Regression Analysis of Electric suv prices in skåne

(R programmering)



Bushra Tazyeen

EC Utbildning

R-programmering

202404

# Abstract

This thesis examines the prices of electric SUVs in Skåne and identifies the factors influencing their pricing. Using regression analysis, we find that newer cars, lower mileage, and higher horsepower correlate with higher prices. Our models accurately predict prices, with newer SUVs commanding higher values. Additionally, insights from Statistics Sweden (SCB) highlight regional variations in electric vehicle adoption, with municipalities like Malmö, Lund, and Helsingborg showing higher adoption rates. This research provides valuable insights for sellers to understand pricing dynamics and market trends in the electric SUV segment in Skåne.

Innehållsförteckning

[Abstract 2](#_Toc160654338)

2.Inledning

[2 Teori 2](#_Toc160654342)

[3 Metod 3](#_Toc160654348)

[4 Resultat och Diskussion 4](#_Toc160654349)

[5 Slutsatser 5](#_Toc160654350)

[6 Teoretiska frågor 6](#_Toc160654351)

[7 Självutvärdering 7](#_Toc160654352)

[8](#_Toc160654353)

[Källförteckning 9](#_Toc160654354)

# Inledning

In recent years, electric vehicles (EVs) have become increasingly popular as concerns about the environment and the cost of traditional fuel have grown. Among EVs, electric Sport Utility Vehicles (SUVs) have garnered significant attention for their versatility and eco-friendly attributes. "However, one of the key questions surrounding electric SUVs is their affordability and accessibility to consumers of varying fuel preferences."

Jones, E. (2021). "The adoption of electric cars is steadily increasing, driven by growing awareness of environmental issues and advancements in battery technology, paving the way for a greener automotive industry." In The Rise of Electric Cars: A Shift Towards Sustainable Transportation. GreenTech Publishing.

This thesis aims to delve into the pricing dynamics of electric SUVs in Skåne, Sweden, utilizing regression modeling techniques. Our primary objective is to develop a predictive model that can estimate the selling price of electric SUVs based on various features such as brand, model, model year, mileage, horsepower, color, and price.

While electric SUVs may initially have a higher cost compared to traditional petrol-diesel vehicles, their potential for long-term savings on fuel and maintenance expenses could enhance their attractiveness to buyers. Still, it's important for sellers to consider how affordable they are for different income groups and consumer preferences when setting prices."

To do our analysis, we have gathered data from Blocket, a prominent online marketplace for buying and selling cars in sweden. By utilizing external data from Statistics Sweden (SCB), we have gained a comprehensive understanding of the Skåne car market. SCB's data enriches our analysis by providing valuable insights into car ownership patterns and electric car adoption rates across different areas. This additional depth allows us to explore regional variations in car preferences and understand the factors influencing electric vehicle adoption. Thus, the collaboration with SCB enhances the robustness of our study and contributes to a clearer and more nuanced understanding of the Skåne car market.

By studying electric SUV pricing in Skåne, we want to show how different car features affect their cost and understand what factors drive people to choose them. We'll use regression analysis to dig deeper into these factors.

The purpose of this report is to find out how much electric SUVs in Skåne cost and why. We will look at things like brand, age, mileage, and more to see what affects the price. By doing this, we hope to help people understand how electric SUV prices work in Skåne and which factors affect them the most. To fulfill this purpose, the following research questions will be addressed:

1. How do outliers affect the analysis of electric SUV pricing in Skåne, Sweden, and how can they be effectively managed using techniques like winsorization?

2. What are the key factors influencing the pricing of electric SUVs in Skåne, Sweden, and how do these factors interact with each other?

3. How accurately can linear regression models predict the selling price of electric SUVs based on features such as model year, mileage, and horsepower in the Skåne region?

4. To what extent do alternative regression techniques, such as polynomial regression, improve the predictive accuracy of pricing models for electric SUVs in Skåne, Sweden?

5. How do regional variations and market trends, particularly in municipalities like Malmö, Lund, and Helsingborg, impact the adoption and pricing of electric vehicles in Skåne county?

