Car Price Analysis

Using R Programming Language



Filip Östlund

EC Utbildning

Car Price Analysis

202404

# Abstract

In this report we gather and analyse car data from Blocket.se. Building a regression model predicting car prices based on various features. We also integrate external data from an API to support our report with additional insights. The dataset from Blocket.se comprises approximately 700 observations, encompassing 15 distinct features. Employing a linear regression framework, we address potential issues such as outliers, high leverage points, and correlations among variables. By leveraging various diagnostic tools and evaluation metrics, including Adjusted R-Squared, BIC, and AIC, we ensure the reliability and accuracy of our models. Our approach results in a model that achieve a Root Mean Squared Error (RMSE) Percentage of 17.8%. This performance metric underscores the efficacy of our models in forecasting car prices with a pretty high degree of precision based on the small data collected.

Table of Contents

[1 Introduction 1](#_Toc165040070)

[1.1 Problem Statements 1](#_Toc165040071)

[2 Theory 2](#_Toc165040072)

[2.1 Data Collection 2](#_Toc165040073)

[2.2 Multiple Linear Regression 2](#_Toc165040074)

[2.3 Model 2](#_Toc165040075)

[2.3.1 Potential Problems 2](#_Toc165040076)

[2.3.2 Stepwise selection 3](#_Toc165040077)

[2.3.3 Model Evaluation 3](#_Toc165040078)

[2.4 Packages 4](#_Toc165040079)

[2.4.1 Dplyr 4](#_Toc165040080)

[2.4.2 FastDummies 4](#_Toc165040081)

[2.4.3 MASS 4](#_Toc165040082)

[3 Method 5](#_Toc165040083)

[3.1 Data Collection and Data Exploration 5](#_Toc165040084)

[3.2 Data Preperation 5](#_Toc165040085)

[3.3 Model training and testing 6](#_Toc165040086)

[3.4 API 7](#_Toc165040087)

[4 Results and Discussion 8](#_Toc165040088)

[5 Conclusion 10](#_Toc165040089)

[5.1 Problem Statements 10](#_Toc165040090)

[5.2 Improvements 10](#_Toc165040091)

[6 Teoretiska frågor 11](#_Toc165040092)

[7 Självutvärdering 13](#_Toc165040093)

[Källförteckning 14](#_Toc165040094)

# Introduction

Regression analysis serves as a foundational technique in statistical modelling, offering a systematic approach to understanding relationships between variables and making predictions based on observed data patterns. Widely utilized across diverse fields such as economics, finance, and social sciences, regression analysis provides a powerful framework for uncovering associations and extracting insights from data.

With cars being the primary mode of transportation for the Swedish population, we have seen an enormous increase in passenger cars year after year. This has propelled the Swedish car market with both the car companies and second-hand sellers always in need of a faster and better way to price their cars according to today’s market. Through this comprehensive analysis, we aim to create a regression model that predicts an average passenger cars price based on the current Swedish car market, using various methods available in the R programming language.

En bild som visar linje, Graf, diagram, sluttning

Automatiskt genererad beskrivning

Figure 1. Cars Per Year (SCB)

## Problem Statements

The main purpose of this report will be to answer the following questions:

1. Can we manually collect enough data from Blocket in two days for it to be enough for a regression model with an acceptable accuracy?
2. What different variables are best to predict the price of a car?
3. Is it possible to develop a regression model with a RMSE of under 20%?
4. Can we use an API to retrieve external data to support the report?

# Theory

## Data Collection

The Dataset was manually collected from Blocket.se. Sweden’s biggest online marketplace with over 5 million visitors each week (Blocket, 2021). The group collecting the data consisted of 6 students. Each student was assigned different parts of Sweden to collect data from to avoid collecting the same observations. Because this project is aimed at average car and not luxury cars and work cars etc, the filter used in the Blocket search was price between 100k-500k and no work vehicles. The data collected for each observation had 15 columns containing text or numeric data. Around 700 observations were collected for the project. Below are a few observations from the collected data.

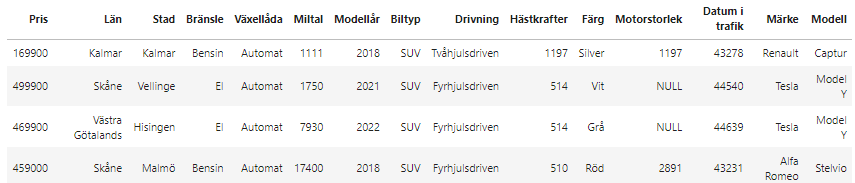


Figure 2: Example of data collected.

