The Price of Green

Understanding Eco-Friendly Volvos in Sweden with Linear Regression



Missi Hansson

EC Utbildning

Kunskapsprov - R

202404

[1 Introduction 1](#_Toc165133014)

[2 Theory 2](#_Toc165133015)

[2.1 Data Exploration and EDA 2](#_Toc165133016)

[2.2 Feature Engineering 2](#_Toc165133017)

[2.3 Regression Models 3](#_Toc165133018)

[2.4 Model Selection Techniques 4](#_Toc165133019)

[2.5 Regularization Methods 5](#_Toc165133020)

[2.6 Model Evaluation Metrics 5](#_Toc165133021)

[2.7 Statistical Inference 6](#_Toc165133022)

[3 Method 8](#_Toc165133023)

[3.1 Data Collection 8](#_Toc165133024)

[3.2 Data Cleaning 8](#_Toc165133025)

[3.3 Proof of Concept 8](#_Toc165133026)

[3.4 Exploratory Data Analysis (EDA) 8](#_Toc165133027)

[3.5 Variable Transformation and Feature Engineering 9](#_Toc165133028)

[3.6 Statistical Modeling and Analysis 9](#_Toc165133029)

[3.7 Model Building 10](#_Toc165133030)

[3.8 Regularization and Refinement 11](#_Toc165133031)

[3.9 Model Selection Criteria: 11](#_Toc165133032)

[3.10 API 11](#_Toc165133033)

[4 Results and Discussion 12](#_Toc165133034)

[4.1 Descriptive Statistics 12](#_Toc165133035)

[4.2 Pricing Patterns for Eco-Friendly Vehicles 15](#_Toc165133036)

[4.3 Mileage Influence on Pricing 15](#_Toc165133037)

[4.4 Key Features Affecting Pricing 15](#_Toc165133038)

[4.5 Model Comparisons 16](#_Toc165133039)

[4.6 Model Selection 17](#_Toc165133040)

[4.7 Interpretation of Statistical Findings 18](#_Toc165133041)

[4.8 Limitations and Further Research 18](#_Toc165133042)

[5 Conclusion 20](#_Toc165133043)

[Appendix A 21](#_Toc165133044)

[6 Sources 27](#_Toc165133045)

Abstract

This study examines the impact of vehicle characteristics on the pricing of used Volvo cars in Sweden, with a particular focus on eco-friendly models. Utilizing data from Blocket.se, the analysis covers 753 observations, assessing how features such as fuel type, mileage, gearbox, drivetrain, and model year influence car prices. Through linear regression and variable transformation, supplemented by diagnostic checks to ensure model validity, the study finds that electric and hybrid vehicles command higher initial prices compared to traditional fuel types. However, preliminary results suggest that these eco-friendly vehicles may depreciate more rapidly, a finding that warrants further investigation due to its implications for consumer purchasing decisions and market valuation. This research highlights the evolving consumer preferences towards eco-friendly vehicles in the Swedish used car market and underscores the importance of advanced features in driving vehicle valuation.

Abbreviations

SCB (Statistiska centralbyrån): The Swedish Statistics Office

Blocket.se: A popular online marketplace for used cars in Sweden

El: Electric cars SCB - Swedish Statistics Office

POC - Proof of Concept

EDA - Exploratory Data Analysis

IQR - Interquartile Range

VIF - Variance Inflation Factor

RSS - Residual Sum of Squares

AIC - Akaike Information Criterion

BIC - Bayesian Information Criterion

MSE - Mean Squared Error

RMSE - Root Mean Squared Error

CI - Confidence Interval

# Introduction

Cars are a huge part of our lives, but these days, it feels like everyone's talking about how they impact the environment and what people are looking for in a car. With more and more people thinking about going green, there's a big shift towards electric and hybrid cars. It makes sense, right? We all want to do our part to help the planet, and these new cars seem like a great way to do that. This report dives into the world of Volvo sales in Sweden; specifically, how much things like gas vs electric, mileage, and features affect how much people are willing to pay.

Sweden, a country known for keeping things clean and green, is no surprise to be a leader in eco-friendly car adoption. Volvo, a Swedish car manufacturer renowned for safety and innovation, is also at the forefront of this movement. This report takes a closer look at Volvo sales on Blocket.se, a popular online marketplace for used cars. We'll explore how features and specs influence what people are willing to pay for these cars.

This report bridges the gap between statistical data and real-world market dynamics by examining two key aspects:

Consumer Trends: We'll analyze data from the Swedish Statistics Office (SCB) on newly registered electric and hybrid cars. This will shed light on how environmental considerations might be reshaping vehicle purchasing preferences in Sweden.

Used Car Market Dynamics: Through Blocket.se, a key platform for used car transactions, we'll delve into the intricacies of eco-friendly vehicle availability and pricing. This will provide a comprehensive view of the used car market landscape.

The study is guided by critical questions that aim to unearth the nuances of consumer behavior and market trends:

1. What pricing patterns emerge for eco-friendly vehicles?
2. Does the mileage of an eco-friendly car affect its price more or less compared to gasoline/diesel cars?
3. What are the key features that most significantly affect the pricing of eco-friendly vehicles compared to gasoline/diesel cars?
4. Can a model be built from this data, where 90% or more of the variation in car price is explained?

# Theory

## Data Exploration and EDA

### Handling outliers

Linear regression assumes a straight-line relationship between independent and dependent variables, but outliers, data points significantly deviating from this trend, can distort analysis. They may stem from data errors or omitted predictors. Outliers compromise reliability, skewing the regression line and inflating coefficient standard errors. Detecting outliers via residual plots and studentized residuals, where values above 3 signal potential outliers, is crucial. Handling outliers involves cautious removal for data errors or addressing model limitations by including missing predictors to enhance accuracy. (Gareth James, 2023)

### Assessing leverage points

High leverage points arise when independent variable values are extreme, which can exert undue influence on the regression model's estimates. They are identified via leverage statistics, where a much higher value than average indicates a potential high leverage point. Cook's distance measures the effect of deleting a given observation from the dataset, with larger values indicating greater influence on the model's parameters. It's essential to determine the source of these points, as they may indicate data errors or considerable variations that require model adjustments. Typically, removal is considered only after thorough analysis and validation. The leverage statistic is calculated as: where is the number of observations, is the value of the independent variable for the observation, and is the mean of . (Gareth James, 2023)

Cook's distance is given by:

where is the value of fitted response when all the observations are included. is the value of fitted response, where the fit does not include observation . MSE is the mean squared error. is the number of coefficients in the regression model. (O'Brien, 2023)

### Normality of residuals

Normality of residuals is essential in linear regression to validate hypothesis tests and confidence intervals. Non-normality may arise from a non-linear relationship, unaccounted variables, or outliers. It can weaken statistical inferences, making tests like t-tests less reliable. Identification typically involves Q-Q plots. (jbstatistics, 2013) Addressing non-normality can include transforming data, adding polynomial terms, or using robust regression techniques, and should always begin with a thorough data review to ensure accuracy and validity.

## Feature Engineering

Feature engineering is a crucial step in model development where raw data is transformed or new data variables are created to enhance model accuracy. It involves creating interaction terms, introducing polynomial features, and designing composite indicators to better capture complex relationships. By leveraging domain knowledge or using statistical analysis, feature engineering can uncover significant predictors that improve model performance. However, it's essential to be mindful of pitfalls like multicollinearity when introducing new features. Proper techniques, such as dimensionality reduction or regularization, ensure the model remains interpretable and robust. (EliteDataScience.com, 2022)

### Dealing with multicollinearity

Multicollinearity occurs when predictors in a regression model are correlated, complicating the isolation of individual effects on the outcome. This issue can lead to unreliable estimates and diminished statistical significance. It's detected through a correlation matrix or by a Variance Inflation Factor (VIF) exceeding 5 or 10. To resolve it, redundant variables may be dropped, correlated ones combined, or regularization methods like Ridge or Lasso applied. Addressing multicollinearity clarifies model interpretation and enhances its predictive stability. (Gareth James, 2023)

## Regression Models

### Non-linearity between dependent and independent

Non-linearity indicates a complex relationship between variables, not adequately captured by a straight line in a linear model. This complexity can be observed in visualizing residuals versus fitted values, where non-random patterns like curves or trends suggest inaccuracies in predictions. Se Figure 1 Example of non-linearity taken from page 94. (Gareth James, 2023).Techniques like these plots or diagnostic plots can help identify non-linearity. Solutions involve either transforming the data (e.g., applying a log transformation) or modeling with non-linear equations to capture the underlying dynamics more effectively. (Gareth James, 2023)

### Autocorrelation

Autocorrelation refers to the phenomenon where residuals from a regression model are correlated over time, leading to misleadingly tight confidence intervals and exaggerated statistical significance. This issue is especially typical in time-series data, where natural trends or periodic patterns exist. Patterns in residual plots, such as successive residuals mirroring each other, indicate autocorrelation. It often occurs due to the time-structured nature of data or relatedness in observations, necessitating careful model adjustment to ensure accurate inference. (Gareth James, 2023)

En bild som visar text, karta, diagram, linje

Automatiskt genererad beskrivning

Figure 1 Example of non-linearity taken from page 94. (Gareth James, 2023)

En bild som visar text, diagram, linje, Graf

Automatiskt genererad beskrivning

Figure 2 Example of Heterodcedasticity taken from page 97. (Gareth James, 2023)

### Homoscedasticity

In linear regression, the assumption of homoscedasticity (constant variance across error terms, symbolized as is fundamental for reliable standard errors and credible confidence intervals and hypothesis tests. Heteroscedasticity, indicated by a funnel-shaped spread in residual plots, compromises this assumption with its unequal error variances. To remedy heteroscedasticity, methods like concave transformations of the response variable (e.g., log or square root) or weighted least squares are effectively employed, ensuring model accuracy and inferential integrity. (Gareth James, 2023)

## Model Selection Techniques

### Forward Selection

Forward stepwise selection is a model-building strategy in linear regression, particularly useful when dealing with numerous potential predictors. This iterative approach begins with no predictors and adds them one by one, based on their statistical significance, such as their p-value or contribution to R-squared. The process concludes when no additional predictors significantly enhance the model, or a maximum model size is reached. This method is less computationally intensive, requiring the fitting of only models. While it’s advantageous in practice and feasible even when , forward stepwise selection doesn't guarantee the discovery of the absolute best model and must retain previously chosen variables in subsequent steps, potentially missing the most optimal subset of predictors. The final model is selected based on criteria like the lowest RSS (Residual Sum of Squares), highest R-squared, or through validation techniques like cross-validation. (Gareth James, 2023)

### Best Subset Selection

Best subset selection is a comprehensive approach in linear regression that strives to identify the optimal model by examining all possible combinations of predictors. This method starts with an empty model and systematically evaluates various configurations. It not only considers adding predictors one by, but also evaluates models where one or more predictors are removed from existing combinations. This ensures the final model doesn't contain unnecessary variables that might not contribute significantly. Statistical metrics like residual sum of squares (RSS) or R-squared are used to compare the fit of each model configuration. The most effective model is then chosen based on criteria like the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mallow's Cp, or adjusted R-squared, or through empirical methods like cross-validation.

