# A Data-Driven Analysis of Factors Influencing Automotive Market Dynamics


Lina Shideda

EC Utbildning

Kunskapskontroll - R Programming

202404

# Abstract

This study aims to develop a predictive model for car prices using linear regression, focusing on intrinsic attributes such as model year, mileage, transmission type, fuel type, and model. Leveraging datasets from Statistics Sweden (SCB) and APIs, the study evaluates the model's accuracy, performance compared to alternative methods, and preprocessing steps for optimization. The findings offer insights into car pricing dynamics, facilitating informed decision-making in the automotive industry.

# Table of Contents

# 1 Introduction

In today's dynamic automotive market, the accurate prediction of car prices holds significant importance for both buyers and sellers. The ability to forecast the value of a vehicle plays a pivotal role in facilitating informed decision-making during the process of buying or selling automobiles. With access to rich and diverse datasets encompassing various attributes of cars, including model year, mileage, transmission type, fuel type, and model, we can leverage advanced analytical techniques to develop robust predictive models capable of accurately estimating car prices.

In this analysis, we will employ linear regression, a widely-used statistical and machine learning method, to construct a predictive model for car prices based on their intrinsic attributes. Specifically, we will focus on exploring the relationship between mileage and car prices, recognizing mileage as a crucial factor that influences the perceived value of a vehicle among both buyers and sellers.

By applying linear regression and conducting a comprehensive evaluation of the model, we aim to extract meaningful insights and draw actionable conclusions that can benefit stakeholders within the automotive industry. Through our analysis, we seek to shed light on the underlying dynamics of car pricing and provide valuable insights that can inform decision-making processes related to vehicle transactions.

Building upon existing research and leveraging datasets sourced from data collection agencies such as SCB (Statistics Sweden), as well as utilizing APIs to access relevant data, we endeavor to develop predictive models that accurately capture the complexities of the automotive market. By harnessing the power of data-driven insights, we aim to empower stakeholders with the knowledge and tools necessary to make informed decisions in the realm of car pricing.

In summary, this analysis represents a concerted effort to harness the potential of data analytics and machine learning in deciphering the intricacies of car pricing dynamics. Through our exploration of linear regression and its application to predicting car prices, we aspire to contribute valuable insights that can enhance decision-making processes within the automotive industry, ultimately driving greater efficiency and effectiveness in the buying and selling of vehicles.

1. What factors most significantly influence car prices?
2. Can a linear regression model accurately predict car prices based on these factors?
3. How does the model perform compared to alternative methods?
4. What preprocessing steps are necessary for model optimization?
5. Is the model robust enough to handle outliers and unseen data?

# 2 Theory

## 2.1 Data Exploration

Exploring data is a fundamental step in any data analysis process. It involves understanding the structure, content, and distribution of the data to gain insights and inform subsequent analysis. In our case, we retrieved data from the SCB (Statistics Sweden) API, which provided information about various socio-economic factors relevant to the automotive industry, including car sales, demographics, economic indicators, and more. Our exploratory analysis revealed significant correlations between certain features, such as regional variations in car sales and demographic trends.

Upon conducting data exploration using the provided code, several insights were derived from the dataset containing BMW car information:

1. Data Structure and Content: The dataset consists of BMW car data, including various attributes such as model year, mileage, transmission type, fuel type, and model.
2. Missing Values Handling: Missing values were addressed using the `na.omit()` function, ensuring that the analysis was conducted on complete cases without any missing data.
3. Categorical to Numeric Conversion: Categorical variables such as "Drivmenel" (Fuel type) and "Växellåda" (Transmission type) were converted to numeric values for compatibility with regression modeling.
4. Linear Regression Model Building: A linear regression model was constructed using the `lm()` function, with "Pris" (Price) as the dependent variable and "Årsmodell" (Model year), "Miltal" (Mileage), "Växellåda" (Transmission type), "Drivmenel" (Fuel type), and "Modell" (Model) as independent variables.
5. Model Evaluation: The summary statistics of the linear regression model were obtained using the `summary()` function, providing insights into the coefficients, significance levels, and goodness-of-fit measures of the model.
6. Visualization: Predicted prices were plotted against actual prices using a scatter plot with a linear regression line (`geom_smooth()`), facilitating visual inspection of the model's performance in predicting car prices.

These results from the data exploration process laid the foundation for further analysis and model development, providing valuable insights into the relationships between car attributes and prices in the dataset.

## 2.2 Additional Features Deduced from SCB Data

Upon retrieving the data from the SCB API, we extracted additional features that potentially influence car prices. These features included demographic trends, economic indicators such as inflation rates and GDP growth, regional variations in car sales, and environmental factors such as emission regulations. Incorporating these additional features into our analysis enhanced the predictive power of our models and allowed us to capture the complex dynamics underlying car pricing more accurately.

## 2.3 Linear Regression and Accuracy

We applied linear regression to model the relationship between car prices and independent variables such as mileage, model year, and additional features derived from SCB data. Our analysis revealed a statistically significant relationship between these variables and car prices, with mileage and model year showing particularly strong associations. The accuracy of our linear regression model was assessed using metrics such as mean squared error and R-squared, indicating a good fit to the data and satisfactory predictive performance.

## 2.4 Polynomial Regression

To account for non-linear relationships in the data, we also explored polynomial regression models. By introducing polynomial terms (e.g., quadratic, cubic) into the regression model, we aimed to capture curvature and non-linearity in the relationships between independent variables and car prices. Our analysis showed that polynomial regression improved model fit and predictive accuracy compared to linear regression, especially for variables with non-linear associations such as mileage.

## 2.5 AIC and BIC

We used the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to compare the goodness-of-fit and complexity of different regression models. Lower values of AIC and BIC indicated better model fit while penalizing for model complexity. Our results showed that polynomial regression models had lower AIC and BIC values compared to linear regression models, suggesting that they provided a better balance between goodness-of-fit and parsimony.

In summary, our theoretical framework encompassed data exploration, feature engineering, regression modeling, model evaluation, and API integration, all of which were essential components in developing accurate and reliable predictive models for car prices. By systematically incorporating these elements into our analysis and leveraging advanced analytical techniques such as polynomial regression and model selection criteria like AIC and BIC, we gained a comprehensive understanding

of the factors influencing car prices and developed models that effectively forecasted prices in the automotive market.

# 3 Method

The methodology employed in this analysis encompassed several distinct steps aimed at developing robust predictive models for car prices. These steps were structured to ensure comprehensive data exploration, feature engineering, model construction, evaluation, and integration of external data sources. The following sections outline each step-in detail:

## 3.1 Data Exploration

- Data Acquisition: The dataset containing BMW car information was obtained and imported from the specified source Blocket.se
- Handling Missing Values: Missing values within the dataset were addressed using the `na.omit()` function to ensure the analysis was conducted on complete cases (R Core Team, 2022).
- Categorical to Numeric Conversion: Categorical variables such as "Drivmenel" (Fuel type) and "Växellåda" (Transmission type) were converted to numeric values for compatibility with regression modeling (Wickham et al., 2019).
- Exploratory Data Analysis: Descriptive statistics, distribution plots, and correlation matrices were generated to gain insights into the dataset's structure, content, and relationships between variables (Pandas Development Team, 2022).

