Homework 1

Machine Learning Main Ideas

Question 1: "[S]upervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs" and "unsupervised statistical learning, there are inputs but no supervising output" (from pg. 1 of book)

Question 2: The difference between a regression model and a classification model is that the regression model has a quantitative response, while a classification model has a qualitative response.

Question 3: The two commonly used metrics in regression are Root Mean Squared Error (RMSE) and Mean Absoulte Error (MAE). The two commonly used Performance metrics for classification problems are Accuracy, and Confusion Matrix.

Question 4:

- Descriptive models are used typically to choose model to best visually emphasize a trend in data.
- Inferential models are aimed to test theories and state the relationship between outcome and predictor(s).
- Predictive models are aimed to predict Y with minimum reducible error, and are not focused on hypothesis tests. (from Lecture Slides Day 2)

Question 5:

- Mechanistic assumes a parametric form for f, and you can usually add parameters, which allows more flexibility. Empirically-driven means that there are no assumptions made about f, which requires a large number of observations. They are similar because both can be overfitting for the data modeled, and both are pretty flexible as well.
- In general, empirically-driven models are easier to understand because they are based on observation rather than theory. Sometimes mechanistic models rely on formulas and theories that may not always accurately depict the data being modeled, especially for the real world.

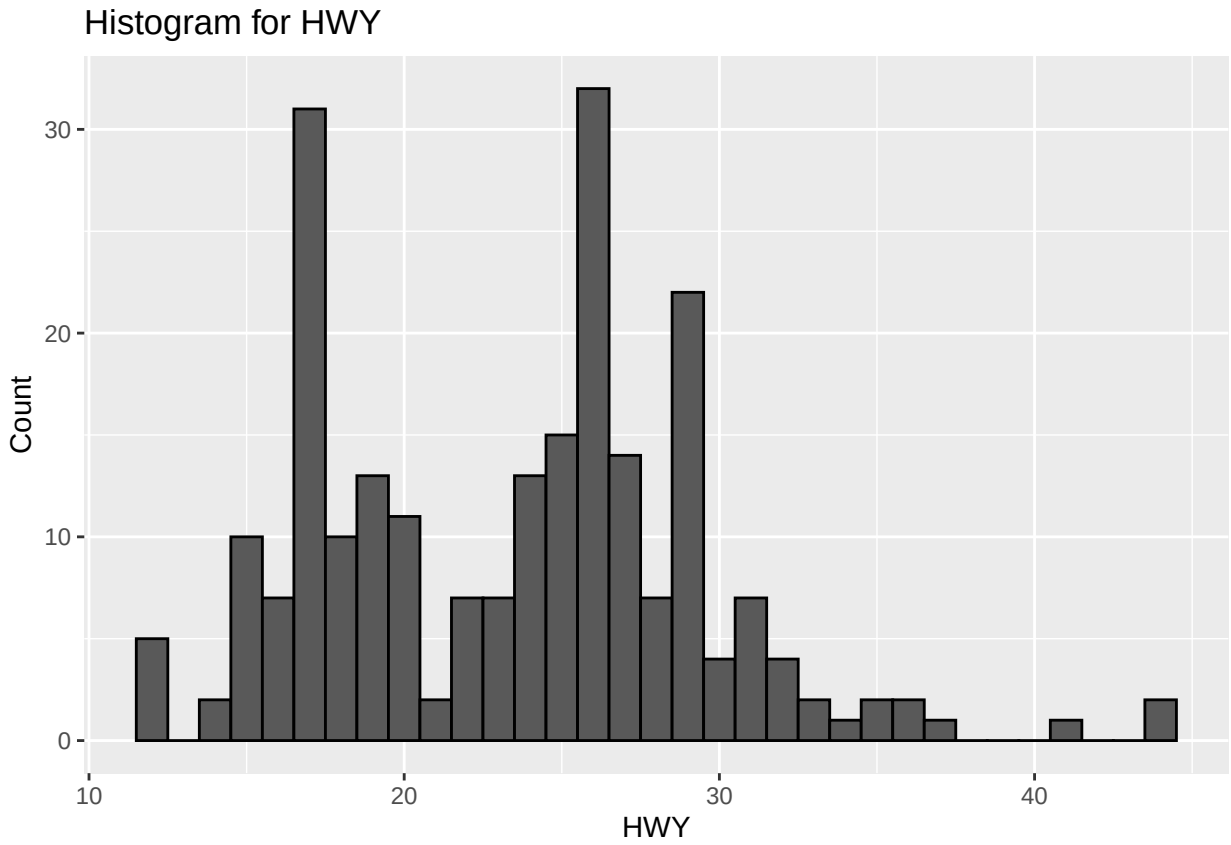
Question 6:

