Rene Guerra - Homework 1

Sunday, October 2th 2022

Machine Learning Main Ideas 1.

Supervised learning demonstrates the existence of a response measurement y_i associated to individual measurement observations x_i, i= 1,. . . ,n of the predictor. Predicts responses that are known.

Unsupervised learning demonstrates the existence of a collection of measurement observations x_i, i= 1,. . . ,n but there is no associated measurement y_i. Results do not reflect stated data.

The difference is that the lack of an associated response in unsupervised learning does not allow to fit a linear regression model. This change also derives data clusters since the unsupervised scenario leads to pattern discovery.

2.

Regression model represents a quantitative response Y such that it describes numerical relationships between variables. Predicts a continuous quantity.

Classification model represents a qualitative response Y such as categorical values. Predicts distinct value labels, discrete.

3.

Two commonly used metrics for regression are Training Mean Squared Error and Mean Squared Error.

Two commonly used metrics for classification are Training Error Rate and Test Error Rate.

4.

Descriptive models interpret patterns in data by utilizing a visual approach such as graphs or scatter plots. This can derive details about trends for the information.

Inferential models take significant features to test theories about a collection of data, and determine a connection between predictors and results.

Predictive models use the best fitting combination of features to describe data. They attempt to predict a response Y with reducible and minimized error, without necessarily testing hypothesis.

5.

Mechanistic modeling utilize theoretical information to predict potential outcomes in an event. Empirically-driven modeling takes already known results from data to derive theories. These models differ in assumption of parametric forms, number of observations, and flexibility. Mechanistic modeling initially assumes a parametric form, it's not exact to the collection of predictors, and it's not flexible with limited parameters. On the other hand, empirically-driven modeling does not assume predictors, requires more observations, and it is flexible. When many parameters are added in the mechanistic model, it derives over fitting similar to empirically-driven data.

Empirically-driven modeling is more flexible since a large number of observations are performed but no assumption of results are made. However, by the bias-variance trade-off, interpretability is inversely related to flexibility. Thus, mechanistic modeling with few parameters is easier to interpret.

Bias-variance trade-off shows the inverse relationship between bias and variance for parameters. It is relevant to mechanistic and empirically-driven models because it derives changes in the models' flexibility/complexity. As the variance increases and bias decreases, complexity for the model increases.

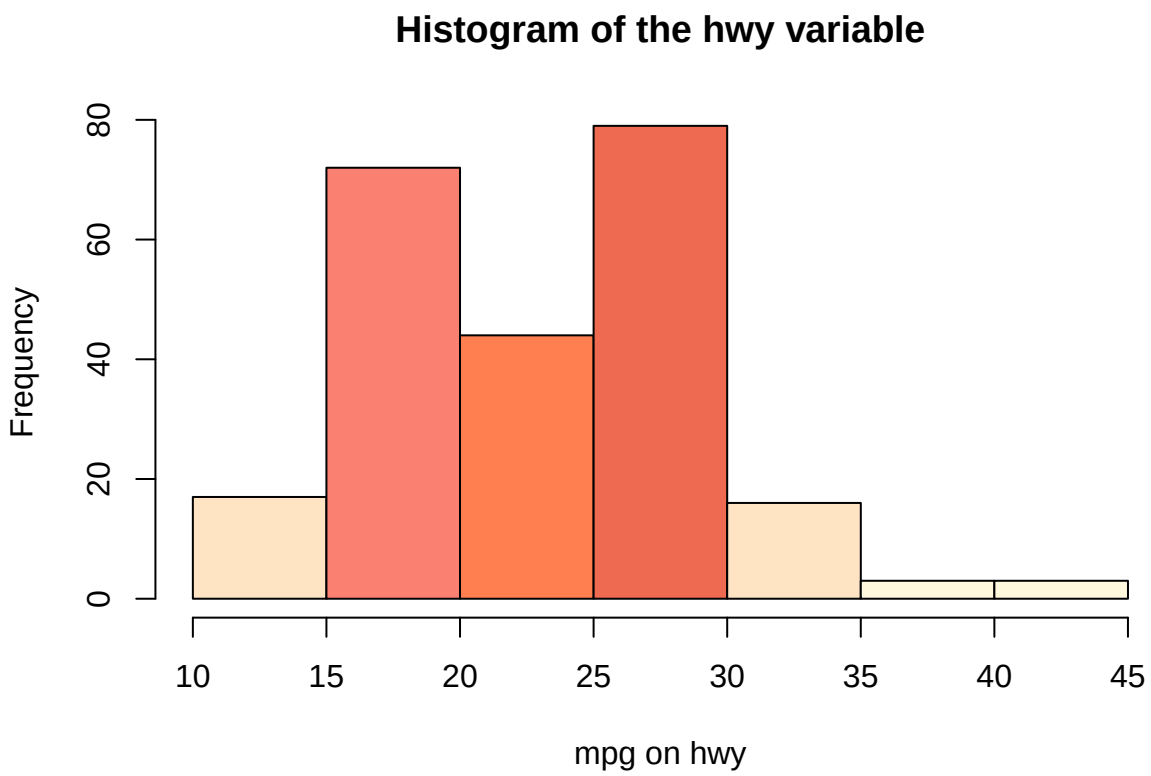
6.

