Petrol cars

Trend and pricing prediction

Dan Heikenberg

EC Utbildning

Examination- R

202404

Abstract

This report will be in two parts. In the first part I´ll try to find any existing trend of the use of petrol cars, by using data from Statistiska Centralbyrån (Statistical Central Office), a Swedish national agency for collecting statistical data from its population. This will be done using “R”, and any necessary plugins. The second part will be a try to train a machine learning model by first web scraping a Swedish popular site for used cars, with the intention of making a model that can be used to predict prices of petrol cars based on predictors like mileage, brand and the year it was made.

Contents

1 Introduction

Petrol and it´s use in cars is a subject that is still relevant. Not only due to its impact on the environment, economic reasons and much more, but because of its potential decline in light of the increasing electric car industry. If there is a visible trend of less registered new petrol cars, that could also have an effect in the second hand market, which could be good knowledge to have for both the buyer and seller.

The purpose of this rapport is to create a starting position for further research into the use of petrol cars, by scraping web data, collecting data from SCB, training models and investigating trends. With this, there can hopefully be a summary about the use and pricing of petrol cars in the future.

1. Can any trend be spotted using data from SCB, concerning the amount of cars using petrol?
2. Can a machine learning model be trained to predict the price of petrol cars using scraped data?
3. Can anything be said about any trend and the pricing of petrol cars?

2 Theory

2.1 Single and multiple Linear Regression Modelling

Linear regression is a model that estimates a linear relationship between a response and a single, or multiple explanatory variables/predictors. The price of a product can be the response variable, and size, material and so on, can be the predictor variables.

2.1.1 Time series object, and Forecasting

Forecasting is predicting using a time series object. To prepare data for forecasting, it needs time values, and it needs to be converted to a special time series object so that the forecasting library can use it to train a model that can look into the future beyond the time values.

2.1.2 Web scraping

Web scraping is a name for the different methods to target and download data from one or more web sites. This is most useful when automated since more data often gives better results. It can be done by using a Python library (addon) such as “Beautiful soup”, or by using special web browser plugins, and so on.

2.1.3 Adjusted R-squared and RMSE

The R-squared value is the proportion of the variance in the response variable that can be explained by the predictor variables in the model (Statology, n.d.). The adjusted R-squared is a modified version that takes into consideration if more predictors are added. A normal R-squared will increase if you add more predictors, which would be bad since those predictors might be completely unrelated to the response variable.

The root mean square error tells us the average distance between the predicted values from the model and the actual values in the dataset. This shows how much error on average the predictions have.

2.1.4 F statistic

The F statistic gives an overall score of how well your model with all predictor variabels provides a fit to the data set. When using a multiple linear regression, it might be misleading to look at the individual predictor values and their importance, since they for example might have an effect on each other.

3 Method

Data concerning registered cars per month was fetched from SCB (Statistiska Central Byrån). This data was imported and transformed in 'R'. The data was then converted to a time series object, to easier see any trend over time. Parameters were tested, and a model was trained to be able to predict 36 months into the future. Data from 2021 to 2024 was used as a test set to previous years which was used as a training set. With this, I got more information about the use of petrol cars.

I scraped data from a popular Swedish second hand buying/selling-site, gathering a collection of about 10´000 cars. This data was later to a great part cleaned by a school colleague of mine, as part of a group assignment, and then I imported it in 'R'. Predictors were evaluated and chosen, and filtered to 'Petrol' cars. Some data was initially skewed, and some outliers were removed. Some predictors were transformed from text to number to make it possible for model training. The goal was to make a model that can predict the price of a petrol car. A multiple linear regression model was trained, with an adjusted R-squared result. This model A test/train-split was also made to explore the root mean square error-value of the model.

Lastly I looked at all data and results to make a summary about the possible future of petrol cars.

4 Result and discussion

With a simple linear regression model using data from SCB and the number of cars registered each month, we can see a down going trend.

Car registrations

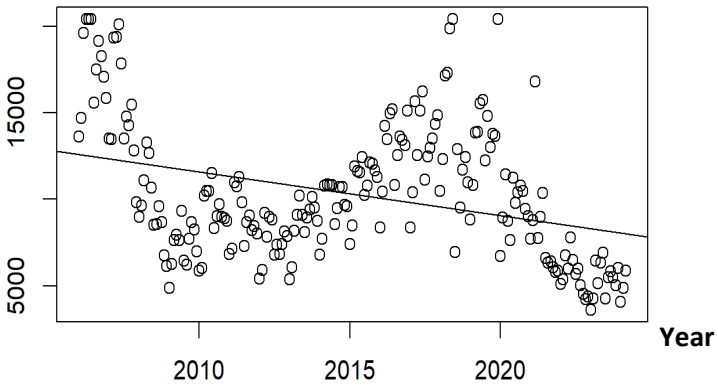

Image 13: A linear regression model with a model plot line, using data from SCB.