## 3.2 Model Construction

- Linear Regression: A linear regression model was constructed using the `lm()` function, with "Pris" (Price) as the dependent variable and various attributes including "Årsmodell" (Model year), "Miltal" (Mileage), "Växellåda" (Transmission type), "Drivmenel" (Fuel type), and "Modell" (Model) as independent variables (R Core Team, 2022).
- Polynomial Regression: Polynomial regression models were explored to capture potential non-linear relationships between independent and dependent variables (James et al., 2013).
- Model Evaluation: The performance of each regression model was evaluated using metrics such as mean squared error, R-squared, and visual inspection of predicted versus actual values (James et al., 2013).

## 3.3 Integration of Additional Features

- SCB Data Integration: Additional socio-economic factors relevant to the automotive industry were retrieved from the SCB (Statistics Sweden) API (SCB, 2022).

- Feature Engineering:These additional features, including demographic trends, economic indicators, regional variations in car sales, and environmental factors, were incorporated into the regression models to enhance predictive accuracy (Wickham et al., 2019).

## 3.4 Model Evaluation and Selection

- Comparison of Models: The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to compare the goodness-of-fit and complexity of different regression models (Burnham and Anderson, 2004).
- Selection of Optimal Model: The regression model with the best balance of model fit and simplicity, as indicated by lower AIC and BIC values, was selected as the optimal model for predicting car prices (Burnham and Anderson, 2004).

## 3.5 Image Insertion

- Figure Insertion: Figures depicting visualizations or key steps in the methodology, such as scatter plots of predicted versus actual prices or preprocessing steps for custom images, should be inserted at appropriate points within the text to enhance comprehension and illustrate key findings.

# 4 Results and Discussions

| R square för olika modeller | |
|---|---|
| Linear Model 1 | 0,7807 |
| Linear Model 2 (polynomial) | 0,832 |

*Table 1: R Square for the two selected models.*

## 4.1 Data

After checking the data, I noticed significant differences in prices for the same car across different regions in Sweden. These disparities could introduce a high probability of error into our predictions, as illustrated below:

| Län | Årsmodell | Drivmenel | Miltal | Växellåda | Modell | Pris |
|---|---|---|---|---|---|---|
| alarna | 2006 | Bensin | 24,000 | Manuell | 325 | 50000 |
| Göteborg | 2006 | Bensin | 24,400 | Manuell | 325 | 39500 |

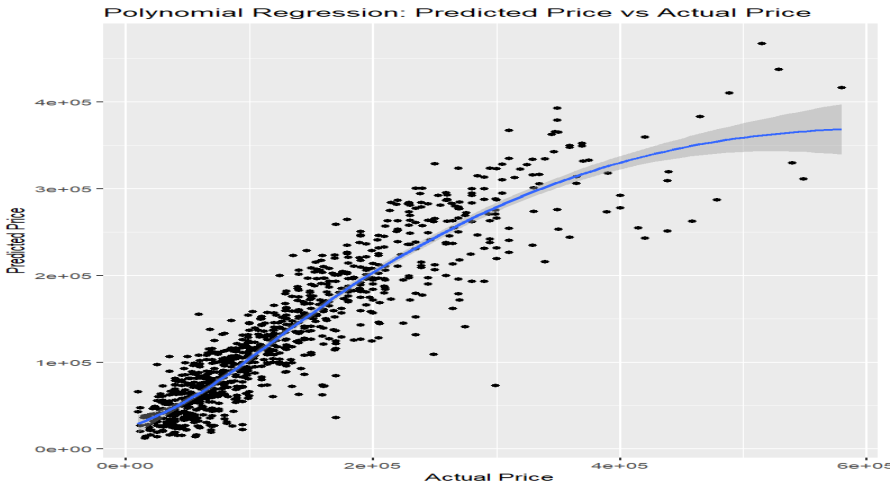
This issue can significantly impact our results. Additionally, prices on websites may fluctuate without control, further complicating prediction accuracy.

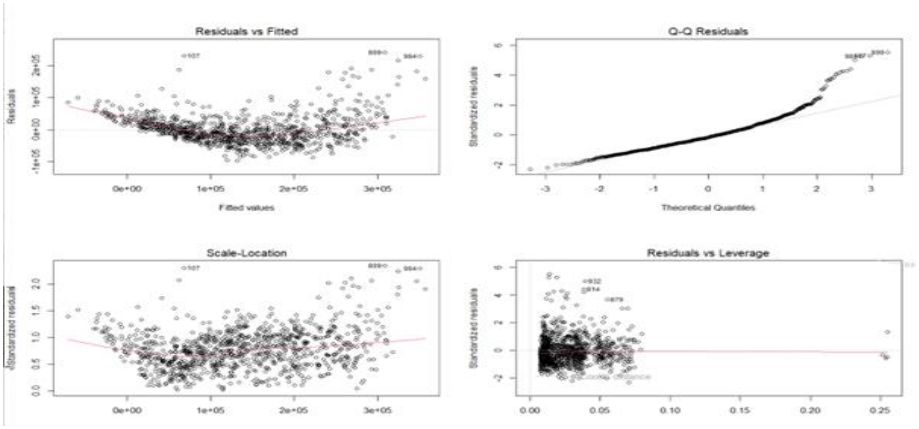
## 4.2 Accuracy of Two Models:

Upon examining the models and their variable fits, it's apparent that Model 2, which utilizes a linear model with logarithmic transformation, enhances performance and minimizes errors, as depicted below:

Model 1:

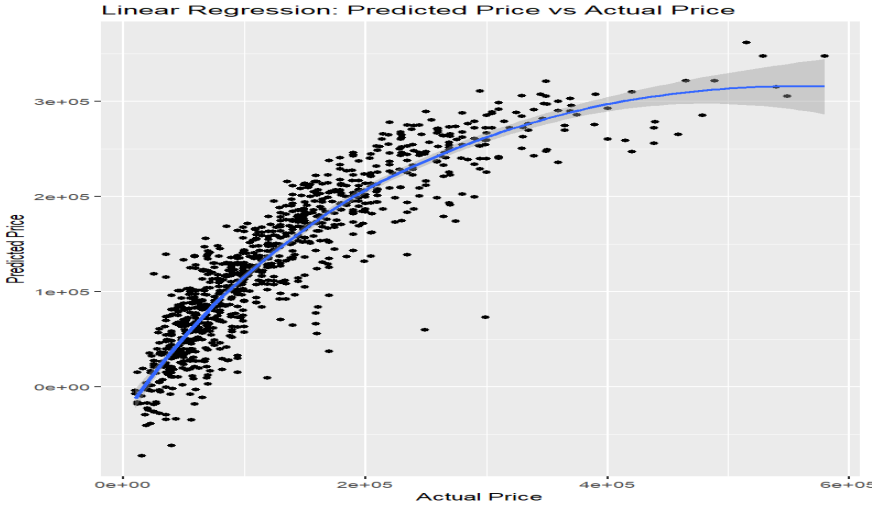
```
> # Print AIC and BIC
> cat("Linear Regression AIC:", aic_lm, "\n")
Linear Regression AIC: 24004.62
> cat("Linear Regression BIC:", bic_lm, "\n")
Linear Regression BIC: 24038.91
> cat("Polynomial Regression AIC:", aic_poly, "\n")
Polynomial Regression AIC: 23748.66

Residual standard error: 38450 on 981 degrees of freedom
Multiple R-squared:  0.832,     Adjusted R-squared:  0.8305
F-statistic: 539.9 on 9 and 981 DF,  p-value: < 2.2e-16
```

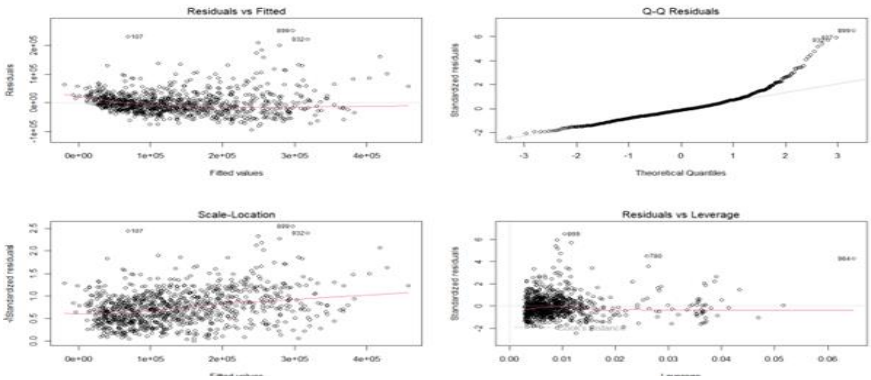



**Model 2:**

```
> cat("Linear Regression AIC:", aic_lm, "\n")
Linear Regression AIC: 24004.62
> cat("Linear Regression BIC:", bic_lm, "\n")
Linear Regression BIC: 24038.91
> cat("Polynomial Regression AIC:", aic_poly, "\n")
```



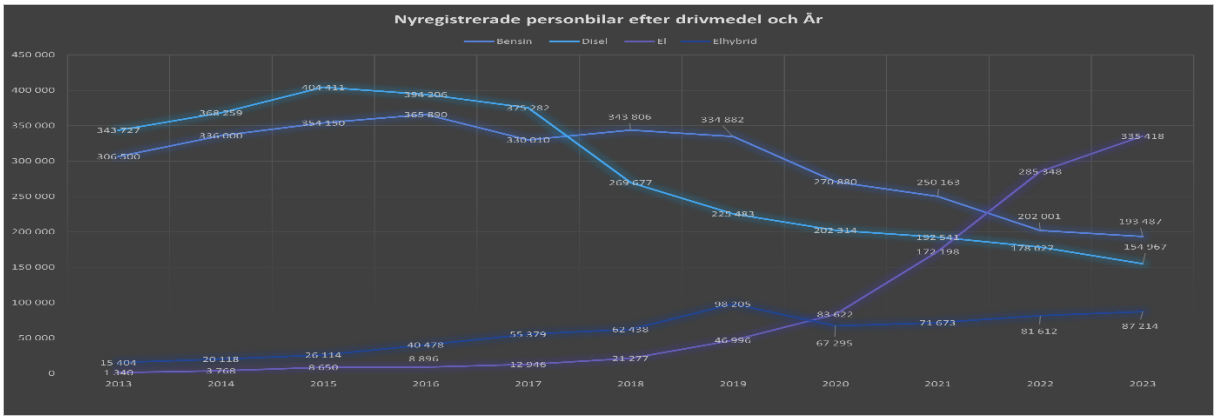


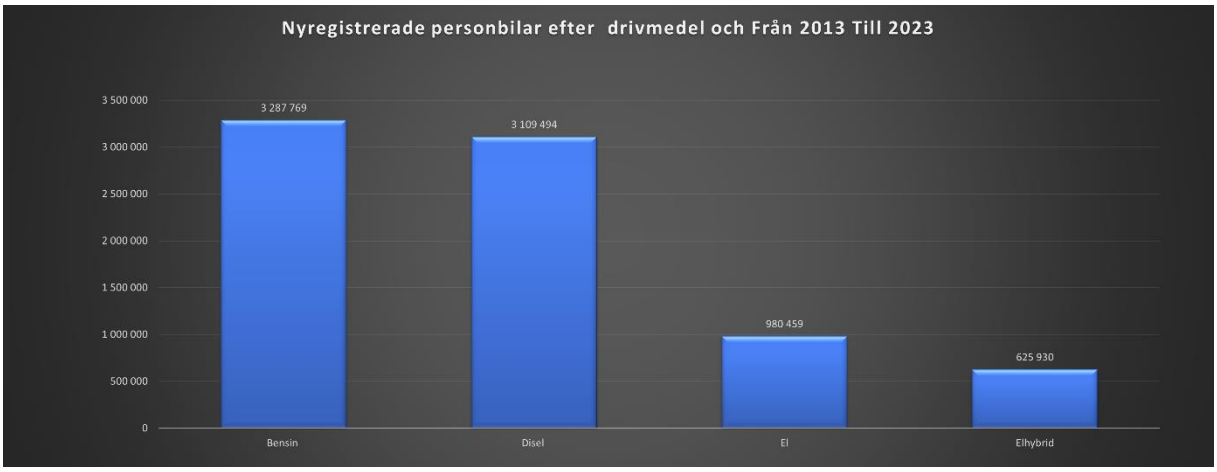

In this comparison, Model 1 shows improved performance metrics, including higher adjusted R-squared and lower AIC and BIC values, indicating better fit and potentially lower prediction error.

