IST_719

Cal Wardell

6/14/2022

Goal

Through visualization, answer various questions about cars being sold in Poland to obtain the best value for the price of a car.

About the data

This data set comes from Kaggle.com (Glotov, 2022). It contains the following information on cars being sold in Poland: make, model, generation, year, mileage, engine size in cc's, engine type, city in Poland, region of Poland, and price of the car in Polish złoty (PLN). My top two questions were: 1. How does mileage affect the price of a car? 2. Which car makes have the cheapest selection of cars?

#Step 1 Clean and prep the data

```
carprice <- read.csv('IST719_InformationVisualization_data.csv', header = TRU
E, stringsAsFactors = FALSE)

#Structure of our data
str(carprice)
```

```
## 'data.frame':    117927 obs. of  11 variables:
##  $ X              : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ mark           : chr  "opel" "opel" "opel" "opel" ...
##  $ model          : chr  "combo" "combo" "combo" "combo" ...
##  $ generation_name: chr  "gen-d-2011" "gen-d-2011" "gen-d-2011" "gen-d-201
1" ...
##  $ year           : int  2015 2018 2015 2016 2014 2017 2017 2016 2014 2015
...
##  $ mileage        : int  139568 31991 278437 47600 103000 121203 119965 20
1658 178666 113000 ...
##  $ vol_engine     : int  1248 1499 1598 1248 1400 1598 1248 1248 1598 1248
...
##  $ fuel           : chr  "Diesel" "Diesel" "Diesel" "Diesel" ...
##  $ city           : chr  "Janki" "Katowice" "Brzeg" "KorfantÃ³w" ...
##  $ province       : chr  "Mazowieckie" "ÅšlÄ…skie" "Opolskie" "Opolskie" .
..
##  $ price          : int  35900 78501 27000 30800 35900 51900 44700 29000 2
8900 34900 ...
```

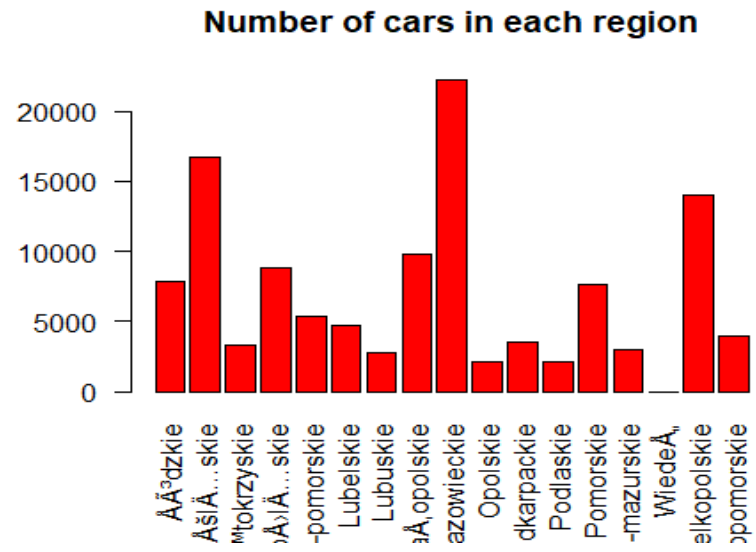
```
carprice <- carprice[carprice$province != '(',] #Saw we had this variable so
we removed it
#Drop car ID column
carprice = subset(carprice, select = -c(X) )
par(mar = c(12,6,2,2)) #Set parameters for the graphs
province_df <- carprice[carprice$province != 'Berlin',]
province_df <- province_df[province_df$province != 'Moravian-Silesian Region'
,]
province_df <- province_df[province_df$province != 'Niedersachsen',]
province_df <- province_df[province_df$province != 'Nordrhein-Westfalen',]
province_df <- province_df[province_df$province != 'Trenczyn',]
province_df <- province_df[province_df$province != 'WarmiÅ"sko-mazurskie',]
```