## Multiple Linear Regression

Simple linear regression is an approach in predicting a quantitative response variable Y based on a single predictor variable X. It operates under the assumption of a linear relationship between X and Y (James et al., 2023, p. 61). If we have more than one predictor, rather than fitting individual simple linear regression models for each predictor, a more effective approach involves extending the simple linear regression model to accommodate multiple predictors simultaneously. This is achieved by assigning a separate slope coefficient to each predictor within a single model. In essence, if we have p distinct predictors, this method allows us to incorporate all of them into the regression model, each with its own slope coefficient (James et al., 2023, p. 72).

This mathematical relationship can be expressed as:

## Model

The model used was the standard R model “lm”. Lm is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (R Core Team, 2021)

### Potential Problems

#### Outliers

An outlier is an observation that significantly deviates from other observations in a dataset. Outliers can arise for a variety of reasons, such as incorrect recording of an observation during data collection. The outliers can skew statistical analyses and distort the interpretation of results, making it important to identify and address them appropriately (James et al., 2023, p. 97)

#### High Leverage Points

High leverage points in regression are observations with unusually extreme values for predictor variables compared to the rest of the dataset. They can substantially affect the regression line's estimation and should be carefully identified to ensure accurate modeling and interpretation (James et al., 2023, p. 99)

#### Collinearity

Collinearity refers to the situation in which two or more predictor variables are closely related. The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effect of each predictor on the outcome variable. It's essential to detect and address collinearity to ensure the reliability and validity of regression model results (James et al., 2023, p. 100)

### Stepwise selection

Stepwise selection is a method used in statistical modelling to automatically select a subset of variables from a larger set of predictors. The two most common stepwise selection methods are forward and backward. Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all the predictors are in the model. At each step the variable that gives the greatest additional improvement to the fit is added to the model. Backward stepwise works the opposite, it begins with the full least squares model containing all predictors, and then iteratively removes the least useful predictor, one-at a-time (James et al., 2023, p. 231).

### Model Evaluation

#### Adjusted

Adjusted R-squared is a modification of the standard R-squared statistic used in regression analysis. While R-squared measures the proportion of variance in the dependent variable explained by the independent variables, Adjusted R-squared adjusts this measure to account for the number of predictors in the model and the sample size. Adjusted R-squared penalizes the addition of unnecessary predictors to the model, preventing it from inflating when adding more variables. A larger Adjusted R-squared indicates a better fit of the model to the data. (James et al., 2023, p. 235).

The formula for Adjusted R-squared is:

Where: is the coefficient of determination (R-squared), n is the number of observations in the sample and p is the number of predictors in the model.

#### BIC

The Bayesian Information Criterion (BIC) is a statistical measure used for model selection, particularly in the context of regression analysis. It provides a way to compare different models based on their fit to the data while penalizing models with more parameters, thereby helping to prevent overfitting. Models with lower BIC values are preferred, indicating a better trade-off between goodness of fit and model complexity. (James et al., 2023, p. 234).

The formula for BIC is:

Where: n is the number of observations in the dataset, RSS is the residual sum of squares, which measures the discrepancy between the observed values and the values predicted by the model and d is the number of predictors in the model.

#### Root Mean Square Error

The root mean square error (RMSE) measures the average difference between a statistical model’s predicted values and the actual values. Mathematically, it is the standard deviation of the residuals. Residuals represent the distance between the regression line and the data points.

RMSE quantifies how dispersed these residuals are, revealing how tightly the observed data clusters around the predicted values. As the data points move closer to the regression line, the model has less error, lowering the RMSE. A model with less error produces more precise predictions (Jim Frost, 2023).

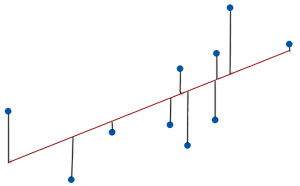


Figure 3. Image depicting the relationship between the residuals and the RMSE.

