Assigment 2

Luis Hinostroza

2023-10-06

Questions about Cars from the "Auto" Dataset.

In this study we used a dataset "Auto" of different cars between 1970's and 1980's from three origins: America (1), European (2), and Japanese (3).

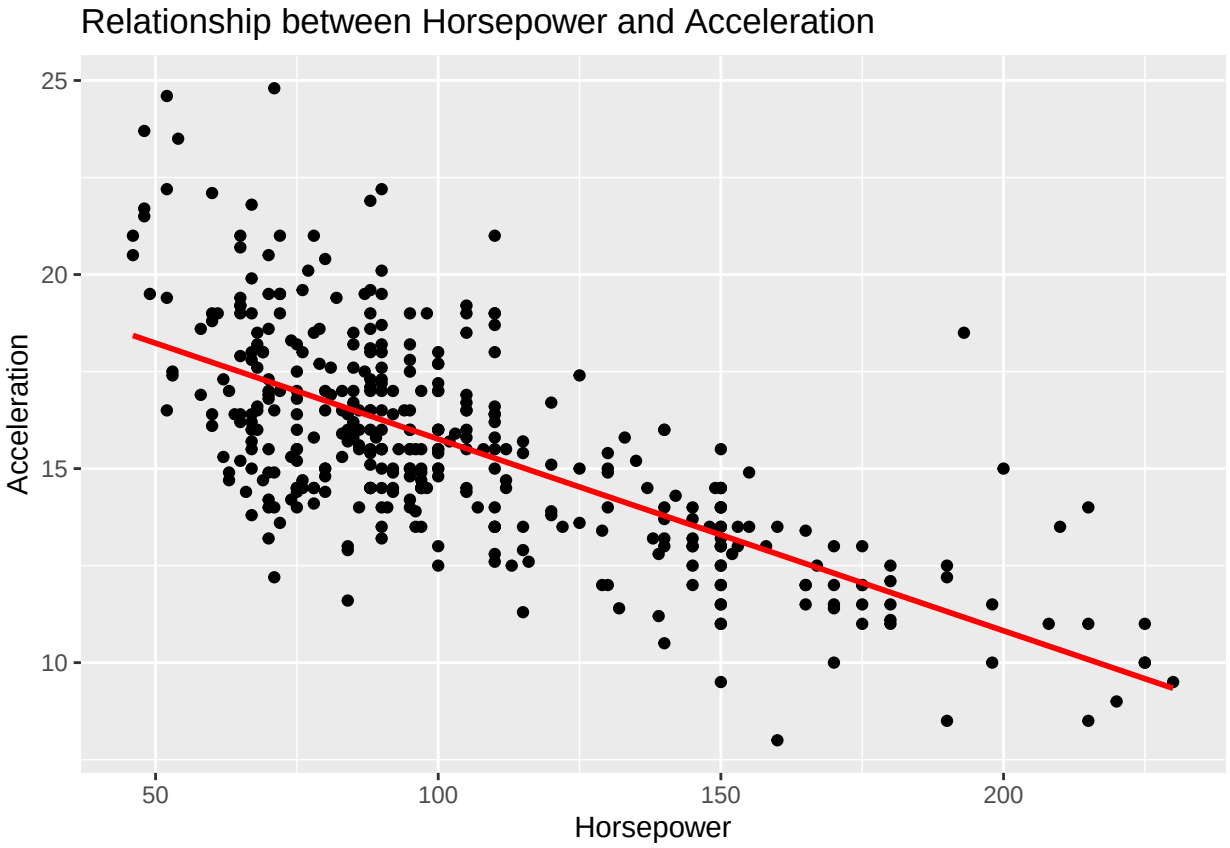
I found that there is a close relationship between all variables. Every variable depend from each other and their values are numerical and categorical. Also, a good source for understanding how and why the mpg, weight, cylinder, horsepower, and displacement can affect the consumption of gas, the time in acceleration, and the size of the engine.

We use different plots like scatterplot, boxplot, mosaic plot, and bar plot. We also use the correlation, chi-squared test, and anova test to understand relationship and if there are statistically significant difference.

Overall, this study is to understand and visualize the data in a more friendly and easy representation of it.

Q#01 - Does more horsepower decreases the acceleration time?

