train_data))
> colnames(train_data) <- gsub("[^[:alnum:]_]", "", colnames(train_data))  # Removes non-alphanu
meric characters
> # Update column names in test_data
> colnames(test_data) <- gsub(" ", "_", colnames(test_data))
```

B.2



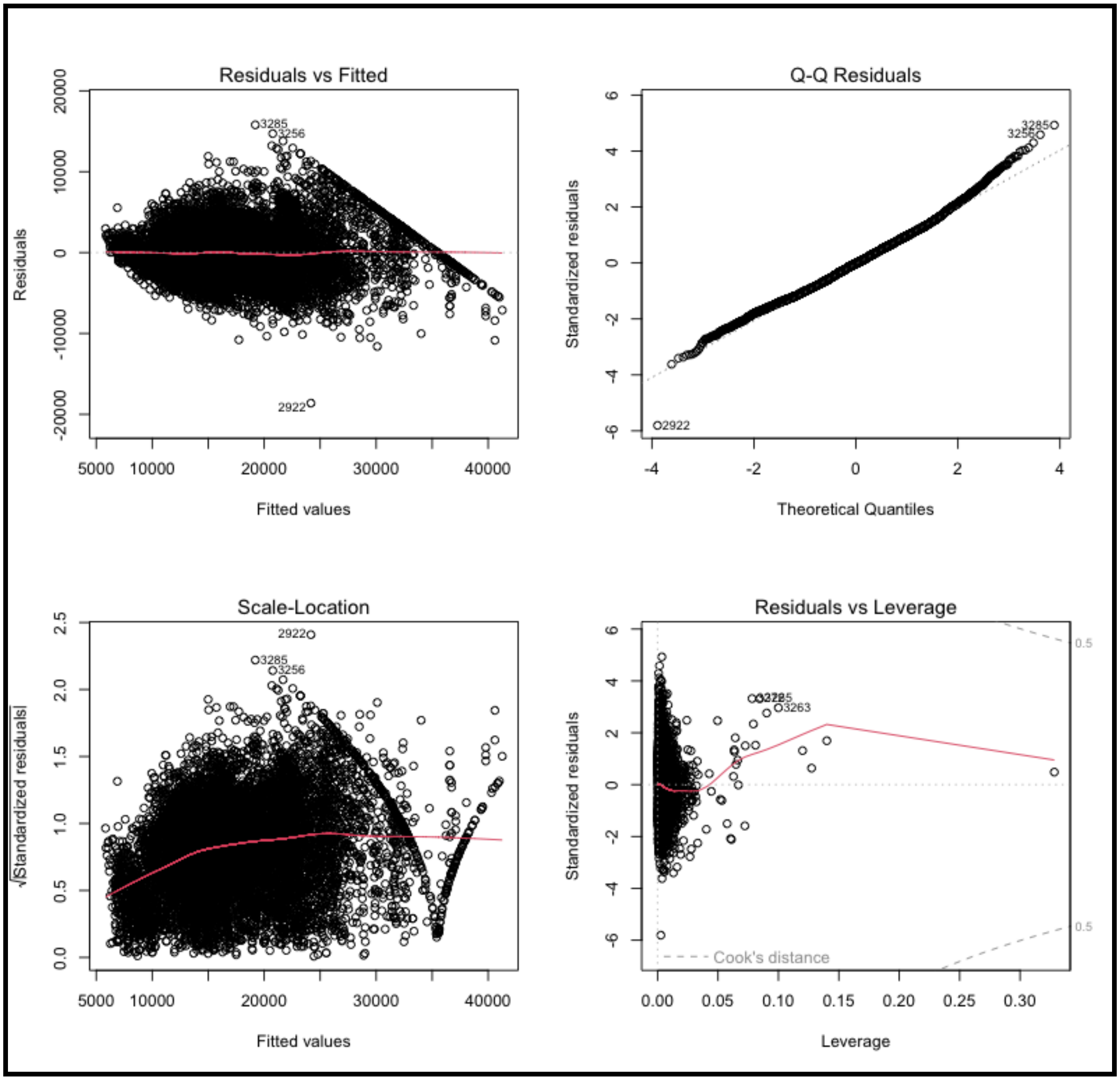
```
# ==============================================================================
# Step 2: Add Interaction and Polynomial Terms
# ==============================================================================
interaction_formula_multi <- price ~
  # Main effects
  km + age + hp_kW + Displacement_cc + cons_comb + Previous_Owners + Weight_kg +