- The first question is inferential because it is using the voter's profile/data to estimate the likelihood that it will vote in favor of the candidate. Also the sampling size is rather large in comparison to the second question. Additionally, the analysis will infer properties from tests and estimates.
- The second question is predictive because it is using a smaller subset of voters to predict whether their personal contact with the candidate would change a voter's likelihood to support a candidate. This analysis focuses more on the past behavior of the candidates to predict their likelihood of support.

Exploratory Data Analysis

```
#install.packages("tidyverse")
library(tidyverse)
library(tidymodels)
library(ISLR)
```

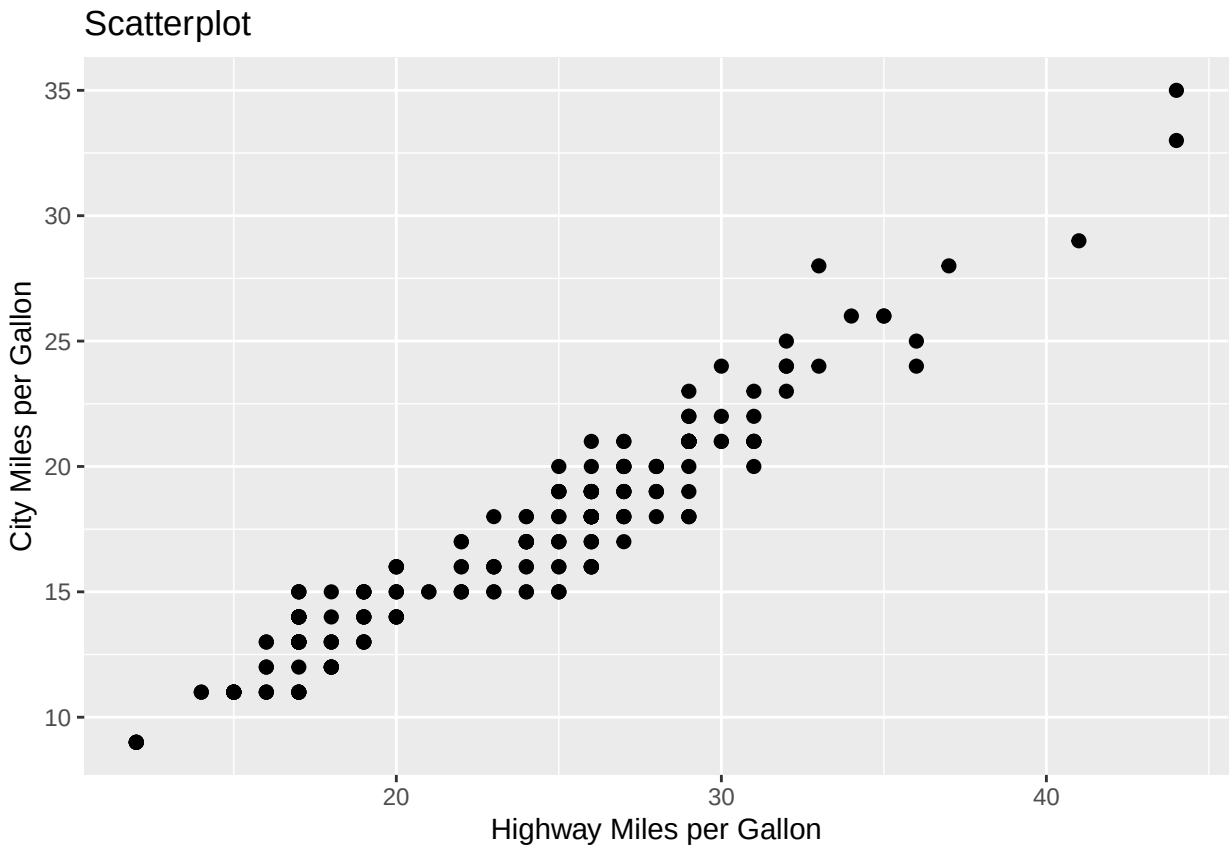
```
# histogram, color
p <- ggplot(mpg, aes(x = hwy)) +
  geom_histogram(color = "black", binwidth = 1) +
  labs(title = "Histogram for HWY", x = "HWY", y = "Count")
p
```



