Analyzing Volkswagen Car Resale Prices Using Blocket Data

Priyadarsini Panda

EC Utbildning

Kunskapskontroll_R

2025 April

Abstract

This report presents a regression modeling analysis aimed at predicting the resale prices of Volkswagen cars, using a dataset extracted from Blocket.se, a popular Swedish online marketplace. The dataset was preprocessed, including cleaning in Excel and outlier removal, before conducting exploratory data analysis (EDA) in R. Multiple regression models were developed to quantify the impact of key factors such as mileage, model year, horsepower, fuel type, and transmission type on selling price. Model diagnostics ensured robustness, and performance was evaluated using Adjusted R^2 and RMSE. Additionally, external data from Statistics Sweden (SCB) was integrated via API to contextualize broader market trends. This study assesses the effectiveness of statistical modeling in estimating used car prices and provides insights to support pricing strategies in the second-hand Volkswagen market.

Keywords

Data cleaning, Exploratory Data Analysis(EDA), Outlier Removal, Correlation Analysis,Statistical Significance, Multiple Linear Regression Analysis, VIF, Cook's Distance, Train-validation-Test Split, ggplot2 Visualization, Model Evaluation (Root Mean Squared Error, Adjusted R^2, BIC), Passenger Cars, Swedish Transport Data , SCB API.

Acknowledgement

I would like to extend my sincere gratitude to Linus and Antonio, my mentor, for their unwavering support and dedication. I also want to thank everyone in my class for their kindness and generosity—it has truly made a difference.

Skapas automatiskt i Word genom att gå till Referenser > Innehållsförteckning.

Innehållsförteckning

1 Introduction

The automotive industry plays a significant role in the economy, influencing consumer behaviour, environmental policies, and technological advancements. Understanding trends in vehicle usage and pricing helps businesses, policymakers, and researchers make informed decisions. This report focuses on passenger car trends in Sweden, utilizing publicly available data from SCB (Statistics Sweden) API to retrieve monthly counts of passenger cars in use from March 2021 to March 2025 and a dataset on Volkswagen vehicle sales to explore key patterns in vehicle ownership and pricing.

In recent years, factors such as fuel efficiency, technological advancements, and economic fluctuations have impacted car sales and pricing. Analysing historical data allows us to detect trends, forecast future developments, and provide insights for consumers and industry stakeholders. This study aims to bridge the gap between market pricing and consumer preferences by evaluating relevant factors affecting car prices, including mileage, horsepower, fuel type, and transmission type.

1.1 Purpose

The purpose of this report is to analyse passenger car trends and determine key factors influencing Volkswagen vehicle pricing in Sweden. To achieve this, the following research questions will be addressed:

1. What are the trends in the number of passenger cars in use in Sweden over time?
2. Which variables have the most significant impact on Volkswagen car prices, and how do they interact?

2 Theory

This study examines volkswagen car pricing in Sweden using data collected from Blocket, a leading online marketplace for used cars. The theoretical foundation integrated automative economics, pricing models and data science methodologies , with supplementary analysis of passenger car trends from Statistics Sweden (SCB) API.

2.1 Market Pricing Theory and Vehicle Depreciation

Cars lose values over time due to aging, wear and demand shifts. That's called depreciation Model. Popular models and fuel-efficient vehicle retain higher resale value and buyer assess mileage, horsepower, transmission type and fuel efficiency before purchasing. Rather than this In Sweden, external factors such as tax incentive for electric vehicles, environmental polocied and intrest rates further influence pricing trends.

2.2 Statistical Foundation:

2.2.1 Linear Regression in Predictive Modeling

Multiple linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variable. The model assumes linear relationship between predictor and outcome. It involves fitting a mathematical model to observed data points, allowing for prediction and inference about the relationships between variables.

The general form of the multiple linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Y = Dependent Variable

$X_{1,\ldots}X_k$, = Independent Variables(Predictors)