With the library 'Forecast', a model with seasonal = false gives the best RMSE: ~ 3902 on its accuracy score, while seasonal = true gives ~ 4175. However, with a future prediction of 36 months, the plot of the seasonal shows a more detailed prediction (image on the left), compared to non-seasonal (image on the right). The shadowed parts are 80% and 95% prediction intervals. The seasonal seems to be repeating, but shows a downward trend.

Car registrations

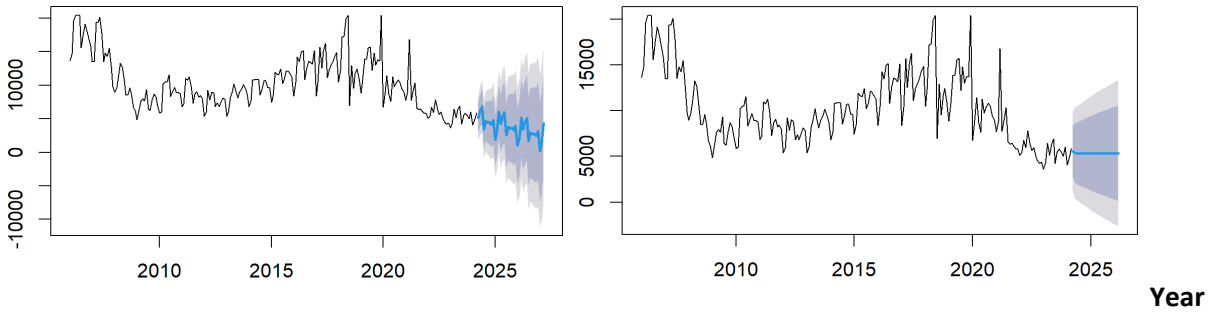

Image 2 and 3: Forecast prediction, seasonal and non-seasonal, using data from SCB.

When looking at the scraped data of second hand petrol cars, we can in this bar plot see the number drop after 2019, but an explanation could be that new cars aren´t as common on second sales markets. Cars around 5 years old might be the largest market.

Petrol cars for sale

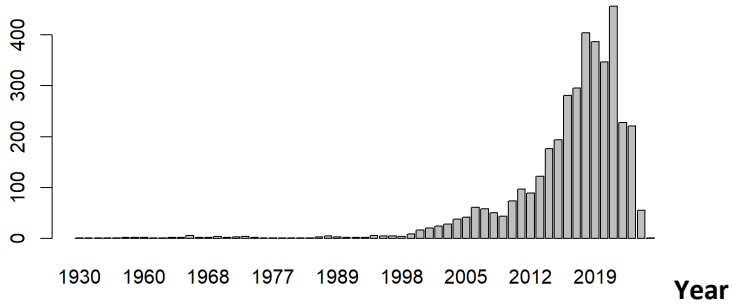

Image 4: Barplot of petrol cars for sale and their production year, using data from Blocket.

Multiple linear regression	
Adjusted R-squared	0.6511
RMSE	~86262

A multiple linear regression model was trained, with a good F-statistic. The adjusted R-squared value was 0.6511, which means that 65.11 % of the total variation in the response variable can be explained by the model. A higher result would be preffered.

A train/test-split was also made to find out the RMSE. This meant that the model didn´t have the test-set to train on, which might have impacted the result, which was a RMSE of ~86262. Root mean squared error is how big the error is compared to the correct value.

These models could be used to predict the price, but they could be better. Most of the results still point to a down going trend in the use of petrol cars, and that might be worth to have in mind while buying and selling petrol cars when deciding on the price. It might be interesting to compare the location of Stockholm to other locations, since there might be more electric charge stations for electric cars, which could make petrol cars cheaper and harder to sell, and also perhaps less common overall.

These results might improve with the inclusion of more predictors, although there were large amounts of missing values in some of these excluded data columns. There could be synergy between for example year and miles, which could be used to possible improve the results, and the Location predictor might help if included. Adjusted R-squared would give a fair result compared to R-squared which would increase just by including more predictors.