While best subset selection offers an exhaustive search for the optimal model, it has limitations. The number of models to evaluate grows exponentially, , with the number of predictors, making it computationally demanding, especially for datasets with a large number of potential predictors. Additionally, evaluating a vast number of models increases the chance of overfitting. (Gareth James, 2023)

## Regularization Methods

Regularization techniques are pivotal in statistical modeling to curb overfitting, enhancing the model's ability to generalize to new data by introducing a penalty term to the loss function. This penalty shrinks the coefficients towards zero, simplifying the model by reducing the model's sensitivity to noise in the training data while maintaining its predictive power. This added robustness is crucial in practical applications where data anomalies and variations are common.

### Ridge Regression

Ridge regression improves upon least squares by imposing a penalty on the size of the coefficients, which helps to reduce model complexity and multicollinearity. The regularization in Ridge is achieved by adding the squared magnitude of the coefficient as a penalty term to the loss function:

where is the regularization parameter that controls the strength of the penalty. λ increases, the impact of the shrinkage penalty grows, effectively reducing the coefficients of the regression model towards zero, which decreases model complexity and variance while the increase of bias is just slight. (Gareth James, 2023)

### Lasso Regression

The Lasso (Least Absolute Shrinkage and Selection Operator) modifies the regularization approach by allowing for the automatic selection of variables in addition to shrinking coefficients. Lasso achieves this by incorporating the absolute values of the coefficients into the penalty:

The Lasso is particularly known for its ability to produce sparse models, effectively performing variable selection. As the lambda (λ) parameter controlling the penalty term increases, coefficients of less relevant variables are driven exactly to zero. This effectively removes them from the model, resulting in a simpler model with fewer predictors. It's important to note that there is a trade-off involved. While reducing model complexity and potentially overfitting is beneficial, driving coefficients to zero can also lead to some loss of predictive power. (Gareth James, 2023)

## Model Evaluation Metrics

Selecting the best model involves evaluating metrics that assess both a model's fit to the training data and its ability to perform well on unseen data (generalizability).

### Adjusted R-squared

R-squared is a popular metric that reflects the proportion of variance in the dependent variable explained by the model. However, it has a limitation; it tends to increase with the addition of more predictors, even if they are irrelevant. This can lead to overfitting. Adjusted R-squared addresses this issue by penalizing models with a large number of predictors:

where is the number of observations, is the number of predictors in the model. (Statistics How To, 2024)

### BIC (Bayesian Information Criterion)

The BIC is a metric that balances model fit and complexity. It penalizes models with a large number of parameters (predictors) while rewarding good fit. A lower BIC score indicates a better model:

where is the number of observations, is the number of predictors in the model, represents the vector of all estimated parameters in the model, represents the likelihood function evaluated at the maximum likelihood. The likelihood function measures how well the model fits the data, with higher values indicating a better fit.

The first term, k log(n), penalizes models with a large number of parameters (k); adding more parameters can lead to overfitting. The second term,, rewards models with a good fit to the data. (Statistics How To, 2024)

### RMSE (Root Mean Squared Error)

RMSE measures the average distance of the errors between the predicted and actual values. It's calculated by taking the square root of the mean squared error (MSE):

Here, is the number of observations, is the actual value for the th sample, and is the predicted value for the th sample. The formula includes squaring the errors. This gives more weight to larger prediction errors. A lower RMSE indicates a better fit, as it signifies that the model's average predictions are closer to the actual values. It is also important to note that RMSE UNITS is calculated in the same units as the target variable. This makes it easier to interpret the magnitude of the errors directly. (Kampakis, u.d.)

## Statistical Inference

### Hypothesis Testing

Hypothesis testing in regression analyzes whether the coefficients associated with variables are significantly different from zero, indicating an effect on the dependent variable. This is typically done using t-tests, where each coefficient's t-statistic is compared against a theoretical t-distribution to determine a p-value. The p-value measures the probability of the coefficient having an effect on the dependent variable. A small p-value (typically less than 0.05) suggests rejecting the null hypothesis in favor of the alternative hypothesis, indicating a significant effect. (Wahlgren, 2015)

### Confidence Intervals

In the context of regression analysis, confidence intervals provide a range of values which likely contain the true value of the regression coefficients. These intervals are key to assessing the precision and reliability of the estimates derived from the sample data. Typically set at the 95% confidence level, these intervals are constructed using the coefficient's standard error and indicate where the true parameter value is likely to fall if the experiment were repeated multiple times under the same conditions. A CI that does not encompass zero suggests that the effect is statistically significant, indicating a real association between the predictor and the outcome variable. The width of the CI provides insights into the estimate's reliability; narrow intervals imply high precision, while wide intervals reflect greater uncertainty. These intervals also show the direction of the relationship, with positive values indicating an increase and negative values a decrease in the outcome variable relative to the predictor. It's important to remember that the validity of confidence intervals relies on the underlying assumptions of linear regression being met (ex. linearity, normality of errors). Violations of these assumptions can affect the accuracy of the CIs.

# Method

This section details the practical work undertaken to achieve the report's objectives. The software used for data manipulation and statistical analysis was R.

## Data Collection

Data was collected from the website Blocket.se. The purpose of this data collection was to gather information on Volvo cars for subsequent analysis and research. To ensure a diverse dataset, each group member was assigned to collect data on 50 Volvo cars, targeting different fuel types. Additionally, one member collected data across all fuel types, starting from a few pages later to avoid duplicates. Furthermore, after confirming that the initial dataset volume was adequate for our analysis needs, we collectively decided to gather additional information. Each group member then proceeded to collect data on an additional 100 Volvo cars, adhering to the same criteria as before, which led to 753 observations.

## Data Cleaning

Data cleaning addressed missing values. Entries with missing engine size were removed. Since electric cars lack traditional engines, including these entries would skew the data. Additionally, Categorical variables were transformed into dummy variables, and potential data duplicates were identified and removed using functions like subset(), ifelse(), mutate(), gsub(), and filter().

## Proof of Concept

A simple linear regression model was constructed using the cleaned dataset. This model, serving as a proof of concept (POC), incorporated Mileage and Model Year to predict the sale price of the cars.

## Exploratory Data Analysis (EDA)

### Data Filtering:

Vehicles manufactured before 2000 and after 2024 were excluded to focus on contemporary market trends. Similarly, car types with low occurrence rates, which could introduce bias or variance issues, were also excluded. Functions like tapply() (categorical analysis) and ggplot() (plotting distributions) were used for this purpose.

### Outlier Management:

Outlier detection was performed using boxplot analysis with interquartile range (IQR) thresholds. Entries with values falling outside the IQR (interquartile range) for sales price and mileage were scrutinized. One observation was found to be questionable and dropped. While high mileage for its model year could be attributed to unusual use, this observation deviated significantly from the overall data pattern. There was no way to validate the accuracy of the observation. Consequently, due to the uncertainty surrounding its validity and its substantial deviation from the data, this outlier was removed from the dataset.

### Categorical Data Analysis:

The frequency distributions of categorical variables like fuel type, gearbox, drivetrain, seller, and car type were visualized using bar plots to assess the diversity and prevalence of different categories within the dataset as well as correlation plots of numeric variables. The tapply () function was used to explore the sales prices across different categories, providing insights into price variations associated with different vehicle characteristics.

## Variable Transformation and Feature Engineering

Significant transformations and feature engineering steps were applied to refine the analysis:

* Combining Categories: Sparse categories within the car types and colors were combined to create more generalized and analytically robust groups.
* Creation of Interaction Terms: Interaction terms like Model\_Year \* Fuel and Mileage \* Fuel introduced because of multicollinearity.

## Statistical Modeling and Analysis

Before delving into the construction of multiple regression models several potential problems that could arise during regression modeling were considered:

* Non-linear relationships between the dependent and independent variables.
* Correlated residuals that violate the assumption of independent errors.
* Heteroscedasticity, indicating non-constant variance in the residuals.
* Non-normally distributed residuals.
* Outliers with the potential to distort model estimates.
* High-leverage points that could overly influence the model.
* Collinearity or multicollinearity amongst the predictor variables.

Each of these aspects was carefully considered and checked throughout the modeling process, particularly for the final model. Tools and diagnostic plots within R, such as variance inflation factor (VIF) calculations and residual plots, were utilized to diagnose and rectify these issues, ensuring the robustness of the models.

### Data Partitioning:

The dataset was divided into training and testing sets to evaluate model performance on unseen data and ensure generalizability.

### Diagnostic Checking:

Residual plots and other diagnostics were used to validate linear regression assumptions, ensuring model appropriateness.

### Univariate Analysis:

Multiple linear regression models were created with each potential predictor variable with the target variable to understand the individual effect of each variable on the sales price. Model summaries are analyzed from each individual model to help assess the strength and significance of the relationship between each predictor and the target variable.

1. Seller
2. Fuel
3. Gearbox
4. Mileage
5. Model Year
6. Car Type
7. Drivetrain
8. Horsepower
9. Color
10. Motor Size
11. Model

### Interaction terms were included to explore the combined effects of variables.

1. Model Year and Fuel: Linear regression with an interaction term between Model Year and Fuel.
2. Mileage and Fuel: Linear regression to explore the relationship between Mileage, Fuel, and Sales Price. Additionally, feature engineering for Mileage Depreciation Rate.
3. Mileage and Model Year: Linear regression with an interaction term between Mileage and Model Year.
4. Horsepower and Fuel: Linear regression with an interaction term between Horsepower and Fuel.
5. Fuel Type and Motor Size: Attempted linear regression with an interaction term between Fuel type and Motor size, but found poor model fit due to motor size's qualitative nature.
6. Fuel Type and Seller: Linear regression with an interaction term between Fuel type and Seller.
7. Drivetrain and Gearbox: Linear regression with an interaction term between Drivetrain and Gearbox.
8. Car Type and Color: Linear regression with an interaction term between Car Type and Color.
9. Gearbox and Mileage: Linear regression with an interaction term between Gearbox and Mileage.
10. Fuel Type and Car Type: Linear regression with an interaction term between Fuel type and Car type.
11. Fuel Type and Gearbox: Linear regression with an interaction term between Fuel type and Gearbox.
12. Model Year and Horsepower: Linear regression with an interaction term between Model Year and Horsepower.

## Model Building

Built three models to assess the impact of different feature sets on sales price prediction.

* Based on the understanding gained from the EDA and a little bit of domain knowledge an initial model was built that included variables believed to be most relevant to sales price.
* Forward Selection Model: Utilized the leaps package to automate feature selection. This method iteratively adds features to the model based on their statistical significance, beginning with an empty model and progressively incorporating the most impactful variables until a stopping criterion is met. It aids in identifying a potentially smaller set of features with strong predictive power. (Miller, 2020)
* Best Subset Selection Model: Employed the leaps package to explore all possible combinations of features and select the model with the best performance, as determined by a predefined criterion such as adjusted R-squared. This method can be computationally intensive but may reveal the optimal combination of features for the dataset.