  # Polynomial terms
  I(km^2) + I(age^2) + I(hp_kW^2) + I(Displacement_cc^2) + I(cons_comb^2) +

  # Numeric-Numeric Interactions
  km:age + km:hp_kW + age:hp_kW + hp_kW:Weight_kg +
  km:Weight_kg + age:Weight_kg + Displacement_cc:Weight_kg +
  cons_comb:hp_kW + cons_comb:km + cons_comb:age +

  # Binary Interactions
  km:Inspection_new + age:cc_Air_conditioning + hp_kW:cc_Air_conditioning +
  cons_comb:Inspection_new + Weight_kg:Previous_Owners +

  # Count Interactions
  cc_Air_conditioning:cc_Navigation_system + cc_Navigation_system:Inspection_new +
  cc_Air_conditioning:Previous_Owners

multi_lm_interactions <- lm(interaction_formula_multi, data = train_data)
summary(multi_lm_interactions)

# Influential points check
influence_metrics_interactions_multi <- influence.measures(multi_lm_interactions)
summary(influence_metrics_interactions_multi)

# Plot influential points: Cook's Distance
plot(multi_lm_interactions, which = 4)
abline(h = 4 / nrow(train_data), col = "red", lty = 2)

# Leverage (hat values) for interaction model
hat_values_interactions_multi <- hatvalues(multi_lm_interactions)
mean_hat_interactions_multi <- mean(hat_values_interactions_multi)
high_leverage_interactions_multi <- which(hat_values_interactions_multi >
                                            2 * mean_hat_interactions_multi)
cat("High-leverage points (interaction model):",
    high_leverage_interactions_multi, "\n")

cat("High-leverage points (interaction model):", high_leverage_interactions_multi, "\n",
    "Number of High-leverage Points:", length(high_leverage_interactions_multi), "\n")

# Plot leverage vs standardized residuals
plot(hat_values_interactions_multi, rstandard(multi_lm_interactions),
     main = "Leverage vs. Standardized Residuals (Interaction Model)",
     xlab = "Leverage", ylab = "Standardized Residuals", pch = 20)
abline(h = c(-2, 2), col = "red", lty = 2)
abline(v = 2 * mean_hat_interactions_multi, col = "blue", lty = 2)
```

B.3 Multiple Linear Regression ANOVA table

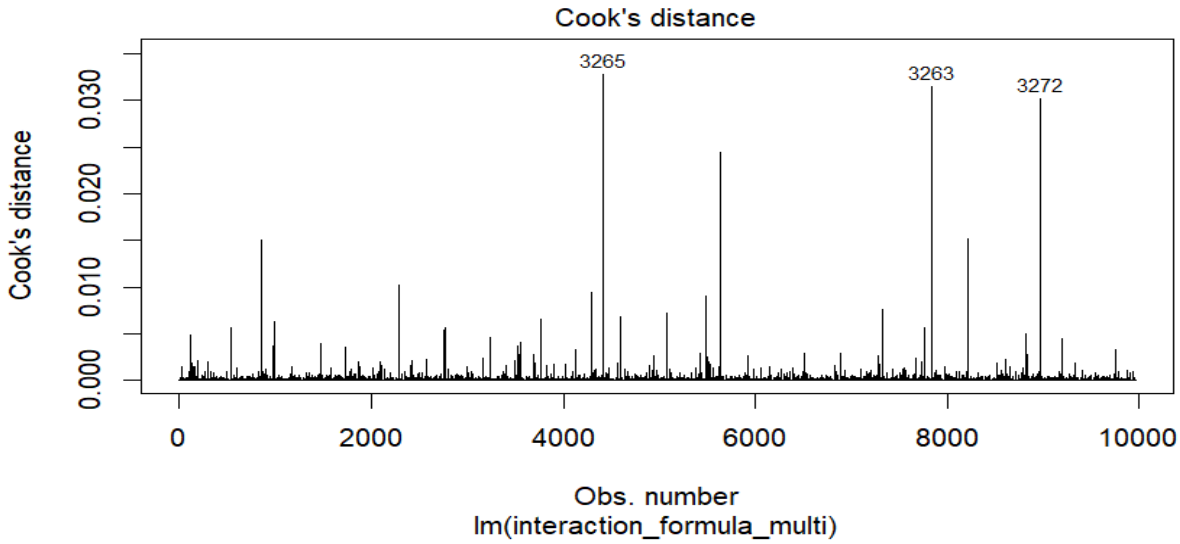


