Prissättning på elbilar

Faktorer som påverkar priset

Natalie Dobrovolska

EC Utbildning

R-programmering

Abstract

The objective of this thesis is to develop a model which can predict prices on electric cars and examine the factors that may have effect on them.

The factors that will be examined are Price, Year, Mileage and Location. The study will use information from Blocket - one of the biggest swedish sales platforms.

Innehållsförteckning

1 Inledning

Marknaden för elbilar är kontant växande. Enligt centrala statistiska byrån har försäljningen av elbilar tredubblats sedan 2020[1].

En av de platserna där man kan köpa elbilar är Blocket. Variationen av annonserna är bred. Det finns bilar i olika prisklasser, färger och modeller.

En fråga som dyker upp är vad de olika priserna kan bero på. Kan det finnas att samband med vilket område de säljs på? Kan priset bero på tillverkningsåret?

Denna rapport kommer att fokusera på prissättning samt kommer att undersöka faktorer som kan påverka priset. Det kommer även att göras en regressionsmodellering för att kunna prediktera priserna.

[1] https://www.scb.se/hitta-statistik/redaktionellt/tredubbling-av-elbilar-pa-tva-ar2/

2 Teori

2.1 Multipel linjär regressionsanalys

För att kunna genomföra arbetet med analys av prissättning och faktorer som kan vara påverkande, kommer en multipel linjär regression att genomföras. Med hjälp av den kommer även prediktion av priser att göras.

Linjär regression är ett användbart verktyg för att förutsäga ett kvantitativt svar på beroende variabeln Y utifrån en eller flera oberoende variabler X. (James, Witten, Hastie, Tibshirani 2023, s.59-73) Det antar att det finns ett linjärt förhållande mellan X och Y. Matematiskt kan detta linjära förhållande skrivas som:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon i$$

där:

- Y är den beroende variabeln.
- X1, X2…, Xp är oberoende variabler.
- β0, β1, β2…, βp är regressionskoefficienter som representerar effekten av varje oberoende variabel på den beroende variabeln.
- ϵ är en slumpmässig felterm som fångar all ospecificerad variation i den beroende variabeln.

I denna rapport är den beroende variabeln Pris. De oberoende variablerna är Plats, Tillverkningsår, Miltal och Märke.

3 Metod

3.1 Datainsamling

1. Vem du har arbetat i grupp med?

Jag jobbade med Daniel H, William, Xiaoyong, Frida, Dan, Khaldoun, Melike, Siarhei.

2. Hur har ni i gruppen arbetat tillsammans?

Under lektionen började vi i gruppen diskutera vad som skulle vara intressant att undersöka. Vi ville begränsa oss i insamlingen av informationen och då dök idén om att undersöka elbilar upp.

Vi började med att kolla Blockets annonser för att se vilken information som fanns tillgänglig om bilar. Där bestämde vi att vi skulle välja dessa parametrar:

- Bränsle, Växellåda, Miltal, Modellår, Biltyp, Drivning, Hästkrafter, Färg, Datum i trafik, Märke, Modell, Pris.

Varje person fick var sitt län att för att undvika överlappningen av datainsamlingen. Vi utgick från de senaste annonserna, bilar sålda av endast företag. Varje person skulle samla in 30 annonser.

När Dan tilldelades i vår grupp, föreslog han web scraping och han ordnade en fil med över 12 400 observationer med all information som vi diskuterade tidigare.

Sierhei tog på sig formatering och rensning av filen, då det fanns mycket som behövdes göras om för att kunna arbeta vidare med informationen.

3. Vad var bra i grupparbetet och vad kan utvecklas?

Bra: Initiativtagande, idéer, engagemang.

Utvecklingsmöjligheter: Kan inte komma på något. Arbetet och diskussioner gick mycket smidigt och bra.

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

Mina styrkor: Engagerad och samarbetsvillig.

Utvecklingsmöjligheter: Fortfarande lite osäker när det gäller vissa statistiska och modellerings moment.

5. Finns det något du hade gjort annorlunda? Vad i sådana fall?

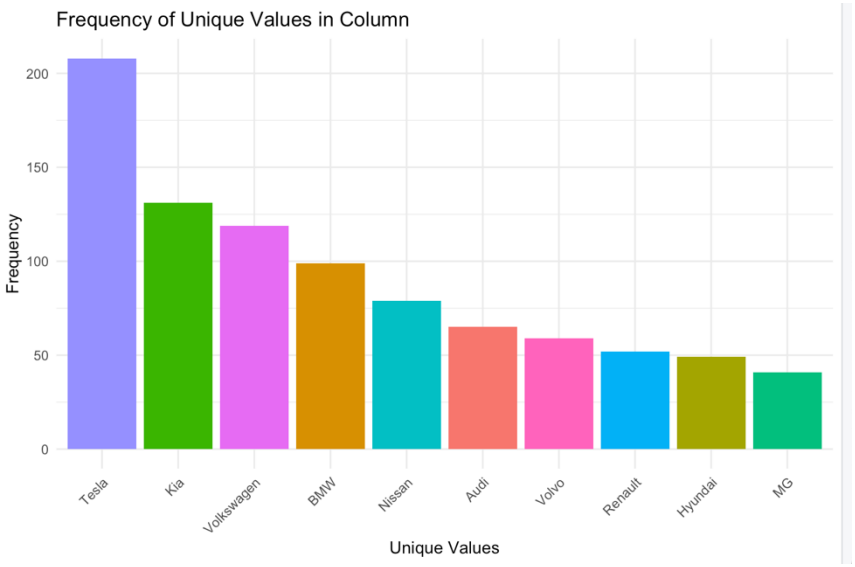
Nej, jag nöjd med hur vårt samarbete gick.

3.2 Databearbetning och analys

I detta arbete användes data från web scraping.

Totalt analyserades 902 observationer om elbilar sålda av företag. Det valdes att begränsa studien till fem variabler – Plats, Tillverkningsår, Miltal, Pris och Märke. Resterande variabler uteslöts.

Den beroende variabeln blev Pris och oberoende – Plats, Tillverkningsår, Miltal, Märke.



Figur 1. Topp tio mest förekommande elbilmärken

Antalet av observationer är ett resultat av en begränsning till topp tio mest förekommande elbilmärken till antal. (Figur 1)

Det gjordes en faktorisering av variabler som tillhörde nominaskala[2]. Dessa variabler blev Plats och Märke.

Majoritet av elbilar som undersöktes såldes i Stockholm, Göteborg och Skåne. Den äldsta modellen var från 2012, den nyaste från 2024. Miltal sträckte sig från 1 till 23 600. Lägsta observerade priset var 2 595 och högsta 1 729 900. De topp 3 märken med flest bilar till försäljning var Tesla, Kia och Volkswagen. (Figur 2)

```
> summary(elcars_top10)
        Location        Year          Miles           Price               Brand
 Stockholm   :350   Min.   :2012   Min.   :    1   Min.   :   2595   Tesla     :208
 Göteborg    :107   1st Qu.:2020   1st Qu.: 1500   1st Qu.: 275750   Kia       :131
 Skåne       :107   Median :2021   Median : 3621   Median : 399900   Volkswagen:119
 Östergötland: 26   Mean   :2021   Mean   : 4413   Mean   : 419601   BMW       : 99
 Kalmar      : 25   3rd Qu.:2023   3rd Qu.: 6224   3rd Qu.: 529900   Nissan    : 79
 Örebro      : 25   Max.   :2024   Max.   :23600   Max.   :1729900   Audi      : 65
 (Other)     :262                                                    (Other)   :201
```

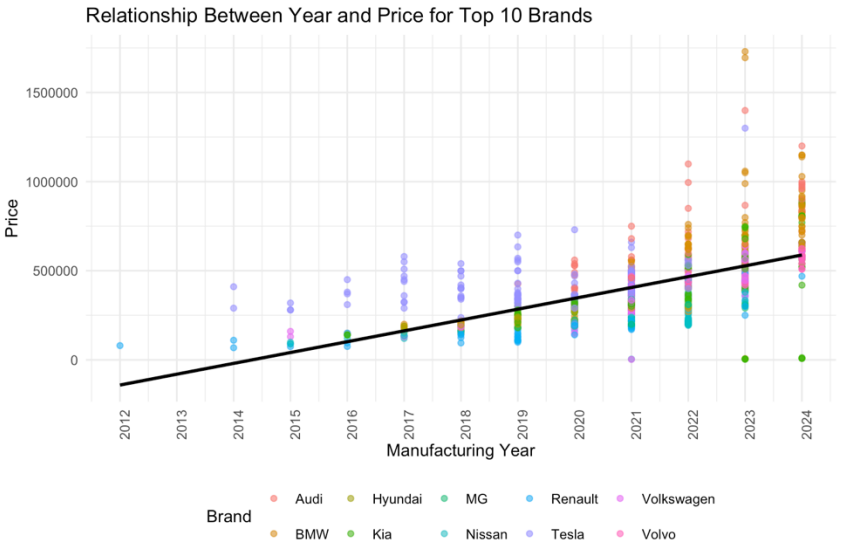
Figur 2. Sammanfattning om variablerna

[2] Den undersökta data som är av en kategorisk karaktär blev omvandlad till en numerisk form.

Vidare undersöktes samband mellan priset och tillverkningsåret. (Figur 3)

Regressionslinjen visade en positiv trend – bilar verkar vara dyrare ju nyare de är. Det fanns dock en stor spridning av priser inom varje tillverkningsår, vilket indikerar att andra faktorer än enbart året kan påverka priset.

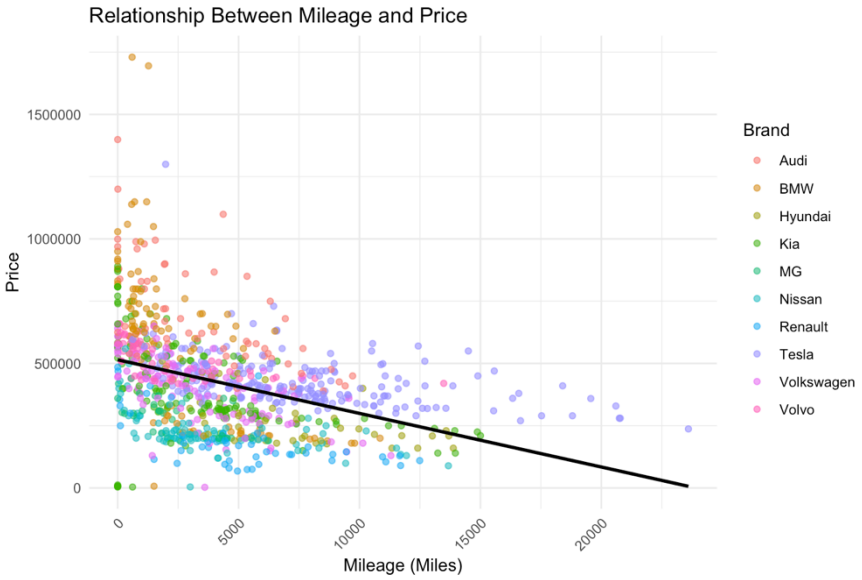
Det fanns också ett antal utomstående punkter, särskilt i nyare årsmodeller. Förklaringen till det kan vara lyxbilar eller modeller med vissa egenskaper som driver upp priset.



Figur 3. Samband mellan pris och tillverkningsår

När man kollade på priset och miltalet så fanns det en tydlig nedåtgående trend som tydde på att bilarnas pris minskade ju högre körsträcka de hade. Det kan ses som en viktig faktor i prissättningen. (Figur 4)

Det fanns dock en stor variation i pris över olika körsträckor och vissa bilmärken. T ex Tesla verkar behålla sitt pris bättre än andra bilmärken även vid högre körsträckor.

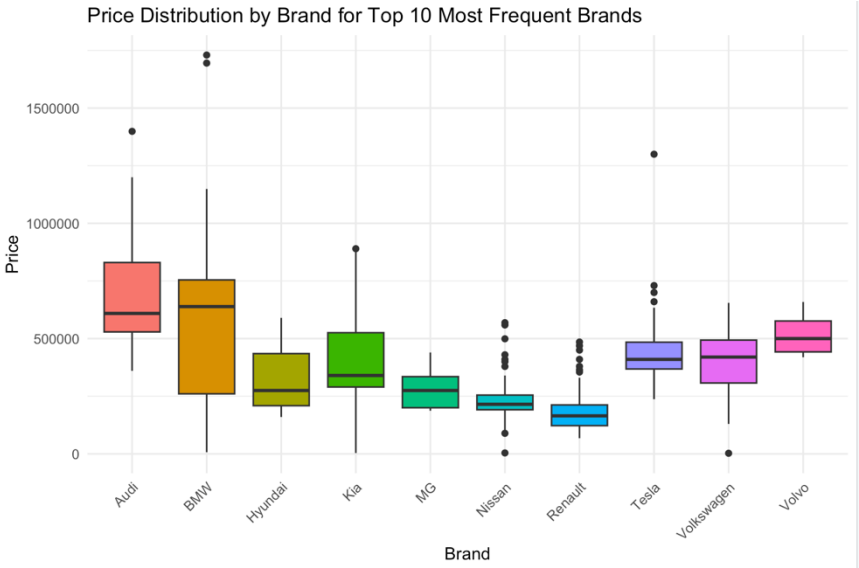


Figur 4. Samband mellan pris och miltal

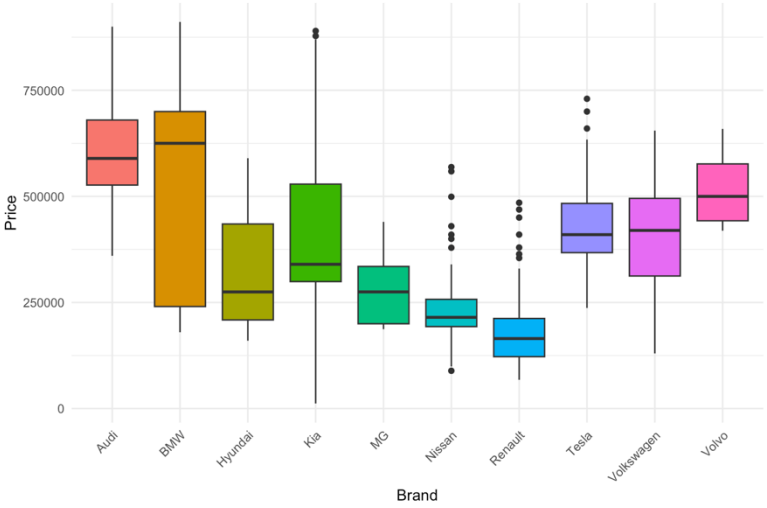
En annan visualisering över pris och bilmärke visade följande: (Figur 5)

Det ser ut som att vissa märken har högre medianpriser och bredare prisintervall (Audi och BMW). Andra märken har tvärtom - ett lägre medianpris och ett mer koncentrerat prisintervall (Kia, Nissan, Renault).

Utifrån visualiseringen kunde man också se att det fanns outliers. Dessa togs bort för att inte påverka regressionsmodelleringen negativt.



Figur 5. Samband mellan pris och bilmärke



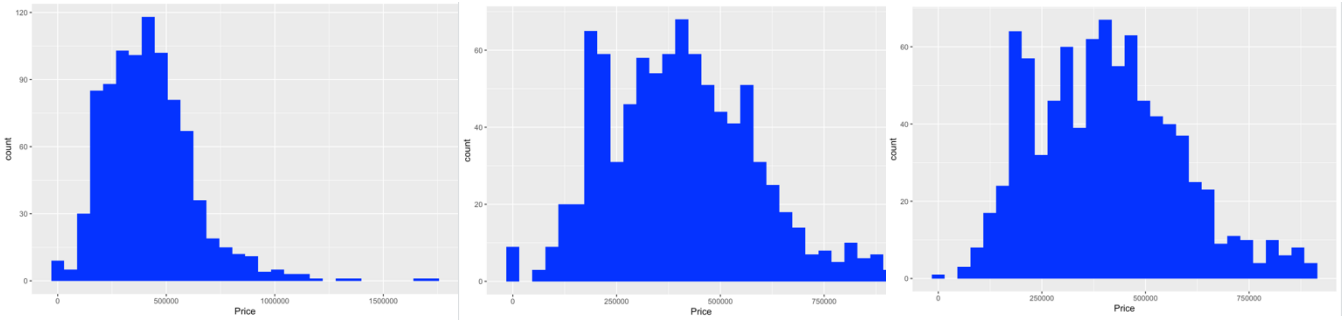
Figur 6. Samband mellan pris och bilmärke efter att outliers rensades bort

Det fanns några extremvärden kvar som behandlades senare. (Figur 6)

Om man kollar på fördelningen av bilpriserna, så finns det en indikation på att den har positiv skevhet. (Figur 7) Man har en högre koncentration av bilar till ett lägre pris och få bilar till ett högt pris. Alternativt ser den ut på det viset pga extremvären/outliers är kvar.

För att åtgärda detta rensades outliers i form av identifiering av första och tredje kvartiler (25:e och 75:e percentiler), beräkning av interkvartilområdet (IQR) genom att använda kvantilerna och sedan beräkning av nedre och övre gränser för outliers genom användning av kvartilerna och interkvartilområdet. En ytterligare åtgärd gjordes, där det filtrerades bort alla värden för den beroende variabeln som är lägre än 10 000.

Resultatet kan man se i nedan histogram. (Figur 7)



Figur 7. Fördelning av bilpriser och resultat efter att outliers blev rensade

Så här ser den uppdaterade sammanfattningen över den undersökta data. (Figur 8)

Man ser ändringen i de flesta kolumner, men viktigast i Pris, då rensning av outliers gjordes.

```
> summary(elcars_top10_newcleaned)
       Location        Year          Miles            Price             Brand       
 Stockholm   :344   Min.   :2012   Min.   :    1   Min.   : 11990   Tesla     :207  
 Göteborg    :104   1st Qu.:2020   1st Qu.: 1655   1st Qu.:278300   Kia       :126  
 Skåne       :104   Median :2021   Median : 3731   Median :399500   Volkswagen:118  
 Östergötland: 26   Mean   :2021   Mean   : 4529   Mean   :406291   BMW       : 87  
 Kalmar      : 25   3rd Qu.:2023   3rd Qu.: 6300   3rd Qu.:524800   Nissan    : 78  
 Örebro      : 24   Max.   :2024   Max.   :23600   Max.   :911000   Volvo     : 59  
 (Other)     :246                                                   (Other)   :198  
```

Figur 8. Sammanfattning av undersökt data

3.3 Multipel linjär regressionsanalys

3.3.1 Test and train

Den observerade data delades upp och Test and Train.

Dessa resultat fick man fram: (Figur 9)

Lågt p-värde i interceptet tyder på att det finns en signifikant relation till den beroende variabeln - Priset.

Tillverkningsår med ett högt t-värde och lågt p-värde indikerar att det finns en positiv relation till Priset. Nyare bilar verkar ha högre priser. Den är också väldigt signifikant för prediktion av priserna.

Miltal har en negativ relation till Priset och med ett lågt p-värde, innebär det att en högre körsträcka leder till lägre priser. Mycket signifikant för prediktion av priserna.

Märken verkar ha signifikant påverkan på priset utifrån det låga P-värdet (under 0,001). Det finns stora skillnader i prissättning mellan märken.

Plats har mindre påverkan på priset. Men det finns några platser som har större påverkan på priset än andra. De billigaste bilarna i genomsnitt säljs i Kronoberg, Stockholm och Halland. De dyraste i Göteborg, Skåne och Gävleborg.

```
Call:
lm(formula = Price ~ Year + Miles + Brand + Location, data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max
-500554  -63214   -6636   49627  377256

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -8.310e+07  5.983e+06 -13.890  < 2e-16 ***
Year               4.141e+04  2.956e+03  14.007  < 2e-16 ***
Miles             -5.459e+00  1.614e+00  -3.383  0.00076 ***
BrandBMW          -9.832e+04  1.961e+04  -5.015 6.80e-07 ***
BrandHyundai      -2.234e+05  2.217e+04 -10.077  < 2e-16 ***
BrandKia          -1.915e+05  1.874e+04 -10.216  < 2e-16 ***
BrandMG           -3.170e+05  2.362e+04 -13.421  < 2e-16 ***
BrandNissan       -3.199e+05  2.036e+04 -15.712  < 2e-16 ***
BrandRenault      -3.131e+05  2.268e+04 -13.805  < 2e-16 ***
BrandTesla        -8.774e+04  1.780e+04  -4.930 1.04e-06 ***
BrandVolkswagen   -1.912e+05  1.870e+04 -10.224  < 2e-16 ***
BrandVolvo        -1.420e+05  2.115e+04  -6.717 3.96e-11 ***
LocationBlekinge   1.021e+05  4.398e+04   2.322  0.02055 *
LocationDalarna    2.176e+04  4.623e+04   0.471  0.63806
LocationGävleborg  4.507e+04  3.945e+04   1.143  0.25362
LocationGöteborg   7.157e+03  3.360e+04   0.213  0.83140
LocationHalland   -3.646e+04  4.102e+04  -0.889  0.37438
LocationJämtland   2.783e+04  5.484e+04   0.507  0.61201
LocationJönköping  4.487e+04  4.132e+04   1.086  0.27791
LocationKalmar     4.339e+04  3.847e+04   1.128  0.25979
LocationKronoberg -9.296e+03  4.752e+04  -0.196  0.84497
LocationNorrbotten 3.275e+04  4.020e+04   0.814  0.41566
LocationÖrebro    -1.857e+04  3.916e+04  -0.474  0.63542
LocationÖstergötland 3.958e+04  3.810e+04   1.039  0.29927
LocationSkåne      7.127e+03  3.377e+04   0.211  0.83290
LocationSkaraborg  3.130e+03  4.072e+04   0.077  0.93875
LocationSödermanland 3.993e+04  4.149e+04   0.963  0.33612
LocationStockholm -5.608e+03  3.236e+04  -0.173  0.86245
LocationUppsala    1.355e+04  3.900e+04   0.347  0.72841
LocationVärmland   4.030e+04  4.500e+04   0.896  0.37084
LocationVästerbotten -3.301e+03  4.380e+04  -0.075  0.93995
LocationVästernorrland 3.265e+04  4.761e+04   0.686  0.49303
LocationVästmanland  2.050e+04  4.091e+04   0.501  0.61641
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 99040 on 668 degrees of freedom
Multiple R-squared:  0.684,     Adjusted R-squared:  0.6689
F-statistic: 45.19 on 32 and 668 DF,  p-value: < 2.2e-16
```

Figur 9. Multipel linjär regression - resultat

Standardfelet för residualerna (RSE) är relativt hög, det tyder på att det finns utrymme för förbättringar av modellen.

Multiple R-squared värdet kan beskrivas att ca 68% av variationen i Pris förklaras av de oberoende variablerna.

Den justerade R-kvadraten är relativt hög 67% och kan tolkas att modellen är robust.

F-värdet tyder på att modellen är statistisk signifikant och effektiv för att förutsäga priserna.

3.3.2 Prediktioner

Efter att modellen tränades gjorde man prediktion på testdatan.

Låt oss kolla på RMSE och R^2.

De resultaten man fått är RMSE = 92 302,32 och R^2 = 0,7364.

I förhållande till medianpriset på bilar (399 500) så är RMSE är ett relativt högt värde. Det indikerar att modellen inte fångar alla aspekter av data och predikterar bilpriserna inte så bra.

R^2 värde som man fått förklarar variansen i priset med 73%. Även här finns utrymme för förbättring.

För att ta reda på vad som kan förbättras i modellen tar man fram VIF (Variance Inflation Factor) värdet. (Figur 9) Med hjälp av den kan man se hur mycket multi-kollinearitet finns i modellen.

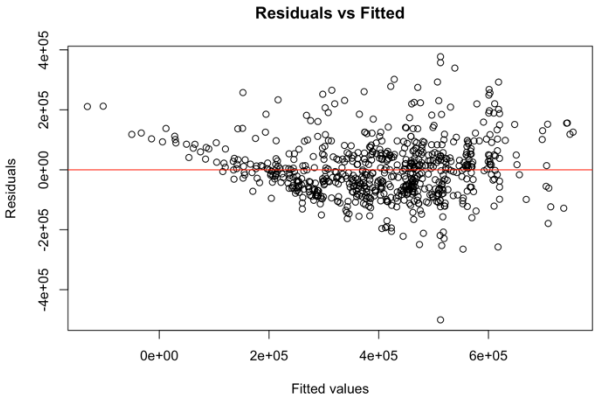
```
> vif(model)
                GVIF Df GVIF^(1/(2*Df))
Year       2.527912  1        1.589941
Miles      2.641391  1        1.625236
Brand      2.930530  9        1.061552
Location 2.078524 21        1.017573
```

Figur 9. Sammanfattning av den undersökta data

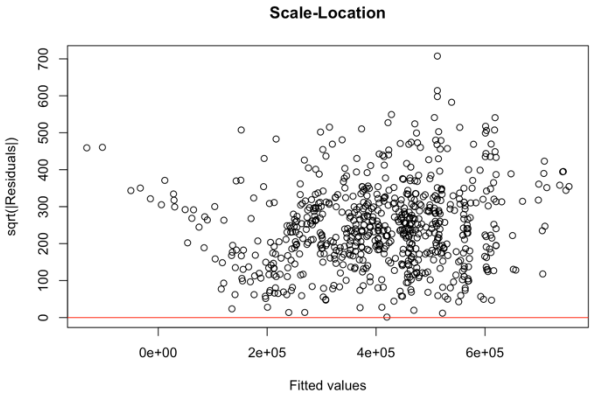
Alla VIF-värden ligger inom acceptabla gränser som kan tolkas att multicollinearitet inte är ett orsakande faktor till sämre prestation av modellen.

Om man kollar på residualerna (Figur 10) så är de ganska jämnt spridda som är bra.
Det finns dock en högre koncentration av punkter i mitten av grafen samt några outliers som ligger längre ifrån noll-linjen.

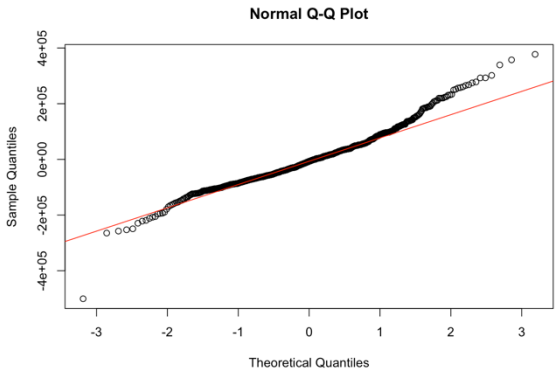
Man ser inte tydliga tecken på heteroskedasticitet, men det finns en ökad spridning av residualer i högra delen av bilden. En ytterligare kontroll för heteroskedasticitet görs lägre ner.



Figur 10. Residualerna



Figur 11. Residualerna



Figur 12. Outliers

Kollar man närmare på spridningen av residualerna så kan man se indikation på en viss heteroskedasticitet. (Figur 11)

Precis som på förra bilden ser man på att spridningen verkar öka något med större förutsedda värden. Det kan vara anledningen till modellens sämre skattningar.

Om man kollar på QQ-plot så ser man att punkterna följer den röda linjen i stor uträckning. Det betyder att residualerna är normalfördelande. Dock finns det avvikelser i båda änden av svansen som tyder på extremvärden/ outliers. (Figur 12)

3.3.3 Justering av modellen och prediktioner

Som nästa steg tränades modellen om med ett antal justerade parametrar för att lösa ovannämnd problematik.

Man rensade outliers genom att bestämma kvantilgränser till 0,05 (nedre gräns kvantil) och till 0,95 (övre gräns kvantil). Heteroskedastisitet var inte tillräckligt tydlig, så det gjordes ingen justering av datan för att åtgärda den. Man kunde annars använt sig av kvadratrotstransformation av logaritmerade priset.

Datan delades i tränings- och testuppsättningar och tränades på nytt. Resultaten som följde visade en annorlunda och något förbättrad modellanpassning jämfört med den tidigare. (Figur 13)

Det som stack ut var Miltal som har fått ett högre P-värde som tyder på att milantalet inte har en stark påverkan på bilarnas pris.

RSE minskade från 99 040 till 82 320 som indikerar på mer exakta prediktioner.

Multiple R-squared värdet visar en liten försämring från 0,684 till 0,6743 och förklarar en något mindre andel av variansen i beroende variabeln jämfört med den första modellen.

Adjusted R-squared fick ett förbättrat värde som innebär att modellen anpassar sig bättre.

F-statistik visar mycket låga p-värden som betyder att modellen är statistisk signifikant. F-värde blev dock lägre i den senaste modellen. Det kan tolkas som att sambandet mellan prediktorerna och den beroende variabeln blev svagare.

Nästa steg var att återigen göra prediktioner med hjälp av den uppdaterade modellen.

```
Call:
lm(formula = Price ~ Year + Miles + Brand + Location, data = training_set_trimmed)

Residuals:
    Min      1Q  Median      3Q     Max
-241617  -55244   -6997   45906  282971

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -8.893e+07  6.289e+06 -14.141  < 2e-16 ***
Year               4.427e+04  3.109e+03  14.240  < 2e-16 ***
Miles             -2.696e+00  1.513e+00  -1.783 0.075167 .
BrandBMW          -7.094e+04  1.903e+04  -3.729 0.000211 ***
BrandHyundai      -1.750e+05  1.963e+04  -8.915  < 2e-16 ***
BrandKia          -1.626e+05  1.715e+04  -9.478  < 2e-16 ***
BrandMG           -2.622e+05  2.198e+04 -11.928  < 2e-16 ***
BrandNissan       -2.848e+05  1.877e+04 -15.171  < 2e-16 ***
BrandRenault      -2.721e+05  2.307e+04 -11.794  < 2e-16 ***
BrandTesla        -3.924e+04  1.637e+04  -2.397 0.016842 *
BrandVolkswagen   -1.374e+05  1.698e+04  -8.094 3.24e-15 ***
BrandVolvo        -8.824e+04  1.884e+04  -4.684 3.49e-06 ***
LocationBlekinge   6.152e+04  4.023e+04   1.529 0.126731
LocationDalarna   -5.746e+03  3.813e+04  -0.151 0.880267
LocationGävleborg  5.787e+04  3.402e+04   1.701 0.089447 .
LocationGöteborg   7.673e+03  2.852e+04   0.269 0.787993
LocationHalland   -3.642e+04  3.651e+04  -0.998 0.318922
LocationJämtland   4.349e+04  5.611e+04   0.775 0.438606
LocationJönköping  1.817e+04  3.801e+04   0.478 0.632784
LocationKalmar     6.086e+04  3.530e+04   1.724 0.085183 .
LocationKronoberg  2.165e+04  3.685e+04   0.587 0.557109
LocationNorrbotten 3.082e+04  3.454e+04   0.892 0.372640
LocationÖrebro    -2.424e+04  3.449e+04  -0.703 0.482440
LocationÖstergötland 3.777e+04  3.283e+04   1.151 0.250397
LocationSkåne     -7.736e+03  2.890e+04  -0.268 0.789035
LocationSkaraborg -6.799e+03  3.604e+04  -0.189 0.850448
LocationSödermanland 3.699e+04  3.888e+04   0.951 0.341761
LocationStockholm -1.603e+04  2.766e+04  -0.580 0.562413
LocationUppsala   -2.029e+04  3.492e+04  -0.581 0.561356
LocationVärmland   3.855e+04  3.809e+04   1.012 0.311940
LocationVästerbotten -2.628e+04  4.173e+04  -0.630 0.529008
LocationVästernorrland 4.949e+04  4.022e+04   1.230 0.219042
LocationVästmanland -3.519e+03  3.603e+04  -0.098 0.922229
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82320 on 597 degrees of freedom
Multiple R-squared:  0.6743,    Adjusted R-squared:  0.6568
F-statistic: 38.62 on 32 and 597 DF,  p-value: < 2.2e-16
```

Figur 13. Multipel linjär regression (efter justeringar)

Resultatet som man fått visade ett lägre RMSE 84 268 jämfört med tidigare 92 302 som kan tolkas att prediktionerna blev mer exakta.

Dock så har R^2 visat en minskning (från 0,7364 till 0,6196). Modellen förklarar en mindre andel av variansen i det beroende variansen jämfört med den ursprungliga modellen.

VIF modellen visar justerade värden, men överlag är de fortfarande ganska låga som betyder att multicollinearitet är inte så allvarlig att den behöver hanteras. (Figur 14)

```
> vif(model_trimmed)
             GVIF Df GVIF^(1/(2*Df))
Year     2.830535  1        1.682419
Miles    2.978966  1        1.725968
Brand    3.315401  9        1.068855
Location 2.355037 21        1.020604
```

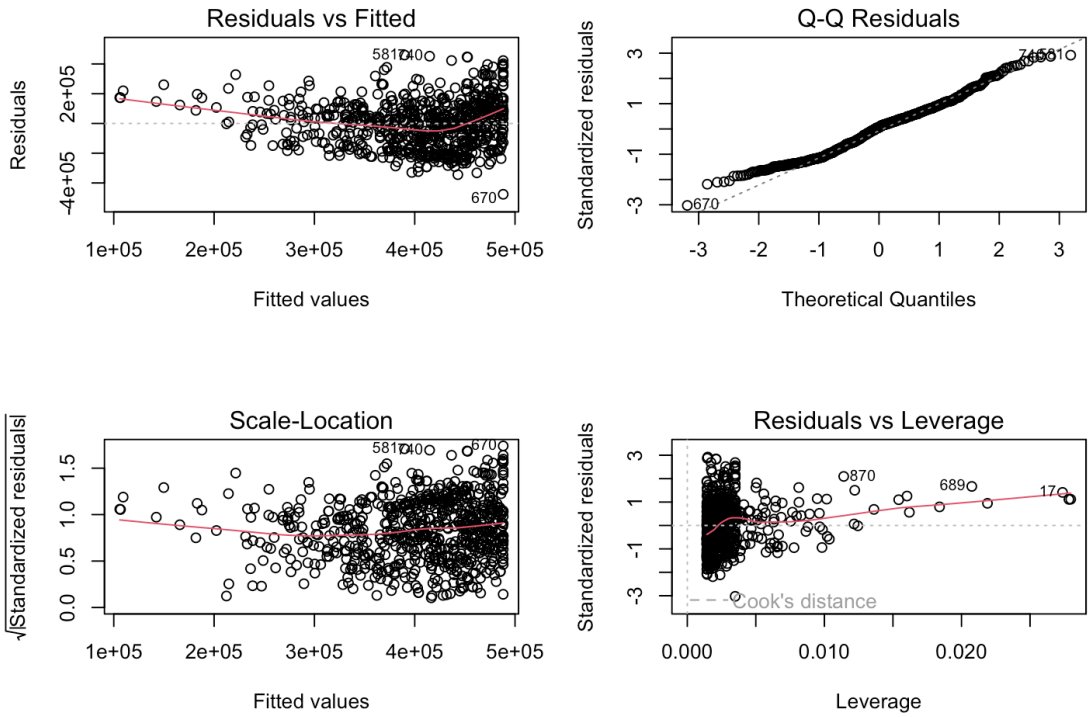
Figur 14. Multipel linjär regression (efter justeringar)

Sammanfattningsvis kan man säga att den justerade modellen presterar bättre när det gäller prediktioner. Dock utifrån F-statistik presterade den första modellen bättre, den var mer signifikant och hade bättre passning av data.

3.3.4 Best Subset Regression

Som ett alternativ testades Best Subset Regression. Målet var att få bättre prediktion av priser.

Resultatet kan man se i Figur 15.
Om man ska beskriva kort så visar residualerna en normalfördelning med några avvikelser vid svansarna, möjligen outliers. Man kan också se att variansen ökar med de anpassade värdena. Det finns en indikation på heteroskedasticitet, då variansen ökar med de anpassade värdena.



Figur 15. Best Subset Regression

När det kommer till RMSE värdet så blev den 160 178 som är mycket högt prediktionsfel.

Adjusted R-squared värdet blev 0,161 som är relativt lågt. Modellen förklarar inte stor del av variationen i den beroende variabeln.

Test_R_squared (Test R-squared) värde blev mycket lågt – 0,181. Det tyder på att modellen inte presterar väl på osett data.

BIC på 18 786,33 är ett högt värde, det tyder på att det kan finnas bättre modeller för datamängden.

Sammanfattningsvis presterade modellen avsevärt sämre än de tidigare. Den måste utforskas mer vad dessa resultat kan bero på, men några möjliga orsaker kan vara överanpassning, multikollinearitet eller för få dataobservationer.

3.3.5 Lasso regression

Det testades även Lasso regression för att se ifall det går att få fram ett bättre prediktionsresultat.

RMSE värde som man fick fram var 95 879. R-squared blev 0,692. Dessa resultat är sämre än de man fick i Multipel linjära regressionen.

4 Resultat och Diskussion

Efter att man använt sig av Multipel Linjär Regression, Best Subset Regression och Lasso Regression kan man se de slutgiltiga resultaten. (Figur 16)

RMSE för olika modeller	
Multipel linjär regression	92 302
Multipel linjär regression (efter justering)	84 268
Best Subset regression	160 178
Lasso regression	95 879
R^2 för olika modeller	
Multipel linjär regression	0,7364
Multipel linjär regression (efter justering)	0,6196
Best Subset regression	0,181
Lasso regression	0,692

Figur 16. Sammanställning av resultaten

Den bäst presterande var Multipel Linjär Regression som resulterade starkare RMSE och R^2 värden.

För att få ännu bättre resultat skulle man behöva utforska datan mer. Det skulle behövas ytterligare kontroller av heteroskedasticitet, residualerna, outliers.

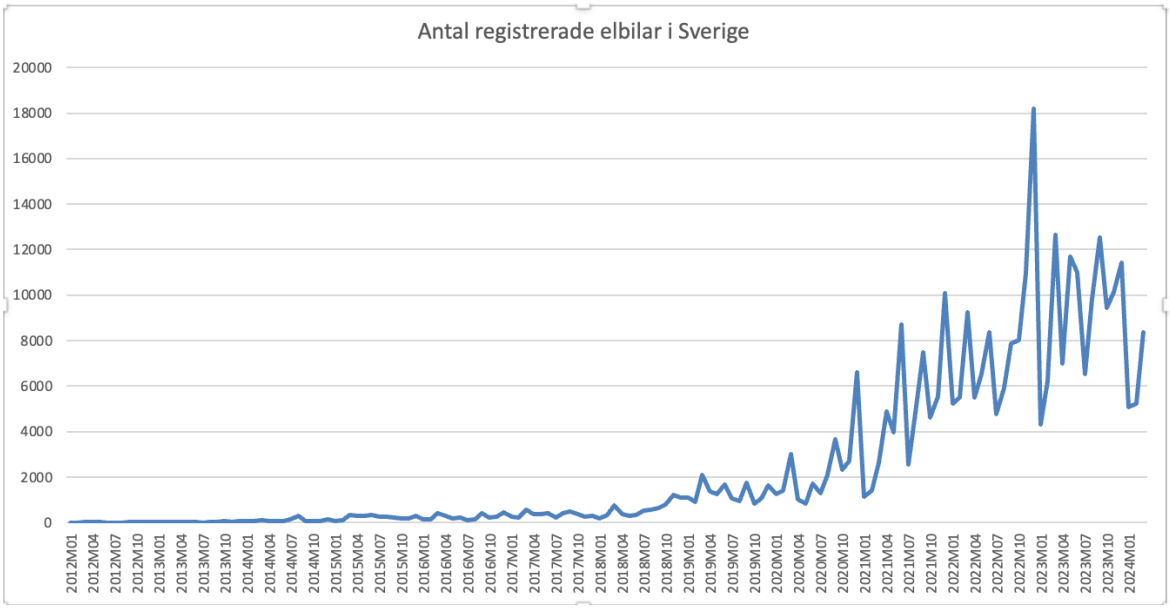
Det som inte nämndes i rapporten var att det genomfördes en logaritmisk transformation av pris pga misstanke om höger skevhet i prisfördelningen. Efter att man tränade modellen klart genomfördes exponentiering av den logaritmiska transformationen för att skapa prediktioner åt Y och inte för logY. Efter det kunde man gå vidare till uträckning av RMSE och R^2.
De resultaten som följde var klart mycket sämre än innan den logaritmiska transformationen.
Det uppfattades då som en felaktig tolkning om höger skevhet och istället utgick man att den fördelningen av priser såg ut på det sättet var pga outliers.
Informationen av den anledningen uteslöts från rapporten. Nu i efterhand kunde den finnas med.

En annan justering som inte nämndes var kvadratrotstransformation av logaritmerade priset pga den misstänkta heteroskedastisitet. Även här blev resultaten mycket sämre och misstanke om felaktig tolkning av data gjorde att den informationen inte togs med i rapporten.

De insikter som man fått genom de gjorda analyserna var att datamängd behövde vara större. Att välja top 10 mest frekventa bilmärken till antal var i efterhand inte ett nödvändigt steg.

Elbilmarknaden är kontant växande enligt SCB siffror och det är aktuellt att undersöka den typ av datan för att få bättre insikter om den typ av marknad[3]. (Figur 17)

[3]https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__TK__TK1001__TK1001A/PersBilarDrivMedel/table/tableViewLayout1/



Figur 17. Antal nyregistrerade elbilar i Sverige

När det gäller de frågorna som ställdes i början av rapporten – om hur de oberoende variablerna – plats, tillverkningsår, miltal och bilmärke påverkar priset så fann man dessa mönster.

Tillverkningsår är signifikant för prediktion. Nyare bilar verkar ha högre priser än bilar som är äldre. Det är ganska logiskt, men samtidigt med möjlighet till uppdatering av mjukvaran borde elbilar behålla sitt värde en längre period än de bilarna som inte kan uppdateras på samma sätt.

Miltal var mycket signifikant för prediktion av priserna. En högre körsträcka leder till lägre priser.

Märken verkar också ha signifikant påverkan på priset, men det finns stora skillnader i prissättning mellan märken.

Det som hade minst påverkan på priset var plats.

5 Slutsatser

Med detta arbete kom man fram till att de faktorer som påverkade priser mest var tillverkningsår, miltal, och märken. Området som bilarna såldes i hade inte någon signifikant påverkan på priset.

6 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.

Quantile-Quantile (QQ) plot är ett grafisk sätt att testa om datan följer en viss fördelning. I fallet som visas i videon kontrolleras det om stickprovsdata följer normalfördelningen. Man jämför den med den teoretiska sannolikhetsfördelningen som är normalfördelad genom att plotta bådas kvantiler mot varandra (x-axeln för den teoretiska normalfördelade datan och y-axeln för stickprovsdatan). Vid normalfördelningen kommer de plottade punkterna ligga ungefär som en rät linje.

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?

Jag skulle svara Karin att det stämmer, att både i Maskininlärning och i statistik kan man göra prediktioner genom att använda modeller på befintlig/historisk data för att sedan kunna prediktera. Användningsområden är mycket breda, men som ett exempel kan man analysera prisutveckling på maten utifrån t.ex. valutakurser, inflation och sedan prediktera priserna vid ändringar i valutakursen och inflationen.

Det som skiljer sig mellan dessa två är att i statistisk regressionsanalys kan man även förstå koppling mellan variablerna och hur de påverkar utfallet. Man kan från ett litet stickprov dra slutsatser om hela populationen. Så i exemplet om matpriserna kan statistisk regressionsanalys även se vad hur inflationen och valutakursändringar påverkar matpriserna, vilken av dessa variabler påverkar mer och samband mellan dem.

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

Om man ska utgå från en regressionsanalys så anger konfidensintervall osäkerheten kring uppskattningen av modellparametrar.

Prediktionsintervall är bredare än konfidensintervall, då det inte bara tar hänsyn till osäkerheten i uppskattningen av modellparametrarna utan också osäkerheten i de enskilda värdena som genereras av modellen.

4. Den multipla linjära regressionsmodellen kan skrivas som:

$Y = \beta 0 + \beta 1x1 + \beta 1x2 + \ldots + \beta pxp + \varepsilon$.

Hur tolkas beta parametrarna?

Varje β parameter representerar effekten av motsvarande oberoende variabel på den beroende variabeln Y.

$\beta 0$ är intercept – den förväntade medelvärdet av den beroende variabeln Y när alla oberoende variabler x är lika med noll.

Övriga β parametrarna mäter den genomsnittliga förändringen i den beroende variabeln Y när den oberoende variabeln xi förändras och de övriga oberoende variablerna hålls kontanta.

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

regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar

mått såsom BIC? Vad är logiken bakom detta?” Vad svarar du Hassan?

BIC används för att ta hänsyn till hur bra en modell passar datan och straffar modeller om den har många parametrar. På det sättet minskar risken för overfitting. Dock så kan det inte svara på hur bra modellen kommer att prestera på osedd data. Därför använder man sig av träning, validering och test set.

Jag skulle fråga Hassan vad som är målet för honom – att välja bland olika modeller eller välja en modell utifrån hur den presterar.

6. Förklara algoritmen nedan för ”Best subset selection”

”Best subset selection” är en teknik som hjälper att hitta den modellen som balanserar bäst mellan modellens komplexitet och passform utifrån de olika möjliga kombinationerna av oberoende variabler i en multipel regression.

M_0 är den grundläggande modellen som inte har några prediktorer. Den predikterar endast det genomsnittliga värdet för varje observation.

För varje antal tillgängliga k prediktorer skapas och passas alla möjliga kombinationer av modeller.

Baserat på det väljer man sedan den bästa modellen M_k som har lägsta RSS eller högsta R^2.

Sedan väljer man den bästa modellen bland alla $M_0, \ldots M_p$ genom att utgå från prediktionsfelen i validerings setet.

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

Förklara vad som menas med det citatet.

George Box menar att alla statistiska modeller är en förenkling av verkligheten. Modellerna har bl a begränsad data, den kan vara skevt fördelad, kvalitén på den kan variera som gör att modellerna inte är perfekta för att återspegla verkligheten.

Samtidigt, tack vare modellerna kan man göra prediktioner, se mönster/samband mellan olika företeelser/variabler, man kan få fram förklaringar till händelser osv. Det är en grund för att kunna gå vidare och skapa en bättre förståelse.

7 Självutvärdering

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

   Många saker tog längre tid för mig än jag förväntat mig.
   T ex fastnade jag när jag försökte förstå vad som orsakade sämre prediktioner, efter att modellerna blev justerade efter utvärderingar. Man förväntade sig att prediktionerna skulle bli bättre, men de blev sämre.

   Jag försökte leta efter svar i boken och på nätet.

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

   Med detta arbete strävade jag efter en VG.

3. Något du vill lyfta fram till Antonio?

   Rapporten tar mycket lång tid att skriva utifrån alla kriterier, struktur och referenser.

   Jag förstår poängen med varför vi behöver utgå från mallen – det är en träning inför examensarbetet. Dock så upplevde jag under arbetets gång att det behövs mycket längre tid än två veckor för att hinna med allt. Bara att koda med ett nytt programmeringsspråk, läsa på och förstå det man gör - det tar två veckor. Rapporten kräver några dagar till.
   Det är ingen kritik mot dig, utan bara en reflektion. Möjligtvis, för framtida klasser kan man överväga att ha ett enklare format på rapporten och sedan introducera ”den riktiga” rapportmallen till examensarbetet.

Källförteckning

https://www.scb.se/hitta-statistik/redaktionellt/tredubbling-av-elbilar-pa-tva-ar2/

James, Witten, Hastie, Tibshirani 2023