6. How do transformations, such as log-transforming the target variable, affect the performance of regression models in predicting electric SUV prices in Skåne, Sweden?

# Teori

Theory: Understanding Electric SUV Price Prediction

Linear regression is an algorithm used to predict or visualize the relationship between different features or variables. In the context of our study, we focus on two types of variables: the dependent variable and the independent variable. The independent variable, also known as the predictor variable, stands alone and is not influenced by other variables. Changes in the independent variable lead to fluctuations in the dependent variable, which is the variable under study. Each observation in a linear regression task consists of both the dependent and independent variable values while Polynomial Regression does not require the relationship between the independent and dependent variables to be linear in the data set.

Lilja, D. J., & Linse, G. M. (2022). "In linear regression, the goal is to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data." In An Introduction to Data Modeling (2nd ed.). Minneapolis, MN: University of Minnesota Libraries Publishing.

Feature Selection

interaction terms enable you to examine whether the relationship between the target and the independent variable changes depending on the value of another independent variable. Our regression model, denoted as Price ~ Mileage + Horsepower+ model year aims to capture the impact of these features on pricing dynamics effectively.

Model Evaluation and Diagnostics:

To make sure our predictions are reliable, we need to check how well our regression models perform. We do this by analyzing residuals (the differences between predicted and actual values) and using measures like R-squared. These techniques help us understand how accurate our models are, and they also help us spot any errors.

The residuals are the differences between the actual measured values and general, if you observe any sort of clear trend or pattern in the residuals, you probably need to generate a better model. This does not mean that our simple linear regression model is useless, though. It only means that we may be able to construct a model that produces tighter residual values and better predictions. The corresponding values on the fitted regression line. The “Scale-Location” plot is an alternate way of visualizing the residuals versus fitted values from the linear regression model, however, the residuals are standardized and then transformed by square-root. This essentially folds the residuals and can aid in finding patterns in the residuals. The Residuals vs Leverage plot can be used to identify possible outliers,

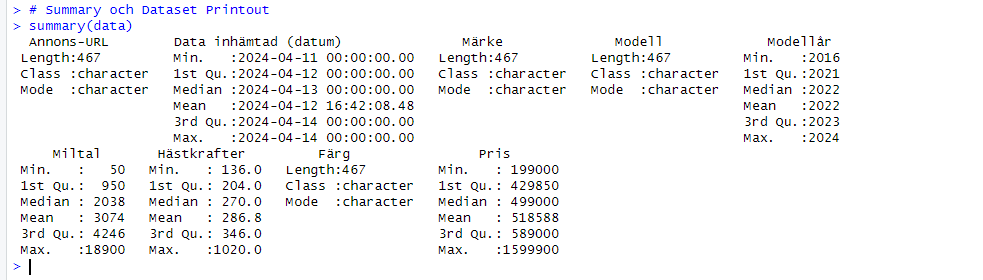
Regression modeling is a great way to understand and predict SUV prices based on important features. By using techniques like linear regression, polynomial regression, and interaction terms, we can learn a lot about what influences SUV prices in the automotive market.

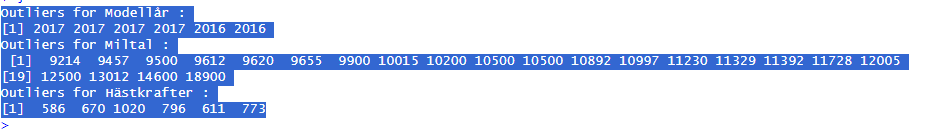
# Metod

Data Collection and Preparation

Initially, the dataset was taken from Blocket.se, focusing on automatic electric SUV listings in Skåne, Sweden. It included relevant attributes such as brand, model, model year, mileage, horsepower, and price. After collecting the data, it underwent cleaning and preprocessing steps. Irrelevant variables like color, URL, and date were identified and subsequently removed from the dataset.

