horizontal line

US Car Prices

March 3, 2023

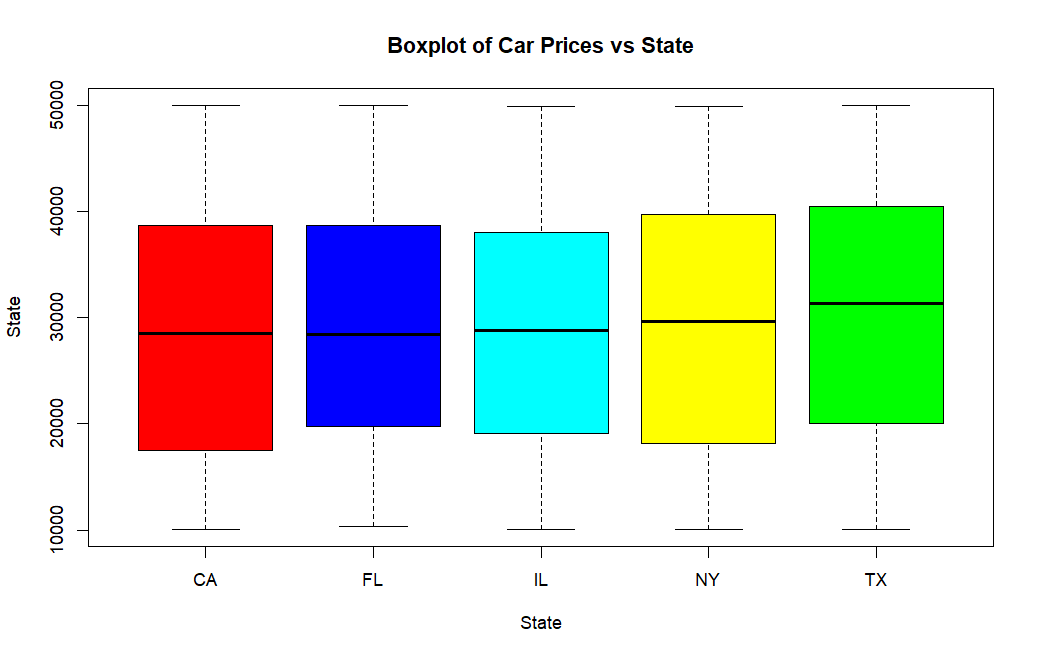
By: Raashi Maheshwari

**────────────────────────────────────**

# Introduction

Hello everyone, my name is Raashi Maheshwari, and I’m currently a freshman majoring in computer science. The COVID-19 pandemic causes chaos in the whole world, causing the world to stop for almost two years. However, now, things have started to go back to a somewhat normal situation, except for one big change. After the COVID-19 pandemic, inflation has been quite absurd. The prices of everything have risen, from things as simple as eggs to cars. I wanted to just compare the prices of cars and analyze if certain brands or models or states have higher prices than others. My dataset is from Kaggle, and it is a collection of 1000 randomly generated data for used car prices. It includes several attributes such as the make, model, year, mileage, price, color, state, and date of each year. While I will not be focusing on all of these aspects, I will choose the ones that I feel would make the most impact on the prices of the cars which would be the state in which state the car is bought, the make of the car, and the mileage it has on it. Although this dataset is not a complete representation of all the brands, makes, models, and states, it will give some insight into what determines the prices of used cars. With the proper analysis, this report might be a source of knowledge for potential buyers of used cars.