5 Conclusions

By finding data through different means, training models and looking at different plots, I can see that there is a possible down going trend in the use of petrol cars. I can also see that there´s a way to predict the prices of petrol cars in the second hand market, while having in mind that the overall price trend could be dropping in the future. There weren´t many new cars on the second hand market, but that could be just because of that they are still new, and not that the trend is taking full effect. However, if the trend has some truth, then there might be fewer second hand cars in the future, and they might be harder to sell which could impact the price.

The model trained on the data from SCB showed a credible down going trend in the time series, but the model trained on the scraped second hand site data, can probably be improved in different ways. It could also be that the data is hard to predict because of the possible down going trend.

The scraped data had several predictors that I left out, due to many missing values that could distort the training, but the Location predictor might provide some additional training. Other improvements could also be tried out, for example by solving the possible synergy between Miles and Year.

A larger web scraping could also be done, with more than 10´000 entries, since not all those 10´000 entries are Petrol cars. These 10´000 car ads include every kind of fuel, from electric and diesel to hybrids.

This is a first cycle of this report, and I might revisit it later to make improvements.

6 Datainsamling

6.1 Vem du har arbetat i grupp med?

W. Blennow, Siarhei, Daniel, Frida m.fl.

6.2 Hur har ni i gruppen arbetat tillsammans?

Bra tycker jag. Jag tog snabbt initiativ till att lära mig webbskrapning eftersom jag insåg det ohållbara med att alla skrapar manuellt, och i värsta fall samma data.

Vi har utnyttjat varandras styrkor. Jag har mest haft kontakt med Blennow som talat med resten av gruppen på dagen, sedan diskuterar jag och han på kvällen, därefter har jag gjort försök till web scraping på natten -tiden då jag arbetar som bäst. Detta har jag laddat upp till Teams till hela gruppen och skrivit om det, sedan har de diskuterat och kollat igenom datan på dagen, och sedan diskuterar jag det på nytt med Blennow på kvällen då han har sammanfattat resten av gruppen, och sedan har jag satt igång med en ny webb-skrapning på natten, o.s.v. tills alla är nöjda. På detta sätt tycker jag att tiden har utnyttjats väl. En grupp måste inte göra allt tillsammans på samma tid, det behöver inte vara det mest effektiva. Jag har också chattat med Daniel och berättat lite om hur web scraping fungerar, och delat med mig av min metod jag bestämde mig för att använda.

Diskussionerna omfattade vad som var möjligt, problem, vilken data som var intressant att skrapa, hur mycket, o.s.v.

Sedan tog Siarhei över med en städning av datan, och sedan var det klart för alla att ladda ner och använda.

6.3 Vad var bra i grupparbetet och vad kan utvecklas?

Jag tyckte det fungerade perfekt. Alla styrkor verkar ha använts, och tidsanvändningen kunde nog inte varit effektivare. Alla fann sina roller och jobbet utfördes.

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

Jag har lätt att ta initiativ om jag känner att det behövs, jag kan även ta ledarrollen, men personligen gillar jag inte gruppuppgifter. Det har varit alldeles för många dåliga erfarenheter, med gruppmedlemmar som inte gör vad de ska, eller inte är lika effektiva. Det är ett stressmoment när det är ens egna betyg/jobb som hänger på det. Jag lär mig hellre alla moment som krävs och gör allt själv. T.e.x så har jag varit med i två spel-projekt där medlemmar plötsligt lämnat projektet. Därför har jag lärt mig programmera, grafik och musik för att kunna göra allt själv. Det kanske inte blir lika bra som en dedikerad grupp där alla har samma vision och samma nivå av effektivitet, men det blir iaf gjort. På en arbetsplats är det större chans att alla gör vad de ska, men på fritiden och på utbildningar där folk kan vara mindre motiverade och även plötsligt hoppa av en kurs, är det värre och mest ett stressmoment.

Jag hade kunnat städa datan också och göra allt själv, jag vill bara få det gjort, men resten av gruppen var väldigt aktiv vilket var roligt. Men jag ogillar fortfarande gruppuppgifter.

6.5 Finns det något du hade gjort annorlunda?

Nej, jag är nöjd. Vår grupp fungerade supereffektivt tyckte jag iaf. Jag vet inte hur de andra känner, men vi jobbade tillsammans och fick fram ett resultat på kort tid.