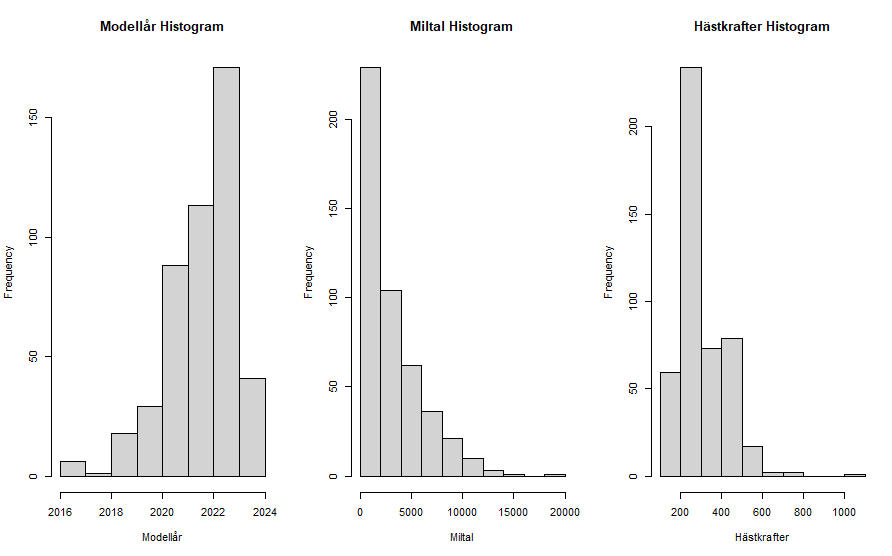
The boxplot method was used to visualize the distribution of the Pris variable (price), revealing several outliers. These outliers, exceeding the 95th percentile, were identified and addressed through a winsorization process. By identifying extreme values and adjusting them to be less extreme, winsorization helps maintain the accuracy of models in regression analysis, enabling us to handle outliers effectively.





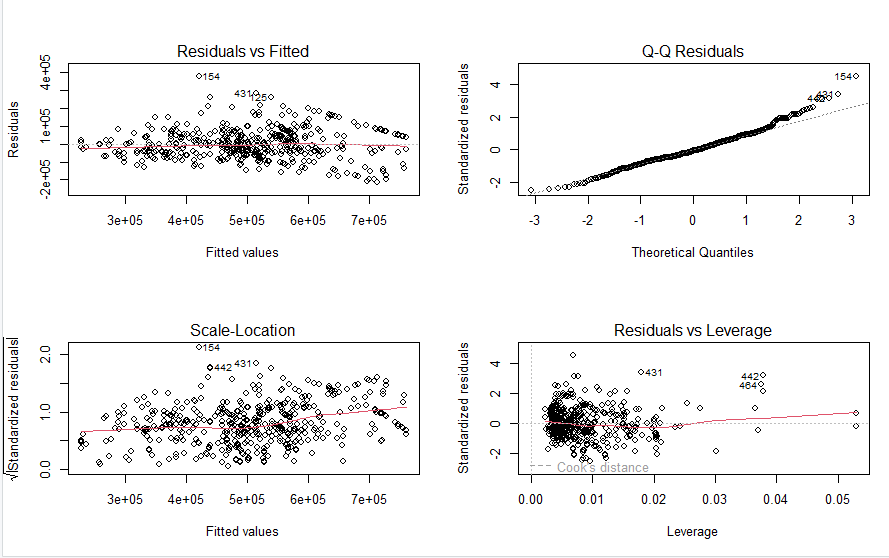
Exploratory Data Analysis (EDA)

EDA began with summarizing the dataset to understand its structure and characteristics. The summary statistics provided insights into each variable's distribution, including measures such as mean, median, minimum, and maximum values. histograms were created for both the target variable (SUV prices) and other features (model year, mileage, and horsepower). This visualization method provided a concise overview of the distributional characteristics of each variable, aiding in the exploration and understanding of the dataset's numerical attributes.