"Given a voter's profile/data, how likely is it that they will vote in favor of the candidate?" is a predictive question because the model wants to predict a response Y from given data with minimal error. It does not attempt to derive hypothesis testing because there is no specific question in the statement derived from the quantitative data.

"How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?" is an inferential question because it wants to test the effects of the predictors on the outcomes. The experimentation's purpose is to find if there is a relationship between both.

Exploratory Data Analysis 1.

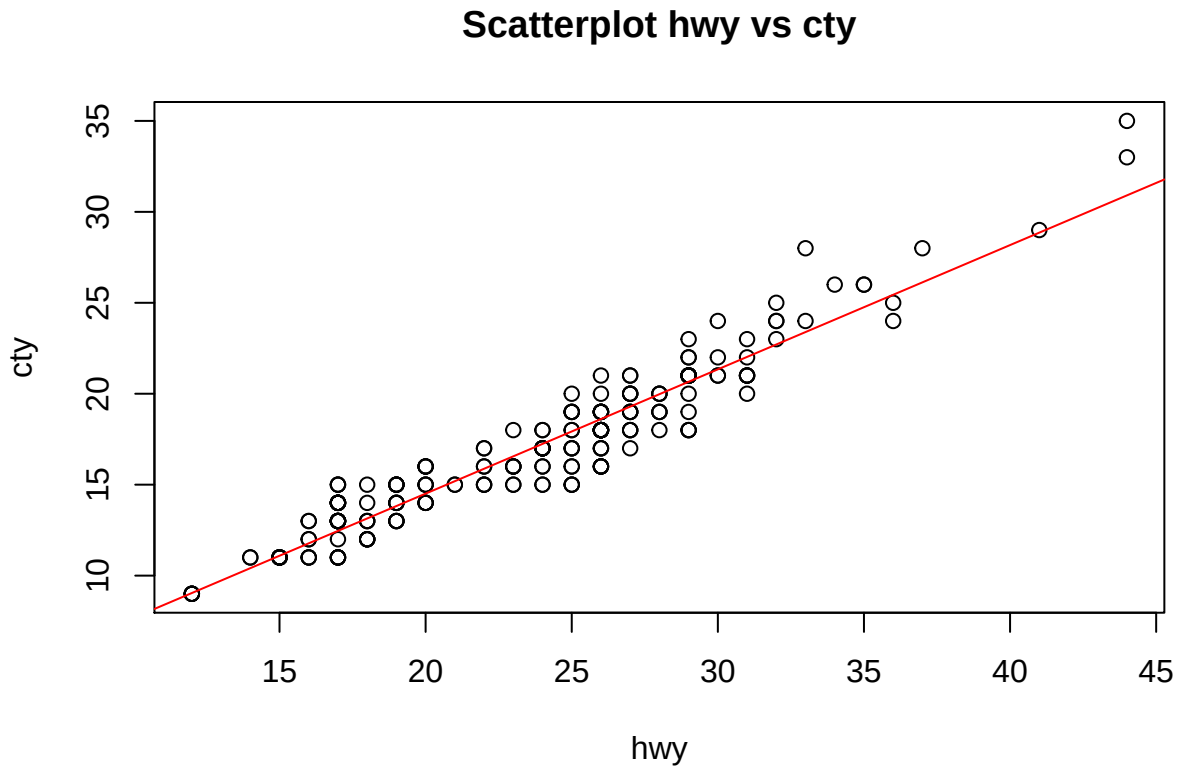
```
library(ggplot2)
library(tidyverse)

data("mpg")
hist(mpg$hwy, main= "Histogram of the hwy variable", xlab= "mpg on hwy",
     col= c("bisque","salmon","coral","coral2","bisque","cornsilk","cornsilk"))
```