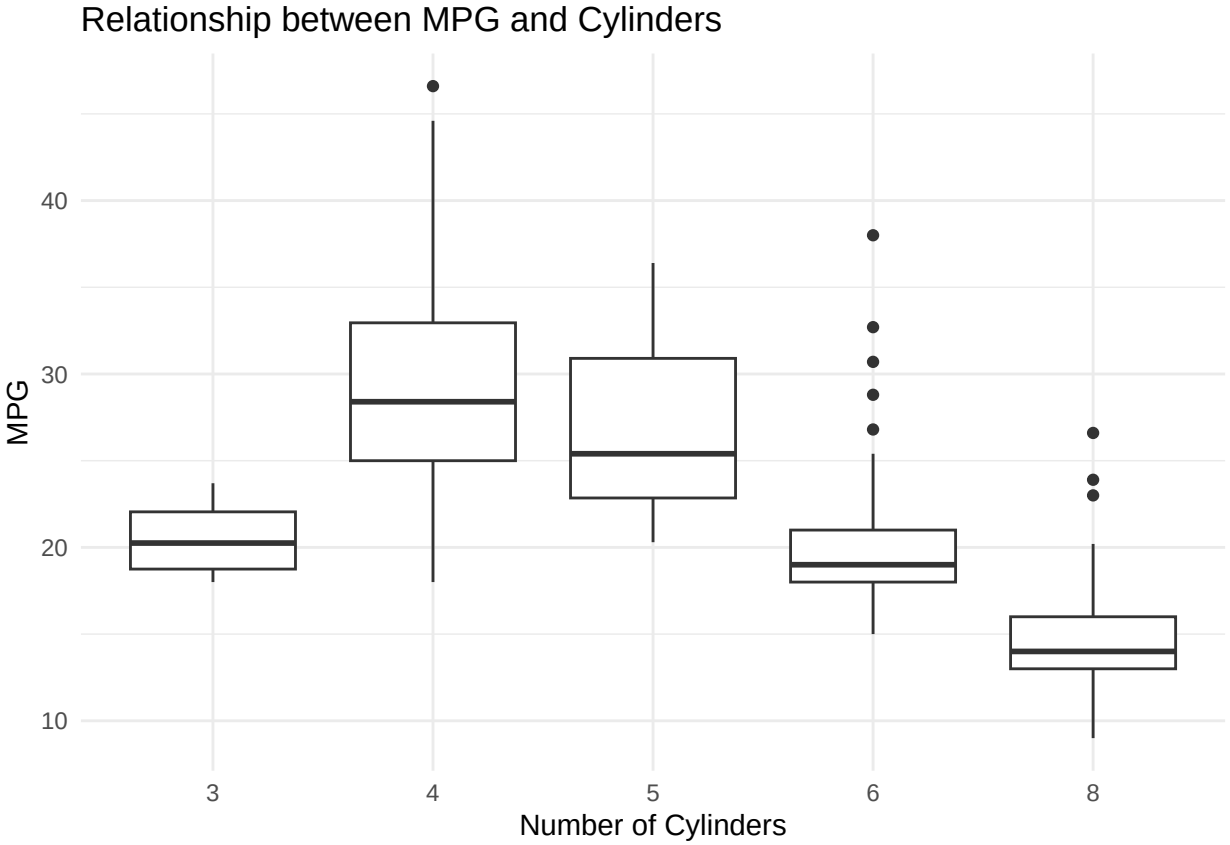
Looking at the graph we can notice there is a relation between hp and acceleration. The more hp the less time accelerating. Also the correlation coefficient represent a pretty strong linear relation.




```
## [1] -0.6891955
```

Q#02 - Does the mpg depends on the amount of cylinders?

The graph shows that the right amount of cylinder can improve the mpg, Four cylinders been the most efficient miles per gallons. The p-value from the ANOVA test shows there's a statistically significant difference in the mean mpg across different cylinders groups.