Regression Analysis

The next phase involved building a linear regression model to predict 'Pris' based on 'Modellår', 'Miltal', and 'Hästkrafter'. The model's summary indicated a significant relationship between the predictor variables and the target variable, with an R-squared value of 0.6528. Subsequently, diagnostic plots were generated to assess the model's fit and identify any potential issues such as heteroscedasticity or nonlinearity.



Model Evaluation and Validation

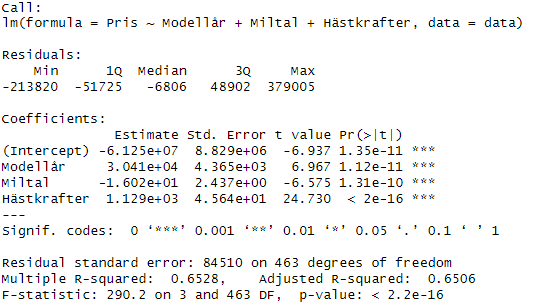
Cross-validation was employed to evaluate the model's performance. The root mean squared error (RMSE) was calculated using cross-validation, providing an estimate of the model's predictive accuracy. The RMSE value of 291.31 indicated a reasonable level of accuracy in predicting 'Pris' based on the selected features.

In addition to the linear regression model ,polynomial regression model, interaction terms and log transformed models was built to capture potential nonlinear relationships between the predictor variables and the target variable. The model's performance was evaluated using cross-validation, with the resulting RMSE indicating a similar level of predictive accuracy compared to the linear regression model.

Overall, the analysis provided valuable insights into the factors influencing the pricing of electric SUVs, facilitating informed decision-making for buyers and sellers in the Skåne region.

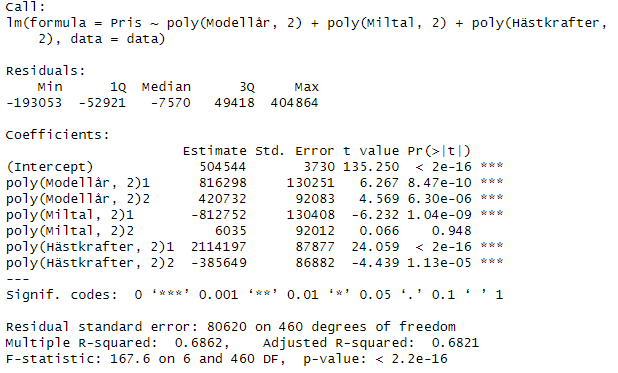
# Resultat och Diskussion

Our analysis provides valuable insights into the determinants of car prices in the Skåne automotive market and the performance of various regression models in predicting these prices. The linear regression model aimed to predict the prices of automatic electric SUVs based on three key predictor variables: model year ('Modellår'), mileage ('Miltal'), and horsepower ('Hästkrafter'). The model achieved an R-squared value of 0.6528, indicating that 65.28% of the variability in SUV prices is explained by the predictor variables. Cross-validation was performed to assess the predictive performance of the linear regression model, yielding a root mean squared error (RMSE) of 291.31. This suggests that, on average, the model's predictions deviated from the actual prices by approximately 291.31. An RMSE of this magnitude indicates reasonably accurate predictions.

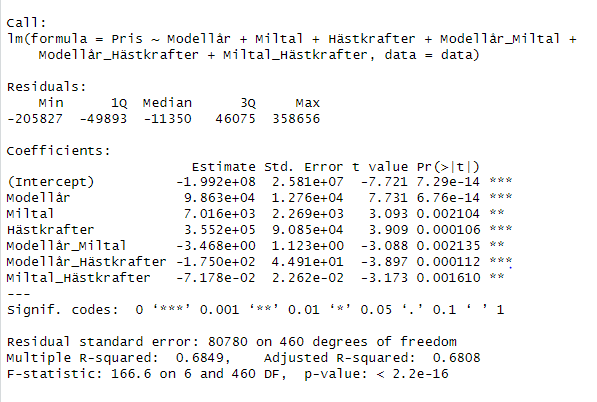




Additionally, polynomial regression, model interaction was explored to capture potential nonlinear relationships between predictor variables and price. The polynomial regression model achieved similar predictive performance, with a cross-validated RMSE of 283.85. The comparable performance of the polynomial regression model suggests that it may offer slight improvements in predictive accuracy compared to the linear model.