7 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.

QQ-plot kan hjälpa att visa om observationerna är normalt distribuerade. Normalfördelning är viktigt inom statistik, och därför kan det vara bra att kunna se om datan är normalfördelad, vilket syns som en rak linje i grafen. Sample Quantiles är de n värden vi har från vår data, och Theoretical Quantiles är n+1 värden jämnt area-fördelade utifrån en standard normal distribution. Grafen med dessa y och x ska nu helst visa en rak linje för att visa att datan är normalfördelad.

7.1 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 kanske fokus faller på prediktioner, men inferens är, enligt James m.fl. (2023, s. 20) viktigt om man vill ha en djupare förståelse av sambandet mellan predictorerna och dess response. Vill man t.ex. veta varför och vad som påverkar prediktionerna, så kan detta vara bra att ha i åtanke när modellen tränas, likväl som företag kanske vill veta vad det är som driver deras vinster, tv- eller radio-reklamen, så att de vet vad de borde fokusera på. När modeller tränas så kan man se, inom vissa konfidensintervall, vilka prediktorer som är med och påverkar responsen, och man kan fatta nya beslut under träningsprocessen.

James m.fl. (2023, s. 20) ger ett fint exempel som lyder: Hur mycket extra kommer huset bli värt om det har en utsikt över floden? Detta är ett inferensproblem. Om man endast är intresserad av att prediktera värdet av ett hus med givna värden, om huset är under elelr över-värderat. Det är et tprediktionsproblem.

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

Om man t.ex har en modell som predikterar priset över hus med x antal sovrum, så skulle man, enligt Statology (Statology, n.d) använda ett konfidensintervall för prisernas medelvärde. Prediktionsintervall är för en enda prediktion, alltså priset för ett enda hus. Ett prediktionsintervall är mer osäkert och intervallet är därför bredare än i konfidensintervall.

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?

B0 är, enligt PennState (PennState, n.d) intercept, med andra ord medelvärdet av response-värdet när alla x är 0, och linjen skär y-axeln. Varje parameter multipliceras med en x-variabel, vilket representerar prediktor-variablerna, och varje beta-parameter representerar medelvärdes-förändringen med utgångspunkt att alla de andra prediktorerna hålls konstanta.

James m.fl. (2023, s. 72) menar även de att beta parametrarna tolkas som den genomsnittliga förändringen på Y för var enhet X, då alla andra prediktiorer också är konstanta.

7.3 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?

Det stämmer. Att testa en modell mot ett test-dataset och beräkna resultatet är en teknik, men BIC är en annan teknik som används för att utvärdera den tränade modellen. Med uppdelning till olika set, kan man träna en modell på ett test-set, sedan validera den mot ett validerings-set genom att låta modellen prediktera med hjälp av prediktorerna, och sedan jämföra med de sanna värdena som hållits undangömda. BIC, AIC m.fl. kan utvärdera en tränad modell utifrån ”probabilistic statistical measures”, (Machine Learning Mastery, n.d), med en ”log-likelihood”-funktion, och ge ett värde på modellens styrka och dess komplexitet. Denna Det finns samtidigt nackdelar med denna metoden, t.ex. att allt för simpla modeller kan favöriseras.

7.4 6. Förklara algoritmen nedan för ”Best subset selection”

Det kan finnas många prediktorer i ett dataset, och alla har olika påverkan på responsen. De kan också ha inbördes påverkan på varandra. Ett sätt att träna en modell när man står inför detta, är ett använda en teknik som kallas subset selection, eller best subset selection enligt James m.fl. (2023, s. 227). Här tränas modeller på en prediktor, och man kan sedan utvärdera vilken prediktor som är bäst, och detta fortsätter med att modellerna tränas på två prediktorer, och man kan utvärdera vilken kombination av två prediktioner som ger bäst resultat, o.s.v. Modeller tränas alltså stegvis utifrån alla möjliga k av p kombinationer, och utvärderas med t.ex. minst RSS-värde. Varje steg kallas Mp och ökar samtidigt antalet prediktioner. I sista steget väljer man den bästa modellen av alla M. Best subset selection kan snabbt växa om man inte sätter en lägre gräns på antalet prediktorer.

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

Modeller försöker bara likna verkligheten, men kan aldrig 100% stämma på grund av den riktiga världens komplexitet (Wikipedia, n.d.). Men en bra modell kan fortfarande vara användningsbar. Det är bra att ha det i åtanke när man använder den.

