BMW Car Price Prediction: Multiple Regression Analysis

Understanding Factors and Developing Predictive Models

Ahmad Zalkat

EC Utbildning

R Programmering For Dataanalys

2024–04

Abstract

This code script conducts a thorough analysis of a dataset featuring BMW cars, aiming to understand the factors influencing their prices and develop a predictive model.

Through stages including data loading, cleaning, preprocessing, modeling, and visualization, the study employs multiple linear regression to predict prices based on key variables.

Results highlight the significant impact of variables like model year and mileage on prices, offering valuable insights for car buyers, sellers, and enthusiasts.

This analysis provides valuable insights to the buyers and sellers' cars.

Förkortningar och Begrepp

- **Exploratory Data Analysis (EDA):**

  EDA is an important first step in any data analysis.

  An analysis approach that identifies general patterns in the data. These patterns include outliers and features of the data that might be unexpected.

- **R-Squared (R^2 or the coefficient of determination):**

  A statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable.

- **F-statistics:**

  Used to determine whether the variance between two normal populations are similar to one another.

- **P-value:**

  Measures the probability of obtaining the observed results, assuming that the null hypothesis is true.

- **NA - Not Available:**

  Short Word in tables and lists for the phrase not applicable, not available, not assessed, or no answer.

- **Std. Error - Standard Error:**

  The approximate standard deviation of a statistical sample population, Measure of the uncertainty in the regression coefficient estimates.

- **Interception - Intercept:**

  Expected value of the dependent variable when all independent variables are zero.

- **Variabler - Variables:**

  Properties or factors used in an analysis to explain variation in another variable.

- **Regression:**

  Predict or explain the variation in one variable based on another variable.

- **Residuals:**

  The difference between observed and predicted values in a regression.

- **Model Summary:**

  View summary of the neural network predictive or classification accuracy.

Contents

1- Inledning

The primary goal of this analysis is to understand how features together affect the prices of BMW vehicles and to create a predictive model that accurately estimates the prices of these vehicles.

The analysis aims to provide valuable information to potential buyers, sellers.

In this report, we delve into analyzing BMW sales data to extract insights and understand trends in the market and we made data cleaning, preprocessing, modeling, prediction, and visualization.

The dataset used in this analysis contains information about BMW vehicle listings, including descriptions, location, year of manufacture, fuel type, mileage, transmission type, model, price, and car information.

Data overview:

The data set includes 995 observations and 10 variables. These variables include:

- **Description:** Description of the car list.
- **Location:** The location where the car is sale.
- **Year of manufacture:** The year the car was manufactured.
- **Fuel type:** The type of fuel the car uses (Bensin, Hybrid, Diesel).
- **Kilometers:** The distance traveled by the car in kilometers.
- **Transmission type:** Transmission type (Automatic, Manual).
- **Model:** BMW series model.
- **Price:** The price from advertisements.
- **Link:** Advertisement link.
- **Date:** The date and time the ad was published.

Number of car models in this report: There are a total of 35 unique BMW models included in the dataset.

Price range: prices from10,000 to high price as 579,900.

The year the car was made: from the years 2000 to 2023, The largest section between 2007 and 2015.

Transmission Types: Manual and Automatic The largest section are automatic cars.

Geographical distribution: The listings cover various locations, indicating the wide availability of BMW vehicles in Sweden.

Conclusion:

BMW car sales data analysis provides valuable insights into current market trends and preferences among buyers and sellers. Understanding car model distribution, pricing trends, and other factors can help both buyers and sellers make pricing.

This analysis serves as a basis for further research and exploration of automobile market price.

The analysis begins with exploratory data analysis (EDA), where we examine the structure, summary statistics and initial patterns within the data set. We implement data cleansing procedures to handle missing values and ensure data integrity. Categorical variables are coded in numerical format for modeling purposes, and unnecessary columns are removed to simplify the data set.

After pre-processing the data, we use a linear regression model to predict car prices by selected independent variables like car model, fuel type, mileage, transmission type and car model. We evaluate regression model performance through summary statistics and visualizations, including scatter plots that compare predicted prices to actual prices, as well as line charts with confidence intervals.

Through this analysis, we provide valuable insights into the factors that influence BMW vehicle prices and provide a reliable predictive model to estimate prices. The analysis assisting car buyers, sellers price decision making for BMW automobile market.

Based on the results of this analysis, the following recommendations are proposed:

Prospective buyers should consider exploring the wide range of BMW models available on the market to find the model that best suits their needs and budget.

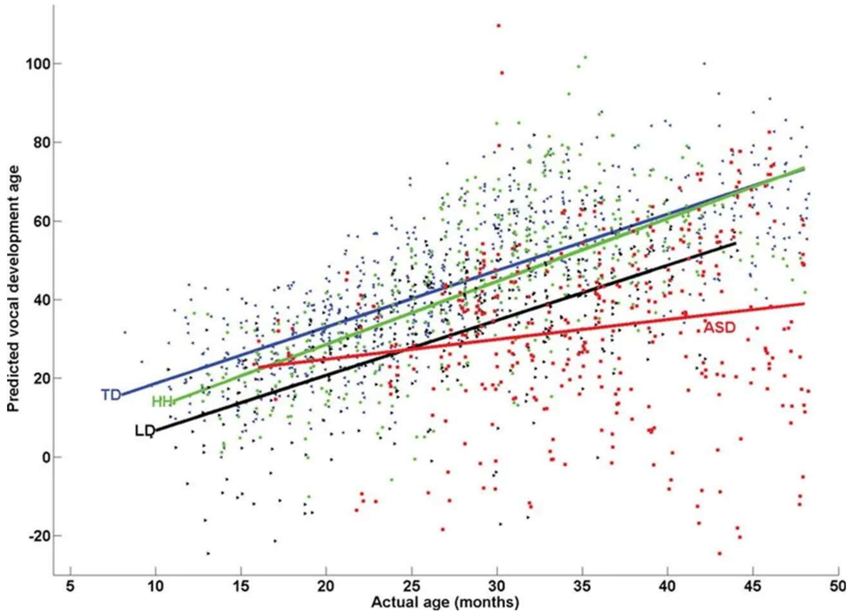
Sellers can use pricing insights to set competitive prices for their listings and attract potential buyers.

Overall, analyzing BMW car sales data provides valuable insights that can assisting the buyers and sellers’ cars.

2- Teori

2.1 Multiple Linear Regression

Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression. The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the dependent variable are known as independent or explanatory variables.



Multiple Linear Regression Formula

$$yi = \beta0 + \beta1xi1 + \beta2xi2 + \ldots + \beta pxip+\epsilon$$

- yi is the dependent or predicted variable.
- β0 is the y-intercept, i.e., the value of y when both xi and x2 are 0.
- β1 and β2 are the regression coefficients representing the change in y relative to a one-unit change in xi1 and xi2, respectively.
- βp is the slope coefficient for each independent variable.
- ϵ is the model's random error (residual) term.