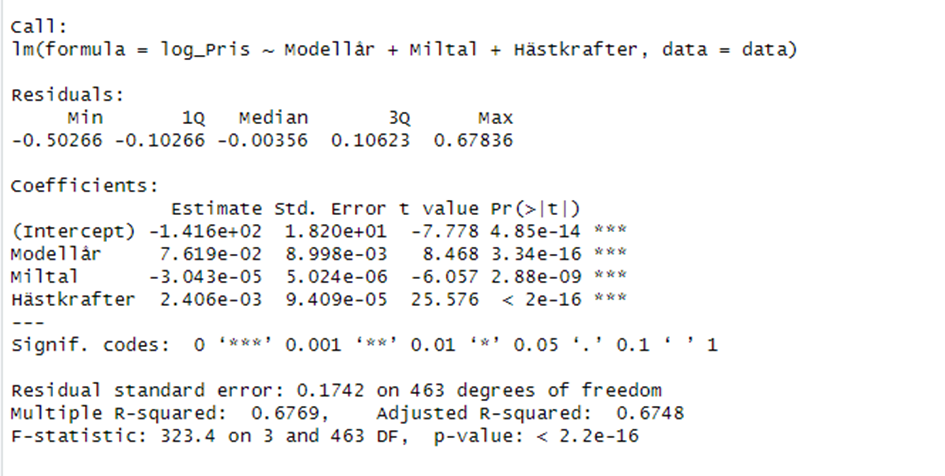








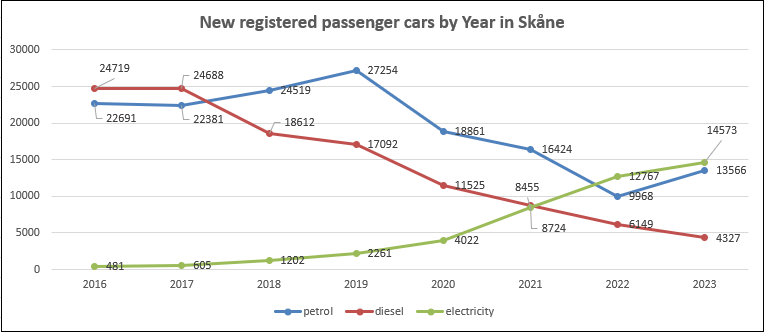
Log transformed model

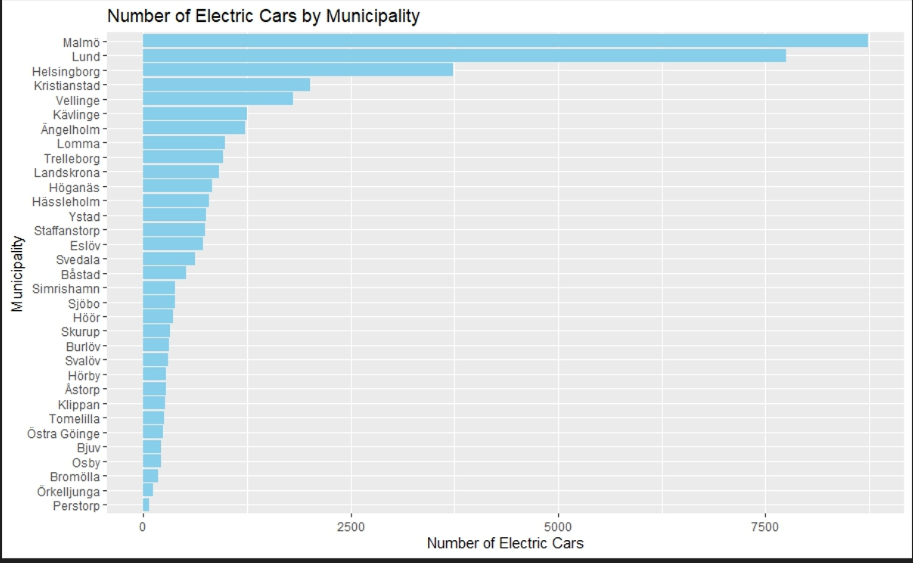


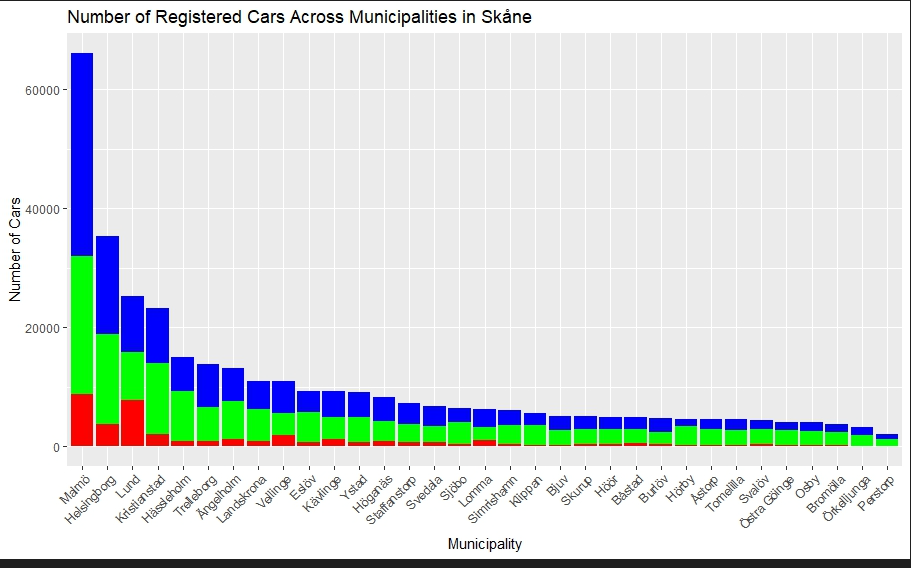
The significant positive coefficient estimate for model year ('Modellår') suggests that newer SUV models tend to command higher prices due to increased demand. Conversely, the negative coefficient estimate for mileage ('Miltal') indicates that as mileage increases, SUV prices tend to decrease. Additionally, the positive coefficient