β0 = Intercept

β1, β2, ...,βk = Coefficients

ε = Error term

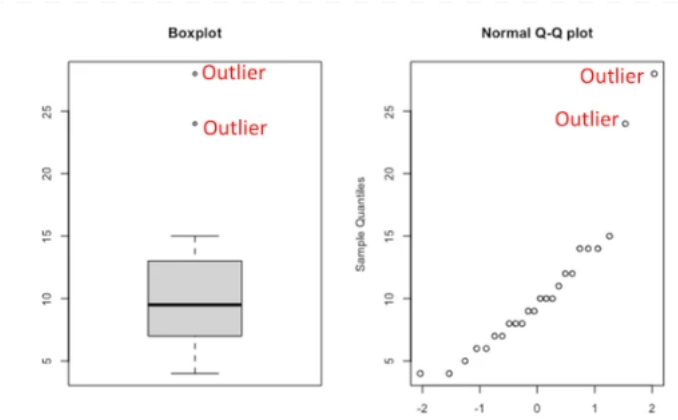
In this study, Linear regression models are applied to analyze Volkswagen car prices(a dependent variable) in relation to key influencing factors, including mileage, model year, horsepower, fuel type, and transmission type(independent variables). To ensure model reliability, multicollinearity analysis is conducted using the Variance Inflation Factor (VIF), which detects correlations between independent variables. Additionally, Cook's Distance and leverage values identify extreme observations that may distort predictions, allowing for the removal of outliers to improve accuracy. Regression coefficients are examined to quantify the effect of each predictor on car pricing, ensuring a precise and interpretable model.

2.2.2 Linearity

The relationship between predictors and the response is linear. To check Linearity Residuals vs. Fitted valued plot should show no clear pattern. Scatterplots for continuous predictors can help visualize linear trends.

2.2.3 Outlier

Outliers are data points or observations where the true value is far from the predicted value. Outliers can distort statistical analysis, affecting mean, standard deviation, and regression models. Handling outliers depends on their cause- some may be removed, transformed or analyzed separately.



(Pic taken from google) outlier

2.2.4 High leverage Point

A high leverage point in regression analysis is an observation with extreme values for the independent variables, meaning it is far from the center of the data distribution. These points have the potential to significantly influence the fitted regression mode.

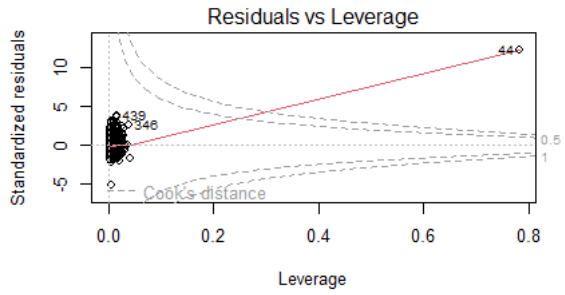
2.2.5 Quantile Quantile plots

The quantile-quantile(Q-Q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution . Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution.

2.2.6 Cook's Distance

It is a measure used in regression analysis to identify influential data points that may significantly affect the model's predictions. Cook's Distance quantifies how much the fitted values change when a

specific data point is removed. It considers both residuals and leverage, meaning points with large residuals or high leverage tend to have higher Cook's Distance.

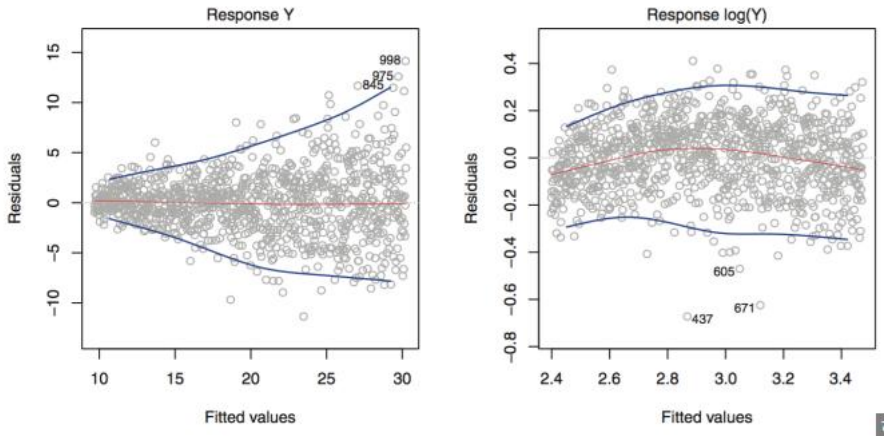


(pic: vizualization from my own model)

In short: Cook's Distance and leverage values identify extreme observations that may distort predictions, allowing for the removal of outliers to improve accuracy.

2.2.7 Heteroscedasticity

Heteroscedasticity occurs when the variance of residuals in regression model is not constant, violating the assumption of homoscedasticity, where errors maintain a uniform spread. In simple term it can be due to a non linear relationship between the dependent and independent variable, and then ant transform of the independent variable such as x^2 or $\sqrt{x}$ can help. It is commonly detected using residuals plots.