3- Metod

The cumulative data collection was done through careful collaboration by a skilled team of six Blocket web site. A targeted selection process was conducted to highlight key variables to determining the determinants of BMW cars pricing, thus facilitating robust predictive modeling for a comprehensive BMW cars evaluation.

3.1 Load and prepare data in R studio:

The analysis begins by loading the dataset from an Excel file using the read_excel function. Data cleaning procedures are applied to handle missing values using the na.omit function. Categorical variables are coded in numerical format to prepare data for modeling.

3.2 Exploratory Data Analysis:

Summary statistics, including the mean, median, and standard deviation, for numeric variables are calculated using the summary function. The structure of the data set is examined using the str function to understand variable types and dimensions. Elementary patterns and relationships between variables are explored through visualizations such as histograms, scatter plots, and box plots.

3.3 Modeling:

A multiple linear regression model was designed to predict car prices based on selected independent variables (Årsmodell, Drivmenel, Miltal, Växellåda, Modell).

The function is used to fit the regression model, and the model summary is obtained using the sum function to evaluate the model performance and its expected significance.

3.4 Evaluation:

Predictions are generated using a regression model fitted to the independent variables in the data set.

Visualizations such as scatter plots that compare predicted prices with actual prices are created to evaluate the predictive performance of the model.

Summary statistics of predicted and actual values are calculated to evaluate the accuracy and reliability of the model.

3.5 Interpretation and Conclusion:

The results of the analysis are interpreted to understand the factors that affect the prices of BMW cars.

Ideas and recommendations are provided based on the results of the analysis.

4- Resultat och Diskussion

Resultat:

4.1 Model Performance:

The regression models show significantly greater variances and variances, with an average Adjusted R-squared 0.7714.

The F statistic shown on the models is significant (F = 669, p < 2.2e-16), indicating that the maximum for the variables is significant.

4.2 Variable Meaning:

The variables "Årsmodell", "Miltal" and "Modell" are all statistically significant, indicating that they have a significant impact on car prices.

The variables “Drivmenel" and "Växellåda" do not show any significant effect on car prices.

Discussion:

4.3 Impact of Variables:

"Årsmodell", "Miltal" and "Modell” are the most important variables in predicting car prices, which is consistent with expectations.

The results suggest that “Drivmenel" and "Växellåda” may be less important factors in explaining variation in car prices.

4.4 Expected Performance:

Despite the higher R-squared, there is still variation in car prices that is not explained by the model. This may be due to other factors not included in the data, or to deficiencies in data specificity.

4.5 Limitations:

Limitations in data availability and quality may have affected model performance.

Research may focus on including additional variables that may the explanatory power of the model, as using more data and validation to evaluate the predictive performance of the model.

In this table showing some of the results we get it, will be discussed separately in the report.

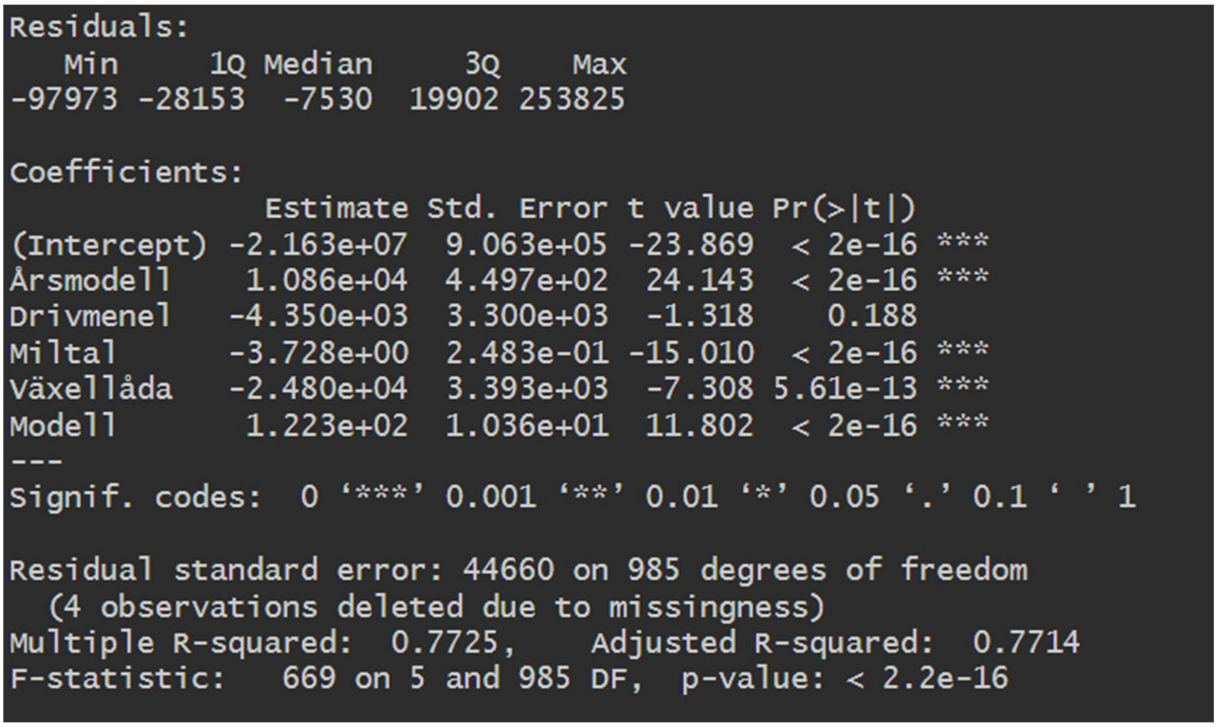
Variable	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-21,630,000	906,300	-23.869	< 2e-16
Årsmodell	10,860	449.7	24.143	< 2e-16
Drivmenel	-4,350	3,300	-1.318	0.188
Miltal	-3.728	0.2483	-15.010	< 2e-16
Växellåda	-24,800	3,393	-7.308	5.61e-13
Modell	122.3	10.36	11.802	< 2e-16

- **Estimate:** represents the estimated coefficients for each variable in the regression model.

- **The Standard Error:** represents the standard errors associated with each parameter estimate.
- **The “t-value”:** represents the t-values calculated for each parameter estimate.
- **“Pr(>|t|)”** represents the probability values associated with each value of t, indicating the importance of each variable in predicting car prices.

These results provide insight into the importance of each variable in explaining variation in automobile prices, as well as the direction and magnitude of their effects.

5- Slutsatser



```
Residuals:
   Min     1Q Median     3Q    Max 
-97973 -28153  -7530  19902 253825 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.163e+07  9.063e+05 -23.869  < 2e-16 ***
Årsmodell    1.086e+04  4.497e+02  24.143  < 2e-16 ***
Drivmenel   -4.350e+03  3.300e+03  -1.318    0.188    
Miltal      -3.728e+00  2.483e-01 -15.010  < 2e-16 ***
Växellåda   -2.480e+04  3.393e+03  -7.308 5.61e-13 ***
Modell       1.223e+02  1.036e+01  11.802  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44660 on 985 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.7725,	Adjusted R-squared:  0.7714 
F-statistic:   669 on 5 and 985 DF,  p-value: < 2.2e-16
```

5.1 Represent:

The differences between actual prices and the prices predicted by regression model:

- Minimum is -97973.
- First quartile is -28153.
- Median is -7530
- Third quartile is 19902.
- Maximum is 253 825.