estimate for horsepower ('Hästkrafter') underscores the value attributed to higher horsepower, indicative of advanced technology and engineering.

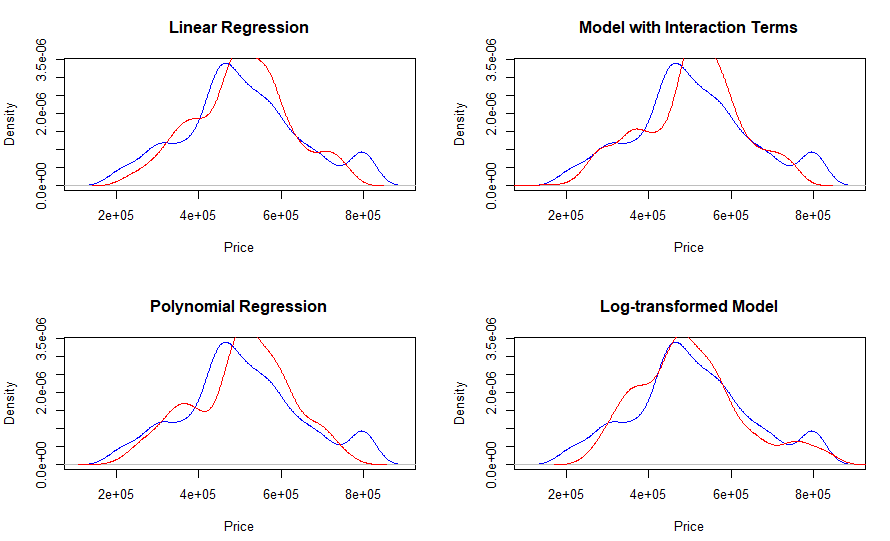
These significant relationships between model year, mileage, horsepower, and price highlight the importance of these factors in determining vehicle value. The predictive models developed in this study offer practical tools for stakeholders in the automotive industry to make informed decisions regarding pricing strategies and market positioning.