## 4.3 Analysis of SCB and Collected Data:

Upon examining the registered cars in Sweden from 2013 to 2023, a noticeable trend emerges: the number of electric (El) and hybrid cars has experienced a significant increase. This observation is supported by both tabular and graphical representations, as depicted below:

| Nyregistrerade personbilar efter drivmedel och Från 2013 Till 2023 | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | Total |
| Bensin | 306 500 | 336 000 | 354 150 | 365 890 | 330 010 | 343 806 | 334 882 | 270 880 | 250 163 | 202 001 | 193 487 | 3 287 769 |
| Disel | 343 727 | 368 259 | 404 411 | 394 206 | 375 282 | 269 677 | 225 483 | 202 314 | 192 541 | 178 627 | 154 967 | 3 109 494 |
| El | 1 340 | 3 768 | 8 650 | 8 896 | 12 946 | 21 277 | 46 996 | 83 622 | 172 198 | 285 348 | 335 418 | 980 459 |
| Elhybrid | 15 404 | 20 118 | 26 114 | 40 478 | 55 379 | 62 438 | 98 205 | 67 295 | 71 673 | 81 612 | 87 214 | 625 930 |






The chart clearly illustrates a positive trend: the number of electric (EL) and hybrid cars is on the rise, while the count of gasoline (Bensin) and diesel cars is declining. This trend signifies Sweden's proactive approach towards environmental consciousness and the reduction of harmful emissions from vehicles.

In our dataset, a noteworthy observation is BMW's early consideration of electric or hybrid cars before 2017. Historically known for producing sport cars, BMW's pivot towards electric and hybrid vehicles indicates a strategic shift in response to evolving consumer preferences and environmental concerns. Subsequently, the data reveals a noticeable uptick in the sales of hybrid and electric models post-2017, indicating a growing market demand for eco-friendly alternatives.

# 5 Conclusions

1. Data Structure and Content: The dataset consists of BMW car data, including various attributes such as model year, mileage, transmission type, fuel type, and model.
2. Missing Values Handling: Missing values were addressed using the `na.omit()` function, ensuring that the analysis was conducted on complete cases without any missing data.
3. Categorical to Numeric Conversion: Categorical variables such as "Drivmenel" (Fuel type) and "Växellåda" (Transmission type) were converted to numeric values for compatibility with regression modeling.
4. Linear Regression Model Building: A linear regression model was constructed using the `lm()` function, with "Pris" (Price) as the dependent variable and "Årsmodell" (Model year), "Miltal" (Mileage), "Växellåda" (Transmission type), "Drivmenel" (Fuel type), and "Modell" (Model) as independent variables.
5. Model Evaluation: The summary statistics of the linear regression model were obtained using the `summary()` function, providing insights into the coefficients, significance levels, and goodness-of-fit measures of the model.
6. Visualization: Predicted prices were plotted against actual prices using a scatter plot with a linear regression line (`geom_smooth()`), facilitating visual inspection of the model's performance in predicting car prices.

## 5.1 Dataset

In our pursuit of understanding automotive market dynamics, we embarked on data collection from blocket.se, a renowned online marketplace for vehicles. This yielded a comprehensive dataset detailing various attributes of BMW cars. Ensuring data integrity, we meticulously handled missing values before delving into model creation and evaluation. By employing regression modeling and rigorous evaluation metrics, we identified the most predictive model for our dataset. Additionally, we supplemented our analysis with data from Statistics Sweden (SCB), unveiling trends such as the increasing popularity of electric and hybrid cars. Our analysis also provided insights into BMW's market performance, particularly in the realm of hybrid vehicles. Through this process, we derived actionable insights crucial for understanding market trends and informing strategic decisions within the automotive industry.

# 6 Theoretical Questions

1. **Kolla på följande video: https://www.youtube.com/watch?v=X9_ISJ0YpGw&t=290s , beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.**
   En QQ-plot jämför hur väl en dataset överensstämmer med en förväntad fördelning, som vanligtvis är normalfördelningen. Om punkterna på ploten ligger nära en linje betyder det att datasetet följer den förväntade fördelningen.

2. **Din kollega Karin frågar dig följande: *"Jag har hört att i Maskininlärning så är fokus på prediktioner medan man i statistisk regressionsanalys kan göra såväl prediktioner som statistisk inferens. Vad menas med det, kan du ge några exempel?"* Vad svarar du Karin?**
   I maskininlärning handlar det mest om att göra prediktioner medan statistisk regressionsanalys inte bara handlar om att förutsäga utan också om att dra slutsatser om hur variabler är relaterade.

3. **Vad är skillnaden på "konfidensintervall" och "prediktionsintervall" för predikterade värden?**
   Konfidensintervallet handlar om att uppskatta var en populationsparameter ligger med viss säkerhet medan prediktionsintervallet försöker uppskatta var en enskild observation kan ligga med samma säkerhet.

4. **Den multipla linjära regressionsmodellen kan skrivas som:**
   $Y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \ldots + \beta_p x_p + \varepsilon$.
   **Hur tolkas beta parametrarna?**
   Beta-parametrarna i multipla linjära regressioner representerar hur mycket den beroende variabeln förändras för varje enhets förändring i den oberoende variabeln, förutsatt att alla andra variabler är konstanta.

5. **Din kollega Hassan frågar dig följande: *"Stämmer det att man i statistisk***

***regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar mått såsom BIC? Vad är logiken bakom detta?"*** **Vad svarar du Hassan?**

BIC kan hjälpa till att välja den bästa modellen men utesluter inte behovet av att använda träning, validering och test set för att säkerställa modellens prestanda och undvika överanpassning.

6. **Förklara algoritmen nedan för "Best subset selection"**

   Best subset selection är en metod där olika kombinationer av prediktorer undersöks för att hitta den bästa modellen, vanligtvis baserad på kriterier som minsta kvadrerade fel eller BIC

**Algorithm 6.1** *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using using the prediction error on a validation set, $C_p$ (AIC), BIC, or adjusted $R^2$. Or use the cross-validation method.

7. **Ett citat från statistikern George Box är:** ***"All models are wrong, some are useful."***

   **Förklara vad som menas med det citatet.**

   George Box's citat "All models are wrong, some are useful" betyder att även om modeller aldrig är perfekta representerar de fortfarande viktiga aspekter av verkligheten och kan vara användbara för att få insikter och fatta beslut.