## Regularization and Refinement

After examining the initial models, regularization techniques were used to address potential multicollinearity among features. Regularization can help improve model generalizability and reduce the impact of correlated features.

Ridge regression was applied to intuitive model, forward selection model and best subset selection model. There after Lasso was applied to intuitive model and the best subset selection model.

Based on the insights from regularization and the initial models, we refined the intuitive model by incorporating the most statistically significant features identified through forward selection and backward selection. This refined model aimed to balance interpretability with statistical significance.

## Model Selection Criteria:

To choose the final model, statistical criteria like adjusted R-squared, BIC (Bayesian Information Criterion), RMSE (Root Mean Squared Error), CI (confidence interval) were compared. These criteria consider both model fit and complexity, help with selecting the model that captures the essential relationships while avoiding overfitting.

By following these steps, the final model that best explains the impact of various features on the sales price of Volvo cars in the dataset. The detailed results and interpretations of this model will be presented in the Results section.

## API

The primary data from Blocket.se was complemented with data on new eco-friendly vehicle registrations retrieved via the Swedish Statistics Office's API using the pxweb package in R. (Måns Magnusson, 2022) This integration enabled an expanded analysis of market trends across various fuel types. Standard procedures ensured data compatibility with the main dataset after access, query, and processing. Visualizations generated using ggplot2 illustrated trends over time, providing broader context for the findings.

# Results and Discussion

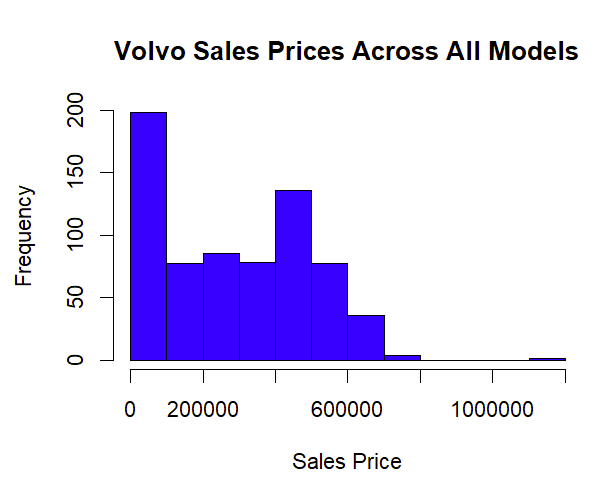


Figure 3 Histogram av Volvo Sales in the dataset.

## Descriptive Statistics

Distribution of Sales Prices

Exploration begins by delving into the characteristics of the used Volvo car listings obtained from Blocket.se. The sales price distribution of used Volvo cars listed on Blocket.se indicates a right-skewed distribution (se Figure 3), with a greater concentration of cars at the lower end of the price spectrum and fewer listings at the higher end. This skewness could be influenced by several factors, including the presence of newer model years, vehicles with lower mileage, or perhaps additional features and trim levels that typically command higher prices. The presence of such newer vehicles can significantly affect the overall price distribution, contributing to the long tail on the right side of the histogram.

In contrast to a normal distribution, which is symmetrical, the observed skewness in the sales price indicates that the median sales price would likely be lower than the mean sales price. If there were multiple peaks in the distribution, it would suggest the presence of distinct groups or segments within the market, such as a clear division between standard and luxury models, or between older and newer cars. However, based on the provided histogram, it appears that there is a single predominant peak, which implies that while there is a range of prices, they are not distinctly segmented within this dataset.

Breakdown by Fuel Type:

The dataset comprises four distinct fuel types: Petrol, Diesel, Electric vehicles, and hybrids. Petrol is represented by 151 observations, Diesel by 158 observations, Electric vehicles by 177 observations, and Hybrids by 170 observations. This representation was ensured through meticulous data collection methods.

Mileage distribution:

Examining mileage across fuel types reveals distinct patterns. Petrol cars has the largest median mileage, as indicated by the boxplot's center line. However, they also display the greatest variation, with their box extending further in both directions. This suggests a wider range of mileage among petrol cars compared to other fuel options. Electric vehicles, on the other hand, show a contrasting picture. Their boxplot exhibits the lowest median mileage, likely reflecting the presence of newer electric models in the data. Additionally, the tighter spread of data points for electric vehicles suggests a narrower range of mileage compared to petrol cars. Hybrids occupy a middle ground. Their median mileage falls between electric and diesel vehicles, and the boxplot indicates a spread that's slightly wider than electric but narrower than petrol cars. Finally, diesel cars exhibit the second-smallest spread in mileage, suggesting less variation compared to hybrids and petrol vehicles. However, it's important to note that there are some outliers for diesel cars that fall outside the interquartile range (IQR), potentially influencing the overall picture for this category. Se **Fel! Hittar inte referenskälla.**

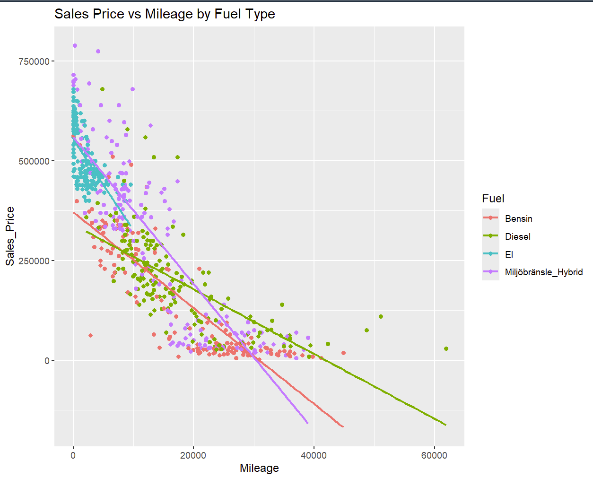


Figure 5 Sales Price vs Mileage by Fuel

Gearbox distribution:

For petrol cars, the chart shows a relatively even distribution between automatic and manual gearboxes In the diesel and hybrid category, automatic gearboxes dominate with a significant majority, while the number of manual gearboxes is less than half the number of automatics. Electric vehicles are exclusively equipped with automatic transmissions; there are no manual electric vehicles represented in this dataset. Overall, the trend indicates a clear preference for automatic transmissions in electric and hybrid vehicles, and to a lesser extent, in diesel vehicles. Petrol cars have the most balanced distribution but still show a slight inclination towards manual transmissions. The complete absence of manual gearboxes in electric vehicles likely reflects the current technological trend and consumer preference for automatics in this category.

Drivetrain distribution:

The analysis reveals distinct trends in drivetrain preference across fuel types. For petrol cars, the distribution heavily favors two-wheel drive over four-wheel drive, with a much higher count for two-wheel drive vehicles. Diesel and electric vehicles exhibit a similar trend, with two-wheel drive being more common than four-wheel drive, although the disparity is less pronounced. Hybrid vehicles, however, show a distinct pattern. They are more commonly found with four-wheel drive than two-wheel drive. This could reflect the desire for more stability and off-road capability in hybrids, or it could be related to the adequacy of two-wheel drive in urban settings for petrol, diesel, and electric vehicles.

Seller distribution:

Company sellers appear to dominate the market on Blocket.se across all fuel types, particularly for electric vehicles. This trend is likely due to electric and hybrid vehicles being a relatively new technology. Buyers might prefer the benefits of purchasing from established dealers, such as warranties and other services, when acquiring these newer types of cars.

Model Year distribution:

The analysis of model years reveals distinct patterns across fuel types. Petrol cars exhibit a relatively even spread across a wide range, with several peaks suggesting a presence of cars from various years. This contrasts with diesel cars, which show a more concentrated distribution around recent model years. This peak in diesel cars indicates that most of them in the dataset are newer models. Hybrid vehicles also follow a trend towards newer models, but their distribution is slightly less concentrated compared to diesel cars, resembling the wider range observed for petrol cars. Electric vehicles stand out with a highly skewed distribution towards the most recent years. This suggests that most electric cars in this dataset are quite new, reflecting the recent surge in electric vehicle production and adoption. This trend likely aligns with the growing market for eco-friendly vehicles in recent years.

Car Type distribution:

Petrol cars are dominated by Hatchbacks and Station Wagons, with Hatchbacks taking the lead. Sedans and SUVs are less frequent in this category. For diesel vehicles, SUVs reign supreme, likely due to the performance demands of diesel engines in larger bodies. Hatchbacks, Station Wagons, and Sedans are present but dwindle in numbers, with Sedans being the least common. Electric vehicles overwhelmingly favor SUVs, with a very high representation. Other car types are scarce or absent, potentially reflecting a consumer preference for larger, more versatile electric vehicles. Hybrids primarily come in Station Wagon form, with a moderate number of SUVs. Hatchbacks and Sedans have an even smaller presence, perhaps due to the balance between space and eco-friendliness that Station Wagons offer.

Color distribution:

The distribution of car colors across fuel types appears scattered. This suggests that there might not be a statistically significant difference in color preference based on fuel type. Silver, black, and white remain popular choices across all categories, with white appearing particularly common for electric vehicles.

Model distribution:

Examining the distribution of car models across fuel types reveals interesting trends. Petrol cars boast a diverse range of models, with the V70 model standing out as the most popular. This suggests a wider variety of consumer preferences within the petrol category. In contrast, diesel car models exhibit a more balanced distribution, with no single model dominating. Electric and hybrid vehicles, on the other hand, display a distinct preference for a limited number of models. Electric cars show a noticeably skewed distribution, with only five distinct models present. Three of these models (C40, EX30, and XC40) hold a significantly higher count than the remaining two. Hybrids exhibit a similar trend, but with less pronounced dominance of the top three models.

Horsepower distribution:

The boxplot reveals distinct patterns in horsepower distribution across fuel types. Hybrid cars generally display the highest median horsepower, accompanied by a wider range of values as indicated by the larger size of their box. In contrast, both diesel and petrol cars exhibit similar median horsepower levels. Electric vehicles also display a wider spread, but their median horsepower falls on the lower end of the distribution. This suggests a potential association between hybrid and electric technologies and higher horsepower capabilities, while diesel and petrol cars tend to cluster around a more moderate horsepower range.

Motor size distribution:

Petrol cars have a wide range of motor sizes, as shown by the box's spread. Electric cars don’t have a motor size because they do not have a traditional combustion engine. Diesel and Hybrid have similar spreads and motor sizes. The median of petrol, diesel and hybrid are similar. Diesel and hybrid vehicles might prioritize fuel efficiency or have a more concentrated range of engine sizes. Electric cars represent a distinct technology with a focus on electric motor power and battery capacity.

Correlation Analysis:

The correlation analysis of Volvo used car data reveals several noteworthy relationships between numeric variables. Sales price and model year have a strong positive correlation (0.90), which is expected as newer models typically command higher prices. Mileage shows a strong negative correlation with both sales price (-0.86) and model year (-0.86), confirming the trend that higher mileage vehicles tend to be older and less valuable.