```
Call:
lm(formula = interaction_formula_multi, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-18619.3  -2296.2    -55.7   2093.1  15791.9 

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                              1.669e+04  1.888e+03   8.839  < 2e-16 ***
km                                      -4.756e-02  1.298e-02  -3.664 0.000249 ***
age                                      9.676e+01  4.777e+02   0.203 0.839486    
hp_kW                                    1.912e+02  1.529e+01  12.505  < 2e-16 ***
Displacement_cc                         -1.028e+01  1.235e+00  -8.324  < 2e-16 ***
cons_comb                               -1.501e+03  4.463e+02  -3.363 0.000773 ***
Previous_Owners                         -8.540e+02  7.693e+02  -1.110 0.266994    
Weight_kg                               -1.295e-01  1.328e+00  -0.098 0.922326    
I(km^2)                                  1.276e-07  2.716e-08   4.699 2.65e-06 ***
I(age^2)                                 4.901e+02  5.793e+01   8.460  < 2e-16 ***
I(hp_kW^2)                              -7.674e-01  4.013e-02 -19.124  < 2e-16 ***
I(Displacement_cc^2)                     4.250e-03  4.854e-04   8.756  < 2e-16 ***
I(cons_comb^2)                          -5.336e+01  5.088e+01  -1.049 0.294318    
km:age                                   1.116e-02  2.656e-03   4.202 2.67e-05 ***
km:hp_kW                                -3.745e-04  8.195e-05  -4.570 4.94e-06 ***
age:hp_kW                               -1.491e+01  2.572e+00  -5.797 6.96e-09 ***