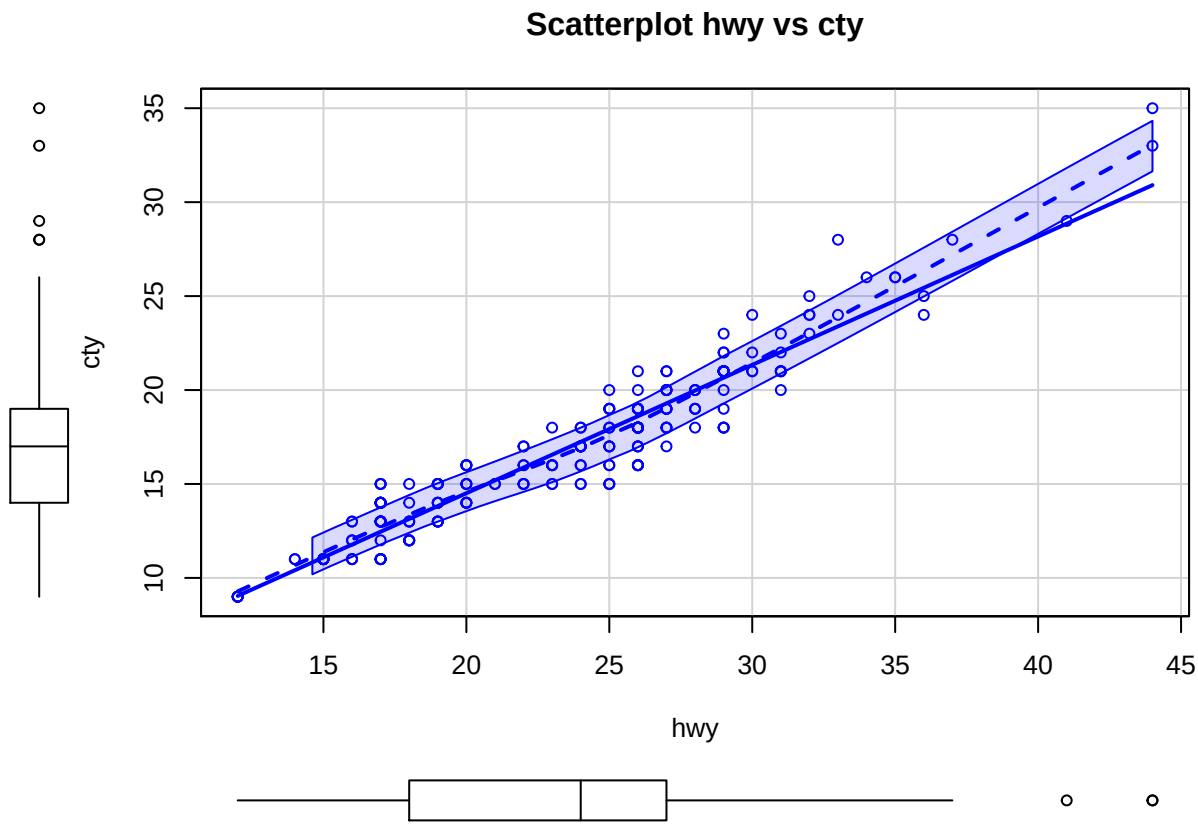
This histogram shows the frequency of miles per gallon on the highway by individual cars in the data set. Highest value of frequency shows that 80 cars have an mgp of about 25-30 on the highway and the lowest value of frequency show that about 4 cars have an mpg of 35-40, and other 4 cars have an mpg of 40-45. The rest of cars are allocated in between these quantities.