(pic taken from google)

2.2.8 Multicollinearity Check using VIF

Multicollinearity occurs when two or more independent variables are highly correlated. This means they are providing overlapping information, which can Inflate the standard error of regression

coefficients, make it difficult to determine the individual effect of each variable and lead to unstable estimates, when making predictions.

Variance Inflation Factor (VIF) is used to detect this issue.
If VIF > 5 or 10, and both limits are commonly used for a variable, then multicollinearity is considered.

2.2.9 Statistical significance

Statistical significance in this context refers to whether each predictor variable contributes meaningfully to explaining the variation in the target variable. The null hypothesis H_0 states that a given coefficient is equal to zero. A p-value below a chosen significance level (typically 0.05) indicates sufficient evidence to reject, suggesting the predictor is statistically significant.

2.2.10 Predictive Model

The predictive model aims to estimate the resale prices of Volkswagen vehicles using key determinants such as mileage, model year, horsepower, fuel type, and transmission type. To evaluate the model's performance, statistical metrics including Root Mean Square Error (RMSE), Adjusted R^2, and the Bayesian Information Criterion (BIC) are employed. RMSE captures the average deviation between predicted and actual prices, indicating the model's predictive accuracy. Adjusted R^2 reflects the proportion of price variation explained by the independent variables, adjusted for the number of predictors, while BIC supports model comparison by penalizing complexity.

Model validation is performed by comparing predicted prices to actual sales figures from the dataset, thereby enhancing the model's reliability and real-world applicability. By uncovering pricing patterns and relationships, the model provides actionable insights into Volkswagen car valuation. These findings support more informed decision-making for both consumers and sellers in the Swedish used car market.

3 Methodology

This section presents the complete methodology used to develop and evaluate predictive models for estimating the resale prices of Volkswagen cars in Sweden. The process involved data acquisition, cleaning, model construction, diagnostics, and the integration of external data to provide contextual insights. Additionally, several R libraries were employed to support data manipulation, visualization, statistical analysis, and predictive modeling. These tools enabled efficient data processing, robust regression analysis, and comprehensive model evaluation, ensuring accuracy and reliability.

3.1 Tools and Libraries Used

All analyses were performed using the R programming language, which is widely known for its statistical computing capabilities. Key libraries included the tidyverse suite for data wrangling and

visualization , the car package for testing multicollinearity and linear regression assumptions. The plotting capabilities of ggplot2 were extensively used to generate both model diagnostics and summary visualizations.

- readxl : Used to import the Volkswagen dataset from an Excel file for analysis.
- dplyr: Provides functions for data cleaning, transformation, and filtering, improving dataset structure.
- tidyr: Helps reshape and organize data for better usability in modeling and visualization.
- Pxweb: For accessing external data via SCB API

3.2 Data Collection and Preparation

The primary dataset was sourced from Blocket.se, a major Swedish marketplace for second hand goods, and it focused specifically on used Volkswagen cars. The dataset included several key variables likely to influence car prices: Miltal (mileage), Modellår (year of manufacture), Bränsle (fuel type), Växellåda (transmission type), Färg(colour), Säljare(seller), Biltyp (car type), Drivning (Drivining), `Hästkrafter (HK)`(horse power), Datum_i_trafik (Date in traffic), which served as the dependent variable.

3.2.1 Data Preprocessing

- **Loading Data**: The dataset was imported using read::_excel(). The raw data underwent several cleaning steps, followed by data cleaning na.omit() to remove missing values.
- **Data Transformation**: Factor conversion was applied to categorial variable using mutate().
- **Exploratory Data Analysis (EDA)**: Functions like str(), summary() and visualization tools such as ggplot2 and corrplot were used to examine relationship among features.

3.2.2 Model Development

Data splitting: The data set was divided into training (60%), validation (20%), and test (20%) sets using a random sampling approach set.seed(123).

3.2.3 Linear Regression Models

To predict sales price, three regression models were developed using selected variables from the dataset:

- Model 1: Included mileage, model year, horsepower, and transmission type to examine their combined impact.
- Model 2: Introduced a categorical variable — fuel type — to explore the potential impact of fuel preferences on car pricing.
- Model 3: A minimalistic model that included both fuel type and transmission type, as well as interaction terms where relevant, to capture more nuanced relationships between predictors.