5.2 Coefficients:

Interpret the values in the output:

- **Intercept:** The intercept represents the estimated price of a BMW car when all independent variables are zero, here we can see the intercept is -2.163e+07, it's no practical because it is outside the reasonable price range.
- **Årsmodell (Model Year):** For each additional year of model, the estimated price of the BMW car increases by 10,860, and the low p-value (< 0.001) and high value of t (24.143) indicate that this variable is highly significant in predicting car prices.
- **Drivmenel (Fuel Type):** The estimated price decreases by 4,350 per type, but this is not statistically significant as the p value is greater than 0.05.
- **Miltal (Mileage):** Estimated price decreases by 3,728 for each additional mile traveled. and the low p-value (< 0.001) and high value of t (15.010) indicate that this variable is highly significant in predicting car prices.
- **Växellåda (Transmission Type):** The estimated price decreases by 24,800 when unit change between (Automat, Manuell), and the low p-value (< 0.001) and high value of t (-7.308) indicate that this variable is highly significant in predicting car prices.
- **Modell (Model):** For each unit change in model, the estimated price increases by $122.3, and the low p-value (< 0.001) and high value of t (11.802) indicate that this variable is highly significant in predicting car prices.

Notation p-value (< 0.001) is a commonly used to indicate that the p-value is less than 0.001. its same (<2e-16) This notation means "less than 2 times 10 to -16," which is effectively zero in the context of hypothesis testing.

Both symbols have the same meaning: the p-value is very small, indicating strong evidence against the null hypothesis.

It's a short way of saying that the p-value is very small and very significant.

5.3 Multiple R-squared and Adjusted R-squared:

- **Multiple R-squared:** (0.7725) indicates that about 77.25% of the variation in car prices can be explained by the independent variables in the model.
- **Adjusted R-squared:** (0.7714) is adjusted for the number of predictors in the model, providing a more accurate estimate of the proportion of variance explained.

5.4 Residual Standard Errors:

The residual standard error (44660) represents the average amount by which observed prices deviate from prices predicted by the model. Lower values indicate a better fit of the model to the data.

5.5 F-Statistics and P-Value:

- The F statistic (669) tests the overall significance of the regression model. A high F statistic and low p value (<0.05) indicate that the model is statistically significant and explains a significant amount of the variance in the dependent variable.
- The P value (< 2.2e-16) of the F statistic indicates that the model is highly significant.

I did several tests before determining the best model, you can check the codes used in code file.

In my report, I focused on the best model and and show the results in a professional statistical.

6- Relevant statistics SCB

Statistics Data from the Statistikmyndigheten website SCB.se

These statistics reveal fluctuations in how many car registrations based on fuel type.

Providing valuable insights into the factors that lead to increased prices and demand for electric cars, as well as decreased registrations for cars that use other types of fuel (diesel and bensin).

The bar chart the fluctuations in demand for cars based on the type of fuel type, we can see the increase and decrease in demand over time.

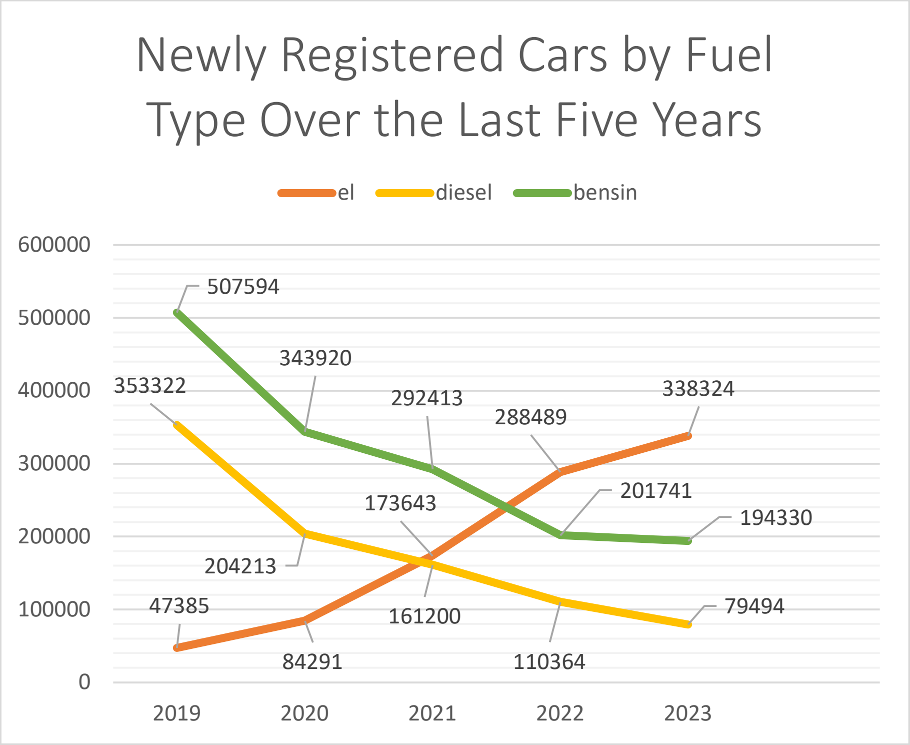


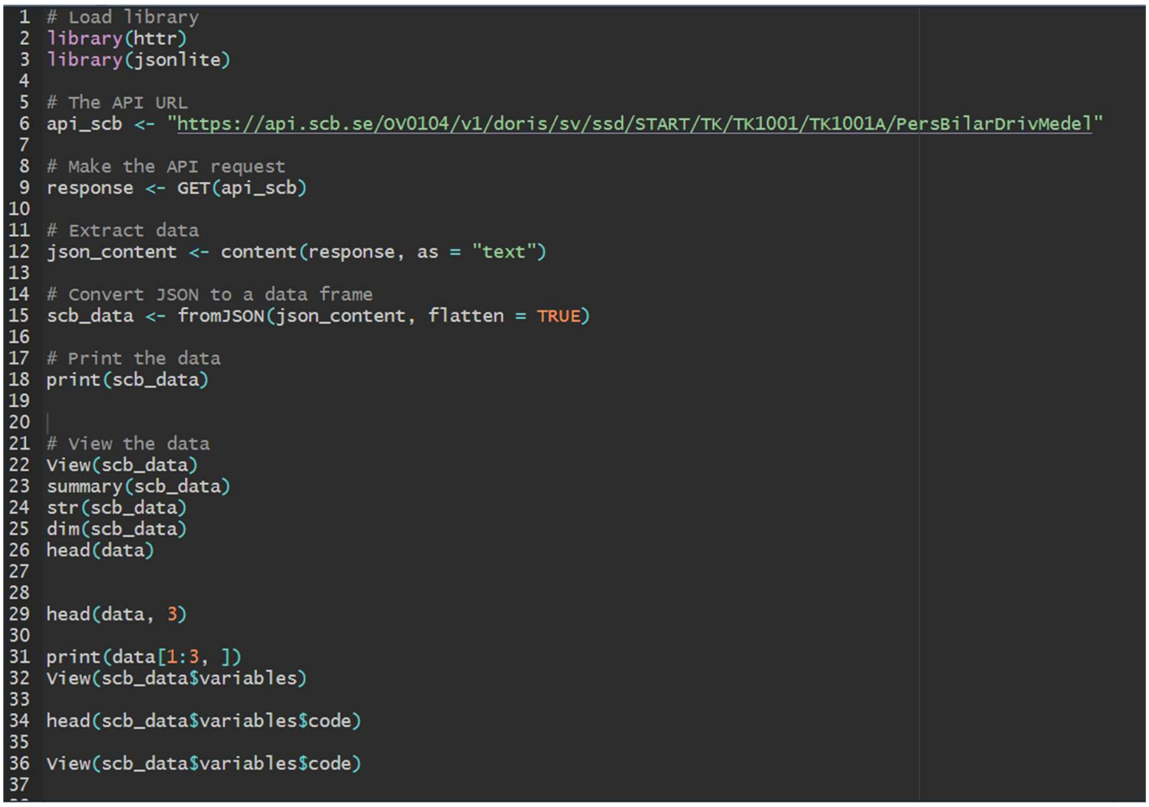
Table Showing Percentage Distribution and actual number of registrations car per year by Fuel Type.