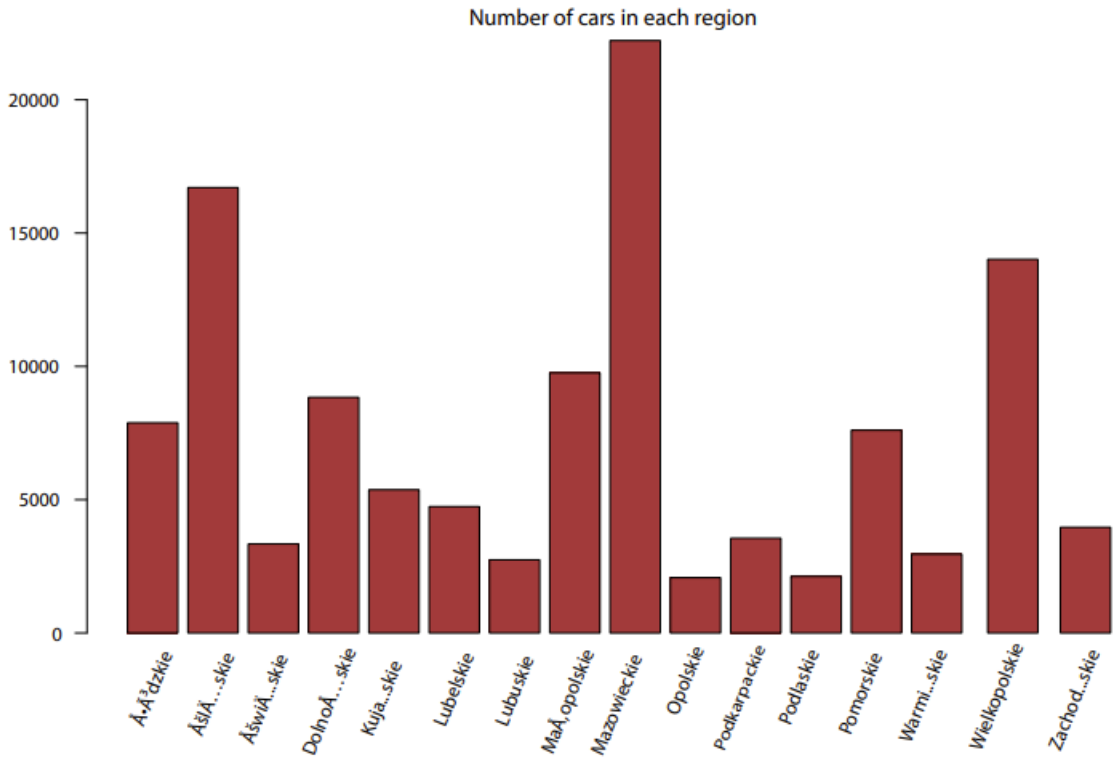
Upon looking at the data, I noticed some bad values; for example, ‘)’ was a location. I filtered these bad values out. I also wanted to focus on the regions of Poland that had the greatest selection of vehicles, because more options provide a better probability of getting a good deal. So, I filtered out the regions that had a very low selection of cars.

Step 2: Create plots

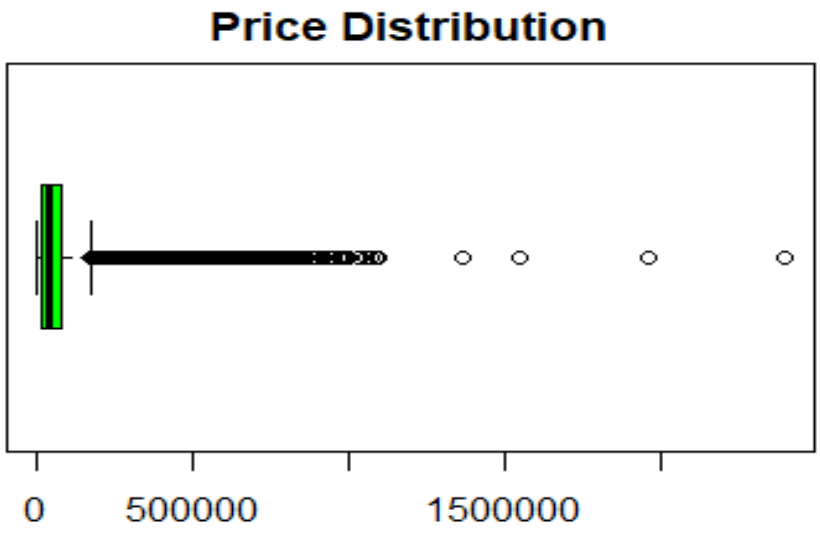
In this class we learned how to enhance visuals through Adobe Illustrator. So after each plot I will show the plot after I improved them in Adobe.