Horsepower has a strong positive correlation with sales price (0.75), indicating that vehicles with higher power are priced higher, likely due to performance or luxury features associated with more powerful engines. There is a moderate negative correlation between horsepower and motor size (-0.41), suggesting that as engines become more efficient.

## Pricing Patterns for Eco-Friendly Vehicles

Electric and hybrid cars are noticeably more expensive than traditional gasoline and diesel vehicles. Electric vehicles have the highest regression coefficient, indicating that they are priced, on average, 359,719 SEK higher than gasoline cars. This significant premium is supported by a p-value less than 0.0001 and a confidence interval of 327 712 SEK to 391 726 SEK (95% confidence level). While the confidence interval is wider, it excludes zero, suggesting a strong likelihood that electric vehicles cost more than gasoline cars on average.

Hybrid vehicles also show a higher price point compared to gasoline cars, with an additional cost of 153 139 SEK. This premium is statistically significant, as indicated by a p-value less than 0.0001. However, there is also some uncertainty associated with this estimate. Based on the confidence interval, hybrid vehicles are likely priced between 120 830 SEK and 185 448 SEK higher than gasoline cars, on average (95% confidence interval).

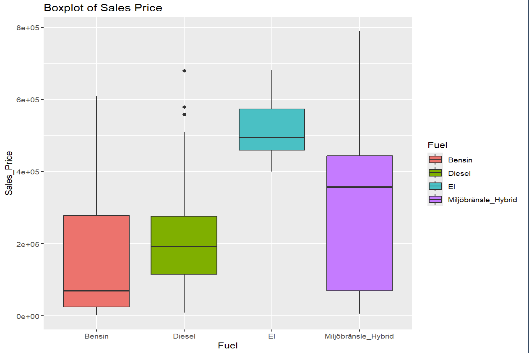


Figure 4 Boxplot of Sales Price across Fuel type

On the other hand, diesel vehicles demonstrate a smaller increase of 46 295 SEK over gasoline cars, baseline. Although this increase is less pronounced than that for electrical cars and hybrids, it remains statistically significant (p-value = 0.00586).

## Mileage Influence on Pricing

Mileage significantly affects used car prices, differing by fuel type. According to this regression models, a unit increase in mileage consistently lowers the vehicle's price. For traditional petrol vehicles, each additional kilometer traveled results in a 16 SEK decrease in the sale price. The effect of mileage on eco-friendly vehicles is more pronounced; the model shows that electric and hybrid vehicles experience an even greater decrease in value per additional kilometer compared to petrol cars. The interaction terms indicate that for electric cars, each additional kilometer results in a price decrease of approximately 9.5 SEK more than petrol cars, while hybrid vehicles show an extra 6.15 SEK decrease. This suggests that eco-friendly vehicles may suffer from faster depreciation due to mileage, challenging the notion that such vehicles inherently maintain their value over time. The adjusted R-squared of 0.8155 indicates a strong fit of the model, explaining over 81% of the variability in sales prices based on mileage and fuel type, with the interaction between these variables being statistically significant (p < 2e-16 for both electric and hybrid vehicles). These nuanced findings shed light on the depreciation dynamics within the used car market for eco-friendly models and emphasize the importance of considering both engineered metrics and real-world data in valuation models.

However, the interpretation of these effects is complicated by the high multicollinearity detected among the variables, particularly between mileage and the fuel types as well as their interaction terms. The Variance Inflation Factors (VIF) for these terms are notably high, with VIF for mileage at approximately 4.52, for fuel types collectively at 31.25, and for the interactions of mileage with fuel types at 34.52. These high VIF values suggest that the predictors are not independent of each other, potentially inflating the standard errors of the regression coefficients and affecting the stability and reliability of the model's predictions.

To mitigate these issues, further steps should be considered. Refining the model by applying regularization techniques such as Ridge or Lasso regression may help reduce the impact of multicollinearity by penalizing the coefficients of correlated predictors, thus enhancing the model's predictive accuracy and interpretability.

## Key Features Affecting Pricing

Beyond mileage, the model year emerges as a prominent factor influencing vehicle pricing.

Model Year: The statistical analysis strongly indicates that the model year holds significant sway over the pricing of vehicles, highlighted by an adjusted R-squared value of 0.7964. This measure reveals that nearly 80% of the variability in car prices can be explained by the year the car was made. Each subsequent model year is associated with an average increase in the sales price of approximately 26 950 SEK, as evidenced by the model's coefficient and its near-zero p-value (< 2e-16). Additionally, the confidence interval for the model year coefficient, ranging from 25 781.74 SEK to 28 144.43 SEK, demonstrates a high degree of precision in this estimate. The diagnostic plots for the regression, including the Q-Q plot and the Residuals vs. Leverage plot, suggest that while model year is a strong predictor, the assumption of homoscedasticity is not met, which could indicate a need for model refinement or the use of alternative estimation techniques to strengthen these findings.

Horsepower: The regression analysis confirms horsepower as a significant factor in car pricing, with an adjusted R-squared of 0.5699. Specifically, each additional unit of horsepower is associated with an average price increase of approximately 1 613.88 SEK, a statistically significant relationship (p < 2e-16). The model's precision is reflected in the tight confidence interval for this estimate, stretching from 1,493 76 SEK to 1 733.99 SEK. Despite a few outliers, the diagnostic plots suggest a reliable model fit.

Car type: Car type emerges as a notable determinant of car prices, particularly for SUVs, which significantly outpace other categories in value. The regression analysis indicates that SUVs are priced on average 317 964 SEK higher than the base category, with the adjusted R-squared at 0.4815, pointing to a considerable portion of price variability being explained by car type. The confidence intervals bolster this finding, especially for SUVs (245 210.50 SEK to 390 718.20 SEK), suggesting a strong association between vehicle type and pricing. Diagnostic plots reveal a relatively even spread of residuals, although some potential outliers indicate the model might benefit from further refinement.

Fuel Type: Fuel type is a significant determinant in the pricing of used cars. Electric vehicles stand out with a substantial higher price tag, as indicated by an additional 356 469 SEK on average compared to petrol cars, and hybrids also have a significant higher price of 151 035 SEK. Diesel vehicles exhibit a smaller yet significant increase in price over gasoline cars by 46 196 SEK. The adjusted R-squared value for the model is 0.4561, which means that around 45% of the variability in sales price can be attributed to fuel type. The confidence intervals provide precise estimates, with electric vehicles showing the tightest range, emphasizing the reliability of the higher price associated with electric cars. Diagnostic plots, including the Residuals vs Fitted and Scale-Location plots, suggest that while the model captures the general trend, there might be heteroscedasticity or non-linearity that could be addressed with further modeling or transformation of variables.

Other Factors: Gearbox, drivetrain, color, motor size, and the specific model also impact pricing but to a lesser extent compared to the above factors. It is noteworthy that manual gearboxes and private sellers tend to fetch lower prices, indicating consumer preferences.

## Model Comparisons

Model 1, intuitive model, has a high adjusted R-squared value of 0.9571742, which suggests the model can explain over 95% of the variance within the sales prices. The Root Mean Square Error (RMSE) stands at 47 810.85 and coupled with a Bayesian Information Criterion (BIC) of 5542.095, the model balances complexity with fit effectively. Additionally, the narrow confidence intervals and the statistical significance of the coefficients further reinforce the model’s reliability.

For Model 2, Forward Selection, the root mean squared error (RMSE) is higher at 73 876.33, indicating less accuracy compared to Model 1. The adjusted R-squared is lower at 0.8580664, suggesting that the model explains less variability in the sales price than Model 1. Its Bayesian Information Criterion (BIC) is significantly higher at 13 393.564, indicating a less favorable balance between the goodness of fit and the complexity of the model and most likely overfitting. Similar for Best Subset Selection, Model 3 exhibits a marginal improvement in the adjusted R-squared to 0.8599266 but retains a high BIC of 13 397.108. The RMSE is 73,789.28, slightly lower than Model 2, but both models inferior.

Ridge Model 1 has an adjusted R-squared of 0.9116305, indicating a strong explanatory capability for the variance in sales prices. The RMSE of 55 370.41, while higher than that of Model 1, is still reasonable given the complexity of the data. The relatively low BIC of 3262.135 suggests that despite regularization, the model avoids overfitting, striking a balance between fit and complexity.

Ridge Model 2 and Ridge Model 3, while utilizing the same regularization technique, do not perform as well. Both models exhibit higher RMSEs (76 837.41 and 75 815.14, respectively) and lower adjusted R-squared values (0.8355465 and 0.8371564, respectively) compared to Ridge Model 1 or Model 1. Their BIC values, while low, are not enough to compensate for the loss in explanatory power, making them less desirable.

Lasso Model 1 closely rivals Ridge Model 1 in terms of statistical metrics, achieving an RMSE of 55 270.17 and an adjusted R-squared of 0.9141887, slightly better than Ridge Model 1. It offers an excellent BIC of 3261.664, the lowest among the models discussed, indicating a model that is well-tuned for prediction while avoiding unnecessary complexity.

Lasso Model 3, however, does not fare as well, with an RMSE of 74 301.12 and an adjusted R-squared of 0.8435954. Although its BIC is competitive at 3338.596, it does not provide enough justification to choose it over Lasso Model 1.

## Model Selection

Model 1 was chosen for its exceptional ability to explain the variability in the sales prices of used Volvo cars. It achieved an adjusted R-squared value of 0.9600, indicating that 96% of the variance within the sales prices is accounted for by the model. This high level of explanatory power, combined with a relatively low RMSE of 47,810.85, suggests that the model accurately predicts prices with minimal error. Furthermore, the Bayesian Information Criterion (BIC) value of 5510.238 reflects a favorable balance between model complexity and goodness of fit, supporting the model's selection over others which showed higher BIC values, indicating potential overfitting

Non-linearity: The "Residuals vs Fitted" plot shows no obvious patterns or systematic structures, suggesting that the relationship between the predictors and the response variable is adequately captured by the model.

Correlated Residuals: There seems to be randomness in the residuals as there is no apparent pattern, which is a good indication that residuals are not correlated.

Non-constant Variance (Heteroscedasticity): The "Scale-Location" plot shows that residuals are spread equally along the ranges of predictors. This is a good sign, as it suggests homoscedasticity, although there is a slight fan shape, indicating potential mild heteroscedasticity.

Non-normally Distributed Residuals: The "Q-Q Plot" reveals that the residuals follow a reasonably straight line, which suggests that the residuals are normally distributed, although some deviation exists at the tails.

Outliers: Outliers were considered during the EDA. Those that are left do not seem to have a strong influence.

High Leverage Points: The "Residuals vs Leverage" plot did show one points really close to the Cook's distance threshold, which would suggest the presence of high leverage for this observation. A closer look at this observation showed that it could be a problem from data collection. The price was exceedingly low as was the mileage for a newer model. The observation could not be verified, and the decision was made to remove it leaving no other observations unduly influencing the model.