3.2.4 Model Evaluation

During the model evaluation step, each regression model was assessed based on its predictive performance and reliability. First, the models were applied to the validation dataset, and their accuracy was measured using **Root Mean Squared Error (RMSE)**, which evaluates how close the predicted values are to the actual sales prices. Additionally, the **Adjusted R-squared** metric was examined to determine how well the model explains the variance in the data while accounting for the number of predictors. In addition, **BIC** was employed to penalize model complexity and prevent overfitting and conducted hypothesis testing and confidence interval estimation for key variables.

To ensure model stability and generalizability, diagnostic tests were conducted. **Cook's Distance** and **leverage values** were analyzed to detect potential outliers that might disproportionately affect the regression results. Influential data points were removed to enhance model robustness. Furthermore, **Variance Inflation Factor (VIF)** was calculated to check for multicollinearity, ensuring that predictor variables were not overly correlated.

3.2.5 Assumption Testing

To validate the use of linear regression, key model assumptions were tested:

To ensure the reliability of Models, assumption testing was performed. Linearity was confirmed through residual plots, homoscedasticity was verified, and normality was assessed using Q-Q plots. Multicollinearity was ruled out with VIF values below 5. Influential observations were identified via Cook's Distance and leverage statistics and removed for model stability.

4 Results and Discussion

4.1 Model Comparison and Selection

Three regression models were evaluated: Model 1 (basic predictors), Model 2 (added fuel type), and Model 3 (added transmission type). Based on Root Mean Square Error (RMSE) values on the validation and test datasets, Model 2 was selected as the best-performing model. Although Model 3 included an additional variable (transmission type), its performance did not significantly improve and showed signs of overfitting. Therefore, Model 2 strikes a balance between simplicity and predictive accuracy.

Model	RMSE(Validation)	RMSE(Test)	Adjusted R^2	BIC
Model 1	39002.93	-	0.8413720	15765.86
Model 2	37016.11	43323.52	0.8483155	15896.18
Model 3	42456.75	-	0.7880944	16931.64

4.2 Model Interpretation

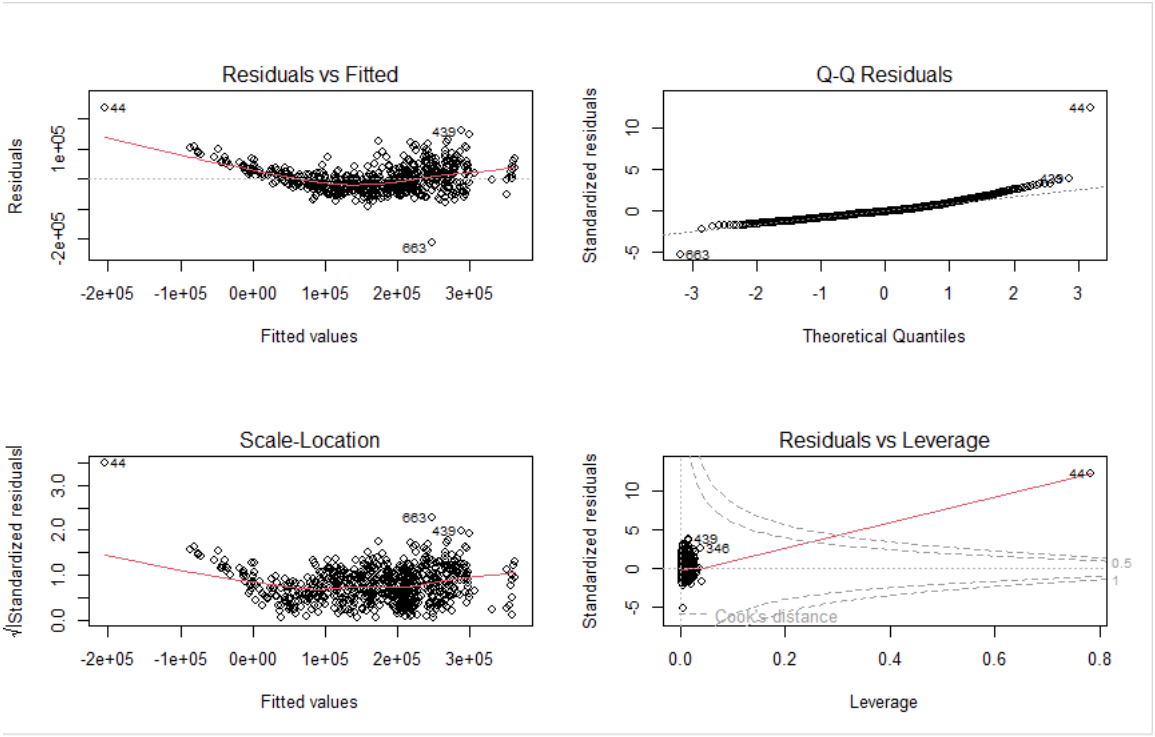
Regression coefficients in Model 2 shows: How different variables affect car prices. As expected, mileage has a negative relationship with price, suggesting that cars with higher mileage are valued lower. Model year shows a positive association, indicating newer cars command higher prices. Fuel type also had a significant impact: diesel cars were priced lower compared to petrol, possibly due to changing consumer preferences and environmental policies in Sweden.