## Packages

### Dplyr

The dplyr package is a popular package in R used for data manipulation. It provides a set of functions that are optimized for speed and ease of use when working with data frames or data tables. One of the most used functions is the select() functions which is used to select specific columns from a data frame (Wickham et al., 2023)

### FastDummies

The fastDummies package is a tool in R designed to create dummy variables from categorical variables. The function used for this is dummy\_cols() which quickly create dummy (binary) columns from character and factor type columns in the inputted data. (Kaplan & Schlegel, 2023)

### MASS

The MASS package is a very popular package in R, often used for statistical analysis. It stands for "Modern Applied Statistics with S," and it contains a wide range of functions and datasets for various statistical analyses. For stepwise selection the stepAIC() function is used which choose the best model by AIC (Venables WN & Ripley BD, 2002). AIC works very similar to BIC with a smaller AIC value indicating a better fitting model. AIC is more suitable when the emphasis is on predictive accuracy, while BIC is often preferred when the emphasis is on model simplicity. (James et al., 2023, p. 234).

# Method

## Data Collection and Data Exploration

The first part of the project was to collect the data. The data collection was covered a bit in the theory. We collected a total of approximately 700 observations from 8 different regions of Sweden. The data was collected in Excel, as well as some data preprocessing was done in Excel. Because we were 6 people the collected data had some values that were supposed to be the same but had been written differently. Same problem was observed with some of the values without the “collectors” doing anything differently, just because the values are put in by the seller and they don’t always type the models the same way with spelling etc.

En bild som visar text, skärmbild, Teckensnitt

Automatiskt genererad beskrivning

Figure 4. Data collected in Excel

## Data Preperation

After importing the data to into R, the data was inspected and then I choose to immediately remove some of the columns that I believed logically would not benefit the model at the level we are going for. The columns removed was:

* Stad (City): Because we already have Län (Region). I personally don’t believe that the cities would affect the price too much. To simplify the model, it was removed.
* Motorstorlek (Motorsize): This one was very clear for me because we have horsepower. Which is a greater measurement, especially when electric cars don’t have any motorsize.
* Datum i Trafik (Date in Traffic): We already have the Model Year which means we already have some measurement of how old the car is.
* Modell (Model): To not overcomplicate the model I choose to remove the model, we already have the brand of the car. There were a lot of models that only showed up on single observations which I believe would not help.

Next step after removing the chosen columns the numeric data was inspected. As we can see some outliers were found and removed. I went back into excel to look at some of the outliers. The most noticeable outliers were in horsepower and model year. The horsepower outlier was a typo from either the person collecting the data or Blocket. The model year outliers were from older cars. Because the average car in Sweden is not as old as these outliers they were removed along with the horsepower outlier. To remove the outliers a z-value method was used were we measure how many standard deviations a data point is from the mean. Below is the data plotted first without outlier removed, and then with outliers removed. A big difference is noticed.

En bild som visar text, diagram, svart och vit

Automatiskt genererad beskrivning

Figure 5. Numeric Data Plotted

Next up was to use FastDummies to create dummy variables for the categorical features. After creating dummy variables, we ended up with 60 variables. Now because we can see a very similar pattern between milage and model year in the earlier plot we checked the collinearity. We set the correlation limit to 0.7 and checked for any correlated variables in our data, and as suspected the variables that was returned were Milage and Model Year. Then we calculated the exact correlation between the two which were -0.753189010746825. Which means they are highly correlated. They were also plotted against each other so we could clearly see the pattern between the two. This information was then taken in account when continuing to create the model because we know something needs to be done about the correlation.

## Model training and testing

Before initiating a model, the data was split into a training and a test set. With an 80/20 split. The first step of the model training was to figure out which one of Milage or Model Year that was the best or if a combination of both would be a better choice. The combination was calculated as Miles Per Year. After training three different models and seeing the results, the Model Year model was the best model based on Adjusted R-Squared and BIC.

Next, we checked the diagnostics plots (Residuals vs fitted, Normal Q-Q, Scale-Location and Residuals vs Leverage). When checking this the first time we got a warning for “not plotting observations with leverage one. This meant that we had some high leverage points that we need to check out. We calculated the high leverage points and then checked them out. I went back to excel to more easily find the high leverage points and see what makes them high leverage. All high leverage points had some value that were singular to them. When the high leverage data points were removed, we could then see on the model summary that there were a few different variables that had “NA” values which meant they had no data points with any values in them. They were 7 variables that were removed.