hp_kW:Weight_kg                          1.109e-01  1.014e-02  10.935  < 2e-16 ***
km:Weight_kg                            -3.705e-05  8.599e-06  -4.308 1.66e-05 ***
age:Weight_kg                           -1.073e+00  2.963e-01  -3.620 0.000296 ***
Displacement_cc:Weight_kg               -2.366e-03  1.073e-03  -2.205 0.027476 *  
hp_kW:cons_comb                          5.153e+00  2.237e+00   2.303 0.021283 *  
km:cons_comb                             7.290e-03  1.829e-03   3.985 6.79e-05 ***
age:cons_comb                           -1.274e+02  6.221e+01  -2.049 0.040519 *  
km:Inspection_new                       -7.329e-04  2.076e-03  -0.353 0.724036    
age:cc_Air_conditioning                  2.784e+01  1.456e+02   0.191 0.848366    
hp_kW:cc_Air_conditioning               -1.853e+01  4.271e+00  -4.339 1.45e-05 ***
cons_comb:Inspection_new                 2.798e+01  2.421e+01   1.156 0.247860    
Previous_Owners:Weight_kg               -2.122e-01  5.080e-01  -0.418 0.676184    
cc_Air_conditioning:cc_Navigation_system  1.082e+03  8.305e+01  13.027  < 2e-16 ***
Inspection_new:cc_Navigation_system     -7.373e+02  1.524e+02  -4.839 1.32e-06 ***
Previous_Owners:cc_Air_conditioning      1.307e+03  3.651e+02   3.580 0.000346 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3212 on 9927 degrees of freedom
Multiple R-squared:  0.7764,	Adjusted R-squared:  0.7758 
F-statistic:  1149 on 30 and 9927 DF,  p-value: < 2.2e-16
```