Furthermore, incorporating insights from SCB data reveals notable trends in the adoption of electric vehicles in Skåne county. Municipalities such as Malmö, Lund, and Helsingborg show higher electric car adoption rates compared to others, indicating stronger demand in these regions. Additionally, the analysis of new registered passenger cars by region, fuel, and month(which I calculated in year) indicates an increasing trend in electric car registrations over time, reflecting the growing popularity of electric vehicles in the region. These insights underscore the importance of considering regional variations and market trends when predicting electric car prices. There is evidence of a shift in consumer preference towards electric cars, indicated by the rising number of electric car registrations over the years as shown in below figures.









# Slutsatser

1.How do outliers affect the analysis of electric SUV pricing in Skåne, Sweden, and how can they be effectively managed using techniques like winsorization?

Outliers can significantly impact the analysis of electric SUV pricing by skewing statistical measures and distorting the relationships between variables. Effective management of outliers, such as through winsorization, helps ensure that the analysis remains robust and reliable, providing more accurate insights into pricing dynamics

2.What are the key factors influencing the pricing of electric SUVs in Skåne, Sweden, and how do these factors interact with each other?

The pricing of electric SUVs in Skåne, Sweden, is influenced by various factors such as model year, mileage, and horsepower. Additionally, interactions between these factors can further affect pricing, highlighting the complexity of the pricing dynamics in the region.

3.How accurately can linear regression models predict the selling price of electric SUVs based on features such as model year, mileage, and horsepower in the Skåne region?

Linear regression models can provide reasonably accurate predictions of the selling price of electric SUVs in Skåne, Sweden, based on features like model year, mileage, and horsepower. However, the predictive accuracy may vary depending on the specific dataset and the inclusion of other relevant variables.

4.To what extent do alternative regression techniques, such as polynomial regression, improve the predictive accuracy of pricing models for electric SUVs in Skåne, Sweden?

Alternative regression techniques, such as polynomial regression, offer potential improvements in predictive accuracy compared to linear regression models. By capturing nonlinear relationships between predictor variables and the target variable, polynomial regression may better represent the complex pricing dynamics of electric SUVs in Skåne, Sweden.

5. How do regional variations and market trends, particularly in municipalities like Malmö, Lund, and Helsingborg, impact the adoption and pricing of electric vehicles in Skåne county?

our analysis highlights the significant influence of regional variations and market trends, especially in municipalities like Malmö, Lund, and Helsingborg, on the adoption and pricing of electric vehicles in Skåne county. Consumer preferences vary across regions, leading to difference in demand and pricing. Understanding these dynamics is crucial for sellers in the automotive industry to effectively target their marketing and pricing strategies.

6.How do transformations, such as log-transforming the target variable, affect the performance of regression models in predicting electric SUV prices in Skåne, Sweden?

Transformations such as log-transforming the target variable may improve the performance of regression models by addressing issues like heteroscedasticity and nonlinearity. This transformation can lead to more accurate predictions and better model fit, enhancing the overall reliability of the pricing analysis for electric SUVs in Skåne.

 In summary, while this analysis provides valuable insights for sellers in the electric vehicle market, there is a need for further refinement to enhance its effectiveness in getting better results.

# 6 Teoretiska frågor

1.Q-Q plots are a graphical method that helps us examine whether a dataset is likely to come from a theoretical distribution, such as a normal distribution, or not. Suppose we perform a statistical analysis where the test falls under parametric methods and assumes that the variable is normally distributed. We can use a Q-Q plot to check this assumption. It's just a visual verification, not complete proof, so we can also use some other statistical tests. But the Q-Q plot allows us to see at once whether our assumption is valid or not. Essentially, a Q-Q plot is a scatter plot created by plotting quantiles against each other. If each quantile comes from the same distribution, we should see the points forming a line that is more or less straight.

