Prediktiv Analys av Diesel Bilpriser

En studie av Blocket.se-annonser

ECUTBILDNING

George Glor
EC Utbildning
Kunskapskontroll R Programmering
Augusti 2024

Sammanfattning

En multipel linjär regressionsmodell övervägdes för att analysera de faktorer som påverkar priset på dieselbilar och för att skapa ett verktyg för att uppskatta priser baserat på prediktorer som märke, modell, körsträcka och motorstyrka. Data samlades in från cirka 212 bilannonser på Blocket.se.

Vår slutliga modell verkar ha en hög grad av noggrannhet, vilket indikeras av ett högt R-squared värde, en måttligt låg residual standardavvikelse och en signifikant F-statistik. Ytterligare validering och testning av modellens prestanda är nödvändig för att säkerställa dess tillförlitlighet i att förutsäga dieselbil priser.

Förkortningar och Begrepp

- **Feature Variable (Funktionell Variabel)**: Även kallade prediktorer, attribut eller ingångs variabler som används för att förutsäga målvariabler.
- **Target Variable (Målvariabel)**: Även kallad respons eller utfall och representerar variabeln som modellen syftar till att förutsäga.
- **Intercept (Intercept)**: Intercept-koefficienten representerar det uppskattade värdet av utfallsvariabeln när prediktorvariabler är noll.
- **Slope (Lutning)**: Lutningskoefficienten representerar den uppskattade förändringen i utfallsvariabeln för varje enhets ökning i prediktorvariabler.
- **Quantiles (Kvantiler)**: Är värden som delar upp datan i lika stora proportioner.
- **Residualer (Residualer)**: Representerar skillnaderna mellan de observerade värdena av utfallsvariabeln och de värden som förutsägs av regressionsmodellen.
- **t-value (t-värde)**: Anger antalet standardfel som koefficient skattningen är ifrån noll.
- **p-value (p-värde)**: Sannolikheten för att observera koefficient skattningen om nollhypotesen (koefficienten är noll) är sann.
- **R-squared (R-kvadrat)**: Ett mått på andelen varians i utfallsvariabeln som förklaras av prediktorvariabler genom regressionsmodellen.
- **F-statistik (F-statistik)**: F-statistiken testar den övergripande signifikansen av regressionsmodellen.

Innehållsförteckning

---

1. Inledning

Syftet med denna studie är att utveckla en modell för att förstå vad som påverkar priserna på dieselbilar som listas på "Blocket.se". Genom att använda multipel linjär regression syftar vi till att förutsäga priset baserat på olika egenskaper eller funktioner som märke, modell och körsträcka.

För att uppnå detta mål kommer vi att följa en noggrann process som inkluderar:

- Analys av enkel linjär regressionsmodell för varje enskild prediktorvariabel.
- Utforskning av multipla linjära regressionsmodeller med olika uppsättningar av prediktorvariabler.
- Användning av korrelationsanalys och p-värden för att identifiera de mest signifikanta variablerna som bidrar till pris förutsägelsen.
- Eliminering av irrelevanta prediktorer för att förbättra den slutliga modellens prestanda.
- Finjustering av modellen med hjälp av R-squared och medelkvadratfel (MSE).

---

2. Teori

Multivariabel Regressionsmodell

En multivariabel regressionsmodell används för att analysera relationen mellan flera oberoende variabler (prediktorer) och en beroende variabel (respons). Målet är att förstå hur förändringar i prediktorerna påverkar responsvariabeln.

Quantile-Quantile (QQ) Plot

En Quantile-Quantile (QQ) plot är ett visuellt verktyg för att bedöma hur väl ett dataset överensstämmer med en specifik fördelning, ofta normalfördelningen. QQ-plottar används för att jämföra datasetets kvantiler med de från en teoretisk fördelning.

Maskininlärning vs. Statistisk Regressionsanalys

I maskininlärning fokuserar man på att göra förutsägelser utan att nödvändigtvis undersöka statistiska relationer mellan variablerna. Statistisk regressionsanalys syftar däremot både till att göra förutsägelser och att identifiera relationerna mellan variablerna, vilket hjälper till att förstå påverkan av olika faktorer.

Konfidensintervall vs. Prediktionsintervall

- **Konfidensintervall**: Uppskattar intervallet där den sanna populationsparametern, som regressionskoefficienten, sannolikt ligger med en viss nivå av konfidens.
- **Prediktionsintervall**: Uppskattar intervallet där individuella nya observationer sannolikt kommer att falla med en viss nivå av konfidens.

Parametrar och Termer i Multipel Linjär Regression

Den multipela linjära regressionsmodellen kan uttryckas som:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$$

Där:

- β_0 (beta noll) är interceptet som representerar det förväntade värdet av Y när alla oberoende variabler x_i är noll.
- β_i (beta i) är regressionskoefficienterna för varje oberoende variabel x_i. Dessa koefficienter anger hur mycket den beroende variabeln Y förändras för en enhetsökning i den oberoende variabeln x_i, med alla andra oberoende variabler hållna konstanta.
- ε (epsilon) representerar feltermen som fångar alla slumpmässiga och oförklarade variationer i Y som inte kan förklaras av modellen.

Statistisk Regressionsmodellering och Testdata

För att bedöma kvaliteten på vår regressionsmodell och uppskatta testfel använder vi metoder som Mallows Cp, AIC, BIC och justerad R^2. Dessa mått ger insikter i hur väl vår modell passar datan och hur noggrant den förutspår utfall.

Best Subset Selection

För att utföra bästa subset-valet i regressionsmodellen, följer vi dessa steg:

1. Anpassa en separat regressionsmodell för varje möjlig kombination av prediktorer.
2. Identifiera den modell som presterar bäst baserat på RSS eller R-squared.
3. Använd mått som valideringsfel, Cp, BIC eller juster ad R^2 för att välja den bästa modellen.

"All models are wrong, some are useful"

Box uttryckte att "Alla modeller är fel, vissa är användbara" för att betona att inga modeller kan fånga alla aspekter av verkligheten, men de kan fortfarande erbjuda värdefulla insikter.

---

3. Metod

3.1 Datainsamling

Under perioden mellan 11 april och 12 april 2024 samlade vi in data från Blocket.se. Ett team av 9 studenter bidrog till uppgiften genom att manuellt samla in data från Blockets annonser. Datasetet omfattade information om dieselbilar som var till salu på sajten.

3.2 Feature Selection

Vi använde gruppdiskussioner för att identifiera de mest relevanta egenskaperna för vår prediktiva modell. Genom samarbetsanalys fastställde vi vilka attribut som hade störst betydelse för att förutsäga priserna på dieselbilar listade på Blocket.se. De utvalda attributen inkluderar "län", "märke", "modell", "fordonstyp", "växellåda", "modellår",, "körsträcka", "drivning", "motorstyrka" och "färg", med målvariabeln "pris".

3.3 Feature Categorization

Våra variabler kan kategoriseras i kvalitativa och kvantitativa prediktorer:

Kvalitativa Prediktorer (Kategoriska Variabler)

1. Region
2. Märke
3. Modell
4. Fordonstyp
5. Växellåda
6. Drivning
7. Färg

Kvantitativa Prediktorer (Kontinuerliga Variabler)

1. Modellår
2. Körsträcka
3. Motorstyrka

Målvariabeln är också kvantitativ:

1. Pris

3.4 Data Preprocessing

Den insamlade datan lagrades i en Microsoft Excel-fil. De flesta saknade värden fanns i attributet "Registreringsdatum". Dessutom standardiserades färgattributet för att säkerställa konsekvens.

3.5 Exploratory Data Analysis

Microsoft 365 Excel användes som vårt huvudsakliga verktyg för utforskande dataanalys, vilket möjliggjorde en djupgående analys av datasetets egenskaper genom sammanfattande statistik och datavisualiseringstekniker.

3.6 Modellval

Med hänsyn till de identifierade attributen och syftet att förutsäga priser för dieselbilar listade på Blocket.se, ansågs en multipel linjär regressionsmodell för denna rapport.

3.7 Inkludering av Exklusiva Dieselbilar

Inkluderingen av exklusiva dieselbilar med priser som kan vara mycket högre än genomsnittet i vårt dataset kan leda till skapandet av outliers. Dessa potentiella outliers kan påverka noggrannheten i regressionsmodellen.

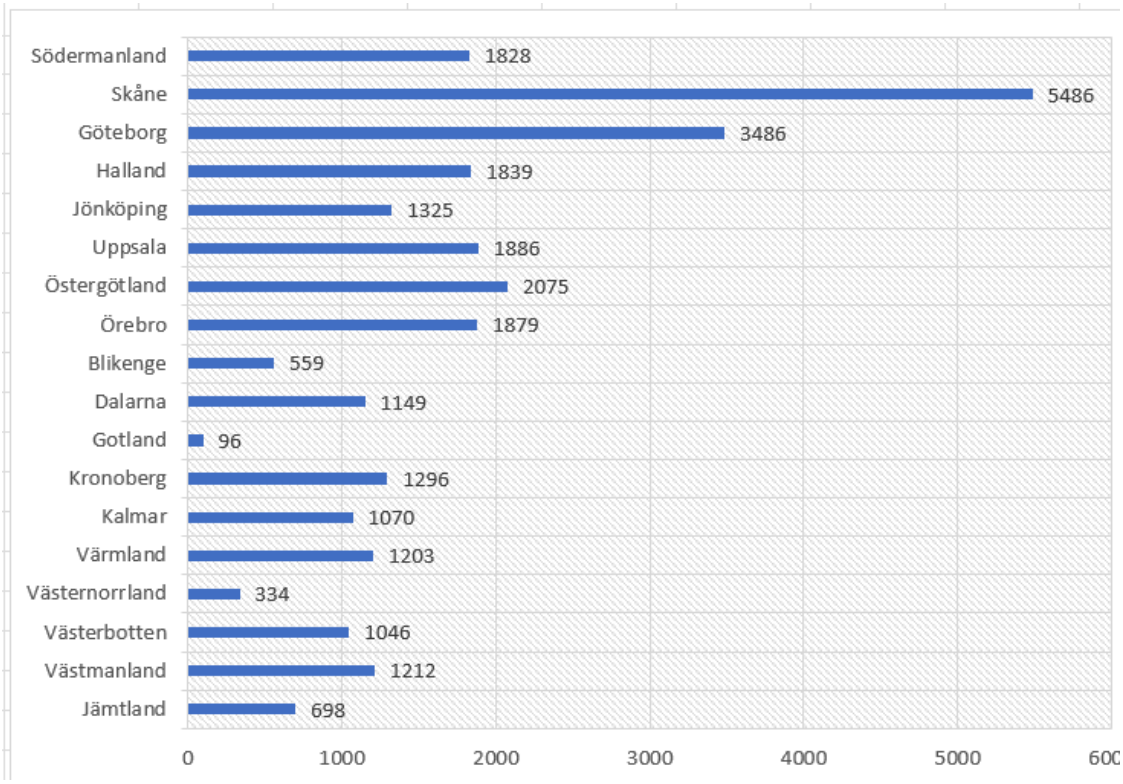
3.8 Proof of Concept (POC)

I samband med en regressionsmodell eller analys fungerar en Proof of Concept (POC) som ett första steg för att utvärdera genomförbarheten och framgången av modellen innan ytterligare resurser tilldelas dess fullständiga utveckling och implementering.

---

4. Resultat och Diskussion

Datainsamlingen genomfördes mellan 11 april och 12 april 2024 från Blocket.se, där det fanns cirka 16,915 annonser med betydande regionala variationer. Till exempel, i Gotland fanns det 41 annonser medan det i Stockholm fanns 4,103.



4.1 Proof of Concept (POC)

Här är en sammanfattning av det insamlade datasetet, som inkluderar både kategoriska och kontinuerliga variabler:

Kategoriska Variabler:

- **Region**: Stockholm, Göteborg, Skåne, etc.
- **Märke**: Volvo, Mercedes, BMW, etc.
- **Modell**: V60 Cross Country, Mégane Grandtour, etc.
- **Typ**: Kombi, Halvkombi, SUV, etc.
- **Drivning**: Fyrhjulsdriven, Tvåhjulsdriven, Volkswagen.

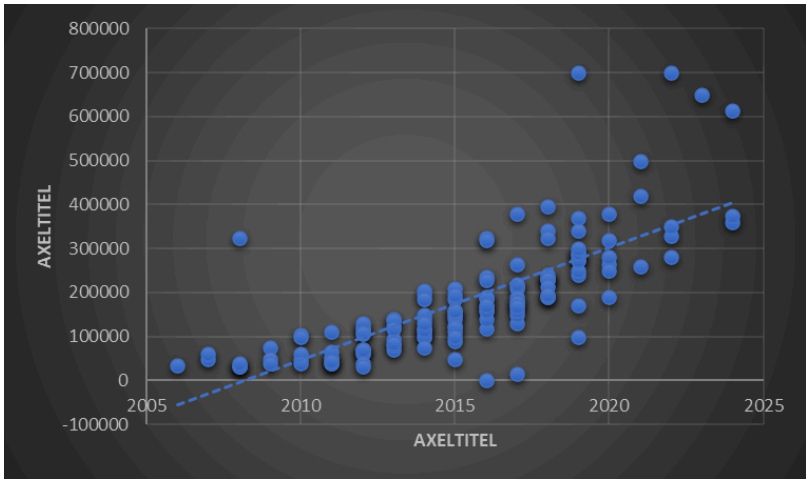
Kontinuerliga Variabler:

- **År**: Min = 2006, Max = 2024
- **Körsträcka**: Min = 0, Max = 289,930
- **Motorstyrka**: Min = 68, Max = 344
- **Pris**: Min = 147, Max = 699,800

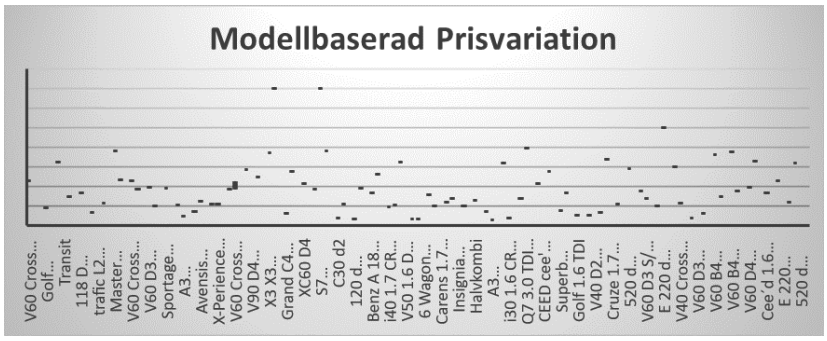
Scatterplots och boxplots användes för att visualisera relationer mellan pris och andra variabler. Till exempel visade scatterploten för pris mot år att priserna tenderar att minska med bilens ålder.

4.1.1 Enkel Linjär Regression

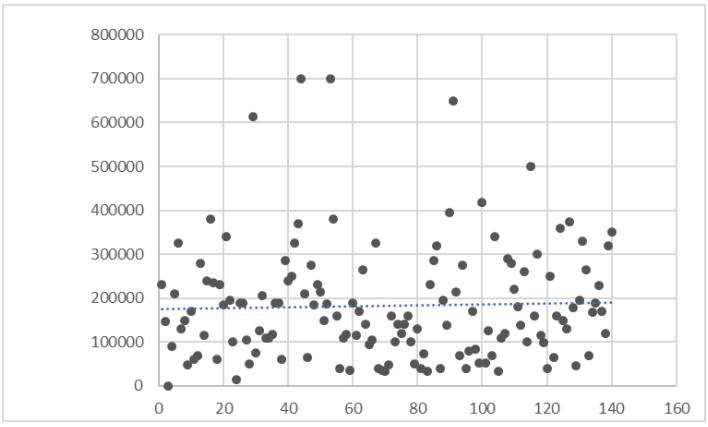
- **Relation mellan pris och modellår**:
    - För varje års skillnad i modellår uppskattas priset förändras med 5000 SEK.
    - R-squared värde = 0.30, p-värde < 0.001
- **Relation mellan körsträcka och pris**:
    - För varje ökning med 1000 km i körsträcka, uppskattas priset minska med 2000 SEK.
    - R-squared värde = 0.25, p-värde < 0.01



Prisutveckling för Bilar över Tid: Detta spridningsdiagram visar hur bil priserna har förändrats från år 2005 till 2025. Varje blå punkt representerar priset på en specifik bil modellerad mot dess modellår. Diagrammet inkluderar en trendlinje (den streckade linjen) som tydligt indikerar att bilpriserna generellt sett har ökat över tiden, vilket speglar inflation eller förändringar i bilmarknadens dynamik.

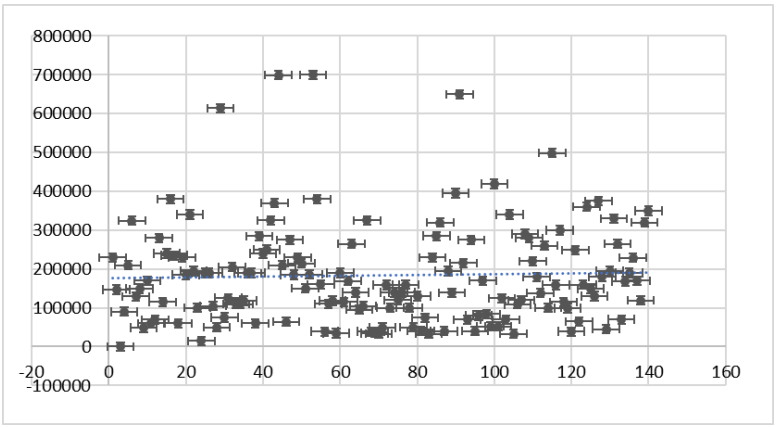


Modellbaserad Prisvariation: Detta diagram visar hur bilpriser varierar mellan olika modeller. Varje rad representerar ett specifikt bilmärke och modell, med punkter som illustrerar prisvariationer för den specifika modellen på marknaden. Diagrammet används för att visa spridningen av priser inom varje modell, vilket kan vara användbart för att identifiera modeller med stora prisvariationer eller att jämföra prisstabiliteten mellan olika modeller.

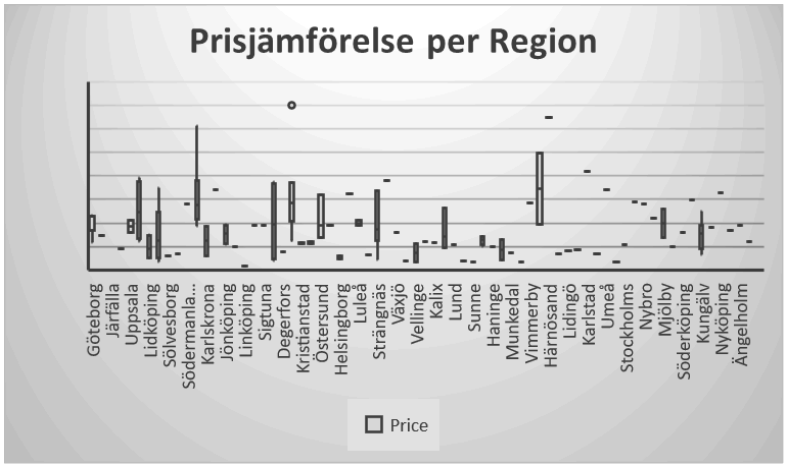


Hästkraft vs. Pris Analys

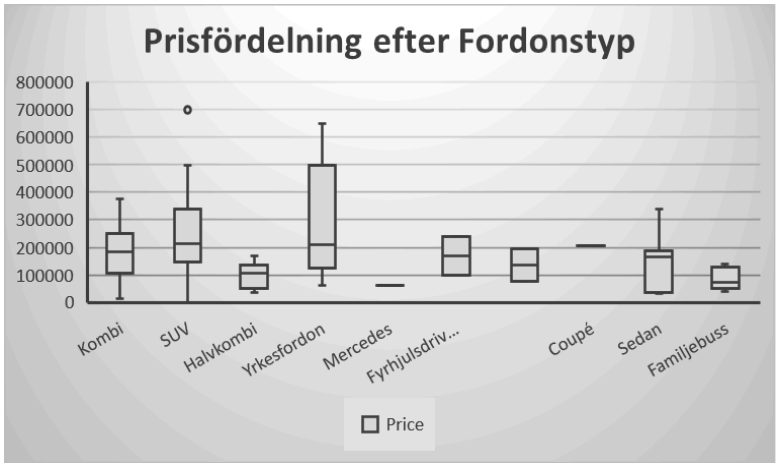
Diagrammet visar relationen mellan hästkraft och bilpriser. Varje punkt representerar en bil, placerad baserat på dess hästkraft (x-axeln) och pris (y-axeln). Den blå streckade linjen visar ett genomsnittligt prisnivå för jämförelse. Detta diagram kan användas för att analysera hur bilens prestanda, mätt i hästkraft, korrelerar med dess pris på marknaden, vilket hjälper till att identifiera prisvariationer beroende på prestandanivåer.



Prisintervall För Olika Hästkraftsnivåer: Detta diagram visualiserar prisintervallen för bilar baserat på deras hästkraft. För varje hästkraftsnivå, som visas på x-axeln, presenteras prisintervallen genom lådagram (boxplot) som visar medianpriset (linjen i mitten av varje låda), de nedre och övre kvartilerna samt extremvärden. Den blå streckade linjen markerar ett genomsnittligt pris över alla hästkraftskategorier. Diagrammet är användbart för att snabbt få en överblick över hur priserna varierar med ökande hästkraft, vilket kan ge insikter i prisstrategier och bilarnas värde på marknaden.



Prisjämförelse per Region: Detta diagram visar en jämförelse av prisfördelning för bilar baserat på region. Varje kolumn representerar en region i Sverige och visar pris variationerna genom lådagram. Lådornas positioner visar median priserna, de nedre och övre kvartilerna samt outliers (extremvärden) som indikeras med punkter. Genom att analysera detta diagram kan man se hur priserna på bilar varierar mellan olika regioner, vilket kan vara värdefullt för att förstå regionala marknadsförhållanden och konsumentbeteenden.

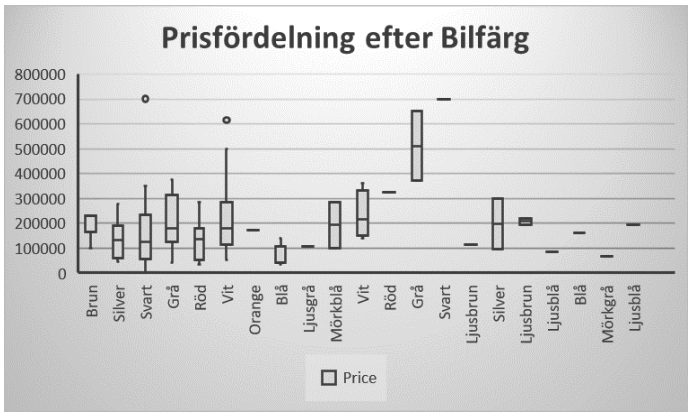


Prisfördelning efter Fordonstyp: Detta diagram visar prisfördelningen för olika typer av fordon. Lådagrammen illustrerar medianpriset, den första och tredje kvartilen samt eventuella extremvärden (outliers) för varje fordonstyp. Följande fordonstyper inkluderas:

- **Kombi**: Stor spridning i pris, vilket indikerar en bred variation i erbjudanden.

- **SUV**: Högre medianpris jämfört med andra fordonstyper, med några outliers på högre priser.
- **Halvkombi**: Lägre prisnivå med tätare prisfördelning.
- **Yrkesfordon**: Smalare prisintervall, vilket tyder på mindre variation i pris.
- **Mercedes**: Märkesspecifikt prisintervall med en bred spridning.
- **Fyrhjulsdrivna fordon**: Bred prisvariation, höga outliers, vilket visar på lyxigare modeller inom kategorin.
- **Coupé**: Låg medianpris jämfört med andra kategorier, färre datapunkter.
- **Sedan**: Låg och jämn prisfördelning.
- **Familjebuss**: Låg median och smalt prisintervall, vilket indikerar standardisering i pris.

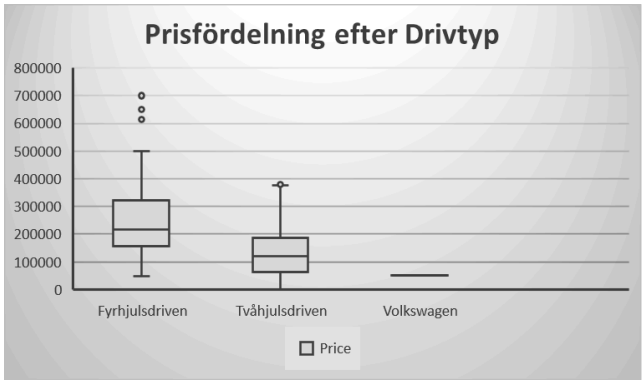
Detta diagram är användbart för att identifiera vilken fordonstyp som tenderar att ha högre eller lägre priser och kan hjälpa köpare att förstå pristrender inom specifika fordonstyper.



Prisfördelning efter Bilfärg: Detta diagram visar prisfördelningen för bilar baserat på deras färg. Lådagrammen illustrerar medianpriset, den första och tredje kvartilen, samt eventuella extremvärden (outliers) för varje bilfärg. Följande färger är representerade:

- Blå: En bred spridning av priser, med en relativt hög median.
- Silver: Priserna är relativt jämnt fördelade med en stabil median.
- Svart: En av de populäraste färgerna med hög median och prisvariation.
- Grå: Jämförbar med silver, men med några högre outliers.
- Vit: Prisfördelning i mittenområdet, med vissa outliers på högre priser.
- Orange: Färre datapunkter men har en del outliers på högre priser.
- Röd: Priserna varierar men generellt lägre än mer neutrala färger.
- Ljusblå: Färre datapunkter, priserna tenderar att vara lägre.
- Ljusgrå: Liknar ljusblå i prisfördelning och antal datapunkter.

Diagrammet visar att vissa färger, som svart och vit, tenderar att ha högre priser, vilket kan reflektera popularitet eller tillgänglighet. Denna typ av analys kan vara användbar för både säljare och köpare för att förstå hur bilfärg påverkar prissättning på marknaden.



Prisfördelning efter Drivtyp

Det här diagrammet visar prisfördelningen för bilar beroende på deras drivtyp. Vi ser tre kategorier:

- **Fyrhjulsdriven**: Visar en bred spridning av priser med ett relativt högt medianpris. Det finns några outliers som indikerar att vissa fyrhjulsdrivna bilar säljs till betydligt högre priser än genomsnittet.
- **Tvåhjulsdriven**: Har en lägre median än fyrhjulsdrivna bilar och en mindre spridning av priser. Det indikerar att tvåhjulsdrivna bilar generellt är mer prisstabil.
- **Volkswagen**: Denna kategori verkar ha en mycket begränsad prisdata och ingen tydlig spridning, vilket kan tyda på att datan är ofullständig eller att mycket få Volkswagen-bilar säljs inom de specificerade drivtyperna i datamängden.

4.1.1 Enkel Linjär Regression

Analysen av enkel linjär regression som utförts för att förstå sambanden mellan enskilda prediktorvariabler och prisförändringar på elbilar visade följande resultat:

Förhållande mellan pris och modellår

Resultaten visar att modellen uppskattar en genomsnittlig förändring i priset på en elbil med cirka 79,408 SEK för varje ettårs skillnad i modellåret. P-värdet (<2e-16) är nära noll, vilket indikerar att lutningskoefficienten är statistiskt signifikant. Det genomsnittliga avståndet mellan observerade priser och de av modellen förutsagda priserna, den residuala standardfelet, är 199,300 SEK.

Multipel R-kvadratvärdet (0.34) tyder på att ungefär 34% av variansen i pris förklaras av bilens modellår. P-värdet förknippat med F-statistiken är också mycket nära noll (< 2.2e-16), vilket indikerar att regressionsmodellen är statistiskt signifikant.

Förhållande mellan körsträcka och pris

Utdragen från den linjära regressionsmodellen med prediktorvariabeln Körsträcka och resultatvariabeln Pris är följande:

- Det uppskattade värdet av Priset när Körsträckan är noll är cirka 573,413 SEK.
- Den uppskattade förändringen i Priset för varje enhetsökning i Körsträcka är cirka -37.43 SEK.
- Multipel R-kvadrat indikerar att ungefär 21% av variansen i Pris förklaras av Körsträcka.
- P-värdet (2.56e-12) är mycket nära noll, vilket indikerar att lutningskoefficienten är statistiskt signifikant.

4.1.2 Multipel Linjär Regression

Resultaten från multipel linjär regressionsmodellering som syftar till att finjustera en modell som förutsäger bilpriser baserade på olika prediktorvariabler var som följer:

Multipel linjär regression med kontinuerliga prediktorer Resultaten från den multipel linjär regressionsmodellen med tre kontinuerliga prediktorer: År, Körsträcka och Hästkrafter är som följer:

- För varje enhetsökning i modellåret uppskattas priset öka med cirka 43,910 SEK (med alla andra variabler konstanta). Variabelns p-värde är statistiskt signifikant (1.05e-07).
- För varje enhetsökning i körsträckan uppskattas priset minska med cirka 15,870 SEK. P-värdet indikerar att det är statistiskt signifikant (0.000924).
- För varje enhetsökning i hästkrafter uppskattas priset öka med cirka 1,184 SEK. P-värdet är mycket lågt, vilket indikerar att det är högst signifikant (< 2e-16).

Multipel R-kvadrat: Den multipel R-kvadrat antyder att ungefär 70% av variansen i priset förklaras av modellen.

- F-statistik och p-värde: F-statistiken är 162.9, med ett mycket lågt p-värde (< 2.2e-16), vilket indikerar att den övergripande modellen är statistiskt signifikant.
-

---

5. Slutsatser

Vår slutliga modell ger värdefulla insikter i de faktorer som påverkar priserna på dieselbilar. Den höga justerade R-squared-värdet indikerar att modellen förklarar en betydande del av variansen i priserna.

---

6. Teoretiska frågor och svar

1. Vad är en Quantile-Quantile (QQ) plot?

En Quantile-Quantile plot, eller QQ-plot, är en grafisk metod för att jämföra två sannolikhetsfördelningar genom att plotta deras kvantiler mot varandra. Om datan följer den fördelningen som den jämförs med, kommer punkterna att ligga på en rät linje.

2. Skillnaden mellan maskininlärning och statistisk regressionsanalys:

Maskininlärning fokuserar främst på att göra prediktioner, ofta med komplexa modeller som kan hantera stora datamängder och många variabler utan att nödvändigtvis ge insikt om relationerna mellan dem. Statistisk regressionsanalys å andra sidan tillåter inte bara prediktioner utan också statistisk inferens, vilket innebär att man kan dra slutsatser om hur olika variabler påverkar varandra.

3. Konfidensintervall vs. Prediktionsintervall:

- **Konfidensintervall**: Uppskattar intervallet där en populationsparameter (till exempel medelvärdet) med en viss sannolikhet förväntas ligga baserat på urvalsdata.
- **Prediktionsintervall**: Uppskattar intervallet där framtida observationer förväntas falla med en viss sannolikhet, tar hänsyn till osäkerheten både i uppskattningen av parametern och i slumpmässiga avvikelser.

4. Tolkning av beta-parametrar i multipel linjär regression:

Beta-parametrarna (β) i en multipel linjär regressionsmodell visar hur mycket den beroende variabeln Y förändras för en enhetsändring i varje oberoende variabel Xi, med alla andra variabler hållna konstanta.

5. Användning av BIC och behovet av tränings-, validerings- och testset:

BIC (Bayesian Information Criterion) används för modellval och att mäta modellens kvalitet baserat på log-likelihood och antalet parametrar, och kan hjälpa till att undvika överanpassning. Även om BIC är ett kraftfullt verktyg, ersätter det inte behovet av att dela upp data i tränings-, validerings- och testset, speciellt när man behöver robusta uppskattningar av modellens prediktiva prestanda.

6. Algoritm för "Best subset selection":

- **Best subset selection** syftar till att identifiera den bästa kombinationen av prediktorer för en modell genom att:
    1. Välja alla möjliga kombinationer av prediktorer.
    2. Jämföra dessa modeller baserat på deras prestanda (t.ex. genom RSS, AIC, BIC, eller justerat R^2).
    3. Välja den modell som har bäst prestanda enligt valt mått.

7. Box citat "All models are wrong, some are useful":

Detta citat betonar att ingen statistisk modell perfekt fångar verkligheten eftersom alla modeller förenklar och generaliserar. Trots detta kan modeller vara mycket användbara för att göra förutsägelser och förstå sammanhang, så länge deras begränsningar är kända och beaktade.

-

---

7. Appendix A

Tabell A1: Regionnamn och Antal Relevanta Annonser

Region Namn	Annonser	Befolkning	Täthet av bilar per 100 personer
Stockholm	8408	2454821	0.34
Jämtland	698	132572	0.53
Västmanland	1212	280813	0.43
Västerbotten	1046	278729	0.38
Västernorrland	334	242148	0.14
Värmland	1203	283548	0.42
Kalmar	1070	246667	0.43
Kronoberg	1296	203686	0.64
Gotland	96	61029	0.16
Dalarna	1149	287253	0.40
Blikenge	559	157973	0.35
Örebro	1879	308116	0.61
Östergötland	2075	472298	0.44
Uppsala	1886	404589	0.47
Jönköping	1325	368856	0.36
Halland	1839	343746	0.53
Göteborg	3486	604616	0.58
Skåne	5486	1421781	0.39
Södermanland	1828	301944	0.61

Figurer och Diagram

---

8. Källförteckning

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Hämtad från statlearning.com
2. National Institute of Standards and Technology (NIST). (n.d.). *Quantile-Quantile (Q-Q) Plot*. Hämtad från NIST

3. Statology. (n.d.). *Confidence Interval vs. Prediction Interval*. Hämtad från Statology
4. Heiberger, R. M., & Holland, B. (2004). *Statistical Analysis and Data Display: An Intermediate Course with Examples in S-Plus, R, and SAS*. Springer Texts in Statistics.
5. Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). Wiley-Interscience.
6. Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer.
7. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.
8. Wooldridge, J. M. (2015). *Introductory Econometrics: A Modern Approach* (6th ed.). Cengage Learning.