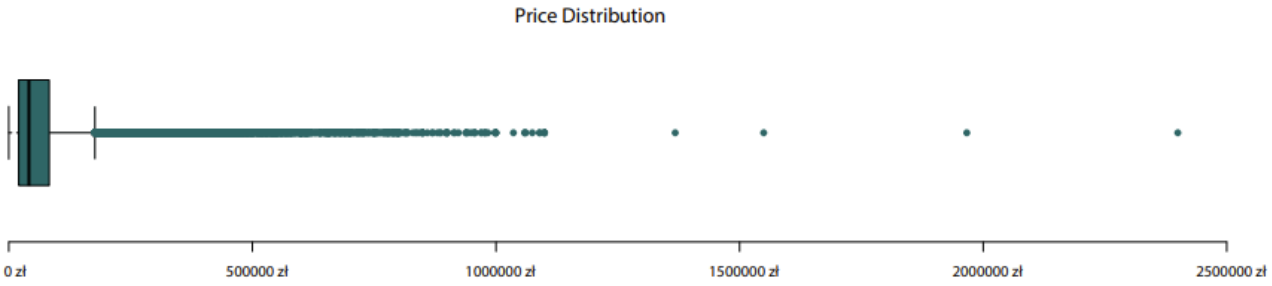
```
#First single dimension visual
barplot(table(province_df$province)
        ,las=2
        , main = 'Number of cars in each region'
        , col = 'red'
        )
```