Now we have a model with no correlation, no high leverage points. To figure out if I need all the variables, I performed a backwards step selection using the MASS Package. The stepAIC chooses the best model that it can find using step backward selection like covered in the theory. Best subset selection would have been optimal but also would have taken too much time. Moving on to the final step, which was model testing. A prediction was made using the test set we set aside earlier, and a Root Mean Square Error was calculated with the actual prices for the test set. Just for fun the model was also tested on my personal car by manually creating a new data frame.

## API

The data that was decided to be retrieved from SCB was total passenger cars in Sweden per year. The API tutorial was that was linked on SCB: s website was followed for an easy step by step route (Kira Gylling, 2019). Importing the SCB class from the pyscbwrapper package made it possible to navigate through the different metadata. But the final data was gathered by navigating to the SCB website, selecting the desired data and then using the provided JSON command and URL to retrieve the data. The data was then processed as normally in Python. The year 2023 was removed due to missing data for the entire year, then the remaining data was plotted to be used in the report. Below is the data in a data frame before plotting.

Year Total Cars

0 2002 4042790

1 2003 4075414

2 2004 4113424

3 2005 4153674

4 2006 4202463

5 2007 4258463

6 2008 4278995

7 2009 4300752

8 2010 4335182

9 2011 4401352

10 2012 4447165

11 2013 4495473

12 2014 4585519

13 2015 4669063

14 2016 4768060

15 2017 4845609

16 2018 4870783

17 2019 4887904

18 2020 4944067

19 2021 4986750

20 2022 4980543

# Results and Discussion

Firstly, the results for the model testing between model year, milage and a combination of both is presented below. As we can see the highest adjusted R-Squared and lowest BIC is the first model with only the Model Year. The milage per year model did the worst, I believe this is because a car can be from 1970 and have the same miles per year as a car from 2022. The idea behind the combination was good so I also tried to keep the Milage Per Year together with the Model Year, but it still only got a little bit worse numbers than the Model Year alone.

|  |  |  |
| --- | --- | --- |
| Model | Adjusted R-Squared | BIC |
| Model Year | 0.7877566 | 11914.45 |
| Milage | 0.7338230 | 12024.27 |
| Milage Per Year | 0.4895101 | 12340.10 |

Table 1. Results of first three models

After removing the high leverage points, I now had a model with 52 variables and an Adjusted R-Squared of 0,7864. Then after performing backward step selection, I ended up with a final model with 27 variables and an Adjusted R-Squared of 0,7991. The backward step selection decreased the variables and simplified the model by removing little bit more than half the variables and increased the Adjusted R-Squared by more than 0,1. Below we can see the final diagnostic plots for the model.

En bild som visar text, diagram

Automatiskt genererad beskrivning

Figure 6. Final model diagnostics plot

As we can see on the first plot. It has a pretty good linear connection. There might be a little bit of a small chance for heteroscedasticity, which could be investigated more, but I choose not to. There is no clear cone shape, the variance is a bit slimmer both in the beginning and the end. Looking at the Normal Q-Q plot there is a little disturbance at the end, but I believe its acceptable for our goal. I removed a single observation that I believe affected the normal distribution the most. The other plots look good enough for the purpose of the report.

Testing the model with the test set gave us the results below. I also tested the model with my own car by manually imputing the values. I asked a friend that works at a car sales company about a valuation for my car and I got a 65-70k kr valuation. The model predicted a price of 65418kr which is almost perfectly if compared to the car company. Fun with such a good result, but its only one car as well. A better way to test would be to get more test cars to evaluate the performance even more.

|  |  |  |
| --- | --- | --- |
| Mean of Price in Test | RMSE | Error Percentage of Mean |
| 236294,2 kr | 42052,7 kr | 17,8% |

# Conclusion

## Problem Statements

Can we manually collect enough data from blocket in two days for it to be enough for a regression model with an acceptable accuracy? Yes, I believe the data was enough to work with and create a model with an acceptable accuracy. More data would have probably made for an even better model or better testing possabilites.

What different variables are best to predict the price of a car? I the end we ended up with 27 variables, but because of some being dummies, we really ended up with only these different categories: Model Year, Horsepower, Region, Gearbox, Car Type, Color and Brand.

Is it possible to develop a regression model with a RMSE Percentage of under 20%? Yes, the final RMSE Percentage was 17,8% which I believe to be a good result for a model with this little data.