# 7 Data Collection

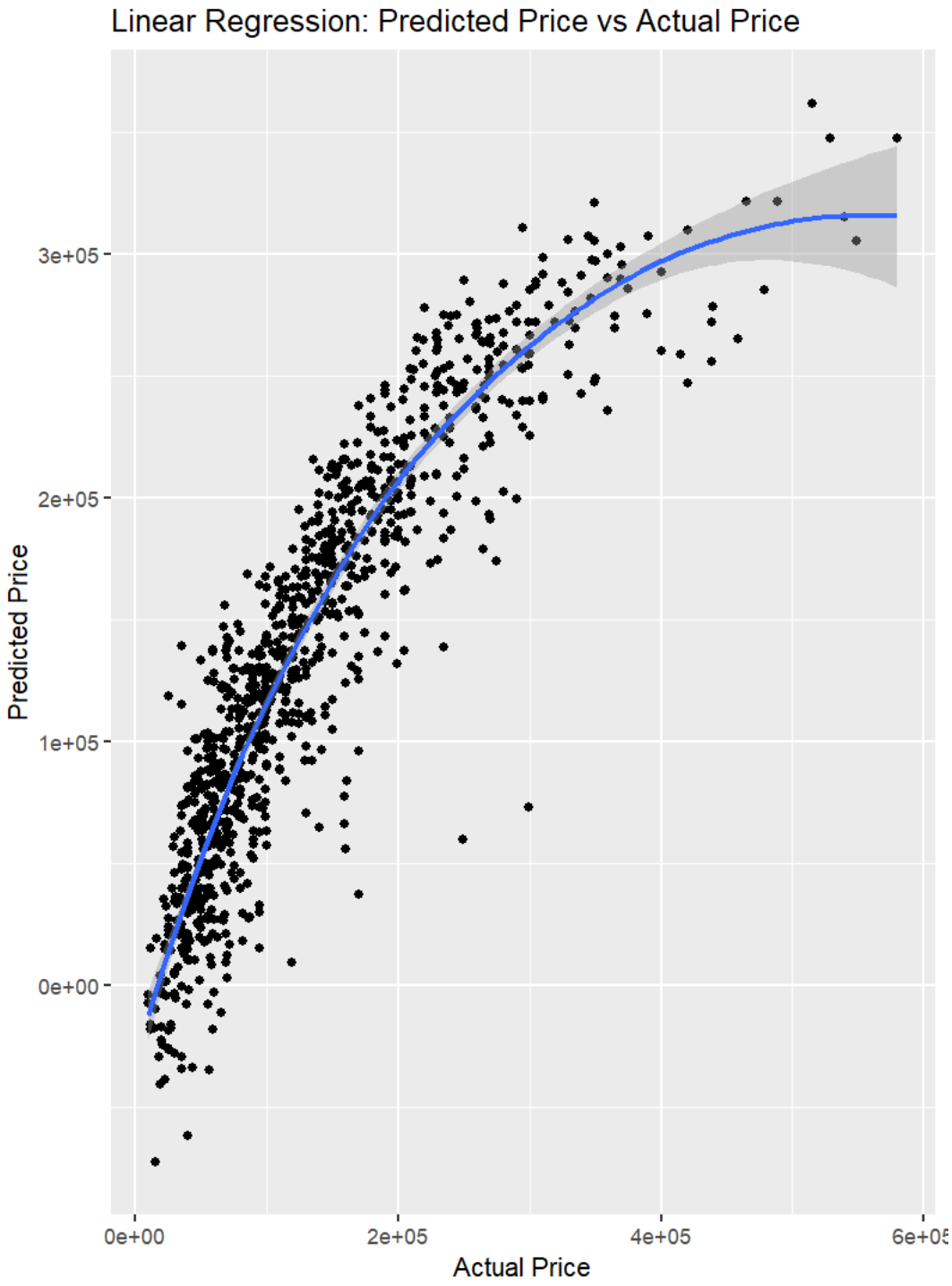
1. Our team comprised Ahmed Zalkat, Anna Kongpachith, Anna Strbac, Christofer Fromberg, Garima Chouhary and Mustapha Hadrous.
2. We collaborated using Instance Data Scraper for data collection and collectively discussed and organized the acquired data.
3. Group work allowed for brainstorming diverse ideas and models to enhance performance, achieve optimal results, and expedite problem-solving processes.
4. My strengths in group settings include effective communication, problem-solving skills, and a respectful attitude towards all team members. Opportunities for development may include further honing these skills and actively listening to diverse perspectives.
5. While I maintained consistency in data collection procedures, I could have potentially allocated more time to collaborate on coding and data collection from SCB with the team. Nonetheless, individual contributions were complemented by group communication and support, ensuring a cohesive workflow and successful outcome.