4.3 Potential Problem (Assumption Testing)

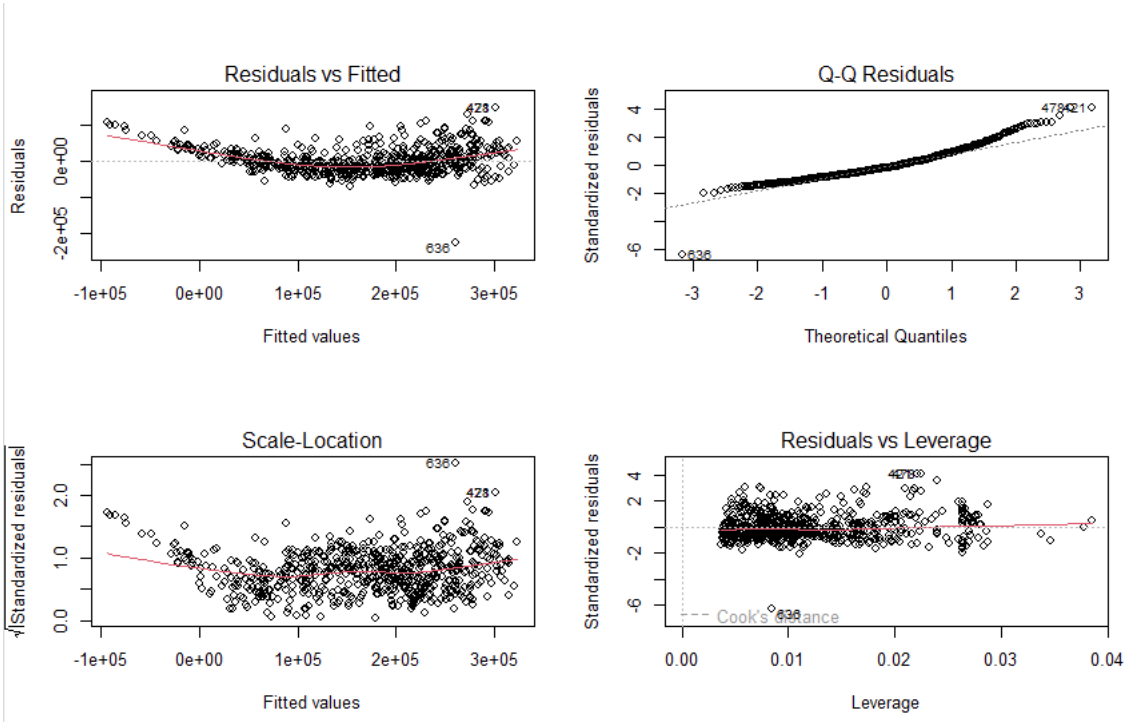
To validate model reliability, several points were checked: Model 2 underwent rigorous assumption testing to validate its reliability.

- Linearity: Residual plots confirmed linear relationships between predictors and sales price.
- Homoscedasticity: Residual variance was consistent, with no major heteroscedasticity detected.
- Normality: Q-Q plots indicated that residuals followed a normal distribution.
- Multicollinearity: VIF values were below the threshold of 5, indicating no serious collinearity.
- Influential Observations: Outlier detection was conducted using Cook's Distance and Leverage values. High-influence points were identified and removed to prevent distortion.