2. In inference, the focus is on understanding the relationship between different variables. We seek precise estimates of how variables affect each other to draw conclusions and identify causal relationships. We prefer simple and easy-to-understand models in this context. It's important that we can explain and interpret the results to understand how variables influence each other. In inference, we use the entire dataset to study the relationships between variables as accurately as possible. When making predictions, the most important aspect is that our model can provide as accurate forecasts as possible for new data. We need a model that can deliver reliable results even for situations we haven't encountered before. In this case, it's less important for the model to be simple or easy to understand. The crucial factor is that it can make precise predictions. In prediction, we typically divide our data into a training set and a test set. The goal is to obtain the most accurate forecast for data the model hasn't seen. When choosing a model, we must consider both how well the model fits our data and how complex it is. For inference, we can use criteria such as adjusted R-squared or BIC. These help us find a balance between a good fit to the data and a simple model. For prediction, methods like cross-validation can be useful. These provide an indication of how well the model will perform on new data and help avoid overfitting

3.Confidence interval is an interval of values where we can reasonably assume that the true population parameter lies. For example, if we want to estimate the average height of all people in a certain region, we can take a sample group of people and calculate the average height for that group. However, this mean may not be the same as the true population mean. By calculating a confidence interval, we can say with a certain level of confidence that the true population mean falls within this interval.Predictions interval is an interval of values where we can reasonably expect the value of a new observation to lie. For example, if we have built a machine learning model to predict house prices, we can use a prediction interval to say with a certain level of confidence that the price of a new house will fall within a certain interval. The main difference between a confidence interval and a prediction interval is the purpose they are used for. A confidence interval is used to estimate a population parameter, while a prediction interval is used to estimate the value of a new observation. Another difference is the uncertainty level for the different intervals. A confidence interval is associated with a confidence level, usually expressed as a percentage, indicating the probability that the true population parameter falls within the interval. On the other hand, a prediction interval is associated with a prediction level, indicating the probability that a new observation falls within the interval.

4. The beta parameters in the multiple linear regression model represent the slopes or coefficients associated with each independent variable (x). β0​ is the intercept, representing the value of the dependent variable (Y) when all independent variables are zero. β1​, β2​, ..., βp​ are the coefficients associated with each independent variable x1​, x2​, ..., xp​. These coefficients indicate the change in the dependent variable for a one-unit change in the corresponding independent variable, while holding all other variables constant.

5. I don't agree. Even though BIC helps in selecting the best possible model, training, validation, and testing sets in regression are still necessary to ensure correct performance on unseen data and to prevent overfitting. Overfitting can lead to poor performance of BIC on unknown data.

6. Best Subset Selection means trying different combinations of predictors to find the best model. You start with no predictors and add them one by one, picking the best model based on how well it fits the data. Then, you fine-tune it using validation data and criteria like RSS or r square

7. George Box's quote means no model perfectly represents reality, but some are still useful for making predictions or understanding valuable insights despite inaccuracies.

# Självutvärdering

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

I faced significant challenges with data cleaning and handling outliers. Understanding statistical concepts within R was crucial but difficult. Most challenging was extracting data through Api.

1. Vilket betyg du anser att du skall ha och varför.   
   I think I should get approved betyg.
2. Något du vill lyfta fram till Antonio?

I have noticed that we are having more discussions but there's less of the detailed explanation of topics and concepts that we used to have before so its becoming a bit challenging.

# Källförteckning

Dawod, Abdaljbbar B. A. "Regression Analysis Using R." King Fahd University of Petroleum and Minerals, Saudi Arabia, December 20, 2015.

Jones, Emma. "The Rise of Electric Cars: A Shift Towards Sustainable Transportation." GreenTech Publishing, 2021.

Lilja, David J., and Linse, Grace M. An Introduction to Data Modeling, 2nd ed., University of Minnesota Libraries Publishing, Minneapolis, MN, 2022.

"ggplot2 Tutorial With R." r-statistics.co. [Link](https://r-statistics.co/ggplot2-Tutorial-With-R.html).

Statistics Sweden (SCB). [Link](https://www.scb.se/).