8 Self evaluation

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

   Webbskrapningen var en utmaning, då jag tror att Blocket gjort det lite svårare att skrapa med beautiful soup nu än det var förr, kanske p.g.a cookies och annat som jag inte är helt van vid att arbeta med, så det tog många timmar och många försök att lösa, men tillslut fann jag ett browser-plugin som jag kunde använda mig av. Det var inte heller helt enkelt, men efter många försök och under en viss tids-press så lyckades jag lära mig det.

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

   Jag förstår R och tycker det är lätt och trevligt att arbeta i, men jag har inte gjort API-delen. Jag gjorde däremot webb-skrapningen om det räknas.

3. Något du vill lyfta fram till Antonio?

   Det har varit bra kursmaterial. Videosarna är alltid intressanta. R var roligare än jag trodde. Namnet R lät så torrt. Boken var bra. Vet inte om jag tycker om datacamp. Att fylla i luckor i annars färdig kod är inte hur min hjärna lär sig, men det kanske passar andra och isåfall är det bra!

9 Appendix

```
# Data imported from SCB
# Nyregistrerade personbilar efter region, drivmedel och månad

###
### Transforming table and data
###

# Transform the wide data set so that the time series is in one column.
library(data.table)
df2 <- melt(setDT(df), id.vars = "Fuel", variable.name = "Month")
# Replace "M" character with "-".
df3 <- df2
df3$Month <- gsub("M", "-", df2$Month)

#Change date column from string to date.
df4 <- df3
df4$Month=paste0(df4$Month,"-01")
df4$Month <- as.Date(df4$Month, format="%Y-%m-%d")

#Filter Fuel column
df5 <- df4
df_bensin <- subset(df5, Fuel == "bensin")
class(df_bensin$Month)
# Scatter plot
plot(df_bensin$Month , df_bensin$value)

###
### Data exploration
###

summary(df_bensin)
# There are no NAs.

# Outlier treatment, atleast one value seems high
summary(df_bensin$value)
#Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#3602    7417    9584   10401   12616   32231
# Mean and median are not very far from eachother, but we´ll try capping.

#find Q1, Q3, and interquartile range
Q1 <- quantile(df_bensin$value, .25)
Q3 <- quantile(df_bensin$value, .75)
IQR <- IQR(df_bensin$value)
#Transform rows in dataframe that have values within 1.5*IQR of Q1 and Q3
# Removing values in a time series may open up new problems.
df_bensin$value[df_bensin$value > (Q3 + 1.5*IQR)] <- (Q3 + 1.5*IQR)
```

```r
# Might aswell check lower values
df_bensin$value[df_bensin$value < (Q1 - 1.5*IQR)] <- (Q1 - 1.5*IQR)
plot(df_bensin$Month , df_bensin$value)
summary(df_bensin$value)
#   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#3602    7417    9584   10300   12616   20416
#Mean and median are now closer

# There seems to be a down going trend, but it isn´t completely linear.
# It might be because of the date column, and possibly a seasonal effect.

# Plot with ggplot
install.packages("ggplot2")
library("ggplot2")
ggplot(df_bensin, aes(x = Month, y = value)) +
  geom_line() +
  scale_x_date(date_labels = "%Y-%m")

###
### Linear regression modeling
###

# Run linear regression, down going trend
fit = lm(value ~ Month, df_bensin)
# Find result
summary(fit)
#Estimate Std. Error t value Pr(>|t|)
#(Intercept) 21637.1315  2150.9493  10.059  < 2e-16 ***
#  Month          -0.6885     0.1297  -5.307 2.75e-07 ***
plot(df_bensin$Month, df_bensin$value)
abline(fit)

###
### Forecast models
###

library(forecast)
# time series object
tsobject = ts(df_bensin$value, start = c(2006, 1), end = c(2024, 3),
frequency=12)
plot(tsobject)

# afc allows us to assess how a time series x relates to its past (datacamp,
n.d.)
acf(tsobject) # It´s a down going trend

# Forecast arima model
ts_model_01 = Arima(tsobject, c(0, 0, 0))
str(ts_model_01)
```

```r
# predictions
avg_forecast = forecast(ts_model_01)
str(avg_forecast)
avg_forecast$mean # mean 10300 all months.
plot(tsobject)
lines(avg_forecast$mean, col = "green")
# 95% / 80% prediction intervall
plot(avg_forecast)