Influential Observation: Before removing outliers(Residuals vs leverage)

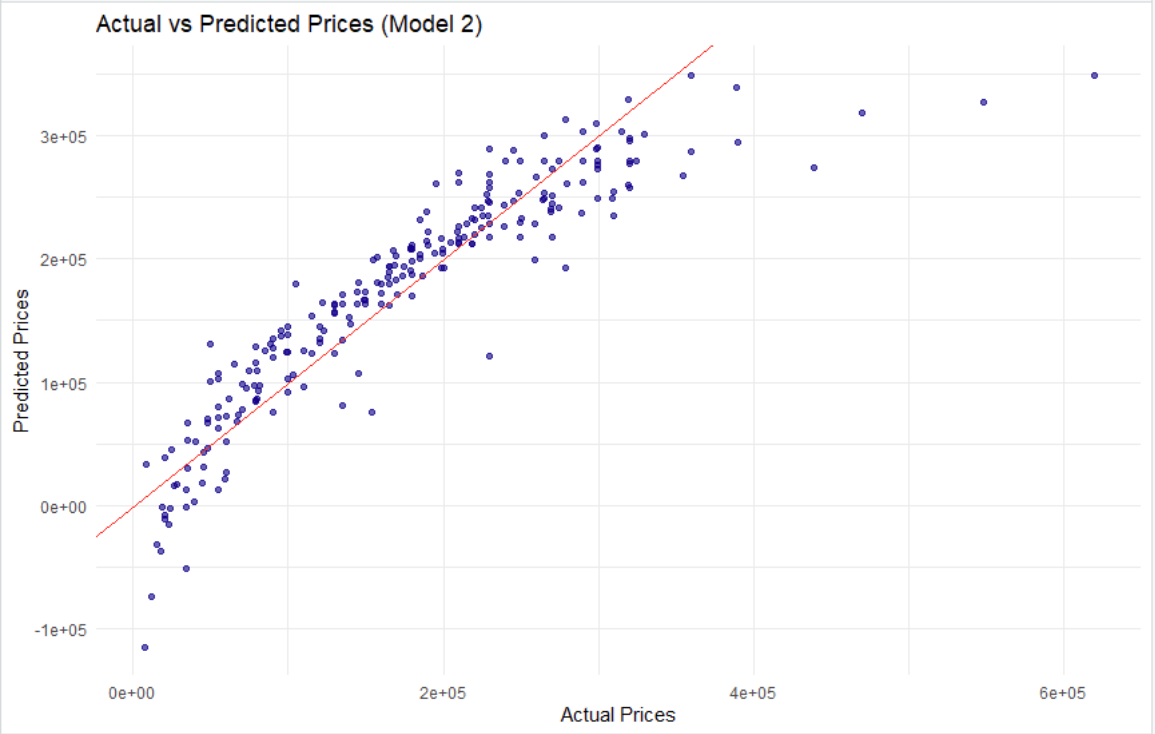


After removing (Residuals vs Leverage)