# Mean Prices Based on State

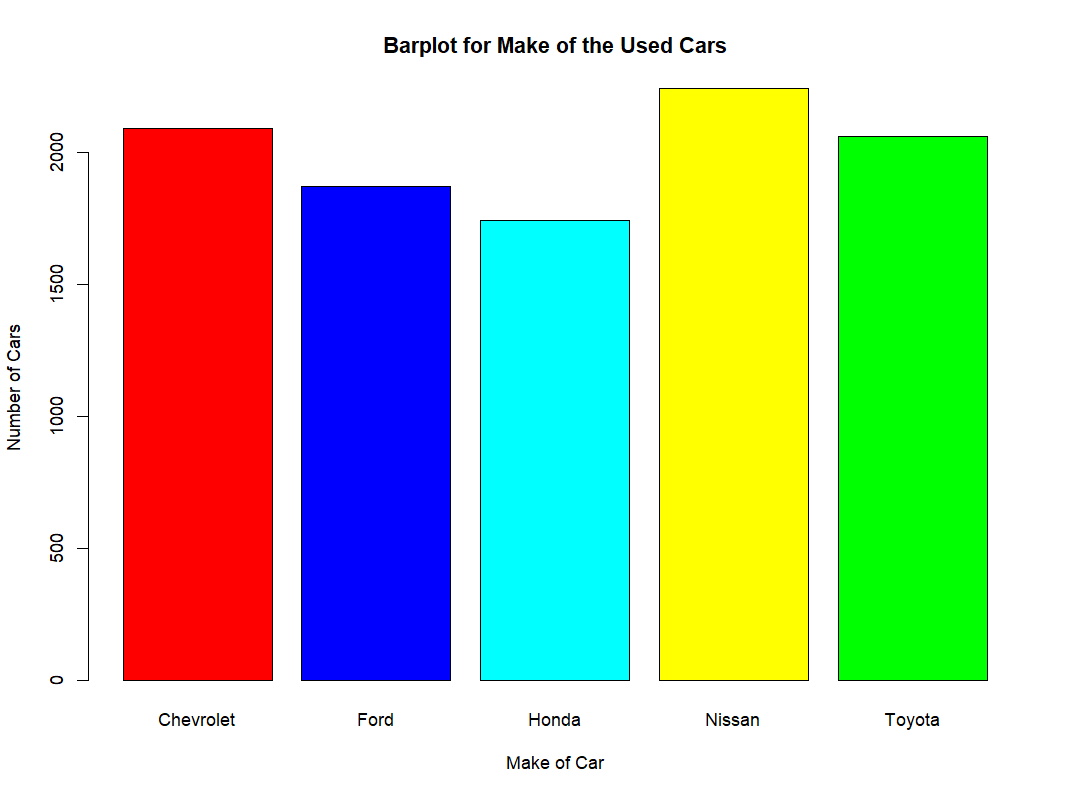


The boxplot above shows the distribution of the five different states in the dataset and the range of the prices of the cars in each state. As the boxplots show above, California seems to have the lowest price for used cars, followed by New York, and then Illinois. On the other hand, Texas seems to have the highest price for used cars, and New York follows.

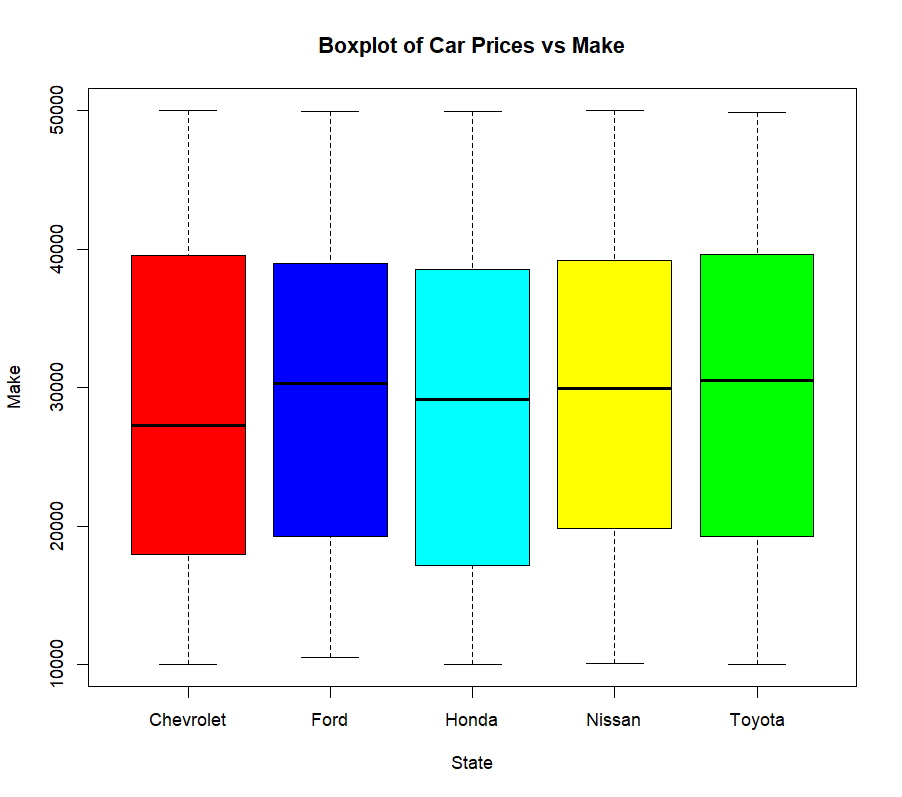
After subsetting the data and using the mean function to find the average prices of cars in every state, I obtained the following results:

* Mean for Florida: $29250.86
* Mean for California: $28424.20
* Mean for Texas: $30604.47
* Mean for Illinois: $28984.39
* Mean for New York: $29587.61

# Distribution of Make of Used Cars

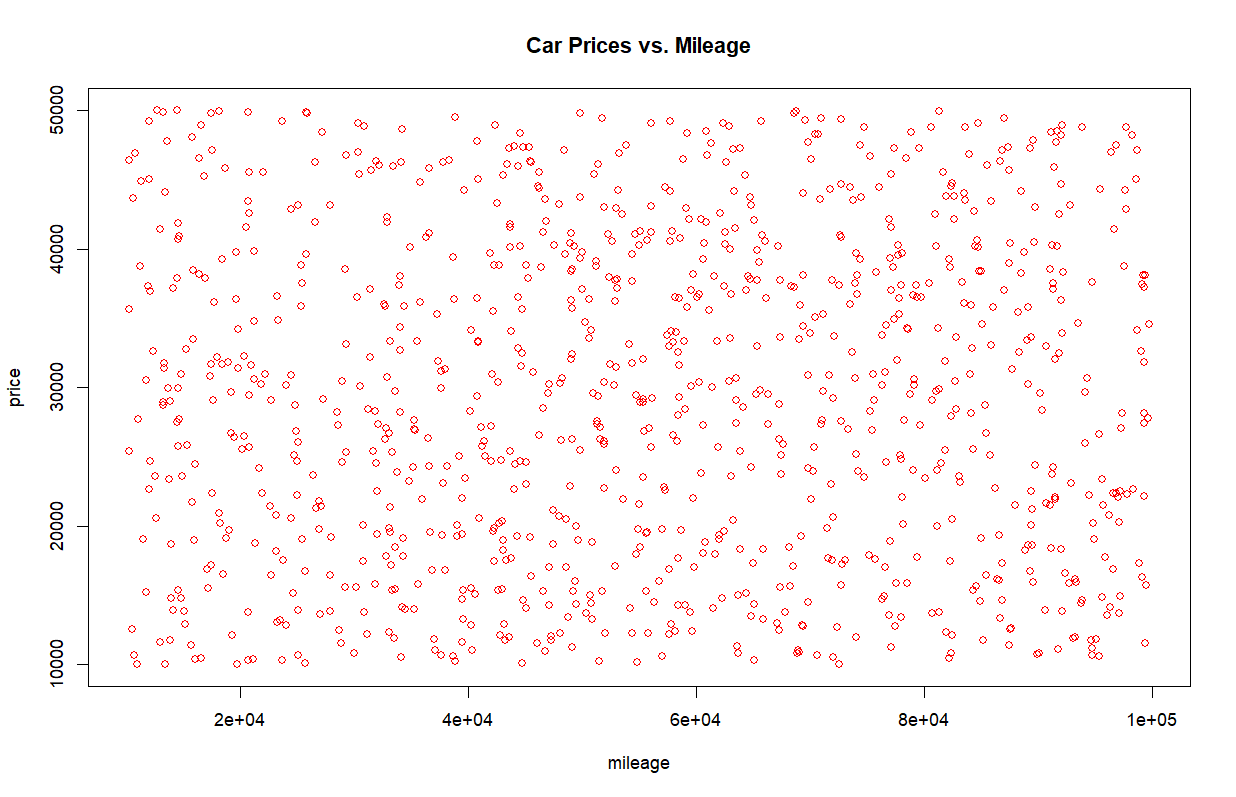
****As seen by the above boxplot, Nissan seems to have the most used cars, with Chevrolet and Toyota following close behind. Ford and Honda seem to have a lower number of used cars for sale.

# Car Prices vs. Mean

****

According to the boxplot above, Honda seems to have the lowest price of used cars followed by Chevrolet. The remaining three makes, Ford, Nissan, and Toyota have approximately the same lowest price. The make of all five brands have almost the same for the highest price of a used car.