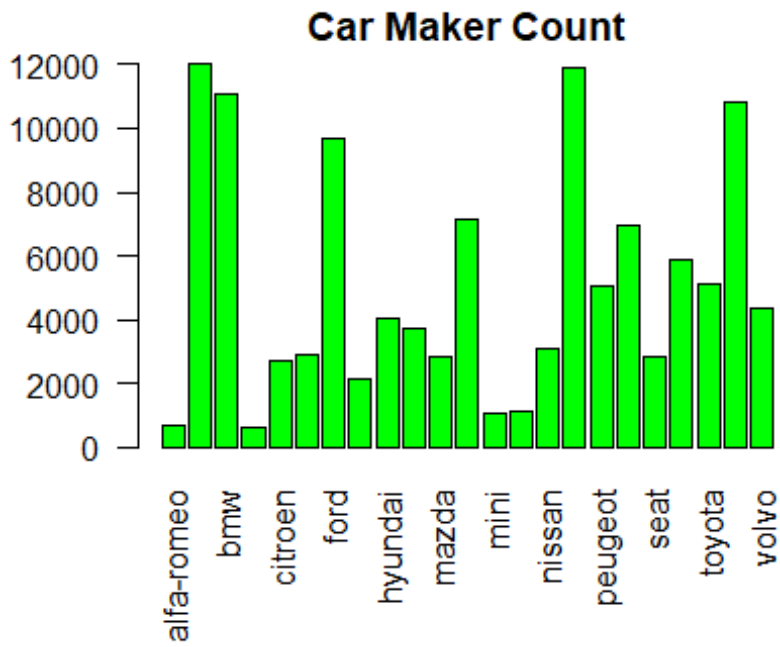


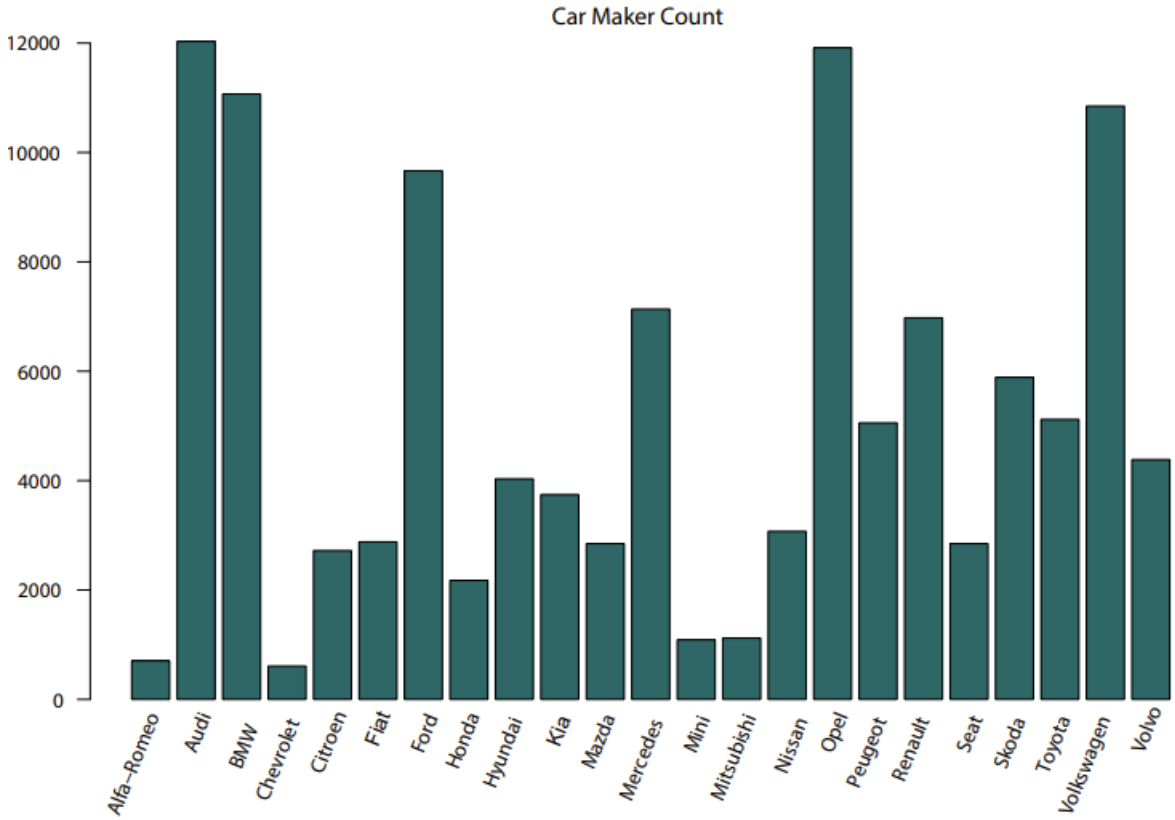

```r
#Second single dimension Visual
par(mar = c(8,6,2,2))
boxplot(x= carprice$price
        #,breaks = 50
        , main = 'Price Distribution'
        , col = 'green'
        , horizontal = TRUE
)
```





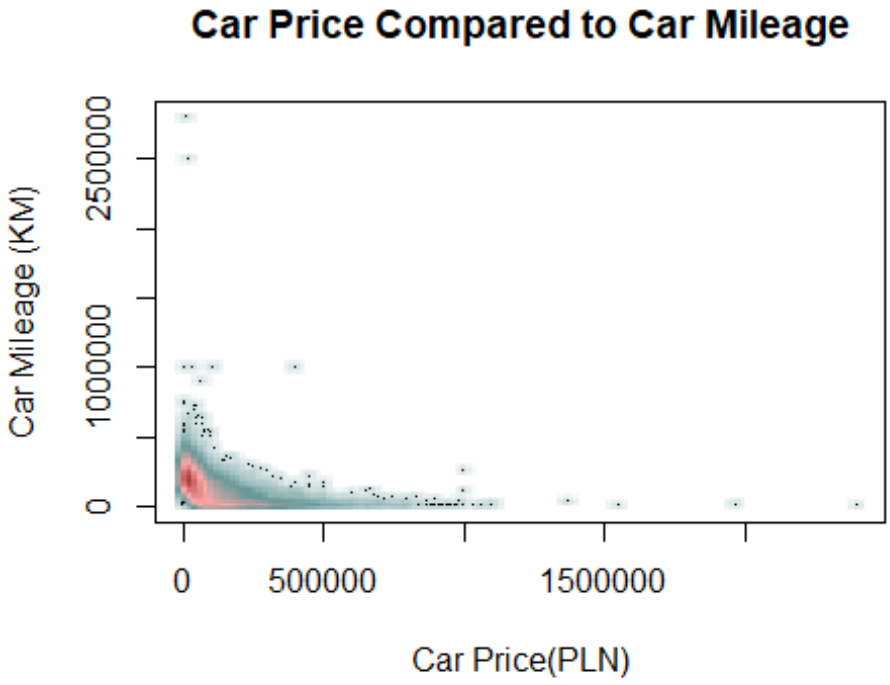

```r
#Third single dimension visual
par(mar = c(8,6,2,2))
barplot(table(carprice$mark)
        ,las=2
        , main = 'Car Maker Count'
        , col = 'green'
)
```