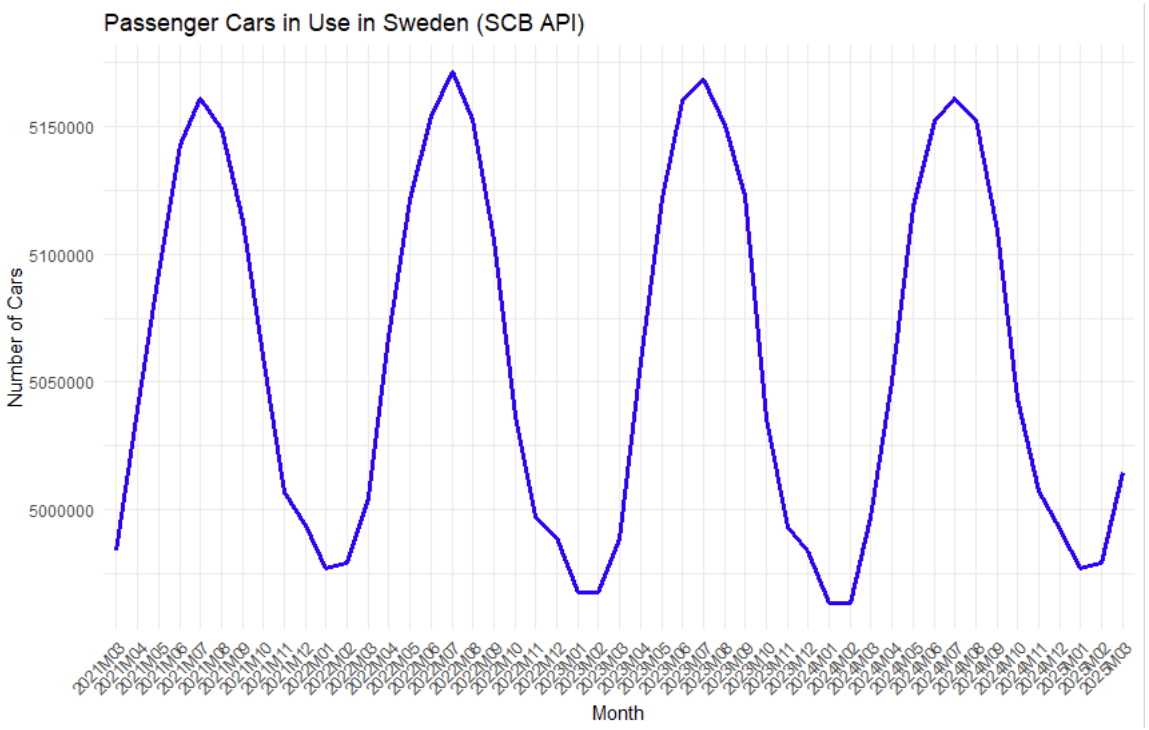
4.4 Prediction Performance

The Actual vs Predicted Prices plot for Model 2 shows that the predicted values align closely with the actual car prices. Most data points lie near the diagonal line, indicating that the model's predictions are reasonably accurate. This supports the model's usefulness for estimating resale values of Volkswagen vehicles in Sweden.



4.5 External Data from SCB API

The external data was retrieved using the PXWEB API provided by Statistics Sweden (SCB). A custom API query pulled monthly data from March 2021 to March 2025, showing the total number of passenger cars in use in Sweden.



5 Conclusion

This study successfully developed and evaluated multiple linear regression models to predict Försäljningspris (sales price) of Volkswagen cars based on key attributes such as mileage, model year, horsepower, fuel type, and transmission type. Through data preprocessing, assumption testing, and model comparison, Model 2 emerged as the best-performing model, demonstrating strong predictive accuracy and generalizability.

1.What are the trends in the number of passenger cars in use in Sweden over time? PXWEB API query retrieved monthly vehicle count data for passenger cars in use from March 2021 to March 2025. According to the visualization, the number of cars in use has generally grown over time with Some months show slight declines.

2.Which variables have the most significant impact on Volkswagen car prices, and how do they interact?

Mileage: More mileage reduces price significantly. **Model Year**: Newer cars increase resale value. **Horsepower**: Higher horsepower correlates with higher price. **Fuel Type**: Fuel choice has an effect but depends on market trends.

The findings confirm that mileage negatively impacts sales price, while newer model years and higher horsepower contribute to increased valuation. Fuel type and transmission type also play significant roles, indicating buyer preferences and market trends. Assumption testing validated the robustness of the model, ensuring reliability in its predictions.

Furthermore, external data from Statistics Sweden (SCB) added valuable context, showing a steady increase in the number of passenger cars in use between 2021 and 2025. This trend reinforces the practical relevance of the model in a growing automotive market.

Overall, the project showcases how statistical modeling can support more informed pricing decisions in the used car market.

6 Part 2: Teoretiska frågor

1. Kolla på följande video: https://www.youtube.com/watch?v=X9_ISJ0YpGw&t=290s , beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

Svar: En kvantil-kvantil (Q-Q) plot är en grafisk metod för att jämföra fördelningen av en datauppsättning med en teoretisk fördelning. Det hjälper till att bedöma om data följer en specifik sannolikhetsfördelning, till exempel en normalfördelning, eller om två datamängder kommer från samma fördelning.

Den plottar sample kvantiler mot teoretiska kvantiler för att kontrollera likheter. Om punkterna bildar en rät linje följer datasetet den förväntade fördelningen. Böjda eller spridda punkter indikerar skevhet, tunga svansar eller avvikelser.

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?

Svar: I Maskininlärning Primärt fokuserad på förutsägelse – med tanke på nya data syftar en ML-modell till att göra korrekta prognoser .Till exempel: att förutsäga huspriser baserat på funktioner. Medan I statistisk regressionsanalys tillåter både förutsägelse och statistisk slutledning – slutledning betyder att förstå sambanden mellan variabler .Till exempel: hur mycket en faktor påverkar ett resultat. ett annat exempel: En regressionsanalys av utbildning och inkomst i Sverige kan visa att varje ytterligare skolår ökar årsinkomsten med i genomsnitt 5 000 kr. Detta förhållande hjälper till att förutsäga inkomster samtidigt som det ger insikt i utbildningens ekonomiska effekter.

3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?

Svar: Konfidensintervall uppskattar intervallet inom vilket det sanna medelvärdet av de förutsagda värdena sannolikt kommer att falla. Den berättar hur exakt du har uppskattat det förväntade värdet för en given ingång.