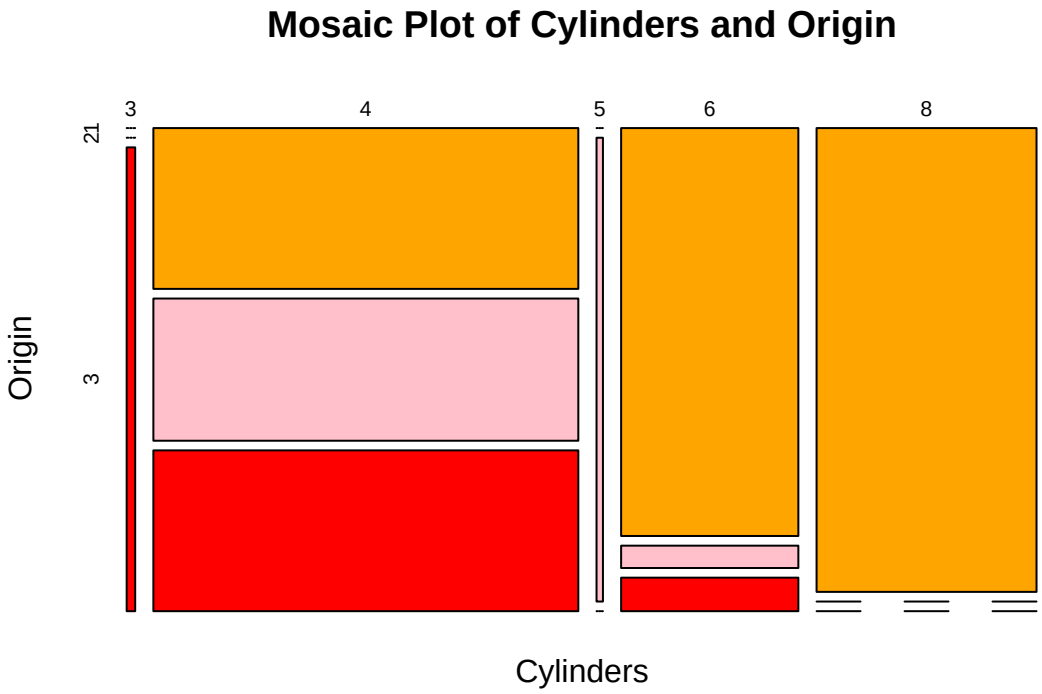



```
##                       Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(cylinders)   4  15275    3819     173 <2e-16 ***
## Residuals            387   8544      22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q#03 - Is the amount of cylinders related to the origin?

We can notice from the graph the distribution of car by number of cylinders. Orange being American (1), pink European (2), and red Japanese (3). The p-value indicates that the two variables are not independent and that there is a significant association between cylinders and origin.

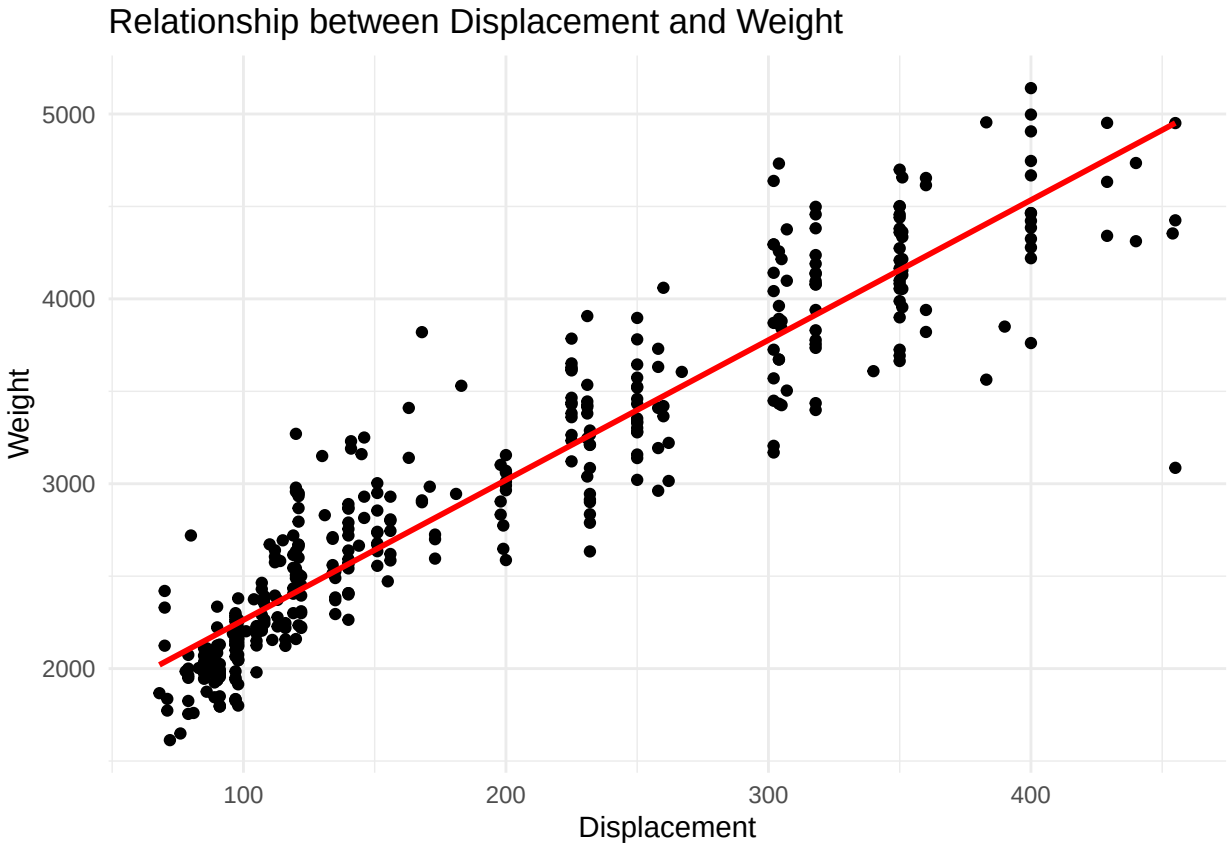
```
##
##       1   2   3
##   3   0   0   4
##   4  69  61  69
##   5   0   3   0
##   6  73   4   6
##   8 103   0   0
```