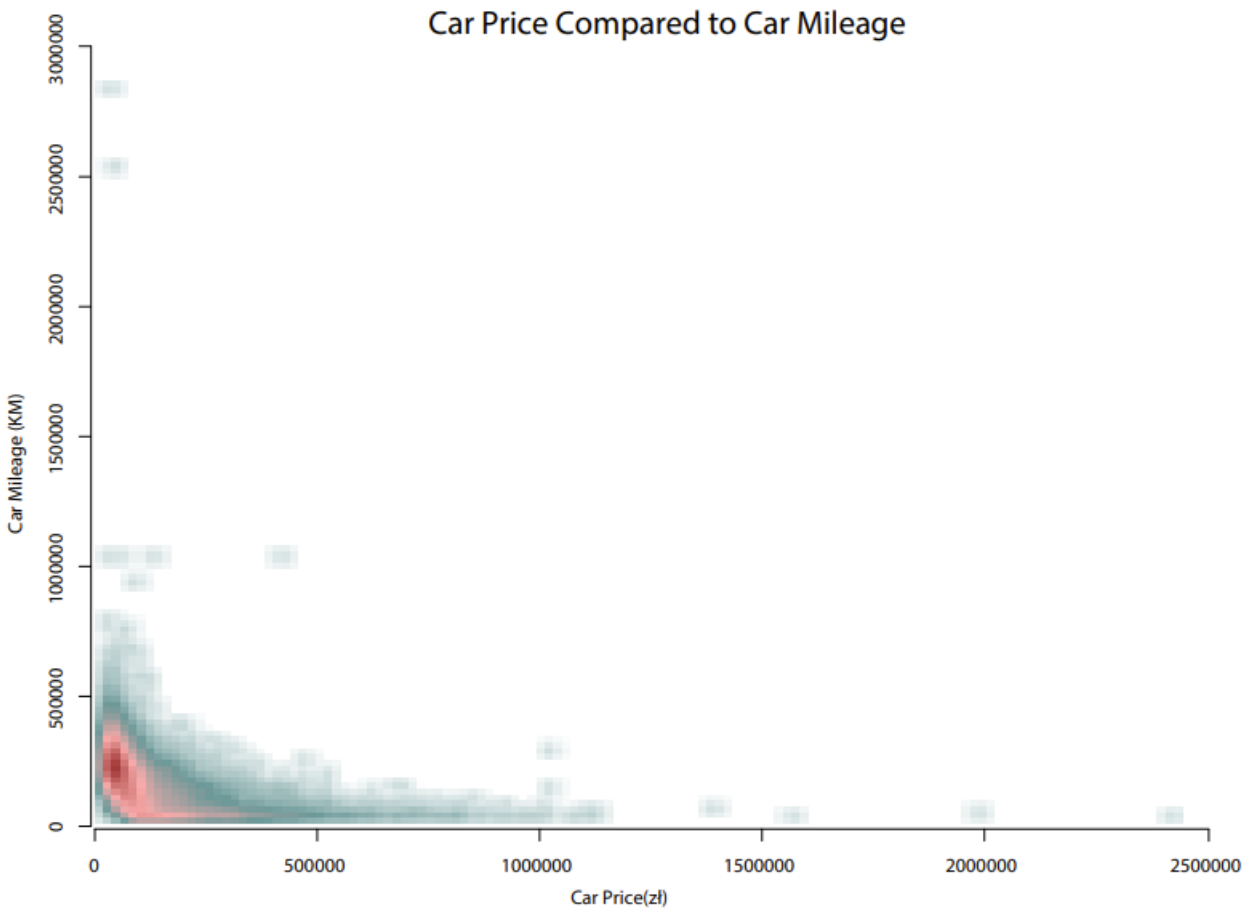

Car Maker Count
12000
10000
8000
6000
4000
2000
0
alfa-romeo
bmw
citroen
ford
hyundai
mazda
mini
nissan
peugeot
seat
toyota
volvo


Car Maker Count
12000
10000
8000
6000
4000
2000
0
Alfa–Romeo
Audi
BMW
Chevrolet
Citroen
Fiat
Ford
Honda
Hyundai
Kia
Mazda
Mercedes
Mini
Mitsubishi
Nissan
Opel
Peugeot
Renault
Seat
Skoda
Toyota
Volkswagen
Volvo