# Car Prices vs Mileage



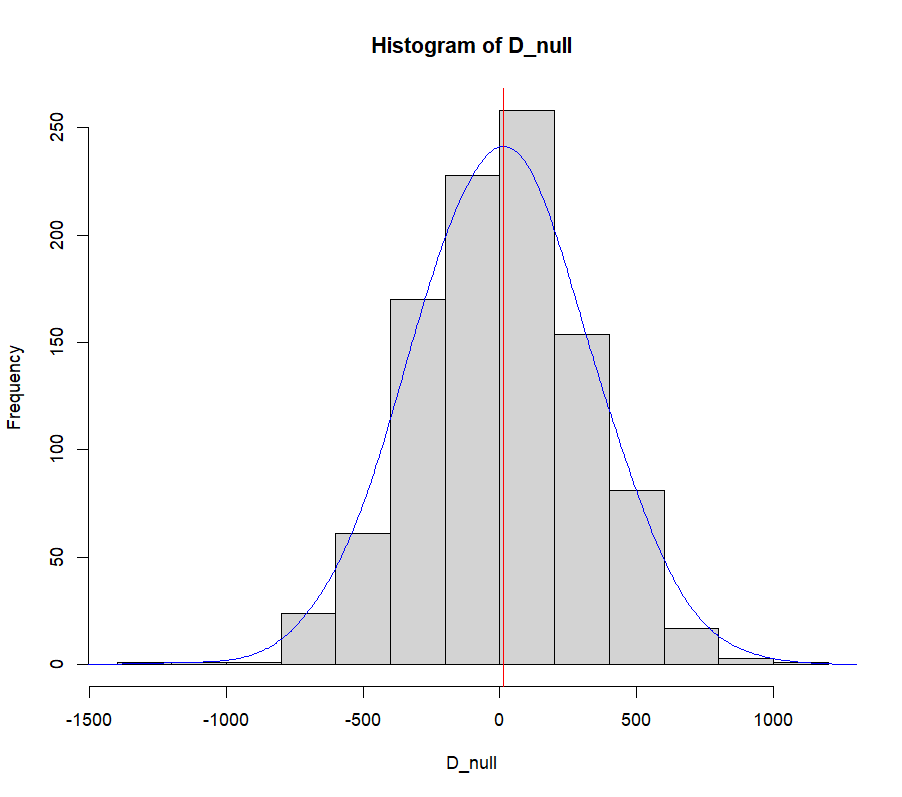
As you can see above, the scatter plot of Car Prices and Mileage is completely scattered, which means that there is no specific pattern or correlation between car prices and mileage. While the common thought is that the higher the mileage on the car, the higher the price, this scatter plot above dismisses that thought. It shows that mileage is not the only factor that plays into the price of a used car.

# 

# Hypothesis 1:

Null Hypothesis: Regardless of whether the car is a Honda or Chevrolet, the price of the used car stays the same.

Alternative Hypothesis: Used Chevrolet cars have a higher price than used Honda cars