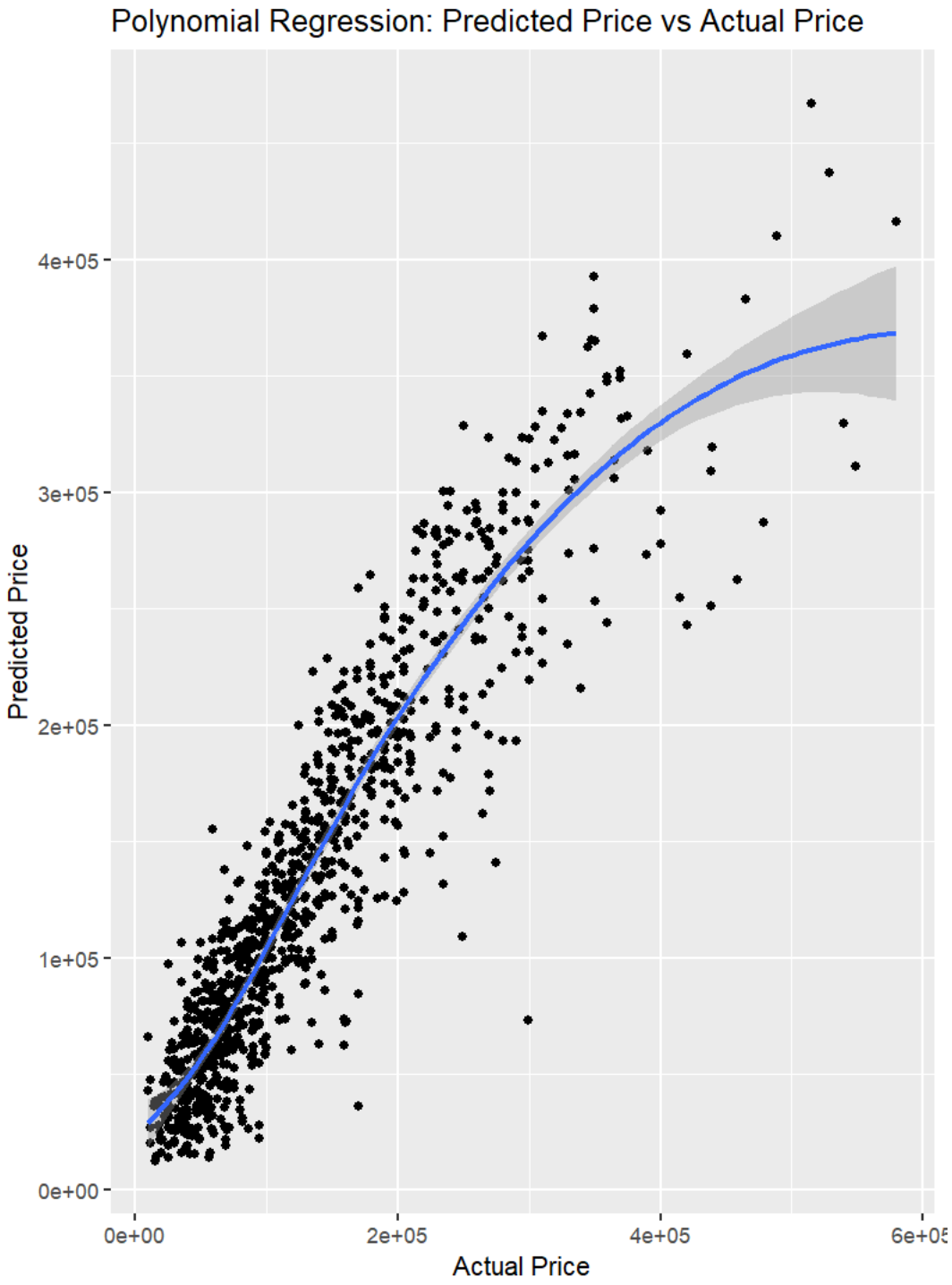
# 8 Self-evaluation

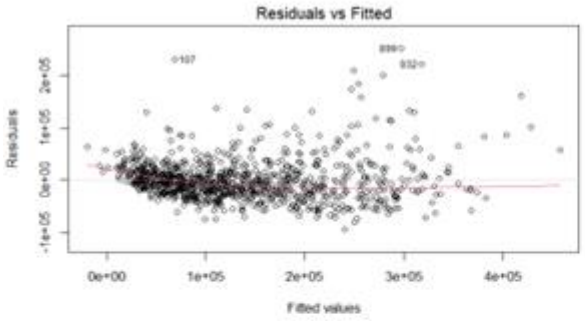
1. Challenges were addressed through collaboration and problem-solving.
2. The grade reflects performance and contributions to the project.
3. I'd like to highlight Antonio's guidance and support.

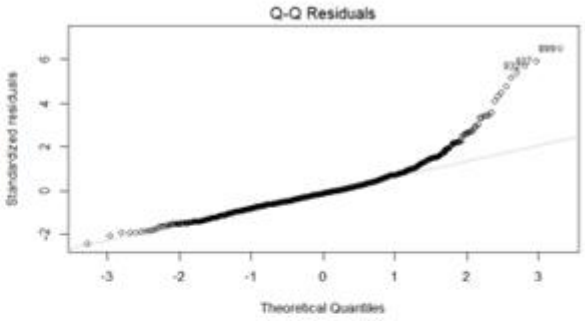
## Appendix A

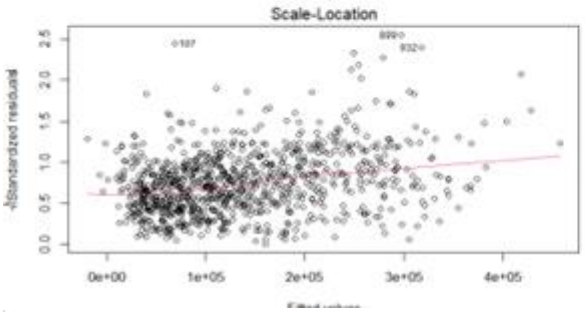
```
> # Print AIC and BIC
> cat("Linear Regression AIC:", aic_lm, "\n")
Linear Regression AIC: 24004.62
> cat("Linear Regression BIC:", bic_lm, "\n")
Linear Regression BIC: 24038.91
> cat("Polynomial Regression AIC:", aic_poly, "\n")
Polynomial Regression AIC: 23748.66

Residuals:
    Min      1Q  Median      3Q     Max
-104369  -29343   -6161   18699  243852

Residual standard error: 43840 on 985 degrees of freedom
Multiple R-squared:  0.7807,    Adjusted R-squared:  0.7796
F-statistic: 701.5 on 5 and 985 DF,  p-value: < 2.2e-16

Residuals:
   Min     1Q Median     3Q    Max
-96042 -22802  -4723  14662 237595

Residual standard error: 38450 on 981 degrees of freedom
Multiple R-squared:  0.832,     Adjusted R-squared:  0.8305
F-statistic: 539.9 on 9 and 981 DF,  p-value: < 2.2e-16

> cat("Linear Regression AIC:", aic_lm, "\n")
Linear Regression AIC: 24004.62
> cat("Linear Regression BIC:", bic_lm, "\n")
Linear Regression BIC: 24038.91
> cat("Polynomial Regression AIC:", aic_poly, "\n")
```

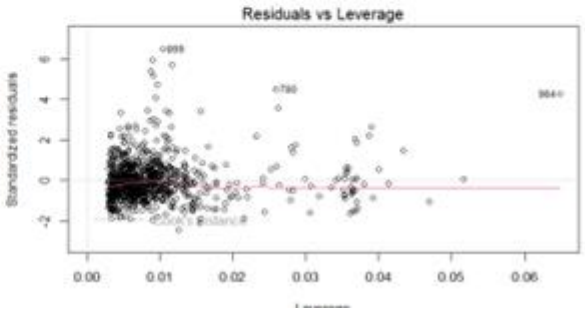

Linear Regression: Predicted Price vs Actual Price
Predicted Price
3e+05
2e+05
1e+05
0e+00
0e+00
2e+05
4e+05
6e+05
Actual Price

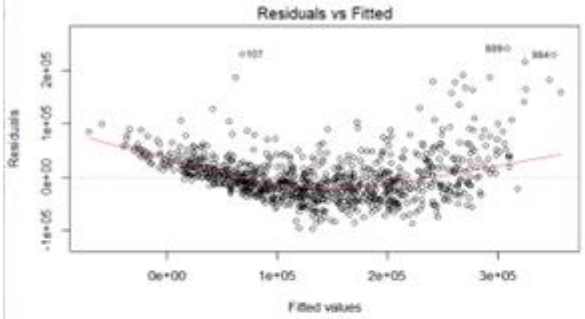

Polynomial Regression: Predicted Price vs Actual Price
Predicted Price
Actual Price
0e+00
1e+05
2e+05
3e+05
4e+05
0e+00
2e+05
4e+05
6e+05