# New try, auto.
# auto will automatically discover the optimal order and parameters (alkaline-ml.com)
ts_model_02 = auto.arima(tsobject, seasonal = FALSE)
forecast_02 = forecast(ts_model_02)
plot(forecast_02)

# New try, auto, seasonal
ts_model_03 = auto.arima(tsobject, seasonal = TRUE)
forecast_03 = forecast(ts_model_03)
plot(forecast_03)

# New try, auto, seasonal, longer prediction interval
ts_model_04 = auto.arima(tsobject, seasonal = TRUE)
forecast_04 = forecast(ts_model_04, h = 36)
plot(forecast_04) # down going
# Seems to be repeating, and more downward trending, still somewhat realistic.

# Train / Test
tsobject_train = window(tsobject, end = c(2021, 2))
tsobject_test = window(tsobject, start = c(2021, 3))

# Seasonal model
ts_model_05 = auto.arima(tsobject_train, seasonal = TRUE)
forecast_05 = forecast(ts_model_05, h = 36)
accuracy(forecast_05, tsobject_test)
#                    ME     RMSE       MAE         MPE      MAPE      MASE
   ACF1
#Training set    58.99293 1583.732  991.7019  -0.1200037 10.3475 0.4449603
0.004447212        NA
#Test set     -3720.82483 4175.221 3995.7719 -68.4865967 70.1225 1.7928368
0.143294221  3.568855

# Non-seasonal model
ts_model_06 = auto.arima(tsobject_train, seasonal = FALSE)
forecast_06 = forecast(ts_model_06, h = 36)
accuracy(forecast_06, tsobject_test)
#                   ME     RMSE       MAE         MPE      MAPE      MASE
 ACF1
```

```
#Training set  -104.2799 2240.135 1564.982  -4.484815 15.38318 0.7021816 -
0.0006434039        NA
#Test set     -3184.9853 3902.533 3658.171 -64.455828 67.47567
1.6413607  0.3130755800  3.381875

# Non-seasonal performed better on RMSE, but MAPE is really high on both, and
so none of these models
# might be very good to make future predictions. Thanks to the linear
regression model we can atleast
# see that the trend seems negative/down going.

###
### Blocket Annonser bensin / Pris
###

# Scraped and cleaned data imported
data1 <- read.csv("skrapning_v3_cleaned.csv", header=TRUE,
stringsAsFactors=FALSE)

###
### Data exploration
###

summary(data1)
# Price and miles looks ok median vs mean.
hist(data1$Miles)
# Miles is  right skewed
install.packages("e1071")
library("e1071")
skewness(data1$Miles) #~1.028
kurtosis(data1$Miles) #~5.310
install.packages("moments")
library("moments")
jarque.test(data1$Miles) # Less than 0.05, nor normal distributed

# Interested in (Price) Fuel -> Bensin, Year, Miles, Gear, and all Brands
# The Bensin filter will exclude models like Tesla.
barplot(table(data1$Gear)) # Something is creating an empty column
barplot(table(data1$Brand)) # Volvo is most frequent
barplot(table(data1$Year)) # Year seems left skewed


# Removing of outliers
#find Q1, Q3, and interquartile range
Q1v2 <- quantile(data1$Price, .25)
Q3v2 <- quantile(data1$Price, .75)
IQRv2 <- IQR(data1$Price)
#Transform rows in dataframe that have values within 1.5*IQR of Q1 and Q3
# Removing values in a time series may open up new problems.
```

```r
data1$Price[data1$Price > (Q3v2 + 1.5*IQRv2)] <- (Q3v2 + 1.5*IQRv2)
pairs(~Price+Miles, data = data1)
# Looks almost linear
summary(data1$Price)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 1  149900  229800  258987  339000  622650

# Remove rows where Gear has no value
data1 <- data1[!(data1$Gear == ""), ]
barplot(table(data1$Gear)) # Empty strings are now removed

# Filter out all except "Bensin"
data2 <- subset(data1, Fuel == "Bensin")
barplot(table(data2$Year)) # Bensin seems to follow the trend



install.packages("ggpubr")
library("ggpubr")

# Miles skewness
data2$MilesLog <- log10(data2$Miles)
ggdensity(data2, x = "Miles", fill = "lightgray", title = "Miles") +
  stat_overlay_normal_density(color = "red", linetype = "dashed")
# Positive skewness

# Remove other columns
#+ #remove by names. All are Bensin, so I can remove Fuel aswell
data3 <- data2[ , ! names(data2) %in% c("X", "Company", "Location", "Fuel",
"Name", "Model", "Engine.Volume", "Horsepower", "PriceLog")]

# Transform dummies for conversion to numeric (nominal)
install.packages("fastDummies")
library(fastDummies)
data4 <- fastDummies::dummy_cols(data3)