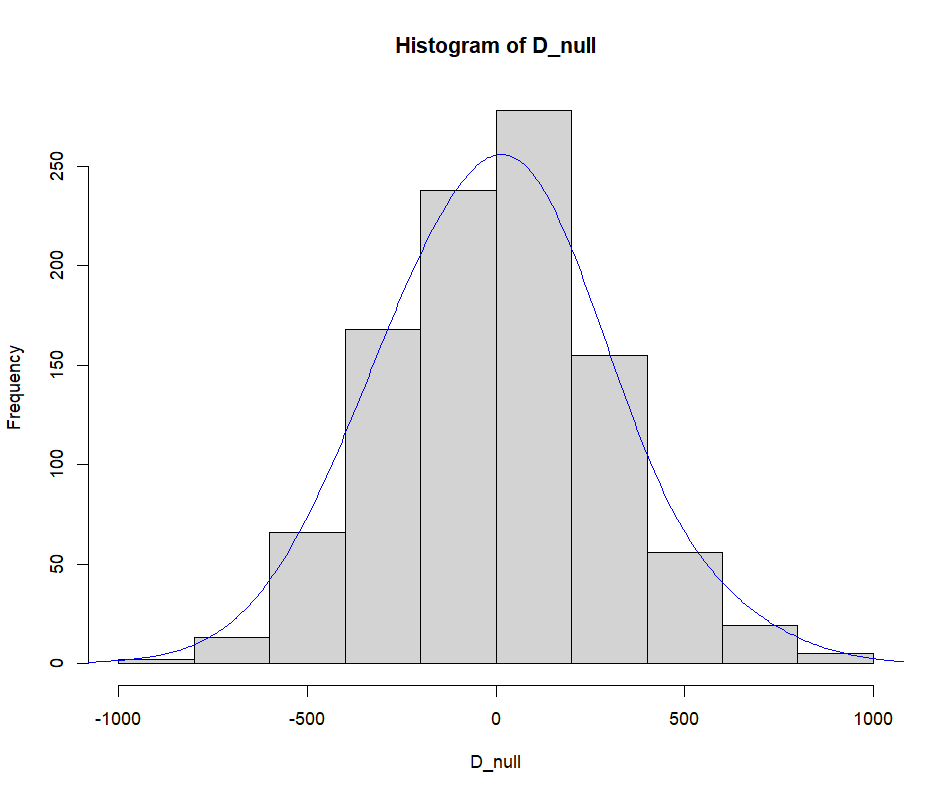


Doing the permutation test for the above null and alternative hypothesis, we get a p-value of 0.499, which is higher than the significance level of 0.05. Since 0.499 > 0.05, we fail to reject the null hypothesis as we have very little evidence against the null hypothesis. This shows that there is a possibility that regardless of whether the car was a Honda or Chevrolet, it has the same price.

# Hypothesis 2:

Null Hypothesis: Used cars sold in Texas and used cars sold in Florida have the same price

Alternative Hypothesis: Used cars sold in Texas have a higher price than used cars sold in Florida



Doing the permutation test for the above null and alternative hypothesis, we get a p-value of 0, which is lower than the significance level of 0.05. Since 0 < 0.05, we have substantial evidence against the null hypothesis, hence we reject the null hypothesis. Thus, we can see that used cars sold in Texas and used cars sold in Florida do not have the same price.

# Conclusion

In conclusion, we can see that there are several things that determine the factor of the price of a car. With this dataset, I analyzed the prices in 5 different states. As seen previously in the report, although there are price differences for each of the different states, they are only slight. However, they are still differences, hence we can still say that buying a car in different states can alter the price of the car. On the other hand, the prices of used cars do depend on the make of the car. Depending on whether the car is a Honda or a Ford, there can be price differences. This conclusion is not only limited to this dataset because there is a high chance that it applies to other models that are not included in this dataset. However, this dataset showed evidence against a common thought: where the higher the mileage, the higher the price of the car. However, with the scatterplot shown above, we can see that there is no correlation between mileage and price. This shows us that there are used cars for sale that have high mileage but can also have lower prices and vice versa. Hence we can conclude that there are several factors that come into play when buying a used car and a potential buyer needs to consider all of these factors before making a decision.

# 

# R Code

car\_prices <- read.csv("C:/Users/msraa/Data101\_Assignments/car\_prices.csv")

NYState <- subset(car\_prices, state == "NY")

CAState <- subset(car\_prices, state == "CA")

TXState <- subset(car\_prices, state == "TX")

FLState <- subset(car\_prices, state == "FL")

ILState <- subset(car\_prices, state == "IL")

> mean(FLState$price)

[1] 29250.86

> mean(CAState$price)

[1] 28424.2

> mean(TXState$price)

[1] 30604.47

> mean(ILState$price)

[1] 28984.39

> mean(NYState$price)

[1] 29587.61

> boxplot(price~state,data=car\_prices,xlab="State",ylab="State", main="Boxplot of Car Prices vs State",col=colors,border="black")

> t <- table(car\_prices$make)

> barplot(t,xlab="Make of Car",ylab="Number of Cars",col=colors, + main="Barplot for Make of the Used Cars",border="black")

>plot(car\_prices$mileage,car\_prices$price,ylab="price",xlab="mileage",main="Car Prices vs. Mileage",col="red")

>boxplot(price~make,data=car\_prices,xlab="State",ylab="Make", main="Boxplot of Car Prices vs Make",col=colors,border="black")

>PermutationTestSecond::Permutation(car\_prices,"make","price", 1000, "Chevrolet", "Honda")

[1] 0.499

> PermutationTestSecond::Permutation(car\_prices,"state","price", 1000, "FL", "TX")

[1] 0

DataSet Link:<https://www.kaggle.com/datasets/at3191/us-car-prices?resource=download>