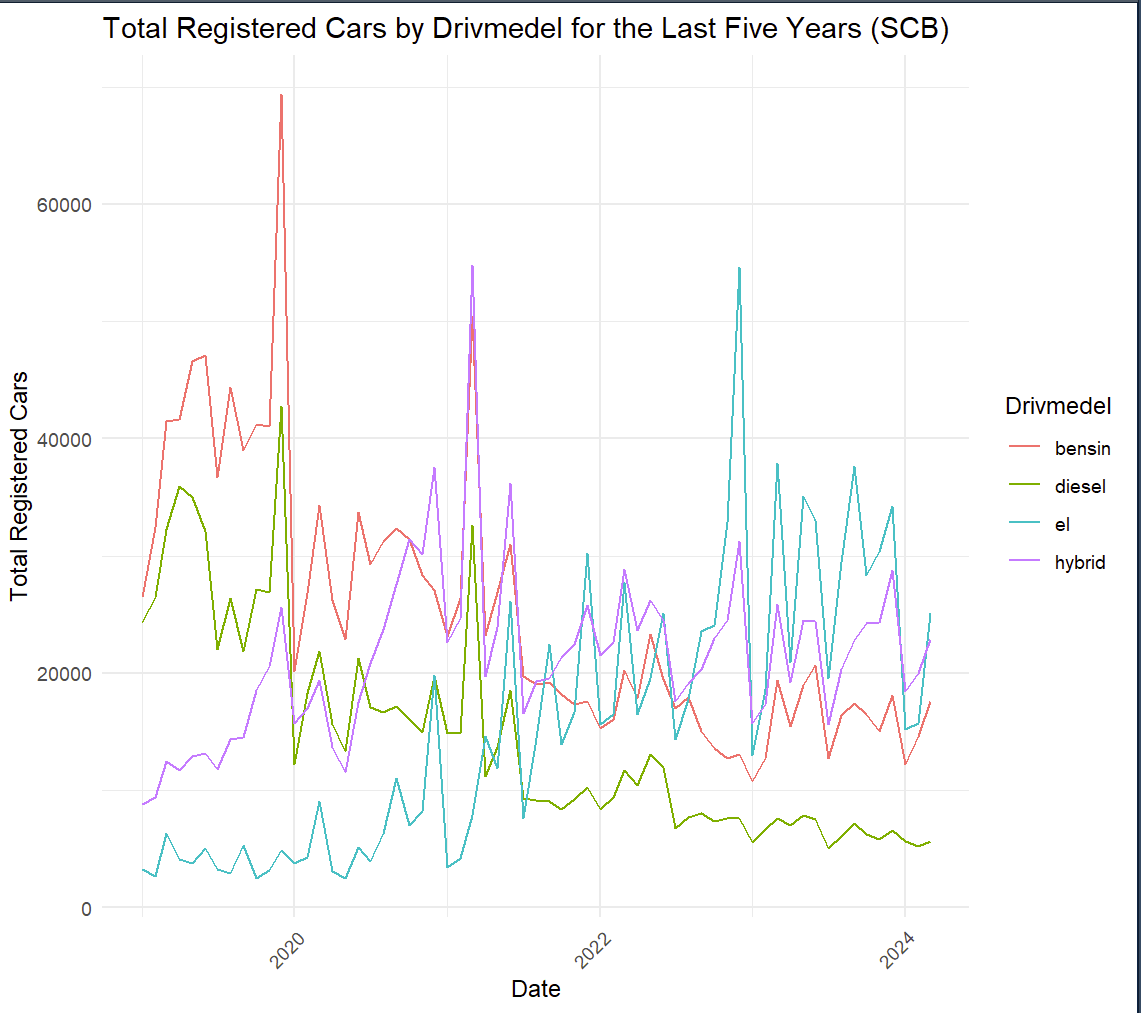
Multicollinearity: To asses multicollinearity Variance Inflation Factors( VIF) was calculated. There does not seem to be a concerning level of multicollinearity all results were under 5.

Given these diagnostics, Model 1 is well-specified, providing a reliable fit to the data without being overly influenced by outliers or high leverage points, making it suitable for both inference and prediction.

## Interpretation of Statistical Findings

The comprehensive analysis conducted via Model 1 explains the factors impacting Volvo cars in Sweden. Notably, SUVs emerge as a standout category, reflecting the increasing consumer inclination towards larger, feature-rich vehicles that harmonize luxury with sustainability. This reflects a significant shift from the traditional notion that eco-friendliness is the domain of compact cars. Instead, advancements in battery technology and energy efficiency are elevating the desirability and feasibility of larger electric and hybrid vehicles.

In this landscape, horsepower retains its relevance, symbolizing the harmonization of performance with eco-consciousness. The data illustrates that consumers expect robust performance, evident in the demand for high-horsepower eco-friendly vehicles, particularly SUVs. These vehicles are not only competing with but also surpassing their conventional counterpart’s petrol/diesel in power and appeal.



A 1 Nu registered cars by fuel the last 5 years. (Statistikmyndigheten SCB, 2024)

The findings from Model 1, with its robust adjusted R-squared value, affirm the growing influence of eco-friendly features alongside traditional attributes like horsepower in dictating car prices. Consumers appear willing to invest in sustainable mobility without abandoning the pleasure of driving, a sentiment echoed by the rise in electric vehicle registrations highlighted by SCB statistics. There is clear evidence of the growing trend the last five years, but particularly the year 2023 stands out as a turning point, signaling a transformative phase in the automotive industry. This phase is characterized by a shift in consumer purchasing decisions, where focus is now on cars that are both eco-friendly and still perform well and feel luxurious. (Statistikmyndigheten SCB, 2024)

However, the analysis also reveals that eco-friendly vehicles, despite their higher initial pricing, may depreciate more rapidly than traditional vehicles. This unexpected finding suggests that factors such as battery degradation, the rapid pace of technological advancements, and perhaps consumer apprehensions about older technology could contribute to faster depreciation rates. The model shows that electric and hybrid vehicles experience a more pronounced decrease in value per additional kilometer compared to petrol cars. These results challenge the notion that eco-friendly vehicles inherently maintain their value over time and suggest that their depreciation dynamics might be influenced by unique factors not fully captured in this study.

Given these insights, there is a clear need for further research to explore why eco-friendly vehicles depreciate faster. This should include a broader dataset and more granular analysis to isolate the effects of technological shifts, market perceptions, and operational longevity on vehicle resale values. Such studies are essential to fully understand the economic and environmental implications of adopting eco-friendly vehicles and to guide consumers, manufacturers, and policymakers in making informed decisions.

## Limitations and Further Research

Limitations:

The limitations of this study are critical to acknowledge for a comprehensive understanding of its scope and the subsequent generalizability of its findings:

Sample Size and Selection: The dataset's starting point was relatively small, with only 750 entries. This limited sample size could affect the robustness of the conclusions drawn, as larger datasets typically provide more reliable insights due to the greater diversity and number of observations.

Single Source of Data: Data was sourced exclusively from Blocket.se, a popular platform in Sweden. While this provides valuable insights, it represents only a portion of the used car market. This limits the generalizability of the findings to the specific demographics and market dynamics of Blocket.se users. Broader trends, such as the overall depreciation rate of electric vehicles compared to petrol cars, might still be evident. However, more nuanced trends, like how price differences vary by specific car model or region within Sweden, might be obscured and require data from additional sources.

Brand Specificity: While the analysis provided valuable insights into the Swedish market by focusing on Volvo, a major player in eco-friendly cars, it's important to acknowledge the limitations of this approach. The entire automotive industry encompasses a wider range of brands with distinct pricing strategies and target audiences. To achieve a more comprehensive understanding of environmentally friendly vehicles, a broader market analysis is necessary.

Condition of the Cars: The dataset did not account for the condition of the cars, such as wear and tear or maintenance history, which can significantly influence used car prices. The omission of these variables might lead to an incomplete understanding of the factors affecting car valuation.

Bias Towards Newer Electric Cars: Given the relative novelty of electric vehicles compared to traditional cars, most of the electric cars in the dataset are likely newer. This could skew the data towards higher valuations for electric cars, as newer cars generally command higher prices regardless of their environmental credentials.

Model Year Limitations: The analysis places significant emphasis on the model year, which may not always be an accurate indicator of a car's technological relevance or its environmental impact, especially in cases where older models have been well-maintained or upgraded.

Market Dynamics: The study does not account for broader economic factors or policy changes that could influence consumer behavior and car valuations, such as fuel prices, taxation policies related to vehicle emissions, or government incentives for buying eco-friendly cars.

By acknowledging these limitations, the study reinforces the need for cautious interpretation of its findings and highlights areas for future research to build upon the groundwork laid here.

Further Research:

In the quest for a deeper understanding of the eco-friendly car market and to extend the findings of the current study, several next steps are proposed:

Expansion of Data Sources: Broadening the dataset to include listings from multiple online marketplaces and dealerships across Sweden will provide a more comprehensive view of the market. This expanded dataset would enhance the generalizability of the findings and potentially unveil nuanced trends that are not apparent from Blocket.se alone.

Longitudinal Study: Implementing a longitudinal study would allow for the observation of trends over time within the eco-friendly vehicle market. Capturing the evolution of consumer preferences and pricing trends over a more extended period could clarify the rate at which newer eco-friendly models are influencing the market. It would also serve to better understand the long-term economic and environmental impacts of this shift towards sustainable mobility.

Consumer Behavior Analysis: Investigating the purchasing patterns and preferences of consumers through surveys or interviews could provide insights into the motivations behind the shift towards eco-friendly vehicles. Understanding the factors that influence consumer decisions, beyond the statistical data, could reveal the psychological, economic, and social drivers that are shaping the market trends observed in the SCB statistics and this study's findings.

# Conclusion

This study provides a detailed analysis of the factors influencing the pricing of used Volvo cars in Sweden, with a specific focus on eco-friendly vehicles. The findings are structured around the research questions posed in the introduction, ensuring a direct response to each inquiry:

1. Pricing Patterns for Eco-Friendly Vehicles: The analysis conclusively shows that eco-friendly Volvo vehicles, specifically electric and hybrid models, command significant price premiums over their petrol and diesel counterparts. Electric vehicles are notably priced, on average, 359 719 SEK higher than petrol cars, affirming a strong market trend that increasingly values sustainability alongside automotive performance.
2. Impact of Mileage on Pricing Compared to Petrol/Diesel Cars: Statistical examination reveals an interesting picture. Specifically, the regression models with interaction terms demonstrate that electric and hybrid vehicles depreciate more rapidly with increased mileage than their petrol/diesel counterparts. This is evidenced by significant negative coefficients for mileage interactions with electric and hybrid fuel types, indicating a steeper decline in value as mileage goes up. Given these insights, it is imperative to conduct further research with a more extensive dataset that considers a broader spectrum of vehicle ages, technological advancements, and battery conditions to accurately assess depreciation patterns.
3. Key Features Affecting Pricing: The study identifies several critical features that significantly affect vehicle pricing, including model year, car type, mileage, and horsepower. Newer models and SUVs, particularly those with electric or hybrid configurations, attract higher prices, reflecting consumer preferences for modern features and larger, versatile vehicle designs. The statistical analysis underscores the importance of these features, with newer model years and SUVs showing a strong positive correlation with higher market prices.
4. Predictive Modeling: The predictive model developed through this research effectively captures over 95% of the variance in car prices, demonstrating its robustness in accounting for key pricing determinants such as fuel type, model year, horsepower, and car type. This high explanatory power validates the model's utility in understanding and predicting used car pricing dynamics.