Exercise 1:

In this histogram, it can be seen that the highway miles per gallon is peaked at 17 and 26. It doesn't seem to follow any particular distribution. there are a couple of outliers that exceed 40.

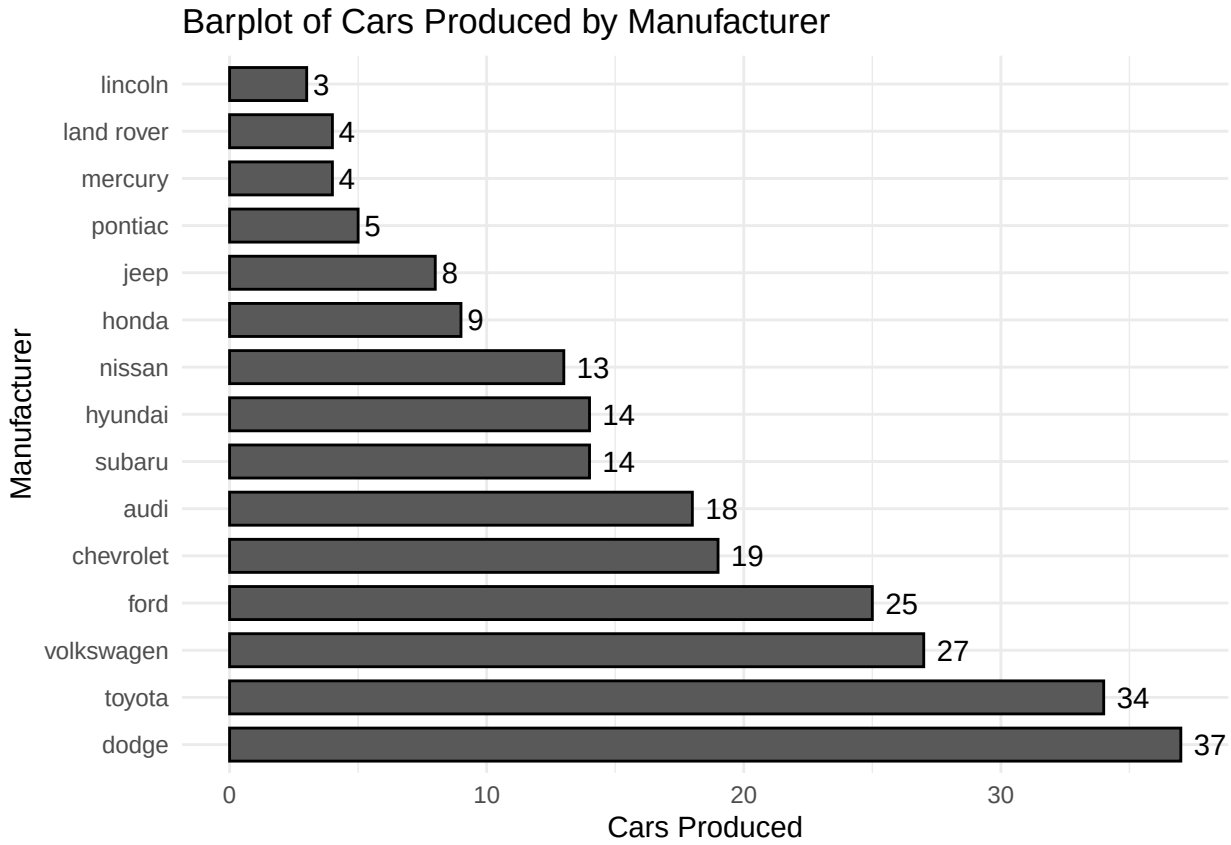
```
ggplot(mpg, aes(x=hwy, y=cty)) +
    geom_point(size=2, color = "black") +
    labs(title = "Scatterplot", x = "Highway Miles per Gallon", y = "City Miles per Gallon")
```