2.

```
library(car)

hwy <- mpg$hwy
cty <- mpg$cty

plot(hwy, cty, main= "Scatterplot hwy vs cty")

abline(lm(cty ~ hwy, data= mpg), col= "red")
```



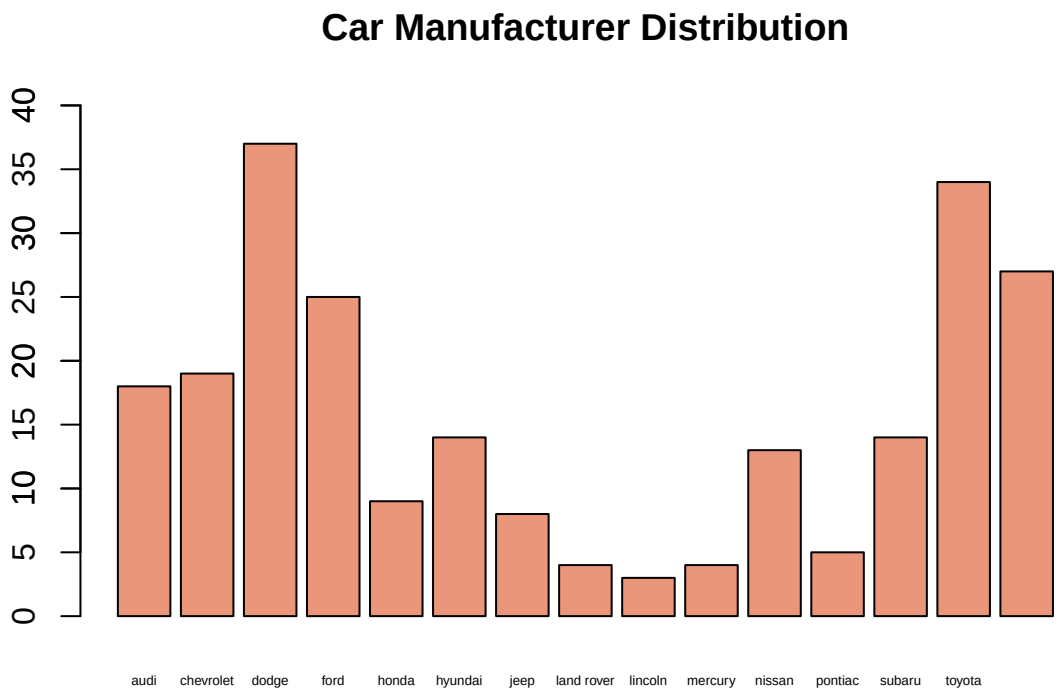

```
scatterplot(cty ~ hwy, main= "Scatterplot hwy vs cty", data= mpg)
```



