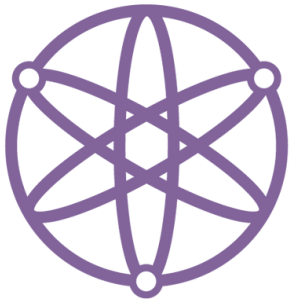


WYDZIAŁ
MATEMATYKI
I FIZYKI STOSOWANEJ
POLITECHNIKI RZESZOWSKIEJ

Statystyczna analiza danych

Badania nad rynkiem samochodowym

Aldona Świrad

Opiekun: **dr Mariusz Startek**

Rzeszów, 2023-05-26

Spis treści

Wstęp

Źródło

Dane pochodzą z ogólnodostępnej platformy Kaggle.
https://www.kaggle.com/datasets/gagandeep16/car-sales

Opis danych

Wybrane dane dotyczą samochodów. Możemy tam znaleźć informacje na temat ich poszczególnyhc cech, takich jak:

- Manufacturer - Marka
- Model - model
- Sales_in_thousands - sprzedaż w tysiącach
- __year_resale_value - roczna wartość odsprzedaży
- Vehicle_type - rodzaj
- Price_in_thousands - cena w tysiącach
- Engine_size - pojemność silnika
- Horsepower - liczba koni mechanicznych
- Wheelbase - rozstaw osi
- Width - szerokokść
- Length - długość
- Curb_weight - masa wlasna pojazdu
- Fuel_capacity - pojemnosc baku
- Fuel_efficiency - efektywnosc spalania

Chciałam wybrać temat, który mnie zainteresuje i pozwoli na zdobycie praktycznych umiejętności w analizie danych. Ze względu na moje osobiste zainteresowania oraz wymagania akademickie, analiza danych na temat samochodów wydała mi się interesującym wyborem. Jestem ciekawa zależności między różnymi parametrami samochodów, takimi jak moc silnika, czy cena oraz jakie wnioski można wyciągnąć na podstawie analizy statystycznej takich danych. Ponadto, widzę praktyczne zastosowanie takiej analizy w przemyśle motoryzacyjnym, co również mnie zainteresowało i skłoniło do wyboru tego tematu do mojego projektu.

Wczytanie danych i ich obróbka

```
# użyte biblioteki
library(e1071)
library(tidyverse)
library(hrbrthemes)
library(viridis)
## wczytywanie danych
setwd('D:/IiAD sem_4/Statystyka/Projekt_SAD')

cars <- read.csv('Car_sales.csv', header = T, sep = ',')

cars <- na.omit(cars)
colnames(cars)[3] <- "Sales"
colnames(cars)[6] <- "Price"
cars$Sales <- cars$Sales*1000
cars$Price <- cars$Price*1000
cars <- cars[, -c(15,16)]
```

Opis wybranych bibliotek

1. library(e1071): Biblioteka “e1071” jest jednym z najważniejszych pakietów w języku R dla analizy danych i uczenia maszynowego. Zapewnia różnorodne narzędzia i funkcje do klasyfikacji, regresji, analizy skupień i wiele innych technik statystycznych. W szczególności, biblioteka e1071 dostarcza implementacje popularnych algorytmów uczenia maszynowego, takich jak maszyny wektorów nośnych (SVM), metody klasyfikacji Bayesa i wiele innych. Jest to niezwykle przydatne narzędzie dla badaczy i analityków danych, którzy chcą wykorzystać zaawansowane techniki analizy danych w języku R.

2. library(tidyverse): Biblioteka “tidyverse” to zestaw kilku powiązanych pakietów R, które są wykorzystywane do manipulacji, wizualizacji i analizy danych. W skład tidyverse wchodzą popularne pakiety, takie jak ggplot2, dplyr, tidyr i wiele innych. Dzięki tidyverse możesz łatwo przeprowadzać zaawansowane operacje na danych, takie jak filtrowanie, sortowanie, grupowanie, łączenie danych z różnych źródeł itp. Biblioteka ta jest ceniona za spójność i czytelność kodu, co ułatwia pracę z danymi w R.

3. library(hrbrthemes): Biblioteka “hrbrthemes” to pakiet R zawierający zestaw tematów graficznych (themes) do wykorzystania w pakiecie ggplot2. Dostarcza wiele estetycznych i profesjonalnie wyglądających tematów, które można stosować do tworzenia wykresów o wysokiej jakości. Hrbrthemes oferuje różnorodne opcje kolorystyczne, układy i style czcionek, które mogą być dostosowane do konkretnych potrzeb wizualizacji danych. Jest to przydatne narzędzie dla osób, które chcą poprawić estetykę swoich wykresów i prezentacji danych.

4. library(viridis): Biblioteka “viridis” to pakiet R dostarczający zestaw wysokiej jakości kolorowych map gradientowych, które są szczególnie przydatne przy tworzeniu wizualizacji danych. Kolorowe mapy gradientowe viridis są zaprojektowane tak, aby były czytelne również dla osób z deficytami wzroku, dzięki

czemu są popularne w dziedzinie wizualizacji naukowych i statystycznych. Biblioteka viridis oferuje różne palety kolorów, które można łatwo stosować w pakiecie ggplot2 lub innych narzędziach do wizualizacji danych w R.

Charakterystyki liczbowe

Średnia cena wzgledem marki

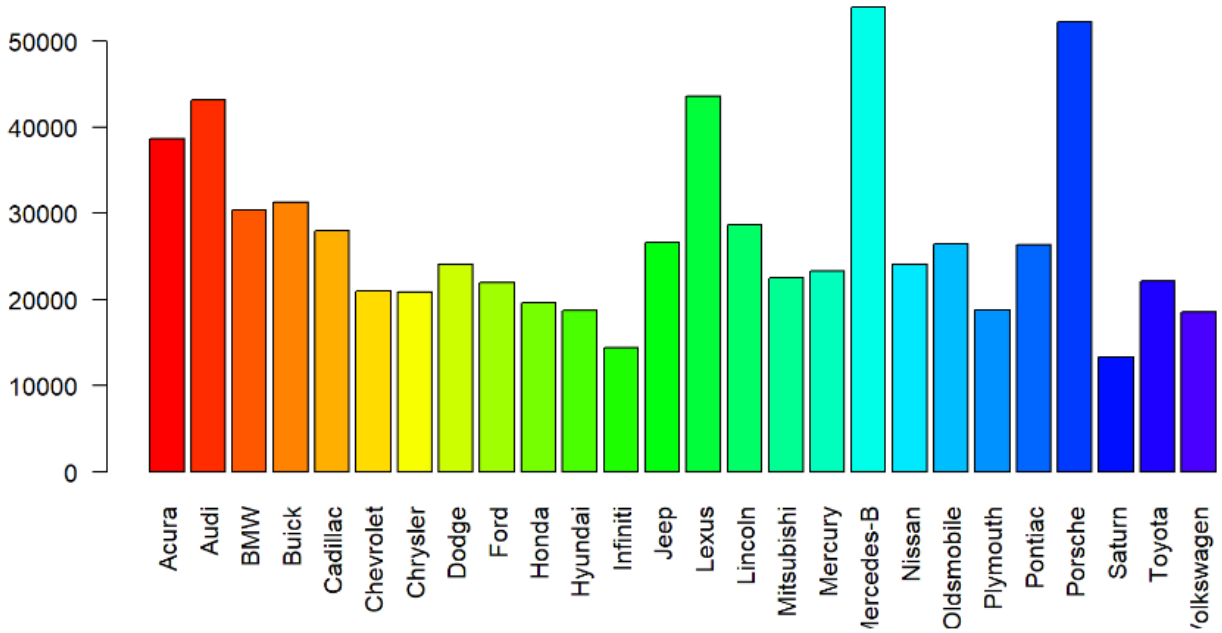
```
# obliczanie sredniej
#srednia wzgledem marki (grupowanie)
vprices <- c()
vcounter <- c()
price <- cars$Price[1]
counter <- 0



for (i in 2:length(cars$Price))
  if (cars$Manufacturer[i-1]==cars$Manufacturer[i]){
    price <- price + cars$Price[i]
    counter <- counter + 1
    if(i==length(cars$Price)){
      price <- price + cars$Price[i]
      counter <- counter + 1
      vprices <- append(vprices, price)
      vcounter <- append(vcounter, counter)
    }
  }else{
    price <- price + cars$Price[i]
    counter <- counter + 1
    vprices <- append(vprices, price)
    vcounter <- append(vcounter, counter)
    counter <- 0
    price <- 0
  }

#srednie
avg_by_manu <- vprices/vcounter
```

Wykres słupkowy średnich cen



Odchylenie standardowe średnich cen względem marek

```
# odchylenie
stand_dev <- sd(avg_by_manu)
```

```
## [1] 10658.29
```

Dominanta dla rozmiarów silników

```
# dominanta
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
getmode(cars$Engine_size)
```

```
## [1] 2
```

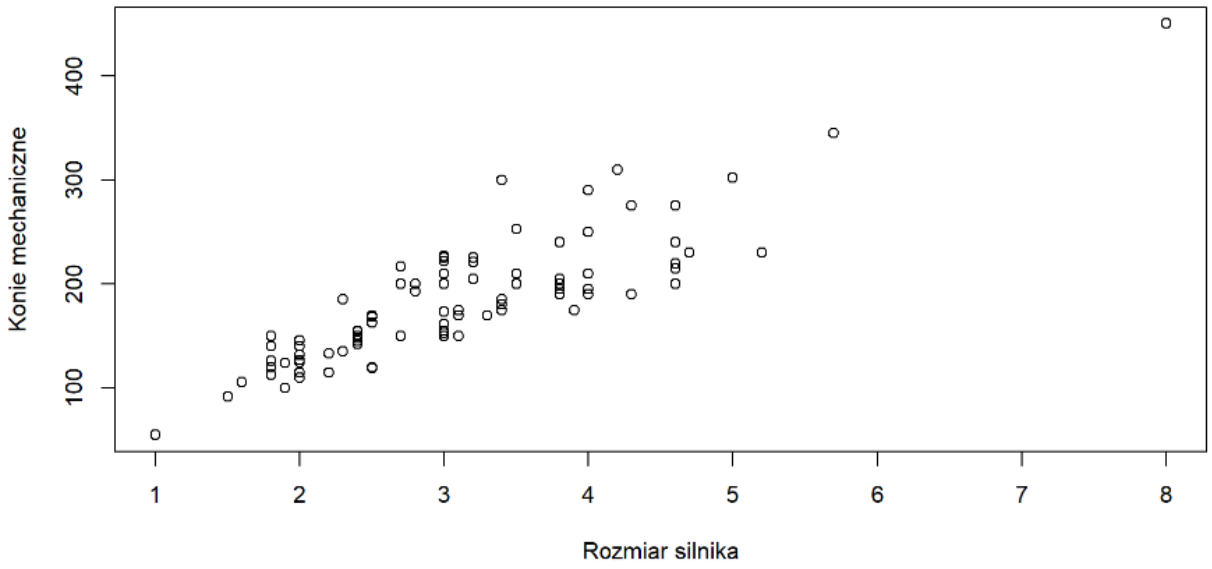
Kwantyle wśród rozmiarów silników

```
# kwantyle
kwantyle <- quantile(cars$Engine_size)
```

```
##   0%  25%  50%  75% 100%
##  1.0  2.2  3.0  3.8  8.0
```

Korelacja między wielkością silnika, a jego mocą

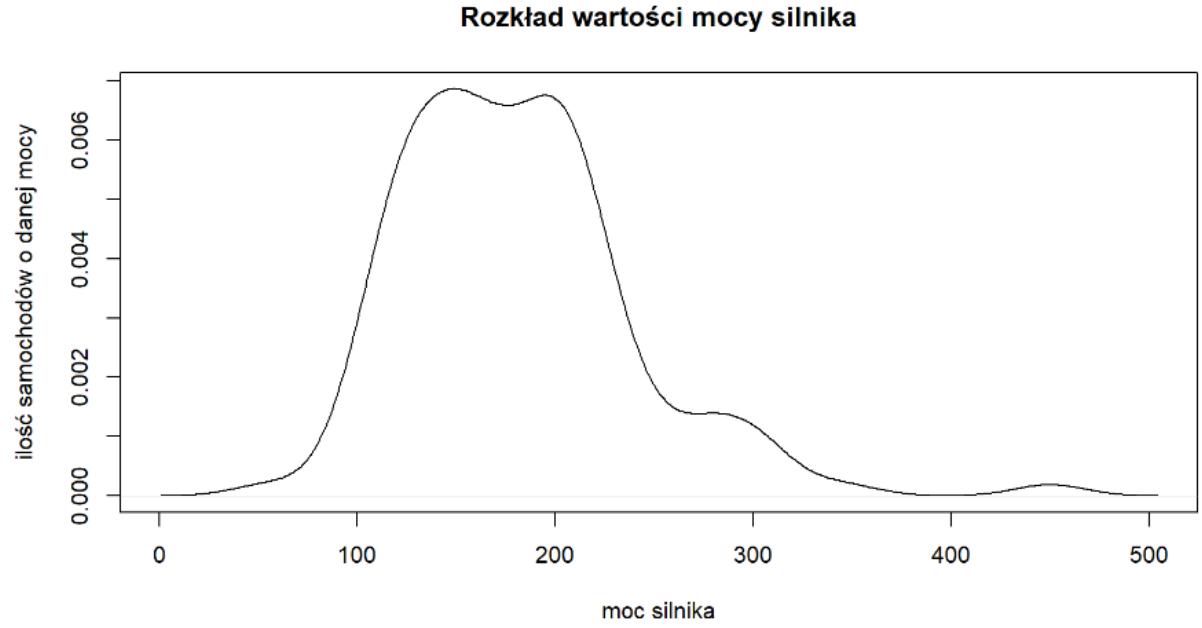
```
#korelacja
plot(cars$Engine_size, cars$Horsepower)
```




```
## [1] 0.8616183
```

Rozkład wartości mocy silnika

```
#rozklad
rozklad_hp <- plot(density(cars$Horsepower))
```



Współczynnik zmienności

```
#wspolczynnik zmiennosci
cv <- sd(cars$Horsepower)/mean(cars$Horsepower)
```

```
## [1] 0.3232079
```

Współczynnik asymetrii

```
as <- 3*(mean(cars$Horsepower)-median(cars$Horsepower))/sd(cars$Horsepower)

## [1] 0.3216518
```

Współczynnik splaszczenia

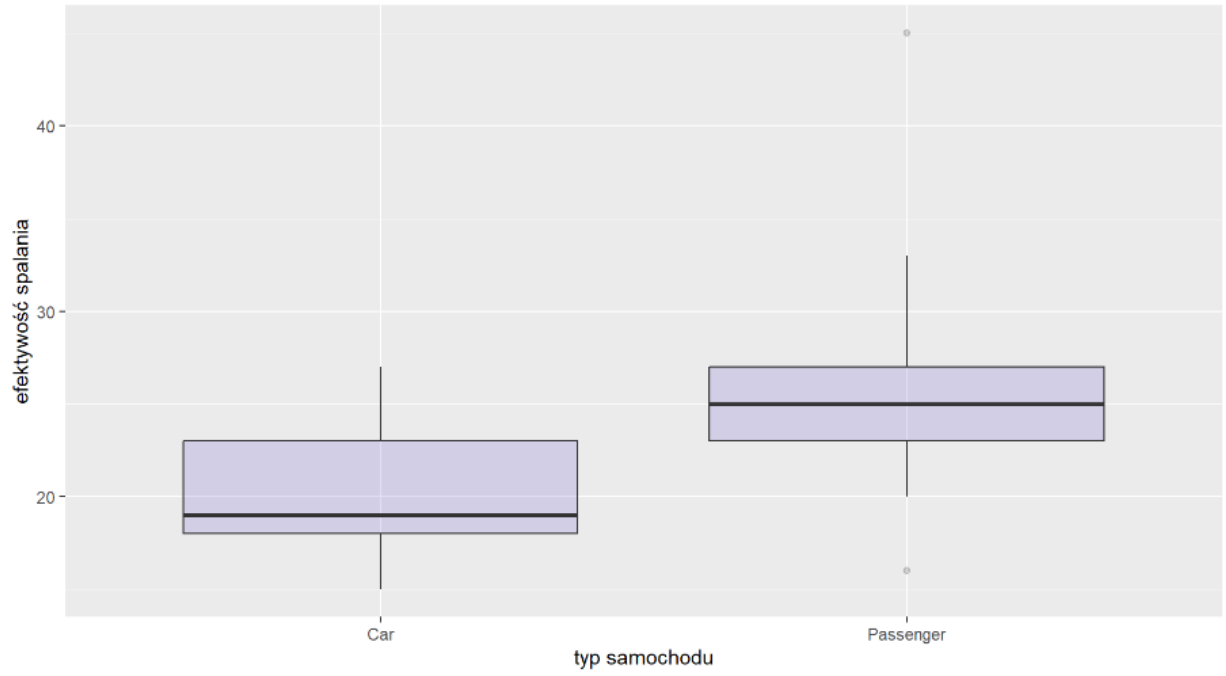
```
kurtoza <- moment(cars$Horsepower, order=4, center=TRUE)/(sd(cars$Horsepower)
^3)

## [1] 345.8519
```

Wykres pudełkowy efektywności spalania

```
#wykres pudełkowy

ggplot(cars, aes(x=as.factor(cars$Vehicle_type), y=cars$Fuel_efficiency)) +
    geom_boxplot(fill="slateblue", alpha=0.2) +
    xlab("Vehicle_type")
```



Passenger – samochód osobowy

Car – samochód inny niż osobowy

Podsumowanie wyników charakterystyk

Srednia cena samochodu	59112.32
Odchylenie standardowe średnich cen względem marek	10658.29
Dominanta dla rozmiarów silników	2
Kwantyle wśród rozmiarów silników	0% 25% 50% 75% 100% 1.0 2.2 3.0 3.8 8.0
Korelacja między wielkością silnika, a jego mocą	0.8616183
Współczynnik zmienności	0.3232079
Współczynnik asymetrii	0.3216518
Współczynnik splaszczenia	345.8519

Hipotezy statystyczne

Średnia cena samochodu to 60 000

- H0-Średnia cena samochodu to 60 000
- H1-Średnia cena samochodu nie jest równa 60 000
- poziom istotności α - 0,1

```
t.test(cars$Sales, mu=60000, alternalive="less", conf.level = 0.1)
```

```
##
##  One Sample t-test
##
## data:  cars$Sales
## t = -0.12792, df = 116, p-value = 0.8984
## alternative hypothesis: true mean is not equal to 60000
## 10 percent confidence interval:
##  58238.42 59986.22
## sample estimates:
## mean of x
##  59112.32
```

```
#Średnia cena samochodu
mean(cars$Sales)
```

```
## [1] 59112.32
```

Wartość funkcji testującej należy do obszaru krytycznego, więc H0 odrzucamy. Średnia samochodu dla populacji nie wynosi 60000.

Cena samochodów ma rozkład normalny

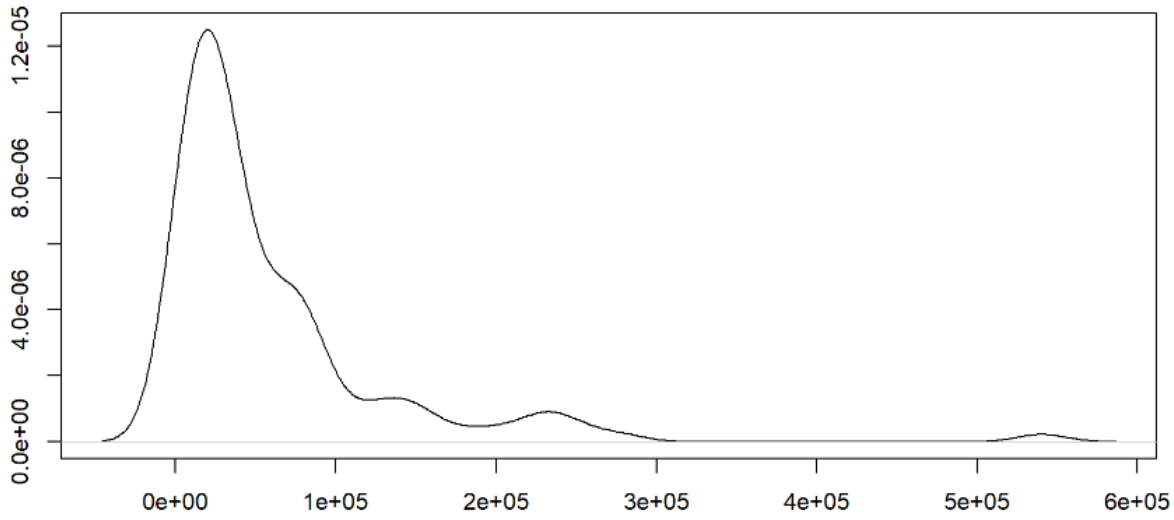
- H0-Cena samochodów ma rozkład normalny
- H1-Cena samochodów nie ma rozkładu normalnego
- poziom istotności α - 0,1

```
shapiro.test(cars$Sales)

##
##  Shapiro-Wilk normality test
##
## data:  cars$Sales
## W = 0.67779, p-value = 1.123e-14
```

Wartość p < 0.05, więc H0 odrzucamy. Cena samochódów nie ma rozkładu noramalnego.

```
plot(density(cars$Sales))
```



Rozkład pojemności silnika i mocy samochodu jest podobny

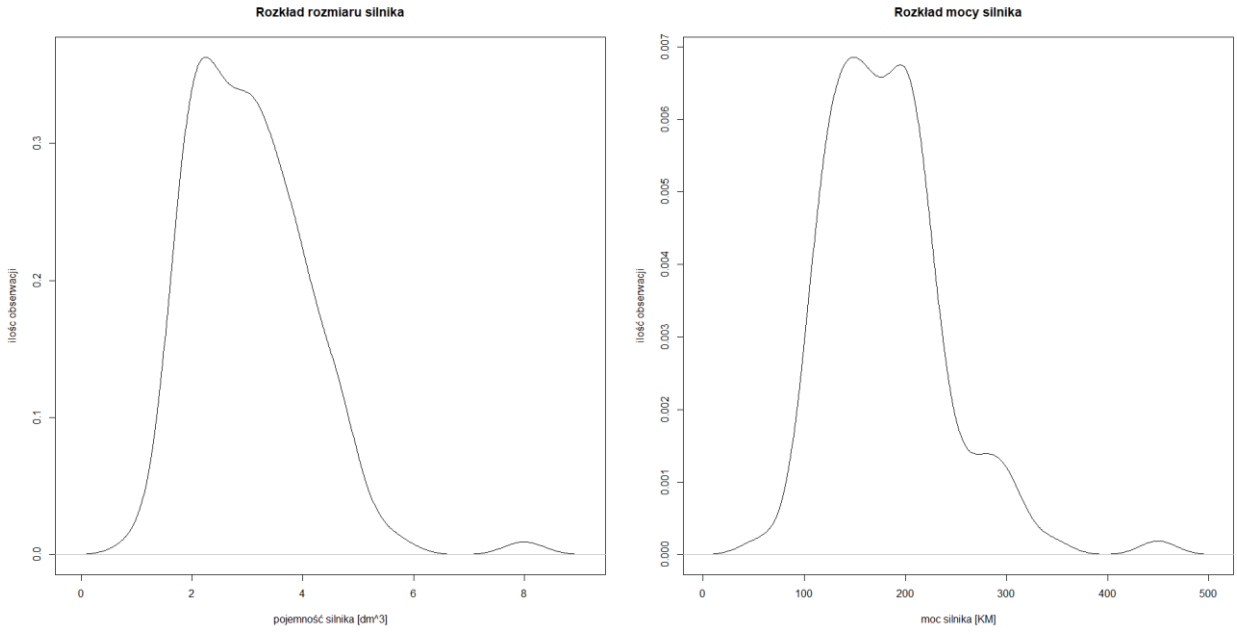
- H0- Rozkład pojemności silnika i mocy samochodu jest podobny
- H1- Rozkład pojemności silnika i mocy samochodu nie jest podobny
- poziom istotności α - 0,1

```
ks.test(cars$Engine_size,cars$Horsepower)

## Warning in ks.test.default(cars$Engine_size, cars$Horsepower): wartość
## prawdopodobieństwa w obecności powtórzonych wartości będzie przybliżona

##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  cars$Engine_size and cars$Horsepower
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Wartość p < 0.05, więc H0 nie odrzucamy. Rozkład pojemności silnika i mocy samochodu jest podobny.



Listing komend

Biblioteki:

stats

- na.omit() – usuwa wiersze z wartością NA
- sd() – oblicza odchylenie standardowe
- quantile() – oblicza kwantyle (0%, 25%, 50%, 75%)
- density() – oblicza gęstości w wskazanych punktach
- mean() – oblicza średnią
- median() – oblicza mediane
- t.test() - służy do przeprowadzania testów t-Studenta w celu porównywania średnich dwóch grup.
- shapiro.test() - służy do przeprowadzania testu Shapiro-Wilka w celu sprawdzenia, czy dane pochodzą z populacji o rozkładzie normalnym.
- ks.test() - służy do przeprowadzania testu Kołmogorowa-Smirnowa w celu porównania dwóch rozkładów danych.

e1071

- moment() – oblicza kurtoze

base

- unique() – znajduje wartości nie powtarzające się
- plot() – generuje prosty wykres z podstawowymi ustawieniami

tidyverse

- geom_box() – generuje wykres skrzynkowy - forma graficznej prezentacji rozkładu cechy statystycznej

ggplot2

- ggplot() – generuje bardziej zaawansowane wykresy

viridis

- rainbow() – nadaje wykresowi kolorów z pełnej gamy kolorów, tak, że każdy słupek może otrzymać inny kolor

`hrbrthemes`

- labs() - służy do dostosowywania etykiet osi w wykresie. Jest to skrót od "labels" (etykiety) i często używana w połączeniu z funkcjami tworzącymi wykresy, takimi jak plot() czy ggplot2.

Wnioski

Z przeprowadzonej szczegółowej analizy wynika, że badanie rynku samochodowego za pomocą narzędzi statystycznych dostarcza cennych informacji i wniosków. Projekt pozwolił mi zgłębić się w tematykę statystyki i wykorzystać ją w praktyce, a także poszerzyć umiejętności korzystania z języka R i środowiska RStudio.

Przede wszystkim projekt pozwolił mi lepiej zrozumieć trendy na rynku samochodowym, oraz zależności między danymi cechami samochodów. Dzięki analizie danych udało się także zidentyfikować czynniki wpływające na ceny samochodów.

Narzędzia statystyczne wykorzystane w projekcie umożliwiły mi odkrycie nie tylko oczywistych zależności, ale także ukrytych faktów i tendencji na rynku samochodowym. Dzięki temu mogłam zgłębić temat i uzyskać bardziej kompleksowe spojrzenie na badany obszar.

Dodatkowo, poprzez praktyczne zastosowanie narzędzi statystycznych i analizę danych w projekcie, zdobyłam praktyczne umiejętności, które mogą być bardzo przydatne w przyszłej pracy. Umiejętność korzystania z języka R i środowiska RStudio otwiera nowe możliwości analizy danych i umożliwia bardziej zaawansowane badania w przyszłości.

Podsumowując, przeprowadzony projekt na temat rynku samochodowego za pomocą statystycznej analizy danych dostarczył cennych wniosków, poszerzył moją wiedzę i umiejętności z zakresu statystyki oraz umożliwił dalsze rozwijanie się w tej dziedzinie. Pozyskane umiejętności mogą okazać się niezwykle przydatne w przyszłej pracy, szczególnie w obszarze analizy danych.

Dzięki szczegółowej analizie jesteśmy w stanie poznać badany przez nas temat, np:rynek samochodowy oraz wyciągać wnioski i odszukać nawet nieintuicyjne fakty. Projekt pomógł mi w głębszym zaznajomieniu się z używaniem narzędzi statystycznych. Poszerzyłam także swoją znajomość języka R oraz środowiska RStudio. Powyższy projekt jest podstawą do dalszego zagłębiania się w statystykę od jej strony prkatycznej, a poznane umiejętności mogą okazać się bardzo przydatne w przyszłej pracy.