# Now we can remove the non-numeric columns
data5 <- data4[ , ! names(data4) %in% c("Gear", "Brand")]

# Linear modelling
#lets try all variables
multiple_model <- lm(Price~.,data=data5)
summary(multiple_model)
# F statistic is very small, p-value: < 2.2e-16, which shows good relationship
# between predictor and response variable.

# Adjusted R-squared:  0.6907
```

```r
# Adjusted R squared can replace the need for training / test sets.

# Volvo and Manuell shows NA, and that´s probably because every transformed
# fastDummy set should be one column less (n classifications -1). One column
# gear value is enough to explain the other gear column.

# Lets try and remove those columns before continuing, and also miles since we
# have a new MilesLog column.
data6 <- data5[ , ! names(data5) %in% c("Gear_Manuell", "Brand_Volvo",
"Miles")]
multiple_model <- lm(Price~.,data=data6)
summary(multiple_model)
# Many significant p-values
#Adjusted R-squared:  0.6511 which is not very good, but still a result.
# 65.11 %  of the total variation in the response variable can be explained by
the model.
# Worth to know, R-squared will always increase when a new predictor variable
is added to the regression model.
# The adjusted R-squared is a modified version of R-squared that adjusts for
the number of predictors in a regression model.
# https://www.statology.org/adjusted-r-squared-interpretation/

# Train / Test -split
install.packages("caTools")
set.seed(0)
split = sample.split(data6,SplitRatio = 0.8)
training_set = subset(data6,split == TRUE) #all TRUE values get to training
set
test_set = subset(data6,split == FALSE)
# Training of model
lm_a = lm(Price~.,data=training_set)
summary(lm_a) # Adjusted R-squared:  0.6482
# Predicting, with respect to the log conversion.
train_a = predict(lm_a,training_set)
test_a = predict(lm_a,test_set)

# mean squared error, to see how good the model is.
# mean squared error is how big the error is compared to the correct value.
# lets find mean squared error
rmse(test_set$Price, test_a)
# RMSE is 86262.51

# Not a very good result, but it´s a start.

# There are more things one could do, check for multicolliniarity,
# and thereby think about using Lasso, or other models, and include more
# predictors.

# Why I didn´t use/include other predictors from the data:
```

```r
sum(data2$Horsepower == "Unknown") # 1940
# There are too many unknown values in this column. I think it might give a
false picture.
sum(data2$Engine.Volume == "Unknown") # 3199
# Too many missing values in Engine Volume
length(unique(data2$Model)) # 1281
# There are too many unique values in Model for the size of the sample
# Location: I can´t think of why this predictor would change the price,
# except maybe around Stockholm.

# There could be synergy between year and miles.
# An idea would be to make a combined value and
# include the individual values in the model aswell (Hierarchical Principle).
```

References

Blocket. Hämtad 20 mars, 2024, från Blockets ´s sida https://www.blocket.se/

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2023). An Introduction to Statistical Learning with Applications in R. Second edition. Springer.

Machine Learning Mastery. Probabilistic model selection. Hämtad 20 mars, 2024, från Machine Learning Mastery´s sida https://machinelearningmastery.com/probabilistic-model-selection-measures/

Pennstate. The Multiple Linear Regression Model. Hämtad 20 mars, 2024, från Pennstate´s sida https://online.stat.psu.edu/stat501/lesson/5/5.3

Statistiska Central Byrån. Hämtad 20 mars, 2024, från Statistiska Central Byrån ´s sida https://www.scb.se/

Statology. Adjusted R squared interpretation. Hämtad 20 mars, 2024, från Statology´s sida https://www.statology.org/adjusted-r-squared-interpretation/

Statology. Confidence interval vs prediction interval. Hämtad 20 mars, 2024, från Statology´s sida https://www.statology.org/confidence-interval-vs-prediction-interval/

Wikipedia. All models are wrong. Hämtad 20 mars, 2024, från Wikipedia´s sida https://en.wikipedia.org/wiki/All_models_are_wrong