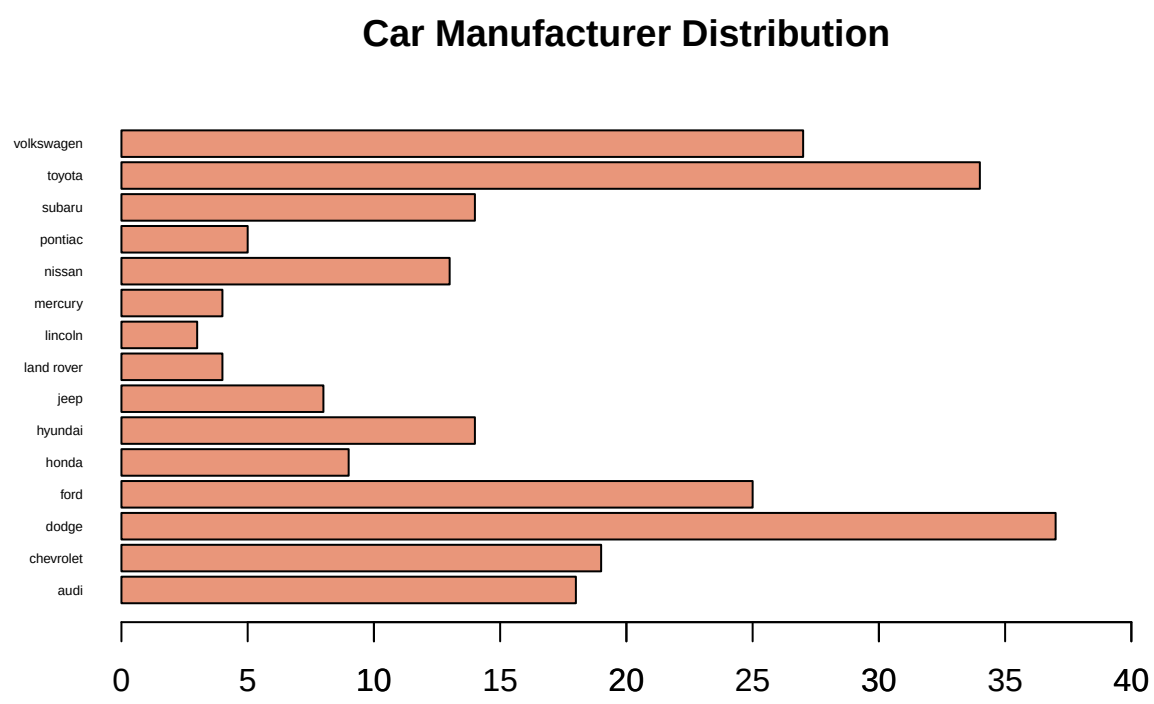
Most of cars have an mpg to the left of 35. The scatter plot graph shows a proportional relationship between mpg in the city and on the highway. If mpg in the city increases mpg on the highway also increases, but it is usually greater.

3.

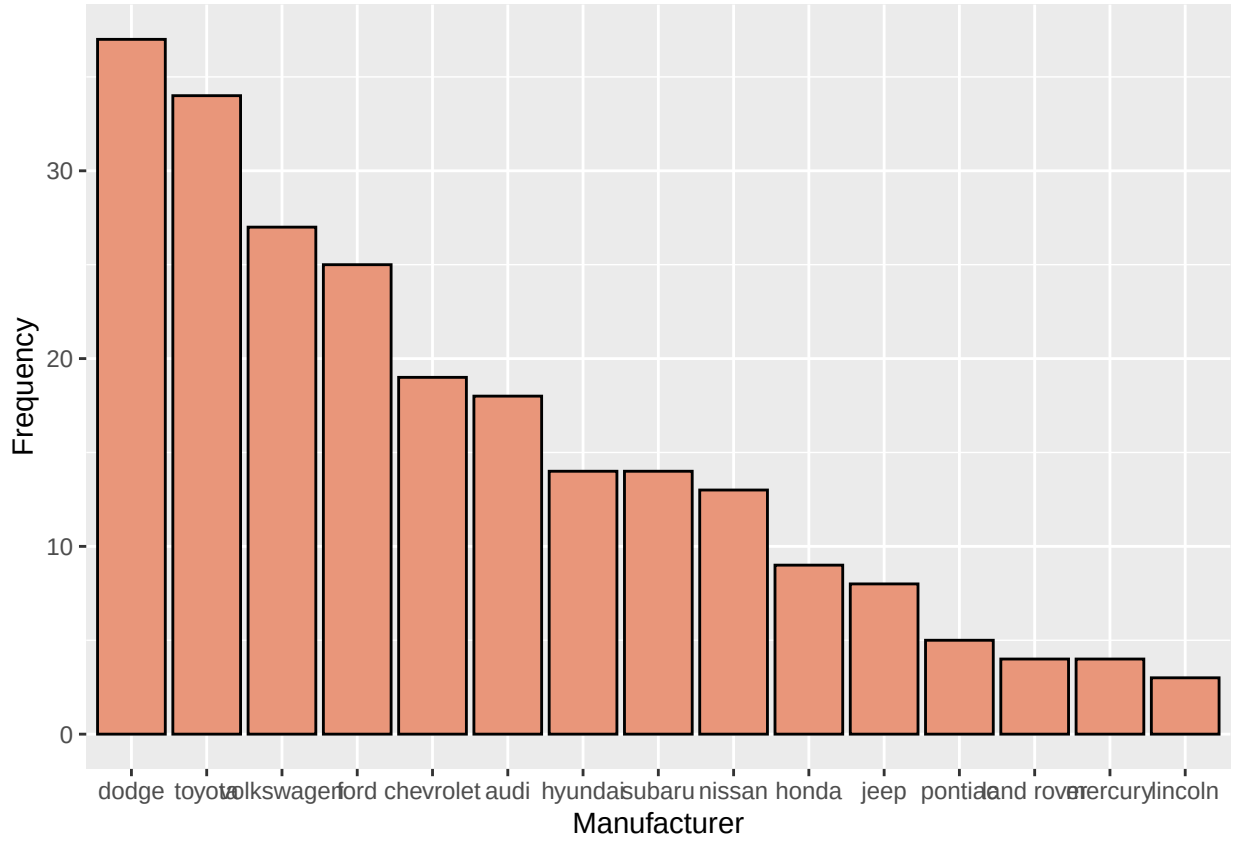
```
make <- table(mpg$manufacturer)
barplot(make, main= "Car Manufacturer Distribution", col= "darksalmon",
        ylim= c(0, 40), cex.names = 0.45 )
axis(side=2, at=seq(5, 40, by=5))
```




```
make1 <- table(mpg$manufacturer)
barplot(make1, main= "Car Manufacturer Distribution", col= "darksalmon",
        xlim= c(0, 40), cex.names = 0.45, horiz= TRUE, las= 1)
axis(side=1, at=seq(5, 40, by=5))
```