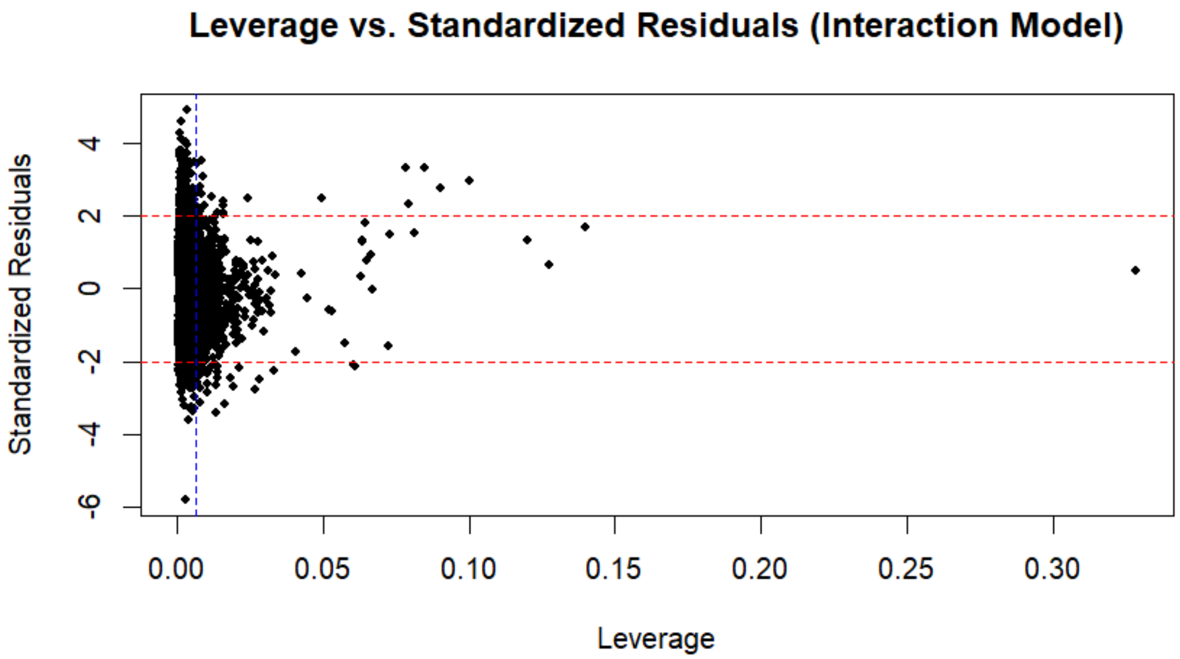
B.4


Residuals vs Fitted
Residuals
Fitted values
20000
10000
0
-10000
-20000
5000
10000
20000
30000
40000
3285
3256
2922
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
6
4
2
0
-2
-4
-6
-4
-2
0
2
4
3285
3256
2922
Scale-Location
√|Standardized residuals|
Fitted values
2.5
2.0
1.5
1.0
0.5
0.0
5000
10000
20000
30000
40000
2922
3285
3256
Residuals vs Leverage
Standardized residuals
Leverage
6
4
2
0
-2
-4
-6
0.00
0.05
0.10
0.15
0.20
0.25
0.30
3285
3263
0.5
0.5
Cook's distance

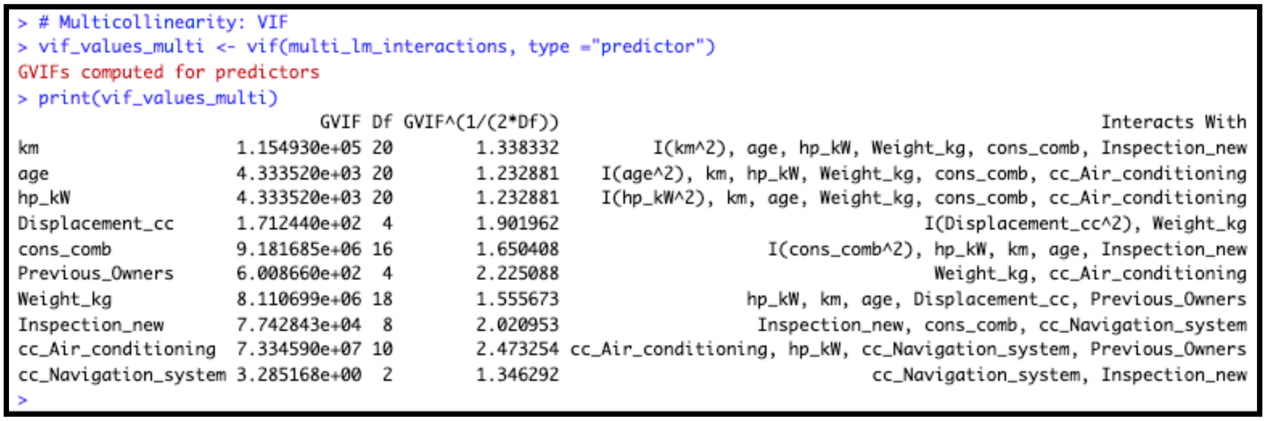
B.5


Cook's distance
Cook's distance
0.000
0.010
0.020
0.030
3265
3263
3272
0
2000
4000
6000
8000
10000
Obs. number
lm(interaction_formula_multi)

B.6

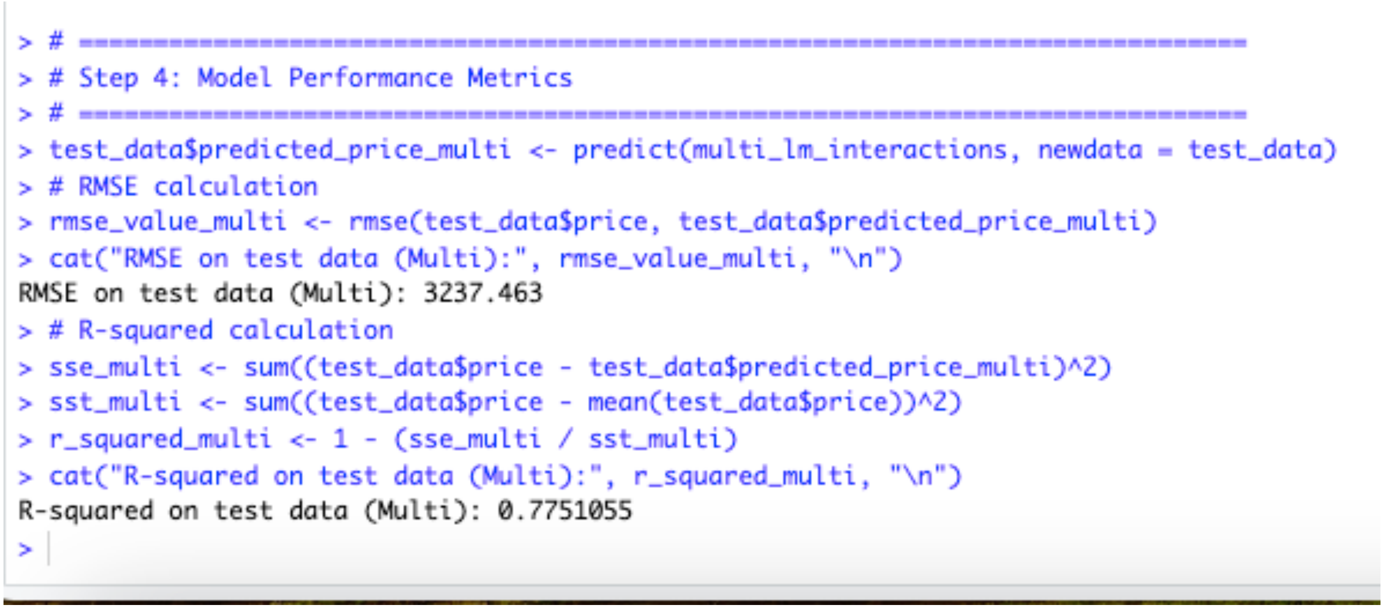

Leverage vs. Standardized Residuals (Interaction Model)
Standardized Residuals
4
2
0
-2
-4
-6
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage

B.7



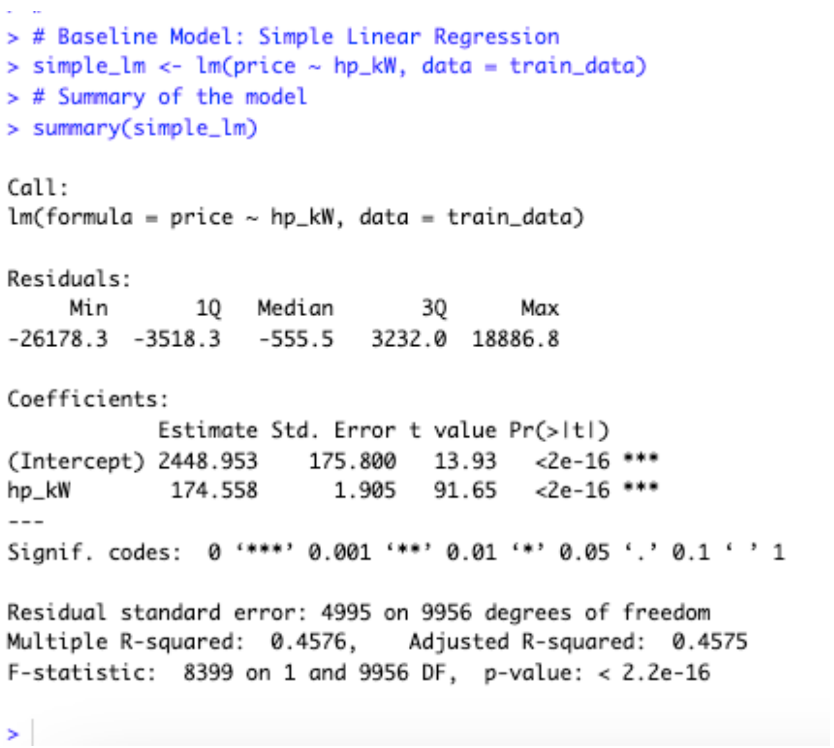
```
> # Multicollinearity: VIF
> vif_values_multi <- vif(multi_lm_interactions, type ="predictor")
GVIFs computed for predictors
> print(vif_values_multi)
                             GVIF Df GVIF^(1/(2*Df))                                                    Interacts With
km                   1.154930e+05 20        1.338332        I(km^2), age, hp_kW, Weight_kg, cons_comb, Inspection_new
age                  4.333520e+03 20        1.232881   I(age^2), km, hp_kW, Weight_kg, cons_comb, cc_Air_conditioning