Can we use an API to retrieve external data to support the report? The API was used to retrieve data for the introduction of the report. Very good to learn how to get started with API:s.

## Improvements

The model is limited by the amount of data gathered, but the results achieved is still acceptable. To gather more data, a web scraper could have been used. As a group we investigated it, but it was a bit to complicated. We wanted to focus more on R.

I would have liked to try best subset selection. I believe it’s possible if I were to leave the computer on for a couple of hours. But I have unfortunately had computer problems the last two weeks with charging problems, so I decided to stay satisfied with backwards step selection.

As seen on the diagnostic plots they are not perfect and of course they will never be perfect, but some improvements could be done. I tried using a Cooks Distance method, to remove a few more “outliers”. The removed outliers increased the Adjusted R-Squared a little bit, but the plots were not drastically improved in any way. It also increased the RMSE which was the main purpose to keep low. Because of this I was satisfied with the result and did not go further investigating more improvement methods.

I also tested a different model (Lasso), but it had a worse result. Therefore, I just went with the normal lm model. But if more time was given, a pipeline could probably be a good idea to try different models to see if there is anyone better than the normal lm model.

# 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 grafisk metod för att bedöma om en uppsättning observationer följer en specifik sannolikhetsfördelning, vanligtvis normalfördelningen. QQ-ploten jämför kvantilerna av den observerade datan med kvantilerna av den teoretiska fördelningen. Om datapunkterna ligger nära den linjära trenden så är datan ungefär fördelad på den fördelningen man testar för.

1. 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 är huvudfokus på att göra prediktioner medan statistisk regressionsanalys även inkluderar statistisk inferens, vilket innebär att bedöma osäkerhet och signifikans i modellen. Till exempel kan vi använda en maskininlärningsmodell för att förutsäga en persons blodtryck baserat på deras ålder, medan statistisk regressionsanalys skulle ge oss både prediktioner och information om hur signifikant och osäker vår modell är.

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

Svar: Ett konfidensintervall är en intervallskattning av ett populationsparametervärde (t.ex. medelvärdet) baserat på ett stickprov. Det ger ett intervall där vi kan vara säkra på att en viss del av gångerna innehåller det sanna parametervärdet. Medan ett prediktionsintervall ger en uppskattning av osäkerheten för en enskild framtida observation av den beroende variabeln.

1. Den multipla linjära regressionsmodellen kan skrivas som: Y = + + + ...+ +𝜀 Hur tolkas beta parametrarna?

Svar: är interceptet eller baslinjenivån av Y när alla oberoende variabler är noll. De resterande beta-parametrarna representerar förändringen i den beroende variabeln (Y) för varje enhets ökning av den tillhörande oberoende variabeln (x). Det vill säga β visar den genomsnittliga förändringen i Y för varje enhets ökning av x, när alla andra prediktorer i modellen hålls konstanta.

1. 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?

Svar: Logiken bakom detta är att mått såsom BIC eller t.ex. Adjusted inkluderar en straffterm som tar hänsyn till modellens komplexitet, vilket hjälper till att undvika överanpassning. Eftersom straffen bestraffar för många parametrar eller för komplexa modeller, behövs inte validering och test set men kan vara bra att ha ändå för att se hur modellen anpassar sig vid nya data.

1. Förklara algoritmen nedan för ”Best subset selection”

En bild som visar text, skärmbild, Teckensnitt, nummer

Automatiskt genererad beskrivning

Svar: Best subset selection är en metod för att välja den bästa uppsättningen av prediktorer för en linjär regressionsmodell genom att systematiskt testa alla möjliga kombinationer av prediktorer. Best subset selection ger alltid den bästa kombinationen av prediktorer men kan vara enormt påfrestande för datorn, speciellt vid en stor mängd prediktorer.

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

Svar: Det han menade var att det finns ingen modell som kan ge en bild av verkligheten, därför är alla modeller fel. Men om man har en modell som är i närheten av verkligheten så kan den vara användbar. T.ex. en modell som predikterar huspriser kan aldrig vara exakt som verkligheten, men om den är ganska nära verkligheten kan den vara användbar för att låt oss säga värdera ett hus utan att lägga ut det till försäljning eller kontakta en mäklare. Då kan man få ett pris som är ungefär eller ett intervall men aldrig ett pris som till hundra procent kommer stämma överens med verkligheten.