Prediction Interval är bredare och tar hänsyn till individuell variabilitet, vilket visar var ett faktiskt observerat värde kan falla för en specifik indata. konfidensintervall är till för att uppskatta befolkningsmedelvärden, medan prediktionsintervall står för variationer i verkligheten i enskilda fall.

4.Den multipla linjära regressionsmodellen kan skrivas som:

Y = $\beta 0 + \beta 1 x 1 + \beta 1 x 2 + ... + \beta p x p + \varepsilon$.

Hur tolkas beta parametrarna?

Svar: $\beta 0$(Interceptet): Värdet på Y när alla x-variabler är noll—det kan representera en basnivå eller startvärde.

$\beta_1, \beta_2, ..., \beta_p$: Varje koefficient anger den genomsnittliga förändringen i Y för en enhets förändring i motsvarande x-variabel, medan övriga variabler hålls konstanta.

ε: Representerar modellens osäkerhet och fångar upp påverkan från faktorer som inte ingår i modellen.

5. Din kollega Nils 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?

Svar: I statistisk regressionsmodellering hjälper BIC till att välja modell genom att straffa mer komplexa modeller som har fler parametrar, vilket minskar risken för överanpassning. Det tillhandahåller ett mått för att jämföra modeller, men det utvärderar inte direkt prediktionsprestanda på osynliga data. Utan ett separat testset finns det ingen garanti för att den valda modellen fungerar bra i verkliga tillämpningar. Även om BIC hjälper till med modellval, måste datauppsättningar fortfarande delas upp i tränings, validerings och testuppsättningar för att bedöma prediktiv noggrannhet.

6. Förklara algoritmen nedan för ”Best subset selection”

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.

Svar: 6

1. Start by fitting the null model, which includes no predictors. It only uses the mean of the outcome variable to make predictions for all observations. This acts as a baseline.

2.Loop over number of predictors for k = 1 to p

a) From the total number of predictors p, compute all possible combinations of k variables and fit a model for each.

b) Out of all models with k predictors, choose the one with the lowest RSS (Residual Sum of Squares) or highest R^2. Call this best model M_k

3. Compare all the best models (one for each number of predictors) using model selection criteria such as Validation set performance, Adjusted R¨2, BIC, AIC or Cross-Validation.

Overall, we use Best Subset Selection to systematically explore all possible combinations of predictor variables and identify the model that best balances predictive accuracy and simplicity. This technique helps avoid overfitting by not relying solely on stepwise methods, which can get stuck in local optima. By evaluating models using metrics like Adjusted R^2, BIC, or validation RMSE, Best Subset Selection ensures that the final model is both statistically sound and interpretable. In practice, especially when

the number of predictors is moderate, it is a powerful approach to find the most informative subset of variables.

7. Ett citat från statistikern George Box är: “All models are wrong, some are useful.” Förklara vad som menas med det citatet.

Svar: George Box’s quote, "All models are wrong, some are useful," påminner oss om att modeller är förenklingar av verkligheten och aldrig kan vara helt korrekta. Det är också värt att betona att en modell är en förenkling av verkligheten och i praktiken tror vi aldrig att en modell och dess antagande alltid är helt uppfyllda. Därför skall man inte bli lamslagen när antaganden inte uppfylls.

Men även felaktiga modeller kan fortfarande ge värdefulla insikter, vägleda förutsägelser och beslut på ett meningsfullt sätt. Nyckeln är att känna igen deras begränsningar samtidigt som de används som verktyg för att förstå komplexa system. En bra modell behöver inte vara perfekt , den behöver bara vara användbar.

7 Självutvärdering

1. Vad tycker du har varit roligast i kunskapskontrollen?
   Svar: Från att samla in Blocket-data till att förbereda och välja en prediktiv modell som ger bra resultat.

2. Hur har du hanterat utmaningar? Vilka lärdomar tar du med dig till framtida kurser?

Svar:Effektiva strategier inkluderar att bryta ner problem i mindre steg, med hjälp av olika inlärningsmetoder. Att koppla teori till verkliga situationer fördjupar förståelsen.

3. Vilket betyg anser du att du ska ha och varför?

Svar: Jag strävar efter att prestera mitt bästa i mitt arbete, och lägger kraft på varje steg.

4. Något du vill lyfta till Antonio?

Svar: Nej!

Appendix A

https://github.com/ppriya23/R_Programming

Källförteckning

- https://www.geeksforgeeks.org/
- An Introduction to Statistical Learning with Applications in R Second Edition by Gareth James, Trevor Hastie, Daniela Witten, Robert Tibshirani