The percentages were based on the years between 2019 and 2023, The percentages differences were calculated over the five-year period.

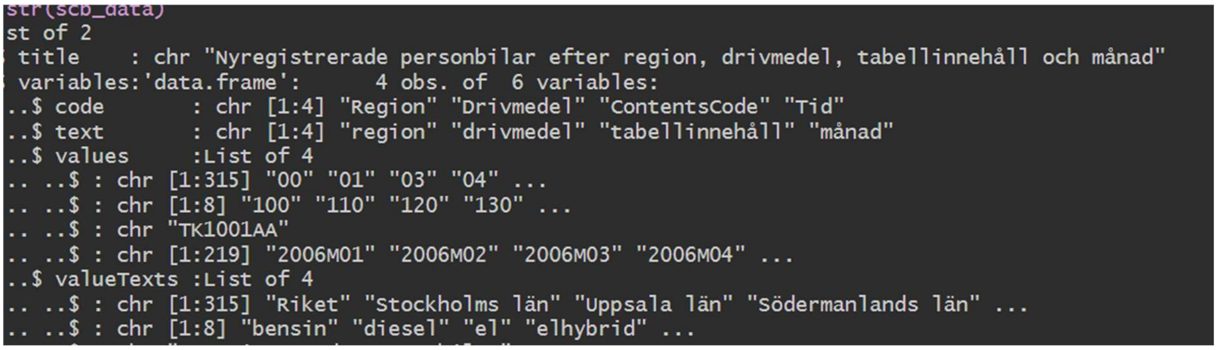
Year	Fuel Type	Number of Registrations	Percentage Distribution (%)
2019	Bensin	507594	56%
	Diesel	353322	39%
	El	47385	5%
	Total Increase for 5 Years	36906	11%
	Total Decrease for 5 Years	149109	36%
2020	Bensin	343920	54%
	Diesel	204213	32%
	El	84291	13%
	Total Increase for 5 Years	89352	26%
	Total Decrease for 5 Years	42986	10%
2021	Bensin	292413	47%
	Diesel	161200	26%
	El	173643	28%
	Total Increase for 5 Years	114846	34%
	Total Decrease for 5 Years	52636	13%
2022	Bensin	201741	34%
	Diesel	110364	18%
	El	288489	48%
	Total Increase for 5 Years	49784	15%
	Total Decrease for 5 Years	28870	7%
2023	Bensin	194330	32%
	Diesel	79494	13%
	El	338324	55%
	Total Increase for 5 Years	49753	15%
	Total Decrease for 5 Years	143119	34%

The data has been imported from SBC.se by (xlsx) extension.

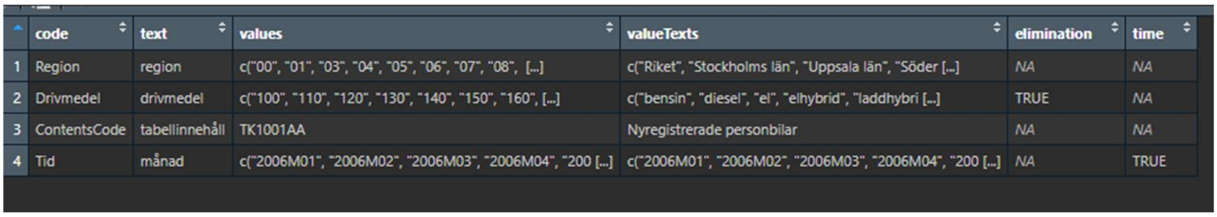
There are alternative methods for importing data and displaying it. Below are some examples of the code used to showing data in R Studio by API (JSON query).



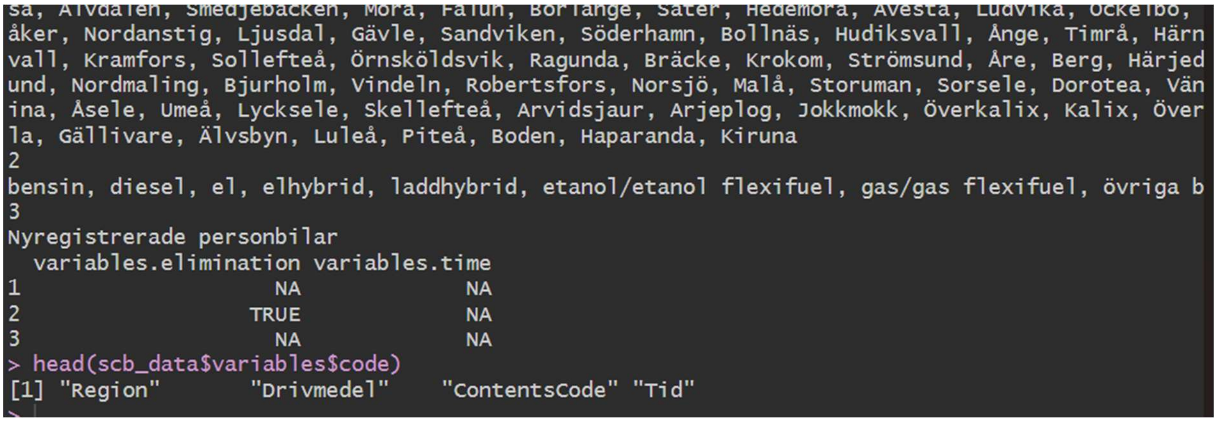
```r
# Load library
library(httr)
library(jsonlite)

# The API URL
api_scb <- "https://api.scb.se/OV0104/v1/doris/sv/ssd/START/TK/TK1001/TK1001A/PersBilarDrivMedel"

# Make the API request
response <- GET(api_scb)

# Extract data
json_content <- content(response, as = "text")

# Convert JSON to a data frame
scb_data <- fromJSON(json_content, flatten = TRUE)

# Print the data
print(scb_data)


# View the data
View(scb_data)
summary(scb_data)
str(scb_data)
dim(scb_data)
head(data)


head(data, 3)

print(data[1:3, ])
View(scb_data$variables)

head(scb_data$variables$code)

View(scb_data$variables$code)