After I gained insight on which models and regions had multiple options to choose from, I created two-dimensional plots to answer my questions. 1. How does mileage affect the price of a car?

```
smoothScatter(carprice$price
              ,carprice$mileage
              ,colramp = colorRampPalette(c("white", "#669999", "#FFAAAA", "#
AA3939"))
              ,main = 'Car Price Compared to Car Mileage'
              ,xlab = 'Car Price(PLN)'
              ,ylab = 'Car Mileage (KM)'
              )
```



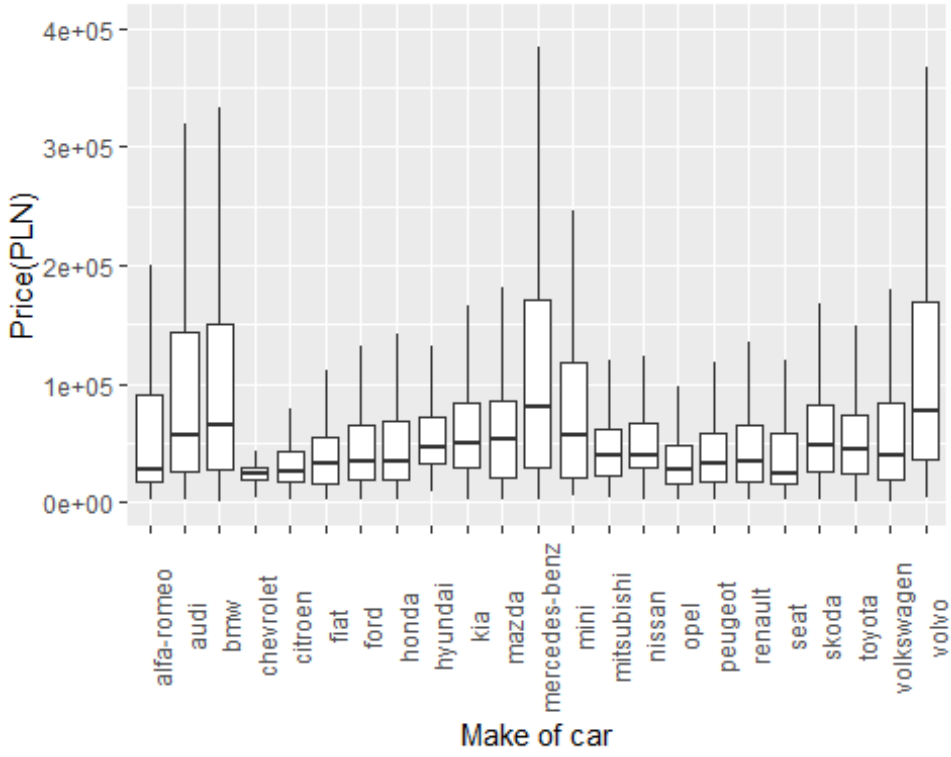


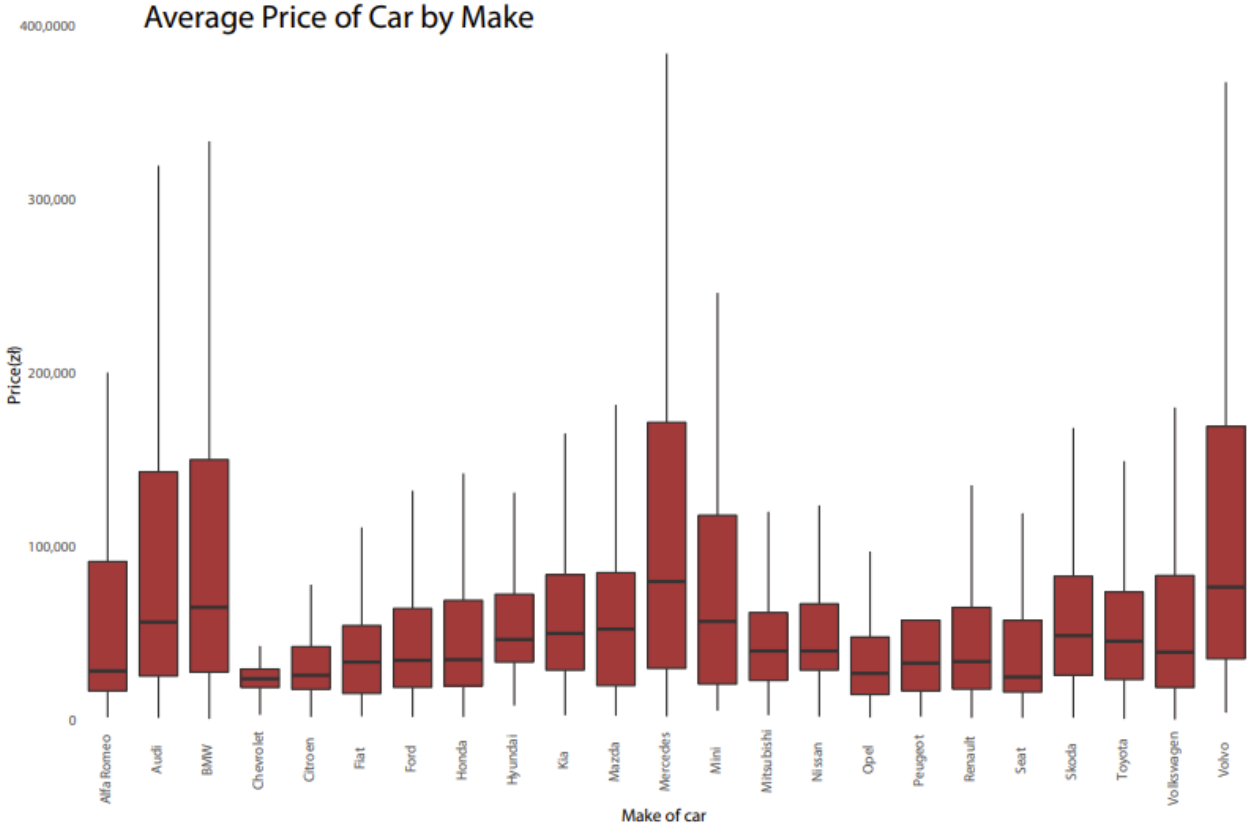
This plot implies that cheaper cars generally have more miles while more expensive cars have fewer miles. Additionally, the plot suggests that the greatest concentration of price and mileage combinations is roughly 100,000 zloty and 25,000 km.

2. Which car makes have the cheapest selection of cars?

```
#maker of the car average price
par(mar = c(8,6,2,2))
library(ggplot2)
ggplot(carprice) + aes(x=mark, y=price) + geom_boxplot(outlier.shape=NA) +
  theme(axis.text.x = element_text(angle = 90)) + ylim(0,400000) + ylab('Pric
e(PLN)')+
xlab('Make of car')

## Warning: Removed 1359 rows containing non-finite values (stat_boxplot).
```


4e+05
3e+05
2e+05
1e+05
0e+00
Price(PLN)
alfa-romeo
audi
bmw
chevrolet
citroen
fiat
ford
honda