The results from this study not only reinforce the market demand for Volvos eco-friendly vehicles but also highlight the significant influence of mileage, vehicle type, and technological advancements on car pricing. These insights are pivotal for stakeholders in the automotive industry, including consumers, manufacturers, and policymakers, as they navigate the evolving landscape of vehicle consumption. Future research should expand the scope of data collection, explore longitudinal trends, and consider additional variables such as maintenance history and regional market differences to enhance the understanding of pricing dynamics in the car market.

# Appendix A

En bild som visar text, diagram, skärmbild, kvadrat

Automatiskt genererad beskrivning

A 2 Boxplot of Sales price

A1: This boxplot visually summarizes the distribution of sales prices. The box represents the interquartile range (IQR), containing the middle 50% of the data. The horizontal line within the box is the median sales price. The whiskers extend to the most extreme data points within 1.5 times the IQR from the quartiles. Data points beyond the whiskers are considered outliers and are plotted individually.



A 3 Table of metrics from different linear regression models.

A2: Contained herein is a tabulated summary showcasing the performance of univariate linear regression models. Each model captures the influence of a single predictor variable on the sales price of used Volvo cars. Detailed in the table are the metrics, indicative of each model’s predictive power and accuracy:

* Model: Denotes the independent variable used in each regression model, such as "seller", "fuel", or "mileage".
* RMSE (Root Mean Square Error): Provides the standard deviation of the prediction errors, quantifying the average discrepancy between the observed actual outcomes and the predictions by the model. Models with lower RMSE values have typically made more accurate predictions.
* R-squared: This metric reveals the proportion of the variance in the dependent variable that is predictable from the independent variable in each model. Models with higher R-squared values are generally considered to have a better fit.
* Adjusted R-squared: Adjusts the R-squared value to account for the number of predictors in the model, thus providing a more accurate indicator for models with a single predictor.

En bild som visar text, skärmbild, Teckensnitt, programvara

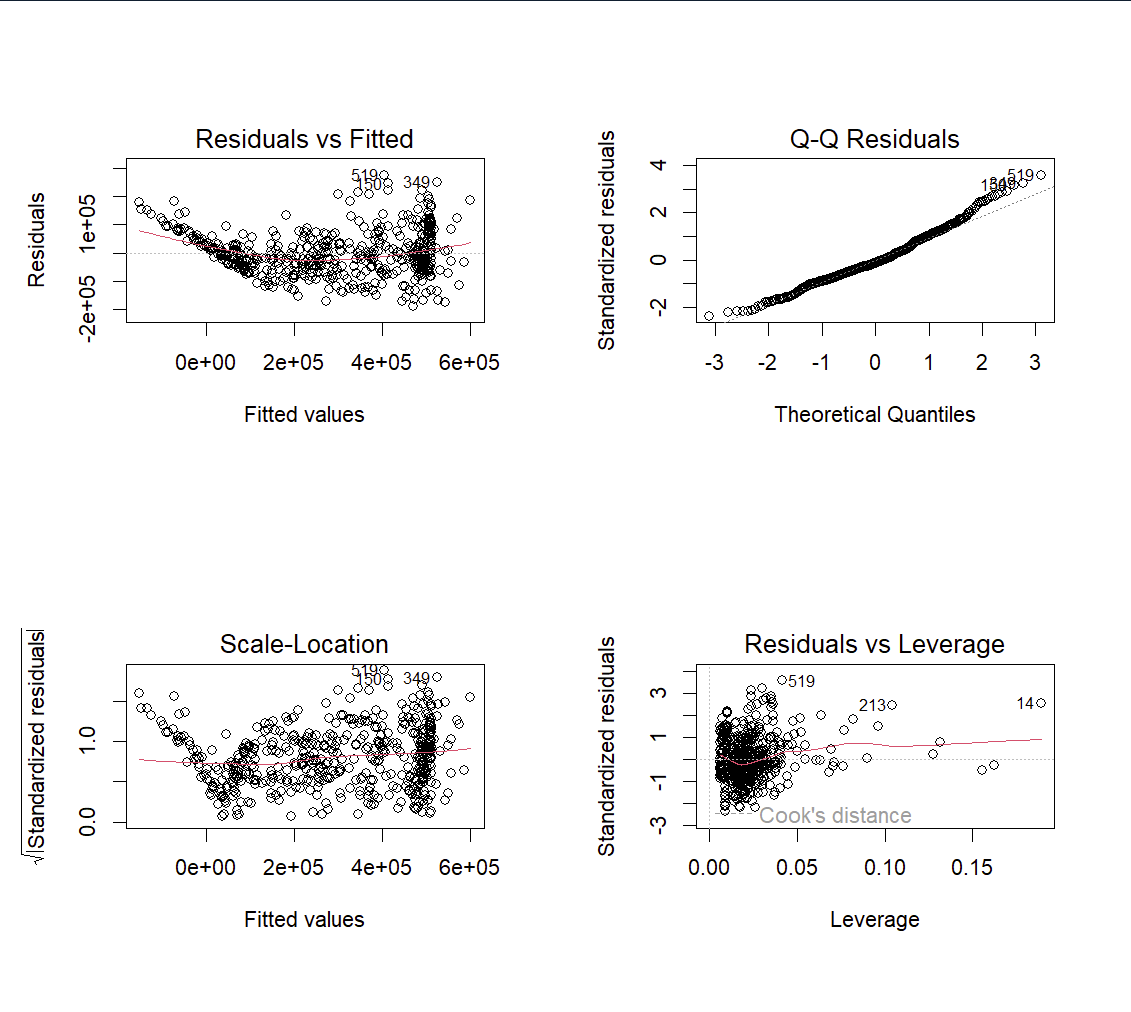
Automatiskt genererad beskrivning

A 4 Table of results from different interactions.

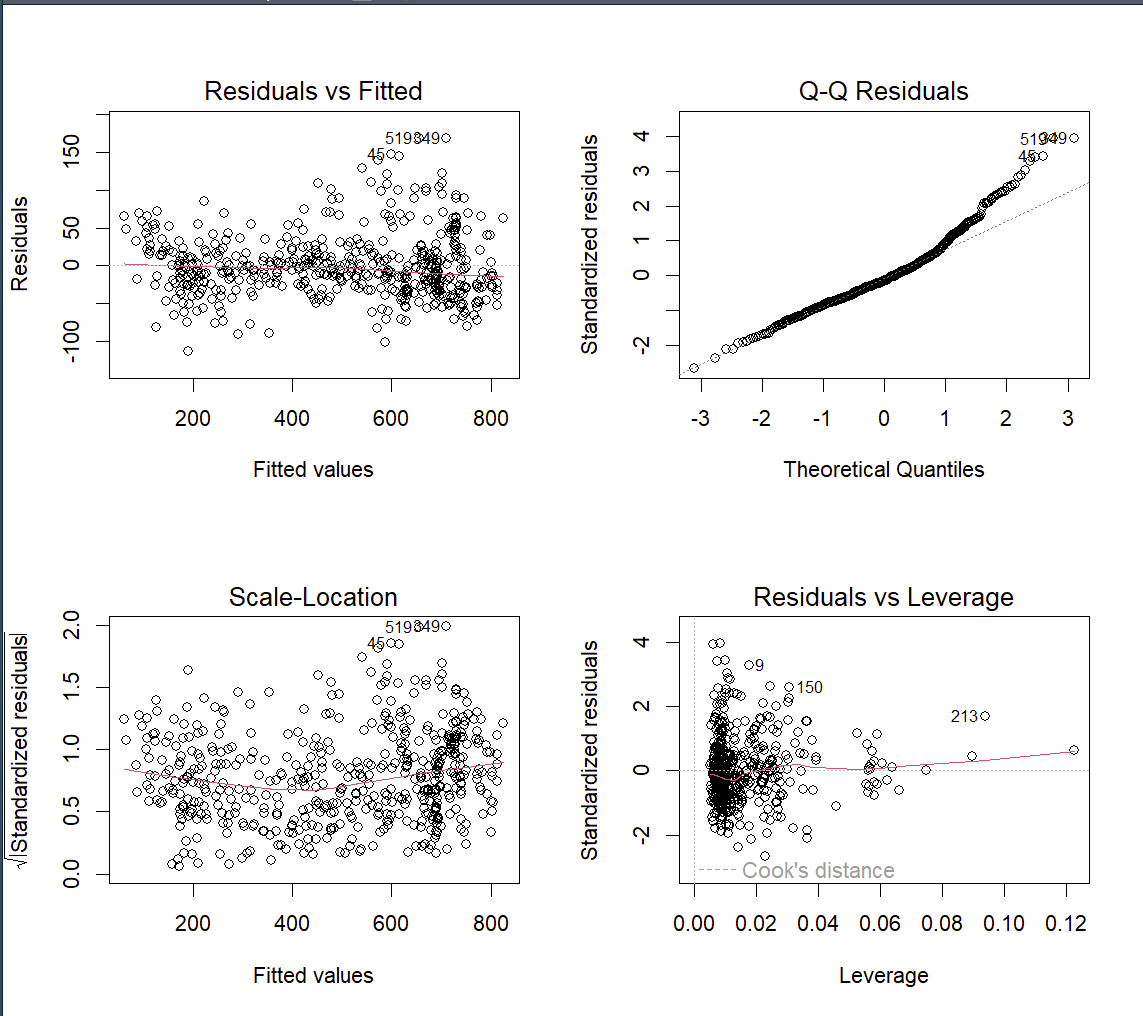
A3: Comprehensive overview of the performance metrics for a series of regression models analyzed in the study. Each model examines different combinations of predictors to determine their impact on the sales prices of used Volvo cars. The table lists the following key metrics for comparison:

* Model: Describes the specific interaction or combination of predictor variables used in each regression model, such as "Model\_Year \* Fuel" or "Mileage + Fuel".
* RMSE (Root Mean Square Error): Represents the standard deviation of the residuals, providing a measure of the average magnitude of the errors between the predicted and actual values. Lower values indicate a better fit.
* R-squared: Shows the proportion of variance in the dependent variable that is predictable from the independent variables. A higher R-squared value indicates a better fit of the model to the data.
* Adjusted R-squared: Adjusts the R-squared for the number of predictors in the model, providing a more accurate measure of model performance, especially for comparison across models with different numbers of predictors.
* Transformer: Indicates whether a transformation was applied to the response variable to meet model assumptions, with "sqrt" denoting a square root transformation and "none" indicating no transformation was used.