Exercise 2:

There seems to be a linear relation between hwy and cty. As hwy increases cty also seems to increase as well. This means that for the vehicles in this dataset, if the mpg within the city is high, the the mpg for highway is also relatively high, and vice versa.

```
ggplot(mpg, aes(x = factor(manufacturer))) +
  geom_bar(stat = "count", width = 0.7, color = "black") +
  theme_minimal() +
  labs(title = "Barplot of Cars Produced by Manufacturer", x = "Manufacturer",  y = "Cars Produced") +
  scale_x_discrete(limits = c("dodge", "toyota", "volkswagen", "ford", "chevrolet", "audi", "subaru", "
  geom_text(stat = 'count', aes(label = ..count..), hjust = -.4) +
  coord_flip()
```

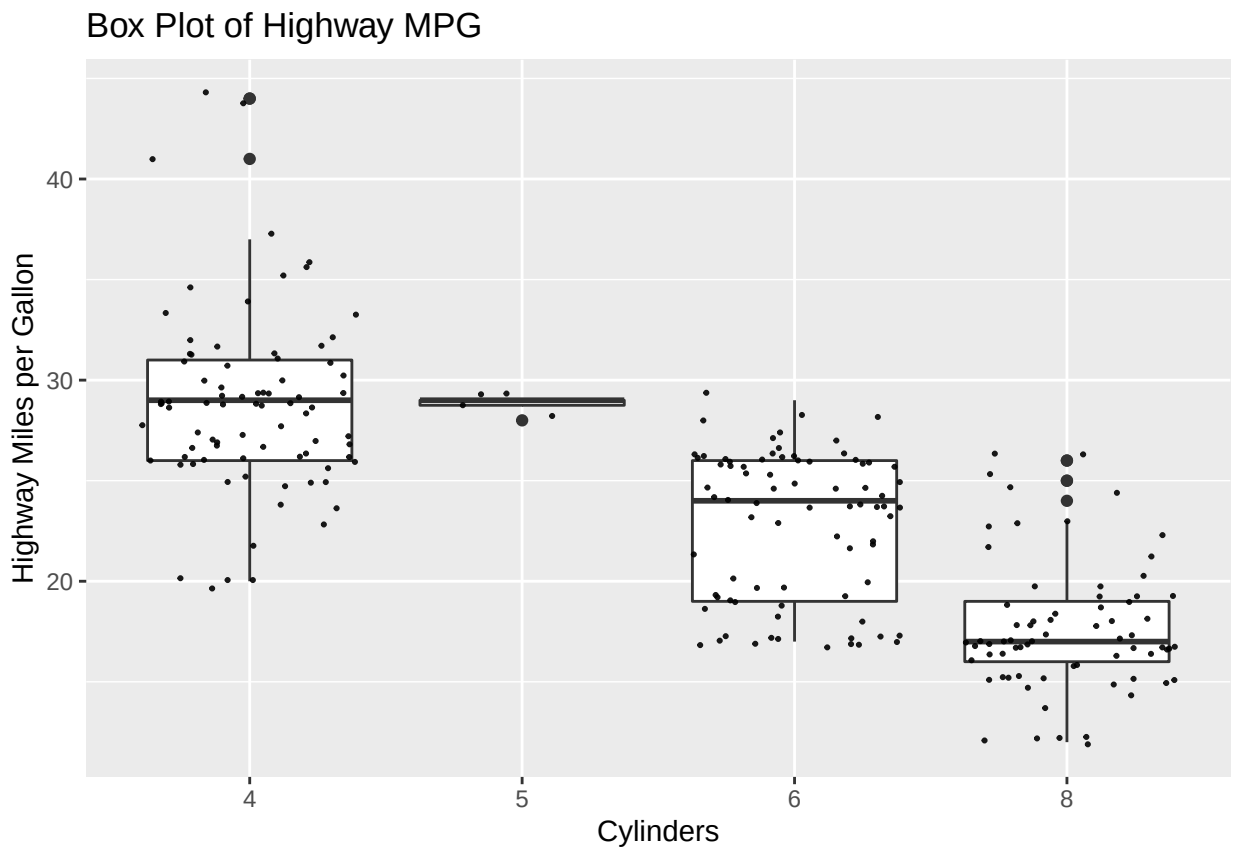


Exercise 3:

The most cars were produced by Dodge with a quantity of 37, and the least cars were produced by Lincoln with a quantity of 3.

Exercise 4: Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?

```
ggplot(mpg, aes(x = as.factor(cyl), y = hwy, group = as.factor(cyl))) +
  geom_boxplot() +