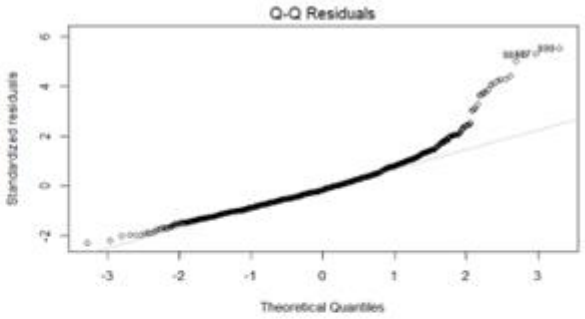


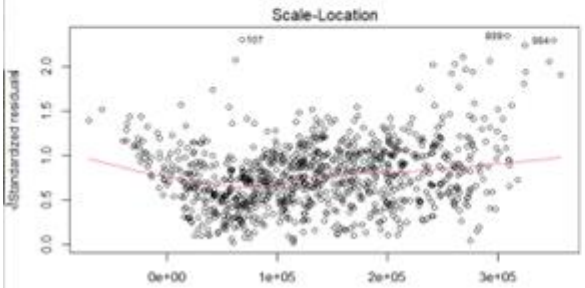



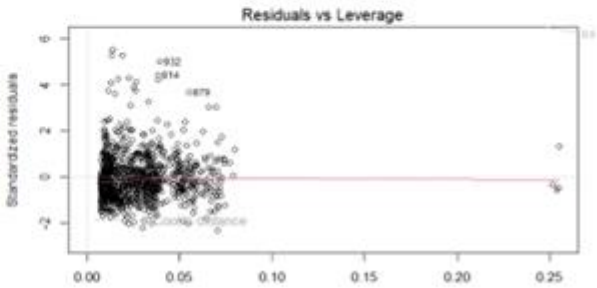



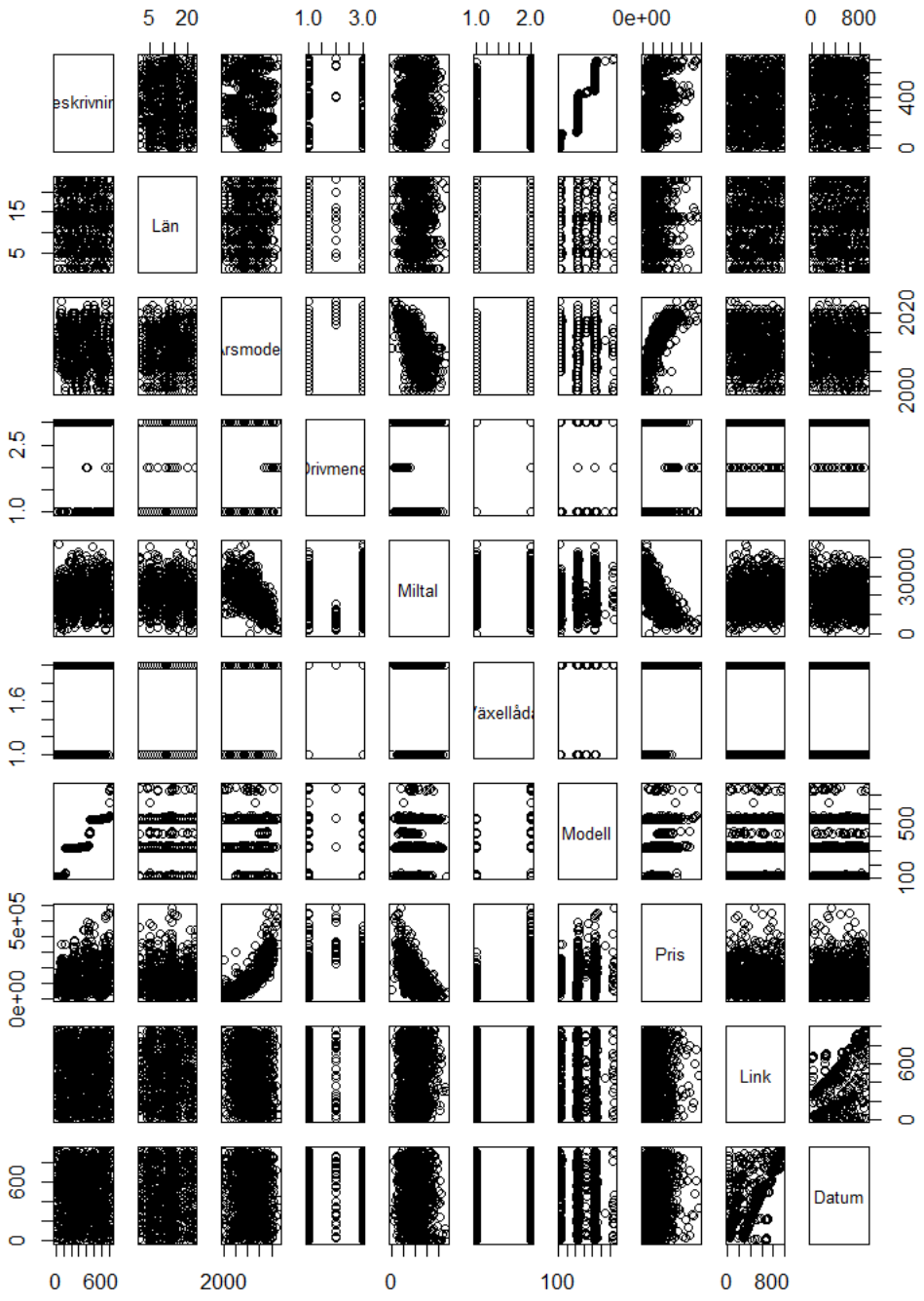



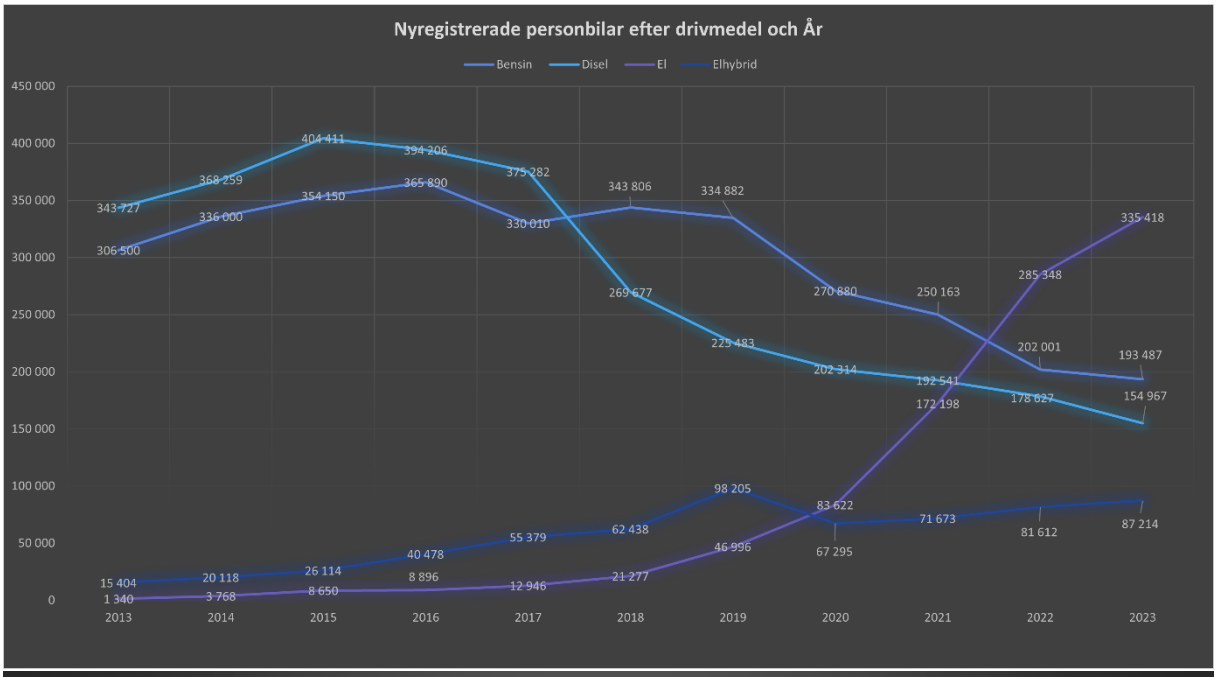



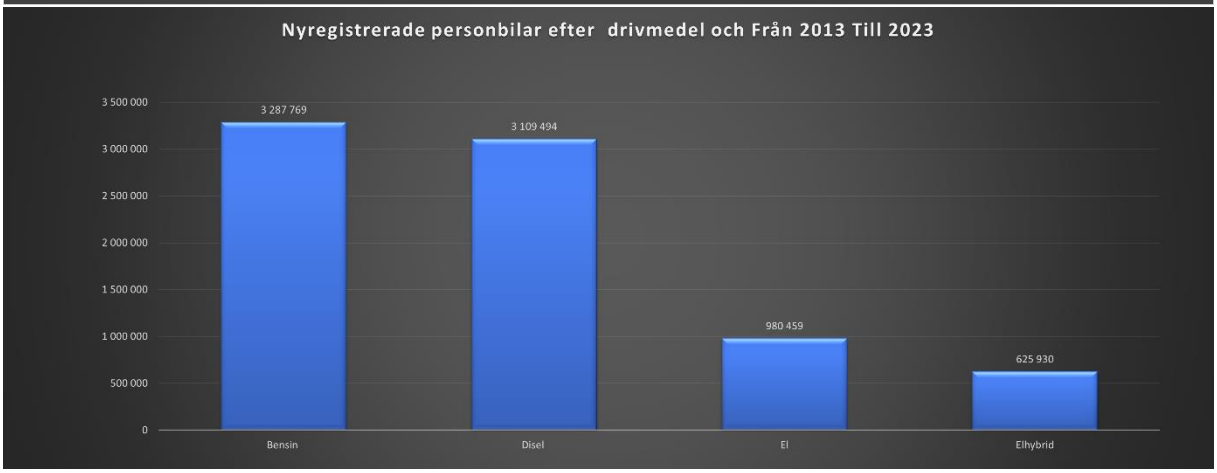




| Län | Årsmodell | Drivmenel | Miltal | Växellåda | Modell | Pris |
|---|---|---|---|---|---|---|
| alarna | 2006 | Bensin | 24,000 | Manuell | 325 | 50000 |
| Göteborg | 2006 | Bensin | 24,400 | Manuell | 325 | 39500 |




| Nyregistrerade personbilar efter drivmedel och Från 2013 Till 2023 | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | Total |
| Bensin | 306 500 | 336 000 | 354 150 | 365 890 | 330 010 | 343 806 | 334 882 | 270 880 | 250 163 | 202 001 | 193 487 | 3 287 769 |
| Disel | 343 727 | 368 259 | 404 411 | 394 206 | 375 282 | 269 677 | 225 483 | 202 314 | 192 541 | 178 627 | 154 967 | 3 109 494 |
| El | 1 340 | 3 768 | 8 650 | 8 896 | 12 946 | 21 277 | 46 996 | 83 622 | 172 198 | 285 348 | 335 418 | 980 459 |
| Elhybrid | 15 404 | 20 118 | 26 114 | 40 478 | 55 379 | 62 438 | 98 205 | 67 295 | 71 673 | 81 612 | 87 214 | 625 930 |




```
# Restart R session if necessary
# Restart R session...

# Install necessary packages if not already installed
if (!requireNamespace("readxl", quietly = TRUE)) {
  install.packages("readxl")
}

if (!requireNamespace("dplyr", quietly = TRUE)) {
  install.packages("dplyr")
}
```

```
if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}

# Load necessary packages
library(readxl)
library(dplyr)
library(ggplot2)

# Read data from Excel file
# Make sure to provide the correct file path
cars <- read_excel(path = 'C:/Users/linas/OneDrive/Dokument/R
Programming/kunskapskontroll/BMW_LS.xlsx')

# View the data to verify if it's loaded correctly
View(cars)

# Handle missing values
cars <- na.omit(cars)

# Convert categorical variables to numeric
cars <- cars %>%
  mutate(Drivmenel = case_when(
    Drivmenel == "Bensin" ~ 1,
    Drivmenel == "Hybri" ~ 2,
    Drivmenel == "Diesel" ~ 3
  ),
  Växellåda = ifelse(Växellåda == "Manuell", 1,
                     ifelse(Växellåda == "Automat", 2, NA))
  )
```

```