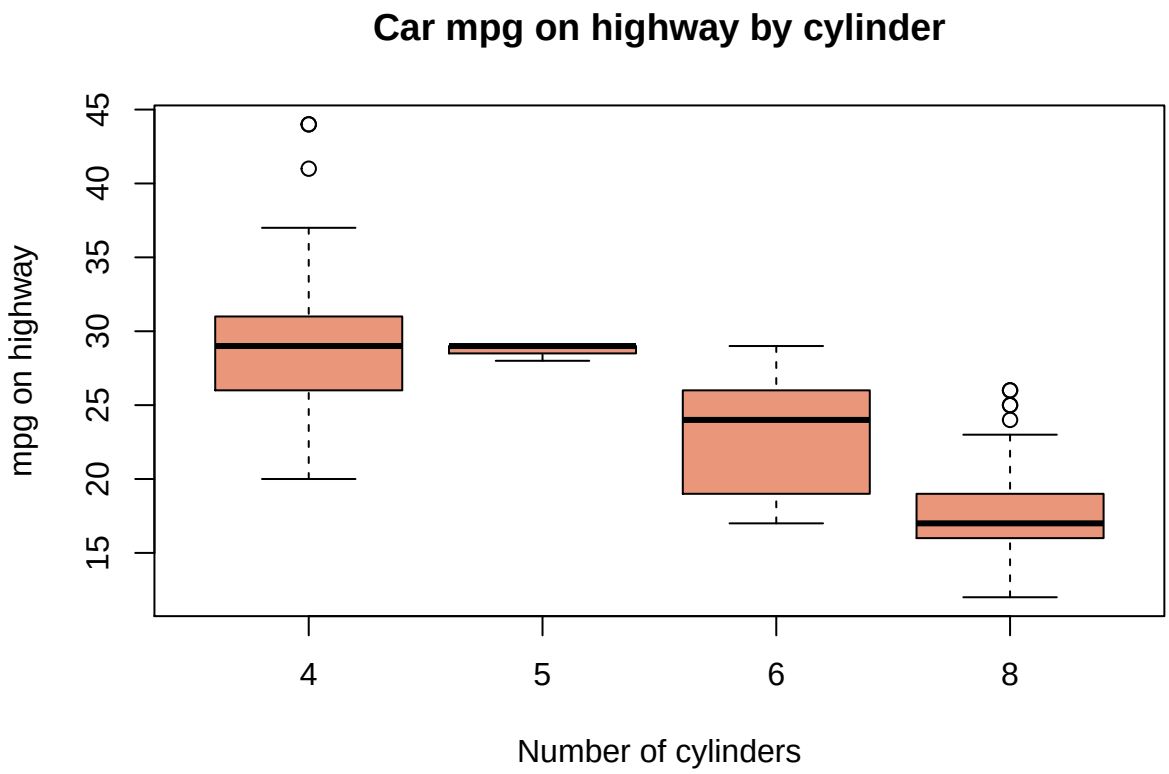

```
make2<-ggplot(mpg, aes(x = reorder(manufacturer, manufacturer,
                                   function(x)-length(x))))+
  labs(y = "Frequency", x = "Manufacturer") +
  geom_bar(fill="darksalmon", colour= "black")

make2
```