```
##
##  Pearson's Chi-squared test
##
## data:  table_cyl_origin
## X-squared = 180.72, df = 8, p-value < 2.2e-16
```

Q#04 - Is there a relation between the displacement and weight?

We can see on the plot that the points follow the red line affirming the linear relation between both variables, proving that weight is related to it displacement. And, the correlation coefficient indicates a strong linear relationship as well.




```
## [1] 0.9329944
```

Q#05 - What is the percentage of cars by origin?

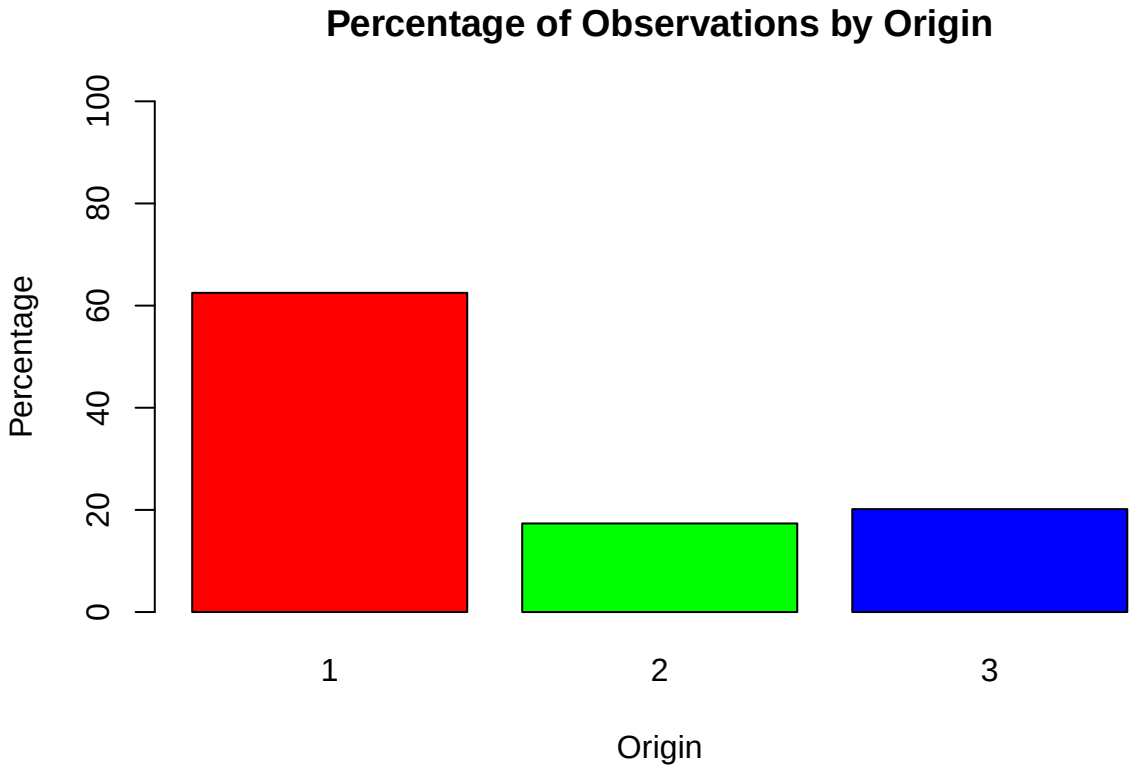
It would be interesting to observe the amount and percentage of cars by origin. The chart shows that mostly all cars come from America (1), following by Japan (3).

Amount of cars by origin.