```



```
str(scb_data)
st of 2
 title    : chr "Nyregistrerade personbilar efter region, drivmedel, tabellinnehåll och månad"
 variables:'data.frame':      4 obs. of  6 variables:
..$ code       : chr [1:4] "Region" "Drivmedel" "ContentsCode" "Tid"
..$ text       : chr [1:4] "region" "drivmedel" "tabellinnehåll" "månad"
..$ values     :List of 4
.. ..$ : chr [1:315] "00" "01" "03" "04" ...
.. ..$ : chr [1:8] "100" "110" "120" "130" ...
.. ..$ : chr "TK1001AA"
.. ..$ : chr [1:219] "2006M01" "2006M02" "2006M03" "2006M04" ...
..$ valueTexts :List of 4
.. ..$ : chr [1:315] "Riket" "Stockholms län" "Uppsala län" "Södermanlands län" ...
.. ..$ : chr [1:8] "bensin" "diesel" "el" "elhybrid" ...
```



	code	text	values	valueTexts	elimination	time
1	Region	region	c("00", "01", "03", "04", "05", "06", "07", "08", [...]	c("Riket", "Stockholms län", "Uppsala län", "Söder [...]	NA	NA
2	Drivmedel	drivmedel	c("100", "110", "120", "130", "140", "150", "160", [...]	c("bensin", "diesel", "el", "elhybrid", "laddhybri [...]	TRUE	NA
3	ContentsCode	tabellinnehåll	TK1001AA	Nyregistrerade personbilar	NA	NA
4	Tid	månad	c("2006M01", "2006M02", "2006M03", "2006M04", "200 [...]	c("2006M01", "2006M02", "2006M03", "2006M04", "200 [...]	NA	TRUE



```
sa, Älvdalen, Smedjebacken, Mora, Falun, Borlänge, Säter, Hedemora, Avesta, Ludvika, Ockelbo,
åker, Nordanstig, Ljusdal, Gävle, Sandviken, Söderhamn, Bollnäs, Hudiksvall, Ånge, Timrå, Härn
vall, Kramfors, Sollefteå, Örnsköldsvik, Ragunda, Bräcke, Krokom, Strömsund, Åre, Berg, Härjed
und, Nordmaling, Bjurholm, Vindeln, Robertsfors, Norsjö, Malå, Storuman, Sorsele, Dorotea, Vän
ina, Åsele, Umeå, Lycksele, Skellefteå, Arvidsjaur, Arjeplog, Jokkmokk, Överkalix, Kalix, Över
la, Gällivare, Älvsbyn, Luleå, Piteå, Boden, Haparanda, Kiruna
2
bensin, diesel, el, elhybrid, laddhybrid, etanol/etanol flexifuel, gas/gas flexifuel, övriga b
3
Nyregistrerade personbilar
  variables.elimination variables.time
1                    NA             NA
2                  TRUE             NA
3                    NA             NA
> head(scb_data$variables$code)
[1] "Region"       "Drivmedel"    "ContentsCode" "Tid"
>
```

7- Teoretiska frågor

Fråga 1 (QQ) Plot

(QQ) plot används för att verifiera antaganden i statistiska metoder eller anpassningen av en datamängd till en viss fördelning.

En grafisk metod för att jämföra fördelningen av två dataset eller en dataset med en teoretisk fördelning

Avvikelser från linjen indikerar skillnader i fördelningarna.

(QQ) plot vanligtvis en normalfördelning.

Fråga 2

Sammanfattningsvis: ML fokuserar på prediktioner, medan statistisk regressionsanalys kombinerar prediktioner med statistisk inferens för att förstå samband och osäkerheter

I maskininlärning handlar det främst om att förutsäga framtida händelser baserat på historiska data.

Exempel. ML-modeller för att förutsäga överlevnadschanser för patienter baserat på dessa faktorer.

I statistisk regressionsanalys gör vi också prediktioner, men vi är också intresserade av statistisk inferens, vilket innebär att vi vill förstå sambandet mellan variabler och testa hypoteser.

Exempel. statistisk inferens:

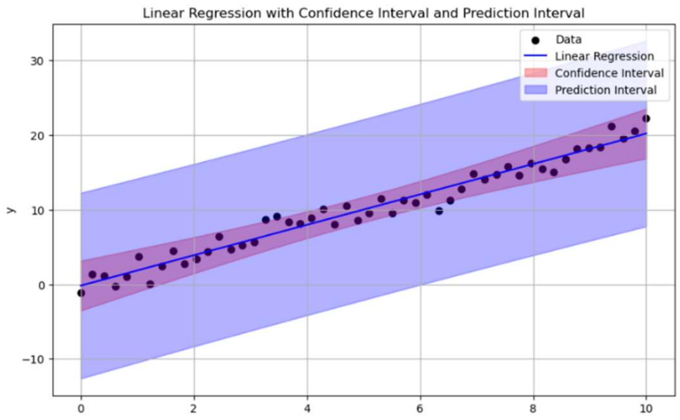
Vi kan testa hypoteser som "Är ålder signifikant för överlevnad?"

Vi kan beräkna konfidensintervall för våra överlevnadsprognoser.
Vi kan förstå hur olika faktorer samverkar för att påverka överlevnad.

Fråga 3 "konfidensintervall" och "prediktionsintervall"

Konfidensintervall används för att skatta var väntevärdet troligen ligger.

Prediktionsintervall används för att skatta var en framtida observation kommer vara för ett visst värde på x



Fråga 4

Beta-parametrarna i den multipla linjära regressionsmodellen representerar koefficienterna som multiplicerar varje prediktorvariabel.

β_0 (Intercept): Det förväntade värdet på responsvariabeln Y när alla prediktorer är noll.

β_1, β_2, …, β_p (Koefficienter för prediktorer): Hur mycket förändringen i varje prediktor påverkar Y, hållande andra prediktorer konstanta.

ε: är felet eller residualen i modellen.

Det representerar skillnaden mellan det faktiska värdet av Y och det förutsagda värdet av Y baserat på modellen.

Exempel. Antag att vi studerar sambandet mellan studieresultat (Y) och studietid (x_1) samt antal lästa böcker (x_2). Vi har följande modell:

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

β_0: Förväntat studieresultat när studietid och antal lästa böcker är noll.

β_1: Förändring i studieresultat per enhet ökning i studietid, hållande antal lästa böcker konstant.

β_2: Förändring i studieresultat per enhet ökning i antal lästa böcker, hållande studietid konstant.