hyundai
kia
mazda
mercedes-benz
mini
mitsubishi
nissan
opel
peugeot
renault
seat
skoda
toyota
volkswagen
volvo
Make of car


Average Price of Car by Make
400,0000
300,000
200,000
100,000
0
Price(zł)
Alfa Romeo
Audi
BMW
Chevrolet
Citroen
Fiat
Ford
Honda
Hyundai
Kia
Mazda
Mercedes
Mini
Mitsubishi
Nissan
Opel
Peugeot
Renault
Seat
Skoda
Toyota
Volkswagen
Volvo
Make of car

Mercedes, Volvo, Audi, and BMW are the car makes with the largest distribution of pricing, making them less likely to offer a cheaper car. Ford, Honda, and Opel have a greater chance at providing more affordable cars because the distribution of their prices is much smaller and lower.

Conclusion

Finding a good deal on a car can be difficult, but visualizing the data offers quick and efficient insight to help our decisions. The recommendation I would make to an individual searching for a good deal on a new car would be to look at a Ford in the Mazowiekie region of Poland. This make has a low-price range and has many cars for sale. In addition, the Mazowiekie region has the largest selection of vehicles in Poland, providing customers with a better chance at finding a good deal. Finally, because the data suggests that low mileage correlates with higher cost, I would recommend looking for a car with relatively low mileage but not at a very high price.