```
##
##   1   2   3
## 245  68  79
```

Percentage of cars by origin.

```
##
##        1        2        3
## 62.50000 17.34694 20.15306
```



Limitations of Report:

- The dataset contains a limited set of variables, which may not capture all factors influencing car characteristics and performance. For comprehensive analysis, we may need more variables or features.
- The origin variable is represented as numerical values (1 for American, 2 for European, and 3 for Japanese). This is very misleading as they are categorical values.
- The dataset is relatively old, as it shows cara from the 70s and 80s. The relationships and trends observed in this dataset may not be representative of modern cars due to new technologies, changes in manufacturing practices, and shifts in consumer preferences.
- While the dataset provides various attributes for cars, it might lack other definitions or details for each variable, which can lead to misinterpretation or misuse of the data.

- For question #1: We visualize the relationship of the data between horsepower and Acceleration with a scatterplot and a linear fit. We also found the correlation coefficient.

```
library(ggplot2)
library(ISLR)
library(dplyr)
hp_acce <- select(Auto, horsepower, acceleration)
ggplot(hp_acce, aes(x = horsepower, y = acceleration)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Relationship between Horsepower and Acceleration",
       x = "Horsepower",
       y = "Acceleration")
cor(Auto$horsepower, Auto$acceleration)
```

- For question #2: We visualize the relationship of the data between mpg and cylinders with a boxplot. Each box represents the interquartile range of mpg for a specific number of cylinders. We also found the p-value from the ANOVA test to find out if there's a statistically significant difference in the mean mpg across different cylinders

```
library(ggplot2)
library(ISLR)
library(dplyr)
mpg_cyl <- select(Auto, mpg, cylinders)
ggplot(mpg_cyl, aes(x = as.factor(cylinders), y = mpg)) +
  geom_boxplot() +
  labs(title = "Relationship between MPG and Cylinders",
       x = "Number of Cylinders",
       y = "MPG") +
  theme_minimal()

anova_result <- aov(mpg ~ as.factor(cylinders), data = Auto)
summary(anova_result)
```

- For question #3: We created a contingency table to show the frequency distribution of the variables ina matrix format. We also created a mosaic plot to visualize the proportion to the number of cases in each category. We also found the chi-squared test to determine if there is a significant association between the two categorical variables.

```
library(ggplot2)
library(ISLR)
library(dplyr)
library(vcd)
table_cyl_origin <- table(Auto$cylinders, Auto$origin)
print(table_cyl_origin)
mosaicplot(table_cyl_origin, main="Mosaic Plot of Cylinders and Origin", xlab="Cylinders", ylab="Origin
chi_sq_test <- chisq.test(table_cyl_origin)
print(chi_sq_test)
```

- For question #4: We visualize the relationship of the data between displacement and weight with a scatterplot and a linear fit. We also found the correlation coefficient to indicate if there is a strong positive linear relationship.

```
library(ggplot2)
library(ISLR)
library(dplyr)
ggplot(Auto, aes(x = displacement, y = weight)) +
```

```
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Relationship between Displacement and Weight",
       x = "Displacement",
       y = "Weight") +
  theme_minimal()
cor(Auto$displacement, Auto$weight)
```

- For question #5: We plotted a Bar Chart to visualize the percentage of cars by each origin. This observation help us a understand the distribution of all cars by their origin.

```
table_origin <- table(Auto$origin)
print(table_origin)
prop_origin <- prop.table(table_origin) * 100  # Convert to percentage
print(prop_origin)

barplot(prop_origin,
        main = "Percentage of Observations by Origin",
        xlab = "Origin",
        ylab = "Percentage",
        col = c("red", "green", "blue"),
        ylim = c(0, 100))
```