  geom_jitter(color = "black", size = 0.4, alpha = 0.9) +
  labs(x = "Cylinders", y = "Highway Miles per Gallon", title = "Box Plot of Highway MPG" )
```



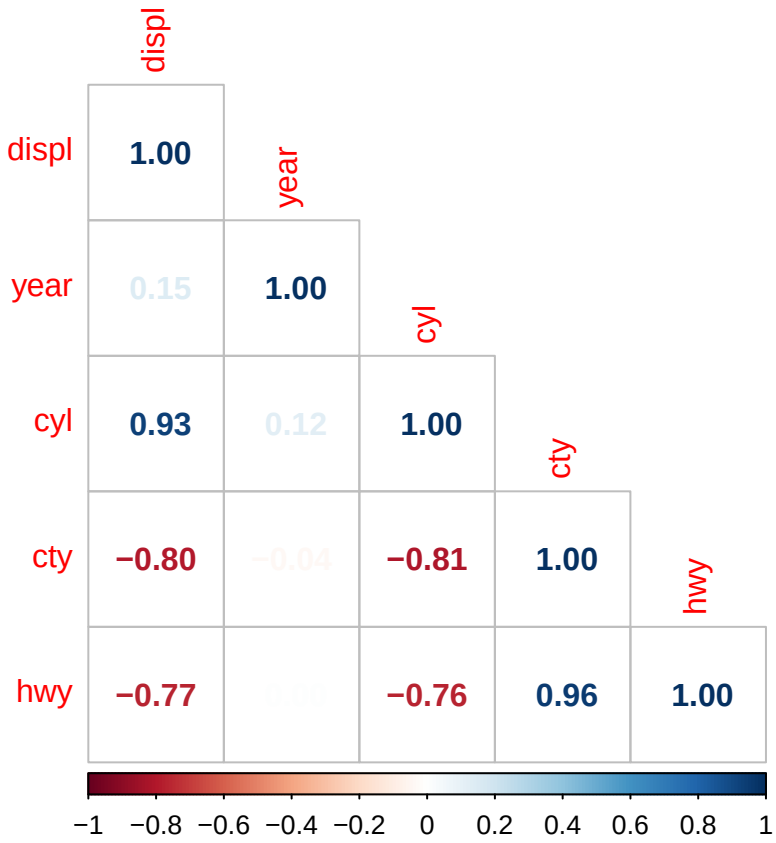
There is a pattern in the the relationship between highway mpg and number of cylinders in the cars. As the number of cylcinders a car has increases, the highway mpg of the car decreases, and vice versa.

```
#install.packages("corrplot")
library(corrplot)
```

Exercise 5:

```
## corrplot 0.92 loaded
```

```
M <- mpg %>%
   select(where(is.numeric)) %>%
   cor() %>%
   corrplot(type = 'lower', diag = TRUE,
            method = 'number')
```



The variables negatively correlated with each other are displ and hwy, displ and cty, cyl and hwy, and cyl and cty. The positively correlated variables are displ and year, cyl and year, and cty and hwy. There are a couple of neutral variables which are year with cty and hwy. For the most part these relationships make sense because it represents how each variable affects the mpg in city and hwy.