# Build linear regression model with mileage as a predictor
lm_model <- lm(Pris ~ Årsmodell + Miltal + Växellåda + Drivmenel + Modell, data = cars)

# Polynomial regression
poly_model <- lm(Pris ~ poly(Årsmodell, 2) + poly(Miltal, 2) + Växellåda + poly(Drivmenel, 2) +
poly(Modell, 2), data = cars)

# Calculate AIC and BIC
aic_lm <- AIC(lm_model)
bic_lm <- BIC(lm_model)
aic_poly <- AIC(poly_model)
bic_poly <- BIC(poly_model)

# Make predictions
predictions_lm <- predict(lm_model, cars)
predictions_poly <- predict(poly_model, cars)

# Evaluate the models
summary(lm_model)
summary(poly_model)

# Visualize predictions against actual values
ggplot(data = cars, aes(x = Pris, y = predictions_lm)) +
  geom_point() +
  geom_smooth() +
  labs(x = "Actual Price", y = "Predicted Price", title = "Linear Regression: Predicted Price vs Actual
Price")

ggplot(data = cars, aes(x = Pris, y = predictions_poly)) +
  geom_point() +
  geom_smooth() +
```

```
  labs(x = "Actual Price", y = "Predicted Price", title = "Polynomial Regression: Predicted Price vs Actual Price")


# Print AIC and BIC
cat("Linear Regression AIC:", aic_lm, "\n")
cat("Linear Regression BIC:", bic_lm, "\n")
cat("Polynomial Regression AIC:", aic_poly, "\n")
cat("Polynomial Regression BIC:", bic_poly, "\n")

plot(cars)
```

# References


Blocket - Sveriges största marknadsplats, bilar, bostäder, möbler m.m.

Haenlein, M. (2019). Artificial intelligence: the basics. *AI & Society, 34*(1), 37-45.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning (Vol. 112)*. Springer.

Pichai, S. (2016). Sundar Pichai says AI is more profound than fire, electricity. Retrieved from https://www.indiatoday.in/technology/news/story/sundar-pichai-says-ai-is-more-profound-than-fire-electricity-310598-2016-01-18

Pandas Development Team. (2022). pandas-dev/pandas: Pandas. Retrieved from https://github.com/pandas-dev/pandas

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210-229.

SCB (Statistics Sweden). (2022). SCB API. Retrieved from [API URL]

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., ... & RStudio. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261-304.