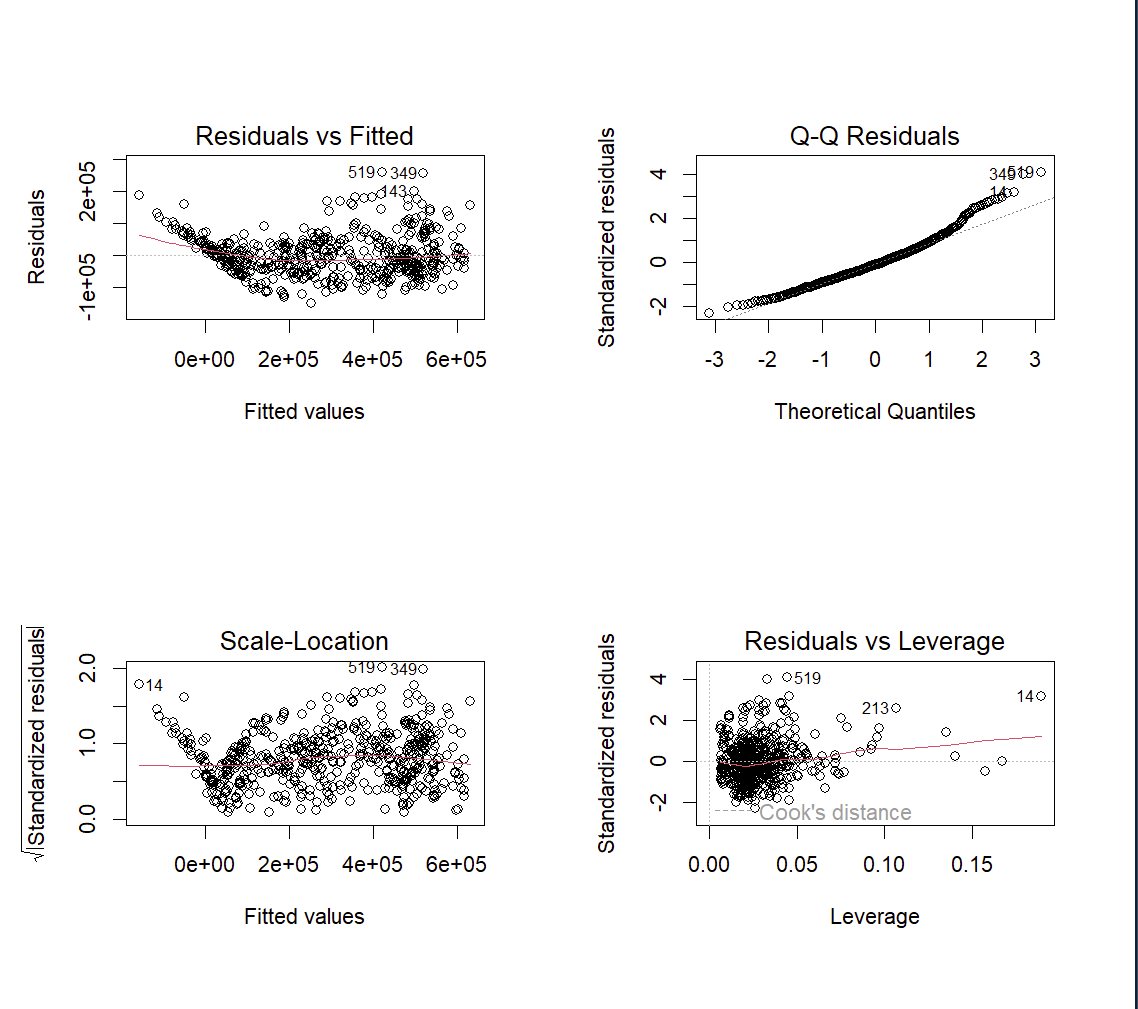
Each row in the table corresponds to a different regression model, with the respective performance metrics enabling a comparative evaluation of how well each model explains the variation in car prices. This table serves as a detailed reference for the models discussed in the report, offering a transparent and quantitative basis for model selection and further analysis.



A 5 Diagnostic plots for Model 3



A 6 Diagnostic plots from Model 1



A 7 Diagnostic plots from Model 2

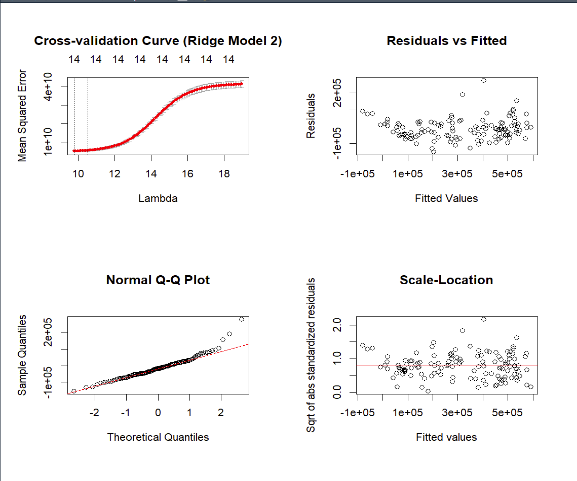
A4-A6 Explanation of Diagnostic Plots:

* Residuals vs Fitted: Checks for non-linearity and homoscedasticity. Points should be evenly dispersed around a horizontal line.
* Q-Q Plot: Assesses the normality of residuals. A linear trend of points along the dashed line indicates normal distribution of residuals.
* Scale-Location: Another evaluation of homoscedasticity, where a random scatter of points suggests equal variance across all fitted values.
* Residuals vs Leverage: Identifies influential data points. Points beyond the Cook's distance lines may unduly influence the model.

These plots are integral to confirming the assumptions underlying linear regression and ensure that the models provide reliable and unbiased estimates.

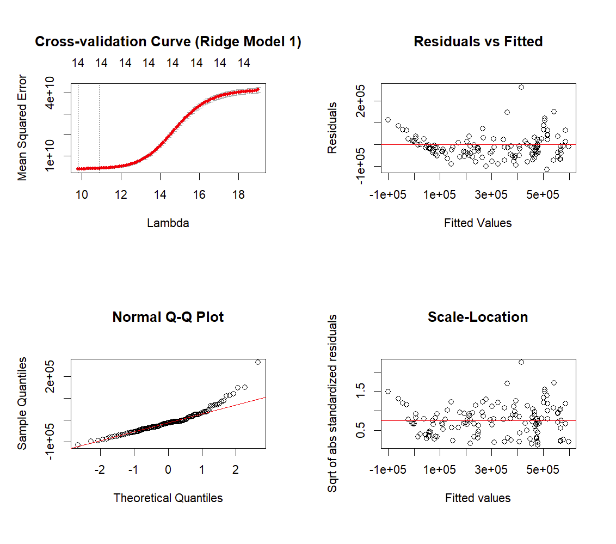
En bild som visar text, diagram, linje, skärmbild

Automatiskt genererad beskrivning



A 8 Diagnostic plots for Ridge 2

A 9 Diagnostic plots for Ridge 3

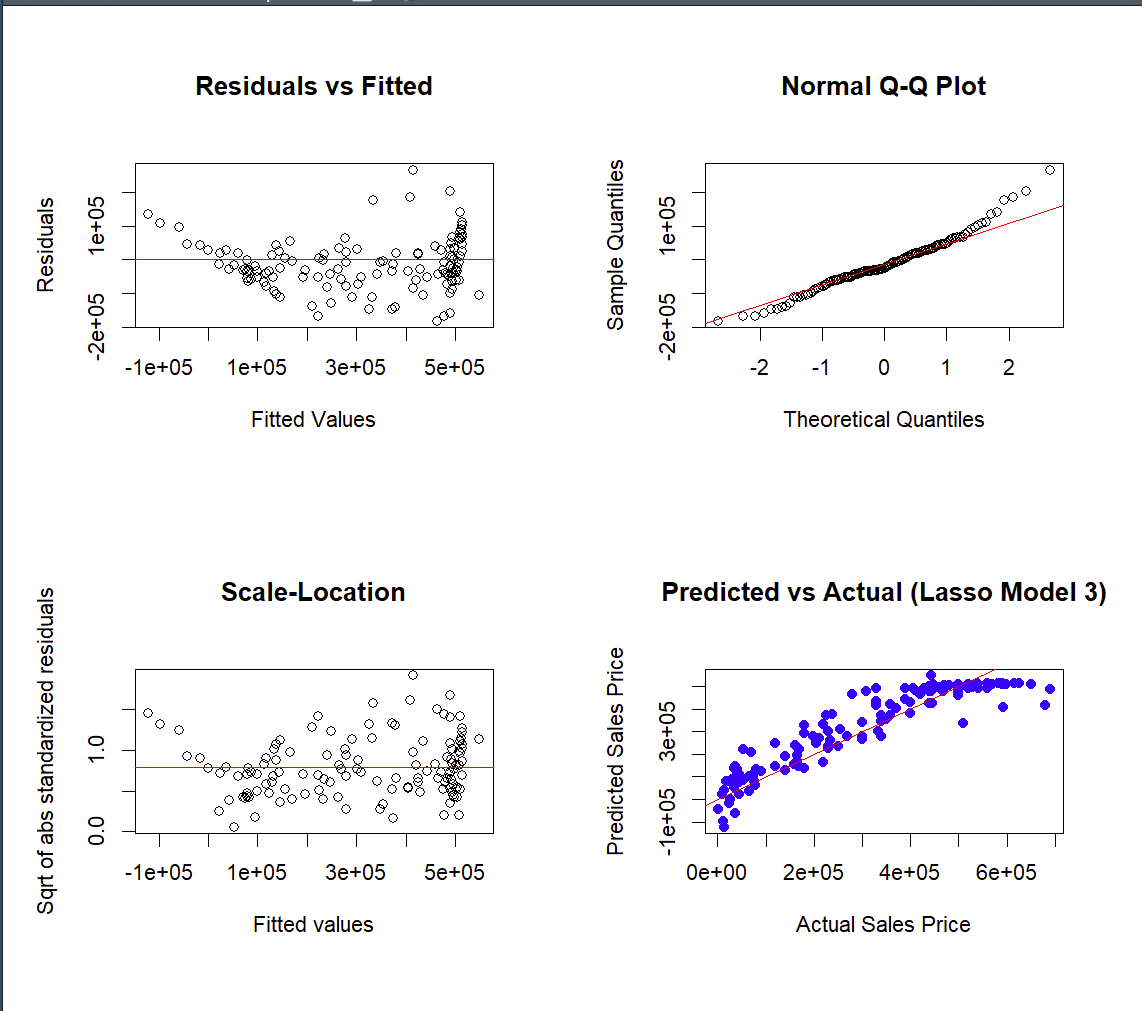


A 10 Diagnostic plots for Ridge 1

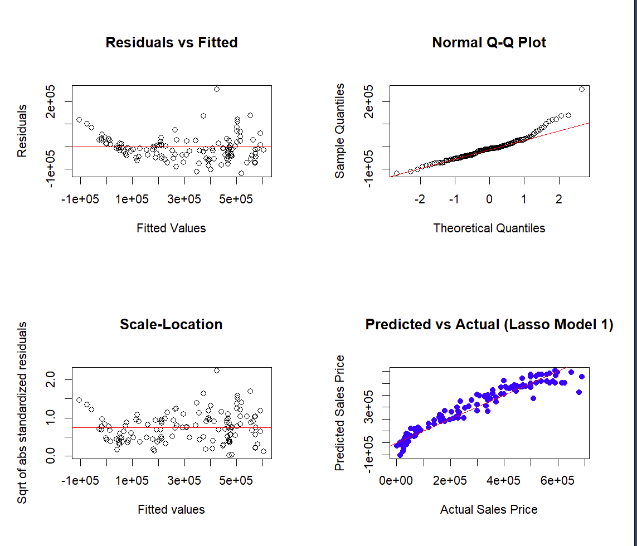
A7-A9 Explanation of Diagnostic Plots:

* Top Left: Cross-validation Curve — Depicts the trade-off between the complexity of the model (as regulated by lambda) and the associated error. The optimal lambda minimizes the error, providing the best regularization strength.
* Top Right: Residuals vs Fitted — Shows how the residuals scatter around the horizontal line representing no bias. Patterns in this plot could indicate problems with the model’s specification.
* Bottom Left: Normal Q-Q Plot — A visual check for the normality of the error terms. Deviations from the diagonal line suggest deviations from normality.
* Bottom Right: Scale-Location — Used to verify the assumption of equal error variances. A horizontal line without a distinct pattern suggests homoscedasticity is met.

Each figure set pertains to a different Ridge Regression model, labeled accordingly. The plots provided enable the evaluation of each model's performance and adherence to the assumptions of linear regression, facilitating an informed selection of the best model for predicting the sales price of used cars. These diagnostics contribute to a deeper understanding of the model's predictive abilities and underlying assumptions.



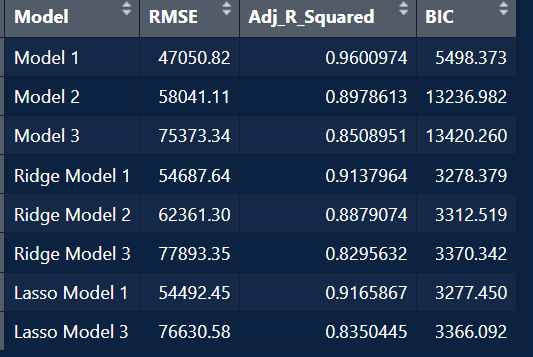
A 11 Diagnostic plots for Lasso Model 3.



A 12 Diagnostic plots for Lasso Model 1

A10-A11: Explanation of Diagnostic Plots:

* Residuals vs Fitted: This plot is a check for any systematic patterns in the residuals that might suggest non-linearity or other model inadequacies.
* Normal Q-Q Plot: Allows for a visual assessment of how closely the residuals match a normal distribution, with deviations indicating potential issues with normality.
* Scale-Location: Verifies that the model's errors exhibit consistent variance across all levels of fitted values, which is important for the reliability of standard errors and statistical tests.
* Predicted vs Actual (Lasso Model 2): Plots the predicted sales prices against the actual prices, with points clustering around the line of equality representing a well-fitting model.

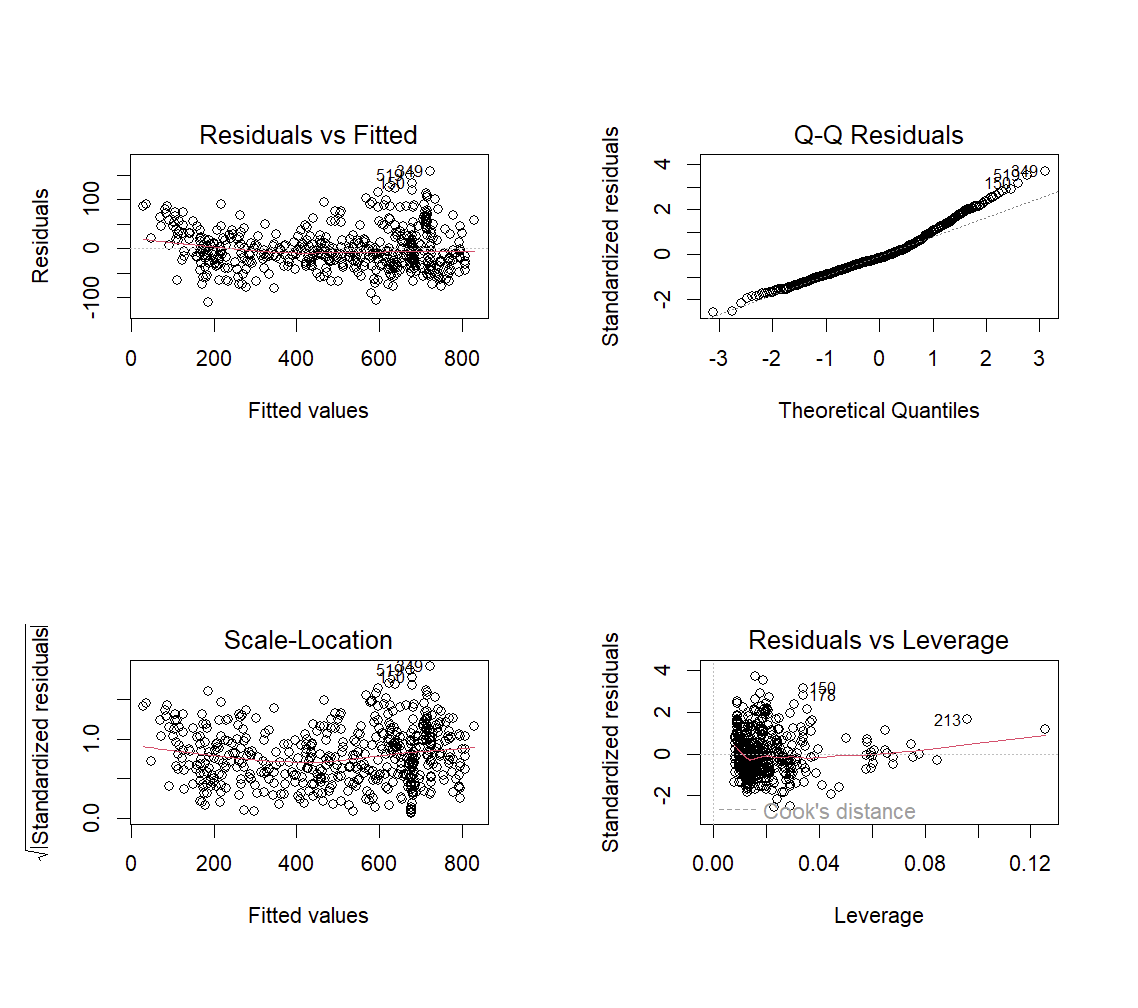


A 13 Table of model comparison.

A 12: The performance metrics of the array of preliminary regression models evaluated prior to the selection of the final model. For each model, key indicators of predictive accuracy and model fit are tabulated:

* Model: Enumerates the various models assessed, including baseline, Ridge, and Lasso regression models.
* RMSE (Root Mean Square Error): Reflects the average difference between the predicted values by the model and the observed actual values. The lower the RMSE, the more accurate the model's predictions.
* Adj\_R\_Squared (Adjusted R-Squared): Indicates the proportion of variance explained by the model, adjusted for the number of predictors. This value accounts for the model complexity and is more indicative of the fit for models with a different number of variables.
* BIC (Bayesian Information Criterion): A criterion for model selection among a finite set of models; the model with the lowest BIC is generally preferred. It is useful for comparing models with varying numbers of parameters as it includes a penalty term for the number of parameters in the model.

This table provides a brief comparison across the models, guiding the selection of the most suitable one for the final analysis. The chosen model, based on these metrics, optimally balances the trade-off between complexity and fit, as demonstrated by its comparative RMSE, Adjusted R-Squared, and BIC values.



A 14 Diagnostic plots for final model

A13 Set of diagnostic plots corresponding to the final regression model chosen for its superior performance in predicting used car sales prices.

* Residuals vs Fitted: This graph is instrumental in evaluating the model's linearity. The random distribution of residuals around the horizontal line suggests that the linear regression assumptions are adequately met.
* Normal Q-Q Plot: This plot tests the normality of the residuals. The alignment of the data points along the reference line indicates that the residuals follow a normal distribution, which is an essential assumption of linear regression models.
* Scale-Location (Spread-Location) Plot: Here, we check for homoscedasticity, meaning the residuals have equal variance across all levels of the independent variables. The uniform scatter of points indicates that the variance of residuals is consistent.
* Residuals vs Leverage: This diagnostic is used to identify influential observations within the data that could unduly affect the regression model’s estimates. Observations with a high Cook's distance might be particularly influential, and while they do not necessarily need to be removed, they should be examined to ensure they do not stem from data entry errors or other issues.

The interpretation of these plots indicates that the final model is well-calibrated, and the assumptions of linear regression are largely satisfied. Observations identified in the "Residuals vs Leverage" plot as having high leverage or a large Cook's distance have been scrutinized to confirm their validity.

# Sources

EliteDataScience.com. (den 6 July 2022). *Elite Data Science.* Hämtat från Best Practices for Feature Engineering: https://elitedatascience.com/feature-engineering-best-practices

Gareth James, D. W. (2023). *An Introduction to Statistical Learning with Applications in R.* Springer.

jbstatistics (Regissör). (2013). *Normal Quantile Quantile Plots* [Film].

Kampakis, D. S. (u.d.). *TheDataScientist*. Hämtat från Performance Measures: RMSE And MAE: https://thedatascientist.com/performance-measures-rmse-mae/

Miller, T. L. (den 16 01 2020). *Regression Subset Selection.* Hämtat från The Comprehensive R Archive Network: https://cran.r-project.org/web/packages/leaps/leaps.pdf

Måns Magnusson, M. K. (den 31 October 2022). *Retrieval and analysis of PC Axis data with the pxweb package. R package version.* Hämtat från Github.com: https://ropengov.github.io/pxweb/

O'Brien, K. (2023). *RPubs*. Hämtat från Cook's Distance for Linear Models: https://rpubs.com/DragonflyStats/Cooks-Distance

Statistics How To. (2024). *Statistics How To*. Hämtat från Adjusted R2 / Adjusted R-Squared: What is it used for?: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/adjusted-r2/

Statistics How To. (2024). *Statistics How To*. Hämtat från Bayesian Information Criterion (BIC) / Schwarz Criterion: https://www.statisticshowto.com/bayesian-information-criterion/

Statistikmyndigheten SCB. (den 03 04 2024). *Statistikmyndigheten SCB.* Hämtat från Nyregistrerade personbilar efter län och kommun samt drivmedel: https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START\_\_TK\_\_TK1001\_\_TK1001A/PersBilarDrivMedel/

Wahlgren, S. K. (2015). *Statistisk Dataanalys.* Lund: Studentlitteratur AB.