hp_kW                4.333520e+03 20        1.232881   I(hp_kW^2), km, age, Weight_kg, cons_comb, cc_Air_conditioning
Displacement_cc      1.712440e+02  4        1.901962                                   I(Displacement_cc^2), Weight_kg
cons_comb            9.181685e+06 16        1.650408                    I(cons_comb^2), hp_kW, km, age, Inspection_new
Previous_Owners      6.008660e+02  4        2.225088                                    Weight_kg, cc_Air_conditioning
Weight_kg            8.110699e+06 18        1.555673                  hp_kW, km, age, Displacement_cc, Previous_Owners
Inspection_new       7.742843e+04  8        2.020953                   Inspection_new, cons_comb, cc_Navigation_system
cc_Air_conditioning  7.334590e+07 10        2.473254 cc_Air_conditioning, hp_kW, cc_Navigation_system, Previous_Owners
cc_Navigation_system 3.285168e+00  2        1.346292                              cc_Navigation_system, Inspection_new
>
```

B.8



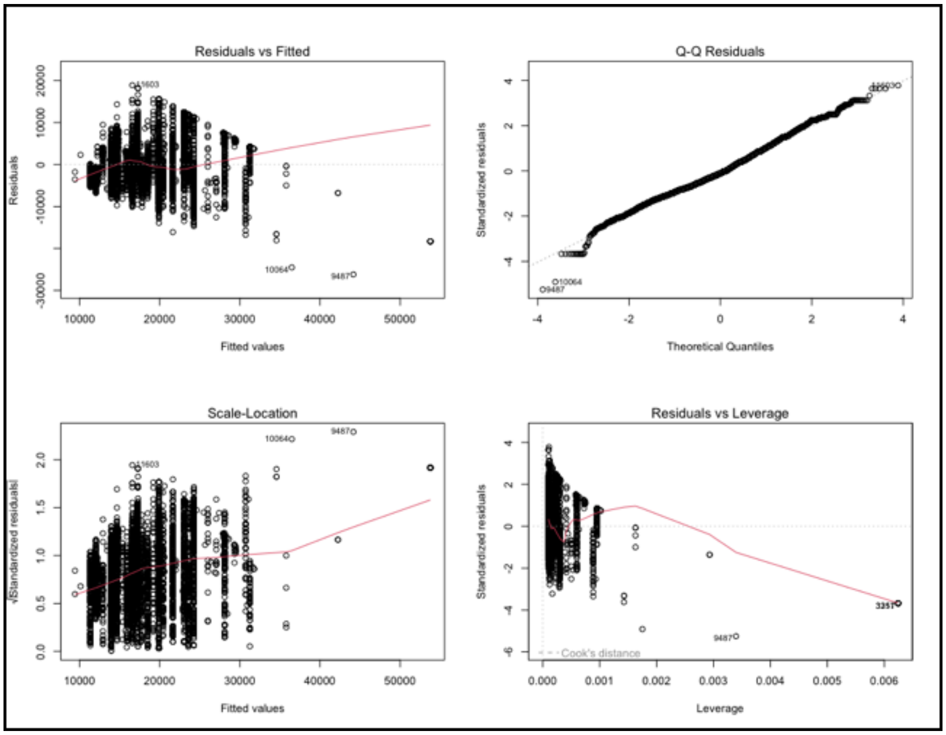
```
> # ==============================================================================
> # Step 4: Model Performance Metrics
> # ==============================================================================
> test_data$predicted_price_multi <- predict(multi_lm_interactions, newdata = test_data)
> # RMSE calculation
> rmse_value_multi <- rmse(test_data$price, test_data$predicted_price_multi)
> cat("RMSE on test data (Multi):", rmse_value_multi, "\n")
RMSE on test data (Multi): 3237.463