Detta är en förenklad tolkning, men den ger en översikt över hur vi kan förstå beta-parametrarna i en multipel linjär regressionsmodell

Fråga 5

Vi kan undvika träning, validering och testset när man använder mått som BIC (Bayesian Information Criterion).

BIC är ett kriterium för modellval, men det ersätter inte behovet av att dela upp data i tränings- och testset.

Träning, validering och testning är fortfarande viktiga steg för att utvärdera modellens prestanda och generalisering.

Fråga 6 Best Subset Selection

Best subset selection

1-M0: Den nollmodell som inte innehåller några prediktorvariabler.

Den förutsäger helt enkelt medelvärdet för varje observation.

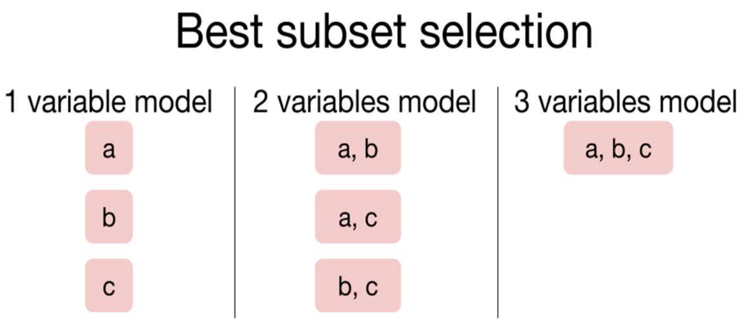
2-För k = 1, 2, … p: Vi passar alla (p/k) modeller som innehåller exakt k prediktorer.

Vi väljer den bästa bland dessa (p/k) modeller och kallar den för Mk.

bäst som den modell med högst R2 eller equivalently lägsta residual sum of squares (RSS).

3-Vi testa ut den bästa modellen från ett utbud av alternativ (M0…Mp) med hjälp av olika metoder som korsvalidering, förutsägelsefel Cp, BIC, AIC, eller adjusted R2.

Exemple: vi har ett dataset med p = 3 prediktorvariabler och en responsvariabel Y.



Vi passar åtta olika modeller:
1-modell utan predictors.

2-modell med predictor x1.

3-modell med predictor x2.

4-modell med predictor x3.

5-modell med predictor x1 och x2.

6-modell med predictors x1 och x3.

7-modell med predictors x2 och x3.

8-modell med predictors x1, x2 och x3.

Sen väljer vi modell som har högst R2

Vi använder vi korsvalidering för att bestämma den slutliga modellen baserat på det lägsta förutsägelse felet Cp, BIC, AIC eller adjusted R2

Vi kan välja den som passar bäst för det specifika problemet.

Genom att tillämpa formler för att beräkna olika metriker, inklusive Cp, AIC, BIC och adjusted R2, får vi en objektiv bedömning av modellens prestanda

Fråga 7 George Box Citatet.

“All models are wrong, some are useful.”

Detta citat att vi bör vara medvetna om metodernas begränsningar och inte säker 100 present lita på dem.

Om modeller är felaktiga kan de ändå vara kraftfulla verktyg för att förstå och hantera komplexitet i vetenskap, teknik och beslutsfattande.

Vi behöver alltid utvärdera deras prestanda, förstå deras antaganden och vara beredda på att justera dem när vi får mer information.

8- Datainsamling

1. Vem du har arbetat i grupp med?
   - Anita Kongpachith
   - Anna Strbac
   - Christofer Fromberg
   - Garima Choudhary
   - Lina Shideda
   - Mustapha Hadrous
   - Ahmad Zalkat
2. Hur har ni i gruppen arbetat tillsammans?

Vi har träffats flera gånger på Teams många att välja vilken data vi kunde använda.

3. Vad var bra i grupparbetet och vad kan utvecklas?

Vi kan utveckla gruppen arbetet genom bättre arbetsfördelning och att organisera möten mer effektivt.

4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp?

Mina styrkor genom grupparbete är att lyssna på alla åsikter och presentera förslag på ett logiskt sätt

5. Finns det något du hade gjort annorlunda? Vad i sådana fall?

Jag tror att när man jobba på ett grupparbete måste dala uppgifter och väljas en person att leda gruppen, organisera möten och sätt en tydlig plan.

9- Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

De utmaningarna under arbetet innefattade svårigheter hantera den bristfälliga data som har som var ojämnt fördelade och finns många Skillnader och prisskillnader, och hitta den Logiska analysen.

Övervinna hinder de utmaningarna

jag började med förstår Data för att se vilken mest relevanta variablerna och faktorerna.

Genom att fokusera på variablerna och genomföra analyser jag kunde extrahera användbara data och insikter från analyser data.

2. Vilket betyg du anser att du skall ha och varför.

jag anser att jag kommer få högt betyg på grund av den omfattande och väl gjorde analysen av predict BMW-bilpriser.

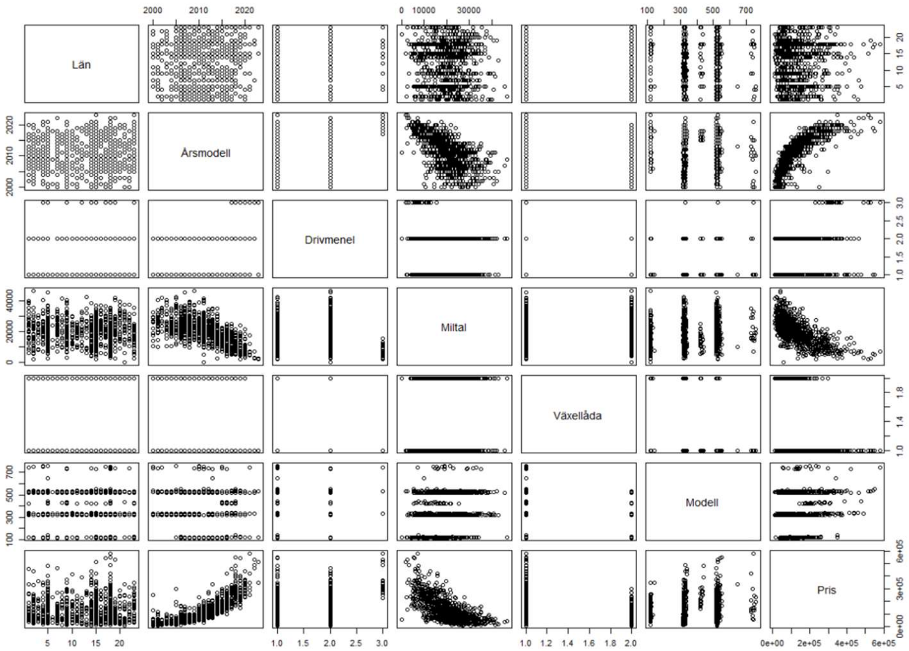
och genom använda statistiska metoder och visa på min rapportresultaten på ett tydligt sätt, jag har gett en omfattande rapport och värdefull analys som kan vara användbar till bilköpare och säljare.

3. Något du vill lyfta fram till Antonio?

Jag kan säga att den här analysen erbjuder om BMW-bilpriser också ger användbara rekommendationer för både köpare och säljare och kan mer informerade beslut.