The data shows that the manufacturer Dodge produced the most cars. The data shows that the manufacturer Lincoln produced the least cars.

4.

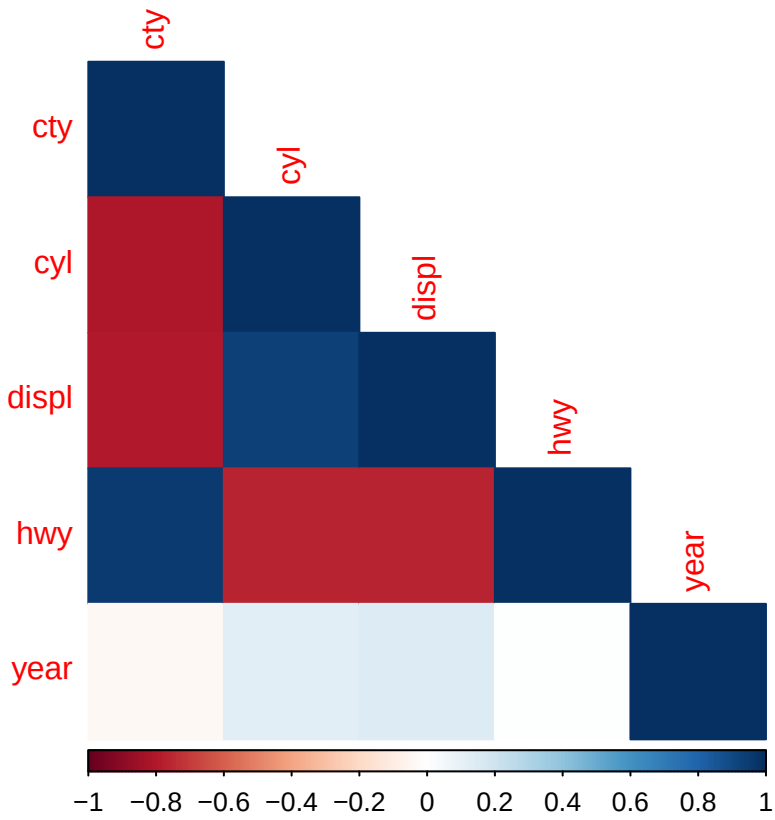
```
boxplot(mpg$hwy ~ mpg$cyl, main= "Car mpg on highway by cylinder",
        ylab= "mpg on highway", xlab= "Number of cylinders", col= "darksalmon")
```



There is an inverse pattern between mpg on the highway and number of cylinders. A car with more cylinders spends more gas, thus their mpg decreases. On the other hand, cars with fewer cylinders have a greater mpg since they don't consume as much gas.

5.

```
library(corrplot)

Matrix <- cor(mpg[,sapply(mpg,is.numeric)], use='complete.obs',
              method= 'pearson')

corrplot(Matrix, method = "color", order = "alphabet", type= "lower")
```



The data shows that there is a positive correlation between hwy-cty and displ-cyl. There is a negative correlation between cyl-cty, displ-cty, hwy-cyl, and hwy-displ.

These relationships make sense to me because a car with less cylinders will achieve a higher mpg in the city and on the highway.

Something surprising is that the car's year doesn't have significant impact on mpg under any factor.