> # R-squared calculation
> sse_multi <- sum((test_data$price - test_data$predicted_price_multi)^2)
> sst_multi <- sum((test_data$price - mean(test_data$price))^2)
> r_squared_multi <- 1 - (sse_multi / sst_multi)
> cat("R-squared on test data (Multi):", r_squared_multi, "\n")
R-squared on test data (Multi): 0.7751055
>
```

C.1 Simple Linear Regression ANOVA Table



```
> # Baseline Model: Simple Linear Regression
> simple_lm <- lm(price ~ hp_kW, data = train_data)
> # Summary of the model
> summary(simple_lm)

Call:
lm(formula = price ~ hp_kW, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-26178.3  -3518.3   -555.5   3232.0  18886.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2448.953    175.800   13.93   <2e-16 ***
hp_kW        174.558      1.905   91.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4995 on 9956 degrees of freedom
Multiple R-squared:  0.4576,	Adjusted R-squared:  0.4575 
F-statistic:  8399 on 1 and 9956 DF,  p-value: < 2.2e-16

> 
```

C.2 Diagnostic Plots Simple Linear Regression


Residuals vs Fitted
Residuals
20000
10000
0
-10000
-30000
11603
10064
9487
10000
20000
30000
40000
50000
Fitted values
Q-Q Residuals
Standardized residuals
4
2
0
-2
-4
11603
10064
9487
-4
-2
0
2
4
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
11603
10064
9487
10000
20000
30000
40000
50000
Fitted values
Residuals vs Leverage
Standardized residuals
4
2
0
-2
-4
-6
3351
9487
Cook's distance
0.000
0.001
0.002
0.003
0.004
0.005
0.006
Leverage