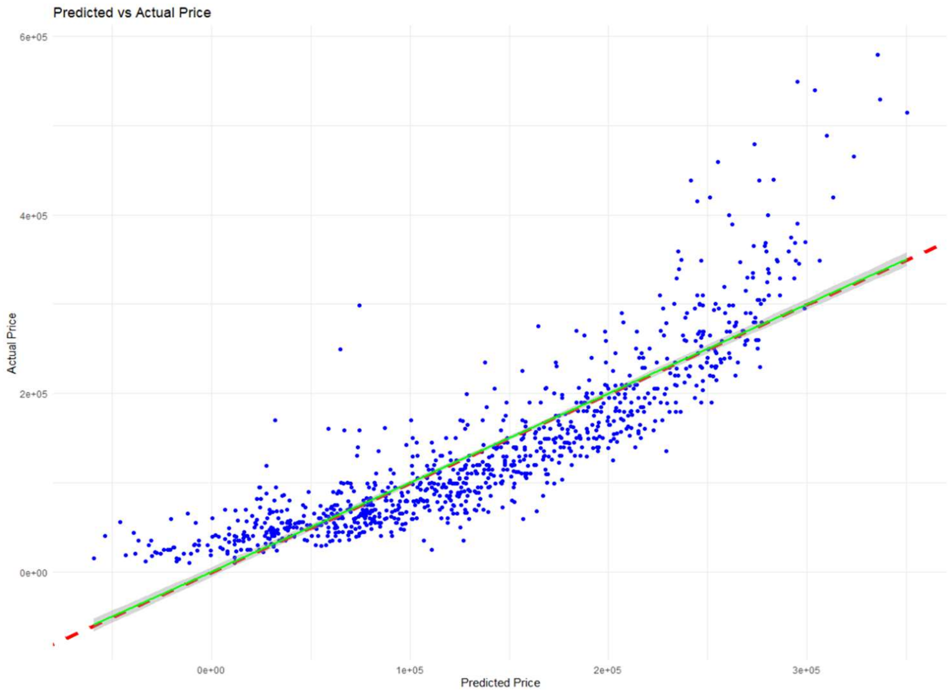
10- Appendix A

In scatter plots we can see variables in a dataset, to see a visual overview of relationships, correlations, distributions, and outliers.



In this scatter plot compares two the predicted prices align with the actual prices.

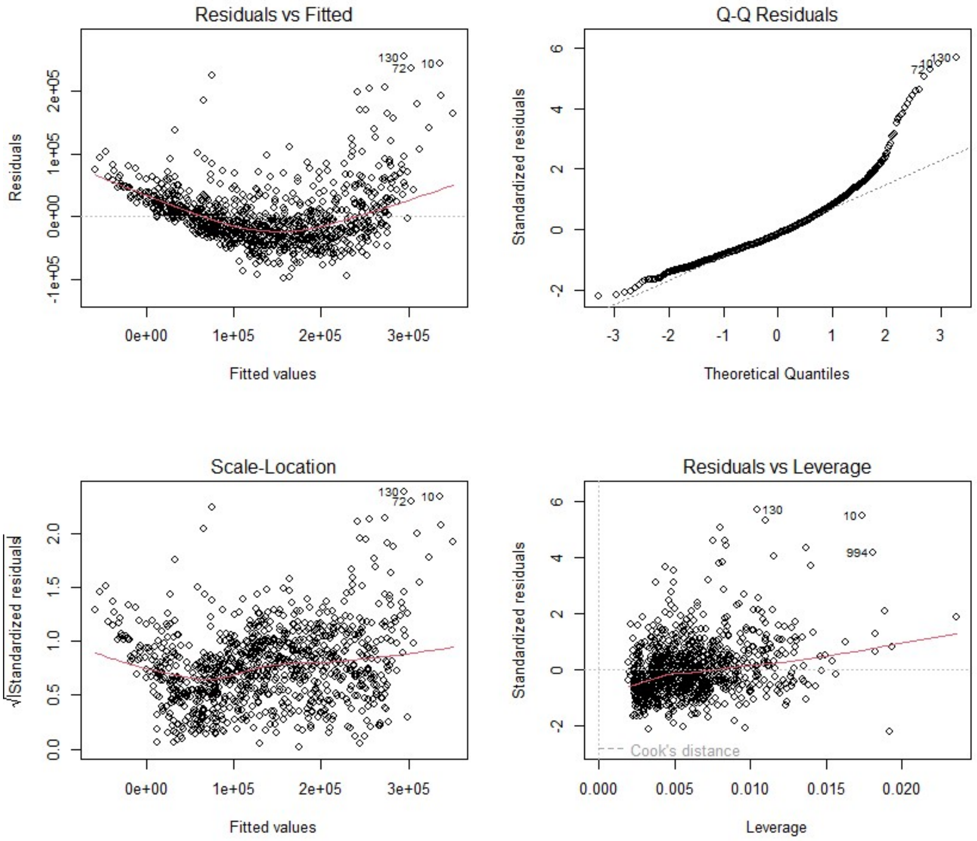
- X represents the predicted prices (Predicted)
- Y represents the actual prices (Actual)
- Blue points represent for individual data points.
- Green line we can see confidence intervals.
- Red dashed line represents perfect prediction.



A plot of distances provides into the performance and assumptions of the model.

These diagnostic plots, including:

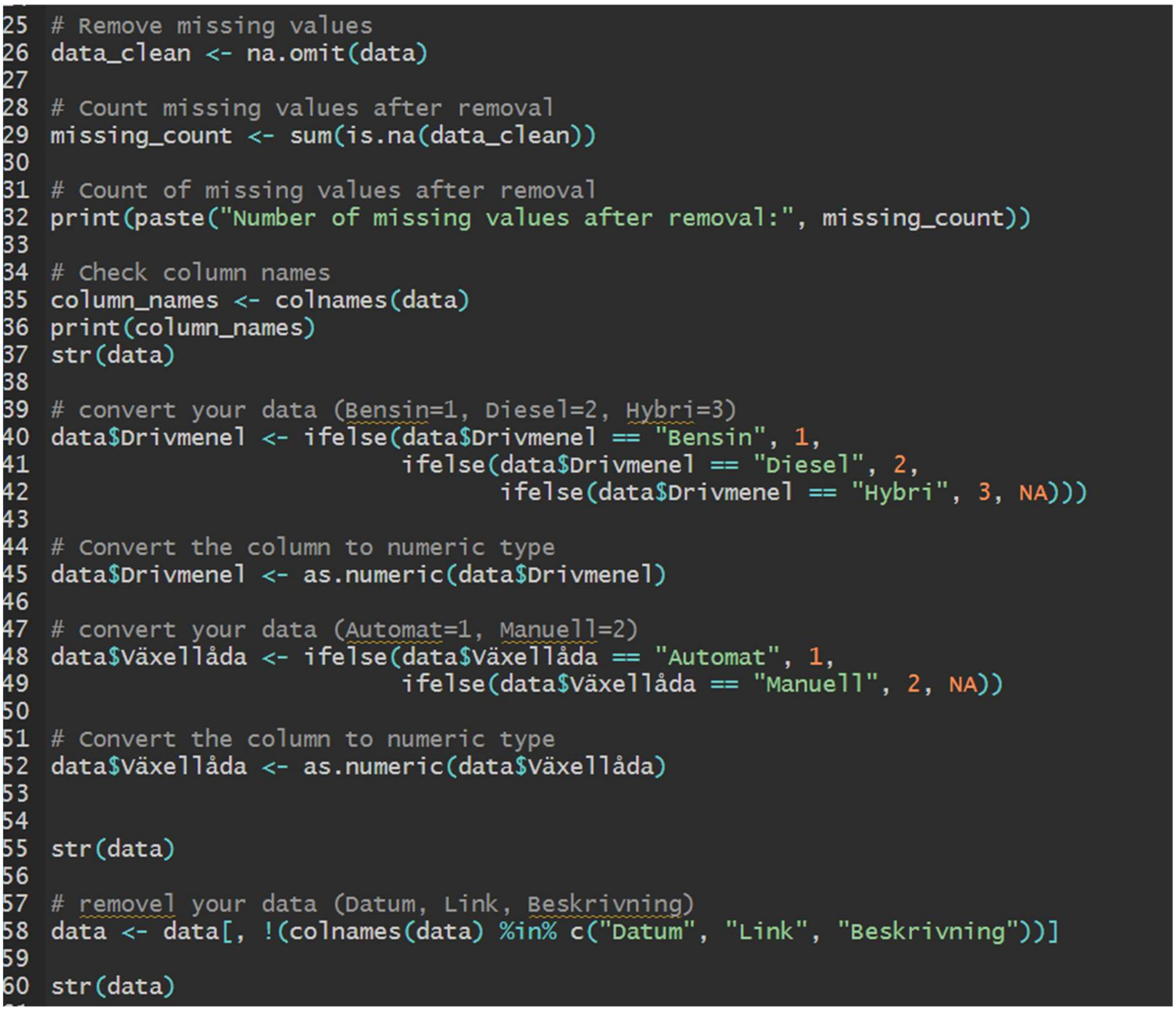
- Residuals vs. fitted.
- QQ plot of standardized residuals.
- Scale-Location
- Plot of residuals vs Leverage.



In The part of code focuses on preparing the dataset for analysis by performing cleaning and transformation steps.

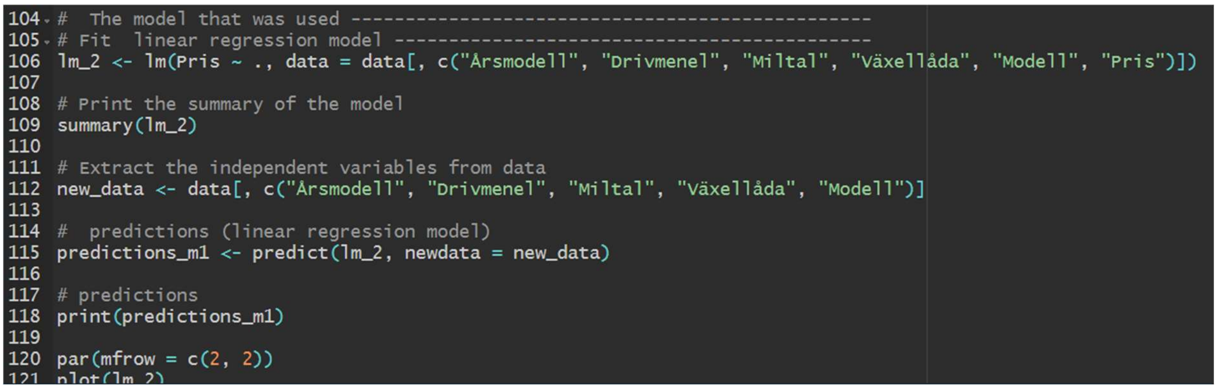
It includes the removal of missing values, conversion of categorical variables into numeric format, and removal of unnecessary columns.

This is a part of my code how the data preprocessing.



```r
# Remove missing values
data_clean <- na.omit(data)

# Count missing values after removal
missing_count <- sum(is.na(data_clean))

# Count of missing values after removal
print(paste("Number of missing values after removal:", missing_count))

# Check column names
column_names <- colnames(data)
print(column_names)
str(data)

# convert your data (Bensin=1, Diesel=2, Hybri=3)
data$Drivmenel <- ifelse(data$Drivmenel == "Bensin", 1,
                         ifelse(data$Drivmenel == "Diesel", 2,
                                ifelse(data$Drivmenel == "Hybri", 3, NA)))

# Convert the column to numeric type
data$Drivmenel <- as.numeric(data$Drivmenel)

# convert your data (Automat=1, Manuell=2)
data$Växellåda <- ifelse(data$Växellåda == "Automat", 1,
                         ifelse(data$Växellåda == "Manuell", 2, NA))

# Convert the column to numeric type
data$Växellåda <- as.numeric(data$Växellåda)


str(data)

# removel your data (Datum, Link, Beskrivning)
data <- data[, !(colnames(data) %in% c("Datum", "Link", "Beskrivning"))]

str(data)
```

In this part include model used and (fitting, summary generation, prediction computation, diagnostic plotting.etc)



```r
#  The model that was used -----------------------------------------------
# Fit  linear regression model -------------------------------------------
lm_2 <- lm(Pris ~ ., data = data[, c("Årsmodell", "Drivmenel", "Miltal", "Växellåda", "Modell", "Pris")])

# Print the summary of the model
summary(lm_2)

# Extract the independent variables from data
new_data <- data[, c("Årsmodell", "Drivmenel", "Miltal", "Växellåda", "Modell")]

#  predictions (linear regression model)
predictions_m1 <- predict(lm_2, newdata = new_data)

# predictions
print(predictions_m1)

par(mfrow = c(2, 2))
plot(lm_2)
```

Källförteckning

Education Topics Explained Channel on youtube

https://www.youtube.com/@EducationTopicsExplained

Normal QQ plot and general QQ plot

https://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Normal_QQ_plot_and_general_QQ_plot

Best subset selection

https://online.stat.psu.edu/stat462/node/197/

"All models are wrong, some are useful."

https://jamesclear.com/all-models-are-wrong

Förstå linjär regression

https://learn.microsoft.com/sv-se/shows/machine-learning-for-beginners/

AIC vs BIC: Difference and Comparison

https://askanydifference.com/difference-between-aic-and-bic/?utm_content=cmp-true

R Tutorial

https://www.statmethods.net/r-tutorial/index.html

Multiple Linear Regression

https://corporatefinanceinstitute.com/resources/data-science/multiple-linear-regression/

Hypotesprövning: Principer och metoder

https://mindthegraph.com/blog/sv_se/hypotesprovning/

Probability and Statistics

https://www.statisticshowto.com/

An Introduction to Statistical Learning

https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf