R-programming

ECUTBILDNING

Wisssam Rateb

EC Utbildning

Examensarbete- Wissam Rateb

202403

Abstract

This report presents an analysis of a car sales dataset obtained from an Excel file. The primary objective is to investigate key aspects of the data, including average prices by car brand, the distribution of cars by fuel type, and pricing differences across countries. Two main research questions guide the analysis:

1. How does the average price vary between different car brands, and what might these differences indicate about their market positioning?
2. How are cars distributed among different fuel types, and what does this reveal about the market's shift toward more sustainable alternatives?

The analysis employs descriptive statistics and data visualization techniques to provide insights into these questions. The findings highlight significant variations in average prices across car brands, a diverse distribution of fuel types, and country-specific pricing trends. These insights offer valuable information for stakeholders in the automotive industry to tailor their strategies in response to observed market trends.

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

Innehållsförteckning

1 Inledning

I denna rapport analyseras dataset innehållande bilförsäljningsdata, vilken laddades från en Excel-fil och SCB. Syftet med analysen är att undersöka olika aspekter av data, såsom genomsnittliga priser per bilmärke, distributionen av olika bränsletyper och prissättningen per land. Genom att tillämpa statistiska metoder och visualiseringstekniker syftar rapporten till att ge en omfattande förståelse för data och dess underliggande mönster. För att uppnå detta syfte ställs följande frågor:

1. Hur varierar det genomsnittliga priset mellan olika bilmärken, och vad kan dessa skillnader indikera om deras marknadspositionering?

2. Hur fördelar sig antalet bilar mellan olika bränsletyper, och vad säger det om marknadens övergång till hållbarare alternativ?

Dessa frågor kommer att senare besvaras genom en rad analyser som beskrivs i rapporten.

2 Teori

Statistisk dataanalys är ett kraftfullt verktyg för att förstå och extrahera information från stora dataset. Genom att tillämpa olika statistiska metoder kan man identifiera trender, förhållanden och avvikelser som kan vara avgörande för beslutsfattande inom olika affärsområden.

I denna rapport används deskriptiv statistik för att sammanfatta data samt grupperingsmetoder för att jämföra genomsnittliga priser och antal observationer mellan olika kategorier.

3. Metod

3.1 Dataimport och inledande undersökning

Data importerades från en Excel-fil med hjälp av read_excel-funktionen i R. För att få en snabb översikt över datasetet inspekterades de första raderna med hjälp av head()-funktionen, vilket möjliggjorde en inledande förståelse för datasetets struktur.

3.2 Deskriptiv Statistik

För att sammanfatta de numeriska variablerna i datasetet användes funktionen summary(). Detta gav en omfattande beskrivning av variabler som medelvärde, median, minsta och största värden, samt kvartiler.

3.3 Genomsnittligt pris per bilmärke

Data grupperades baserat på bilmärket (Märke) och det genomsnittliga priset för varje märke beräknades. Denna analys utfördes med hjälp av group_by()- och summarise()-funktionerna i dplyr-paketet.

3.4 Antal bilar per bränsletyp

För att förstå fördelningen av olika bränsletyper grupperades data efter kolumnen Bränsle och antalet bilar inom varje kategori räknades.

3.5 Analys efter land

Om en kolumn för land (Land) fanns i datasetet, utfördes ytterligare en analys där antalet bilar och det genomsnittliga priset per land beräknades. Resultaten visualiserades med en stapeldiagram genererad av ggplot2-paketet.

4. Resultat

4.1 Deskriptiv Statistik

Den deskriptiva statistiken visade att det finns en stor variation i priserna mellan de olika bilarna. Medelvärdet och medianen av priset ger en inblick i den centrala tendensen av prissättningen, medan spridningen visar på förekomsten av både lågt och högt prissatta bilar.

4.2 Genomsnittligt pris per bilmärke

Analysen visade att genomsnittliga priser varierade avsevärt mellan olika bilmärken. Vissa märken har betydligt högre genomsnittliga priser, vilket kan indikera en premiumpositionering på marknaden, medan andra märken har lägre genomsnittspriser, vilket kan indikera en mer budgetvänlig profil.

4.3 Antal bilar per bränsletyp

Resultaten visade hur bilarna fördelades över olika bränsletyper. Detta är relevant för att förstå marknadens sammansättning, särskilt i en tid då det sker en övergång från fossila bränslen till mer hållbara alternativ.

4.4 Antal bilar per bränsletyp

För de dataset som innehöll en Land-kolumn visade resultaten en variation i både antalet bilar och de genomsnittliga priserna per land. Visualiseringen i form av ett stapeldiagram gav en tydlig bild av dessa skillnader.

5. Teoretiska frågor

QQ(Quantile-Quantile) - QQ-plotten är ett verktyg för att bedöma om datan kommer från en specifik fördelning eller inte. Om punkterna ligger nära den 45-graders linjen indikerar det att datan följer den teoretiska fördelningen. Om punkterna avviker från linjen kan det tyda på att datan inte följer den teoretiska fördelningen

Du har rätt Karin, det stämmer att maskininlärning och statistisk regressionsanalys har vissa likheter, men det finns också viktiga skillnader.

Både maskininlärning och regressionsanalys kan användas för att predicta framtida värden baserat på historiska data. I maskininlärning fokuserar man ofta på att maximera modellens prediktionsförmåga på okänd data, oavsett hur den interna modellen ser ut. Man strävar efter att hitta en modell som kan generalisera väl till nya data.

Regressionsanalys å andra sidan lägger ofta mer vikt vid modellers tolkbarhet och statistiska egenskaper. Man vill inte bara ha en bra prediktionsmodell, utan man vill också kunna förstå hur de olika variablerna i modeller påverkar det förutsagda värdet.

Statistiska slutsatser:

En av de viktigaste skillnaderna är att regressionsanalys kan användas för statistiska inferens, medan maskininlärning generellt sett inte kan det. Statistiska slutsatser innebär att man drar slutsatser om den underliggande befolkningen baserat på ett stickprov av data. Man kan till exempel använda regressionsanalys för att testa hypoteser om sambandet mellan variabler, eller för att beräkna konfidensintervall för modellers parametrar.

Maskininlärningsmodeller, å andra sidan, ger oss inte direkt information om befolkningen. De kan bara ge förutsägelser för individuella datainstanser.

Exempel:

Maskininlärning: En maskininlärningsmodell kan tränas på data om hundbilder för att klassificera nya bilder som antingen hund eller inte hund. Modellens prediktionsförmåga på nya bilder kan utvärderas, men modeller ger inte direkt information om hur troligt det är att en viss hundbild faktiskt visar en hund.

Regressionsanalys: En regressionsanalys kan användas för att studera sambandet mellan en persons vikt och deras längd. Man kan använda modeller för att förutsäga en persons längd baserat på deras vikt, och man kan också dra statistiska slutsatser om hur viktökning påverkar längd i genomsnitt.

Sammanfattningsvis, både maskininlärning och regressionsanalys är värdefulla verktyg för dataanalys, men de har olika styrkor och svagheter. Maskininlärning är ofta bättre att göra precisa förutsägelser, medan regressionsanalys är bättre på att ge statistiska insikter och dra slutsatser om populationer.

3. Konfidensintervall beskriver osäkerheten kring estimerade parametrar i en modell, medan prediktionsintervall beskriver osäkerheten för framtida individuella observationer.

4. visar hur mycket den beroende variabeln (y) förändras för varje enhets förändring i respektive oberoende variabel (x), medan alla andra x-variabler är konstanta.

β*0(Intercept): Förväntat värde på y när alla x-variabler är noll.*

β*1,*β*2,…,*βp*: Hur mycket y förändras för varje enhets förändring i respektive x, medan andra x-variabler är konstanta.*

5. Genom att använda BIC för att jämföra modeller kan behovet av att dela upp data i tränings-, validerings- och testset minska. BIC inkluderar ett straff för överflödiga parametrar, vilket hjälper till att förhindra överanpassning, även med enklare dataset.

Därför kan BIC ofta ge ett pålitligt modellval utan att man behöver använda traditionella metoder för validering som tränings- och testset.

6.

```
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LinearRegression

from sklearn.metrics import mean_squared_error


# Ladda datan

data = pd.read_csv('data.csv')
```

```
# Separera variabler och mål
X = data.drop('target', axis=1)
y = data['target']

# Dela upp datan i tränings- och testuppsättningar
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# tomma listor för att lagra modeller och resultat
    best_models = []
best_scores = []

# Loopa igenom alla delmängder av prediktorer
for k in range(1, X.shape[1] + 1):
    # Skapa alla kombinationer av k prediktorer
    combinations = list(combinations(X.columns, k))

    # Loopa igenom kombinationerna av prediktorer
    for combination in combinations:
        # Skapa en ny delmängdsdataset
        X_subset = X[list(combination)]

        # Skapa en linjär regressionsmodell
        model = LinearRegression()

        # Träna modellen på träningsdatan
        model.fit(X_subset, y_train)
```

```
        # Gör en predict på testdatan
        y_pred = model.predict(X_subset)

        # Beräkna MSE-poäng
        mse = mean_squared_error(y_test, y_pred)

            # Lagra den bästa modellen och poängen
        if not best_models or mse < best_scores[-1]:
            best_models.append(combination)
            best_scores.append(mse)

# Visa den bästa modellen och poängen
print("Bästa delmängd av prediktorer:", best_models[-1])
print("Bästa MSE-poäng:", best_scores[-1])
```

7. Jag tror att George Box menade att alla modeller är ofullkomliga och representerar inte verkligheten perfekt men att trots deras brister kan modeller ändå vara användbara för att ge insikt eller förståelse av komplexa system.

6. Slutsatser

Analysen visade att:

- Bilmärken har olika prissättningsstrategier, vilket återspeglas i deras genomsnittliga priser.
- Det finns en varierande fördelning av bilar baserat på bränsletyp, vilket är viktigt för att förstå trender inom bilindustrin.
- Länder uppvisar olika nivåer av genomsnittliga priser och antal bilar, vilket kan bero på lokala marknadsförhållanden och preferenser.

Dessa insikter kan vara användbara för företag inom bilindustrin för att anpassa sina strategier baserat på de observerade trenderna.

Efter att ha analyserat bilförsäljningsdata kan vi nu besvara de två ursprungliga frågeställningarna:

1. **Hur varierar det genomsnittliga priset mellan olika bilmärken, och vad kan dessa skillnader indikera om deras marknadspositionering?**

Analysen visade att det genomsnittliga priset varierade avsevärt mellan olika bilmärken. Vissa märken hade betydligt högre genomsnittliga priser, vilket kan indikera en premiumpositionering på marknaden, riktad mot kunder som är villiga att betala mer för lyx och kvalitet. Andra märken hade lägre genomsnittliga priser, vilket tyder på en budgetvänlig profil som kan attrahera prismedvetna konsumenter.

2. **Hur fördelar sig antalet bilar mellan olika bränsletyper, och vad säger det om marknadens övergång till hållbarare alternativ?**

Resultaten visade en varierad fördelning av bilar mellan olika bränsletyper. Detta indikerar att även om fossila bränslen fortfarande är dominerande, finns det en ökande närvaro av alternativ som el- och hybridbilar. Denna utveckling tyder på en gradvis övergång till mer hållbara bränslekällor inom bilindustrin, vilket kan vara en respons på både regleringar och förändrade konsumentpreferenser.

7. Självutvärdering

Analysen gav mig värdefull erfarenhet av datahantering och användning av statistiska metoder i R. Jag lyckades strukturera och presentera resultaten tydligt, särskilt genom att effektivt utnyttja verktyg som `dplyr` och `ggplot2`.

Jag insåg dock att en djupare statistisk analys, som regressionsmodeller, hade kunnat ge ytterligare insikter. Dessutom kunde jag ha lagt mer fokus på att behandla outliers och saknade data för att förbättra resultatens noggrannhet.

Sammanfattningsvis stärkte detta arbete mina färdigheter i dataanalys, och jag planerar att inkludera mer avancerade metoder och förbättrad datakvalitet i framtida projekt.

8. Appendix A

Koden som användes för att analysera datan:

```
# Ladda nödvändiga bibliotek

library(readxl)

library(dplyr)

library(ggplot2)


# Ange filväg

file_path <- "C:/Users/LENOVO/Downloads/TK1001AC_20240825-165603.xlsx"


# Läs in data i en DataFrame

df <- read_excel(file_path)


# Visa de första raderna av data

head(df)


# Beskrivande statistik

summary(df)


# Beräkna genomsnittligt pris per märke

average_price_by_brand <- df %>%

  group_by(Märke) %>%

  summarise(Average_Price = mean(Pris, na.rm = TRUE))


print(average_price_by_brand)
```

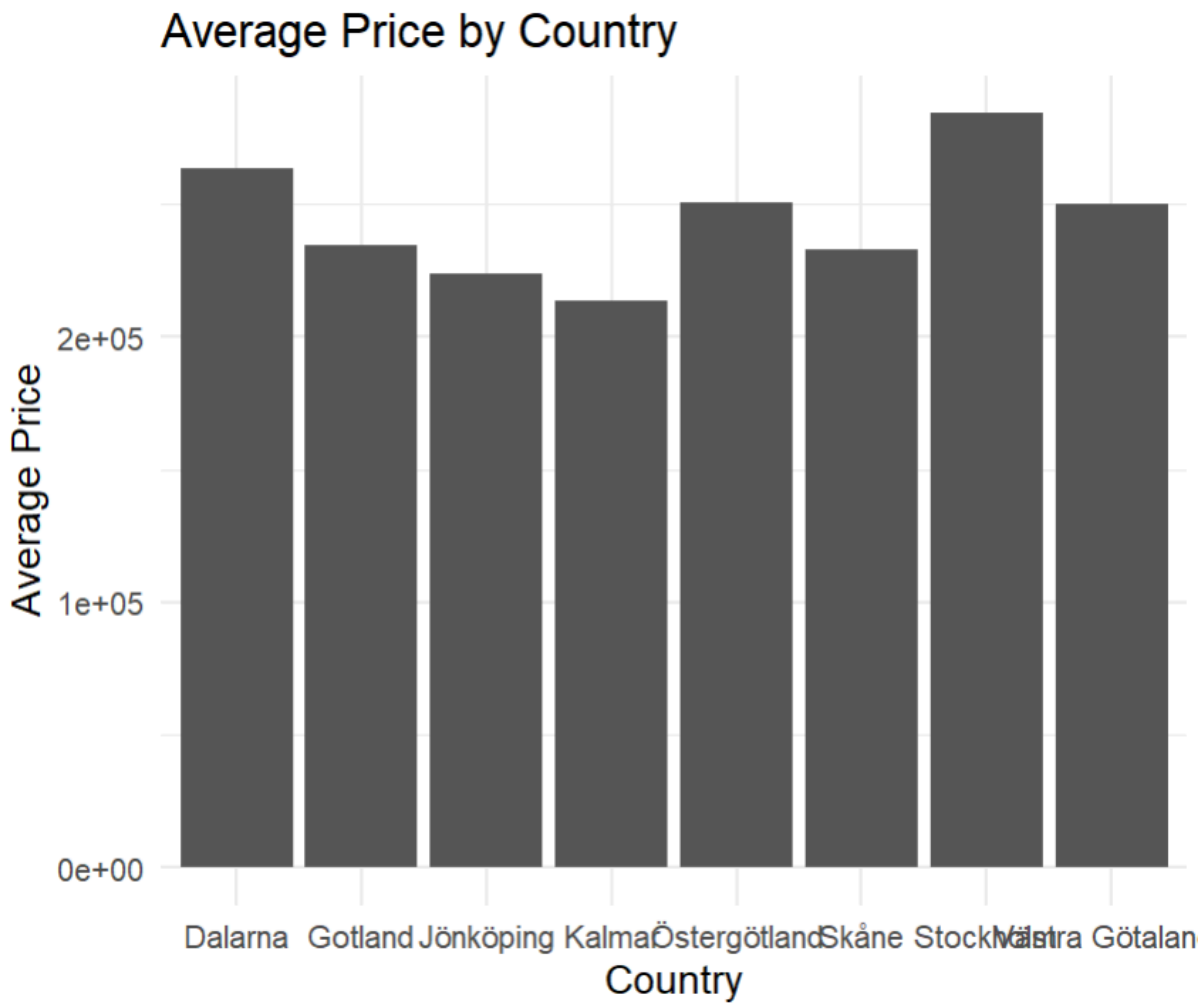
```
# Räkna antalet bilar per bränsletyp
fuel_type_count <- df %>%
  group_by(Bränsle) %>%
  summarise(Count = n())

print(fuel_type_count)

# Gruppanalys efter land och beräkna total antal bilar samt genomsnittligt pris
if ("Land" %in% colnames(df)) {
  analysis_by_country <- df %>%
          group_by(Land) %>%
          summarise(
          Total_Cars = n(),
          Average_Price = mean(Pris, na.rm = TRUE)
          )

  print(analysis_by_country)

  # Visualisering: Stapeldiagram av genomsnittligt pris per land
  ggplot(analysis_by_country, aes(x = Land, y = Average_Price)) +
          geom_bar(stat = "identity") +
          theme_minimal() +
          labs(title = "Average Price by Country", x = "Country", y = "Average Price")
} else {
  print("No 'Land' column found in the dataset.")
}
```


Average Price by Country
Average Price
2e+05
1e+05
0e+00
Dalarna
Gotland
Jönköping
Kalmar
Östergötland
Skåne
Stockholm
Västra Götaland
Country

Källförteckning

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Wickham, H., & Bryan, J. (2023). *Readxl: Read Excel Files*. R package version 1.4.1. URL https://CRAN.R-project.org/package=readxl.