# Självutvärdering

1. **Utmaningar du haft under arbetet samt hur du hanterat dem.** Det var som vanligt svårt att komma i gång men det gick till slut. Jag hade mycket jobb under tidigare delen av kursen vilket gjorde att jag fick ta igen mycket senaste veckorna. Jag valde också att skriva rapporten på engelska för att träna vilket jag inte gjort förut.
2. **Vilket betyg du anser att du skall ha och varför.** Jag anser att jag har uppnått VG då jag under kursen tycker jag fått en väldigt bra förståelse för regressionsanalys och de olika problem som kan uppstå. Jag tycker jag kan implementera regression modeller på ett bra sätt med förståelse för vad jag gör.
3. **Något du vill lyfta fram till Antonio?** Bra kurs, dock tycker jag att du tidigare kurser haft lite mer presentationer live och sedan har vi kunnat diskutera med dig. Nu blev det lite att många lektioner där vi dök upp sen fick vi prata helt själva utifrån det som vi kollat på videos. När vi hade statistik till exempel hade du alltid 1-2h presentation men där du ändå lät folk ställa frågor och vara med vilket jag tyckte saknades lite nu. Vet inte om tanken kanske med flipped classroom att vi ska vara mer i grupper och prata själva men jag tycker att om man hamnar lite i fel grupp så kan lektionen inte ge så mycket medan en annan grupp kan ge hur mycket som helst.

**Datainsamling (Fick bli här efter rapporten)**

1. Vem du har arbetat i grupp med? Arina, Shriya, Shangchanhui, Muhammad, Isabella (Sen även Wissam och Turzo men de kom in sent och vi valde att de inte behövde samla data).

2. Hur har ni i gruppen arbetat tillsammans? Jag tycker samarbetet gick väldigt bra. Vi jobbade på snabbt och kom bra fram till vad vi skulle göra. Alla bidrog med sina åsikter och vad de tyckte skulle funka bäst.

3. Vad var bra i grupparbetet och vad kan utvecklas? Då detta var ett ganska enkelt grupparbete så tycker jag inte att vi hade så mycket att utveckla. Alla gjorde det de skulle och ingen stack ut som att de gjorde mindre än någon annan. Såklart utifrån erfarenhet då alla inte har full koll på bilar, modeller osv. Skulle detta vara ett riktigt projekt kan vissa ha behövt studera lite om bilar innan.

4. Vad är dina styrkor och utvecklingsmöjligheter när du arbetar i grupp? Mina styrkor är att jag kan ta en ledarroll om jag är säker på min sak. Men jag tycker inte om att leda helt själv och lyssnar gärna på vad andra har för åsikter. Utvecklingsmöjligheten är lite samma område. Om det i stället handlar om ett ämne som jag inte har så bra koll på kan jag hamna lite i bakgrunden.

5. Finns det något du hade gjort annorlunda? Vad i sådana fall? Nej jag är fullt nöjd med grupparbetet. Alla bidrog med lika mycket och uppfyllde de mål vi satt.

# Källförteckning

Blocket. (2021). Vad är Blocket?. <https://www.blocket.se/om/vad-ar-blocket>

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2023). An introduction to statistical learning : with applications in R (2nd edition). New York, Springer

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>

Fox, J., & Weisberg, S. (2019). An R Companion to Applied Regression (Third Edition). Sage Publications. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>

Jim Frost. (2023). Root Mean Square Error. Statistics by Jim. <https://statisticsbyjim.com/regression/root-mean-square-error-rmse/>

Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, Davis Vaughan. (2023). dplyr: A Grammar of Data Manipulation. <https://cran.r-project.org/web/packages/dplyr/index.html>

Jacob Kaplan, Benjamin Schlegel. (2023) Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables. <https://cloud.r-project.org/web/packages/fastDummies/fastDummies.pdf>

Venables WN, Ripley BD. (2002). Modern Applied Statistics with S, Fourth edition. Springer, New York. <https://www.stats.ox.ac.uk/pub/MASS4/>

Kira Gylling. (2019). Pyscbwrapper. [pyscbwrapper/pyscbwrapper.ipynb at master · kirajcg/pyscbwrapper · GitHub](https://github.com/kirajcg/pyscbwrapper/blob/master/pyscbwrapper.ipynb)