Homework 1: PSTAT 131

Sabrina Lem

3/30/2022

***ML Main Ideas* Question 1:**
Supervised: Data used in supervised learning is labeled. As a data scientist, we must use this labeled data where we know both the input and output of the desired model to train the model to best map the given information.

Unsupervised: Unsupervised learning uses unlabeled data. Unsupervised learning investigates patterns based on the given input data.

The main differences between supervised and unsupervised learning is that in supervised learning we know the input and output of labeled data. While, in unsupervised learning you only know the input of unlabeled data. Supervised learning branches into classification and regression problems, and unsupervised learning can be separated into clustering and association problems.

Question 2:
In supervised learning, we can categorize the problems into regression and categorical problems based on the nature of the response variable of the model.
Regression is used when the response variable is quantitative. Classification is used when the response variable is qualitative.

Question 3:
Metrics for regression ML problems: 1. Mean Squared Error (MSE) 2. Mean Absolute Error Metrics for classification ML problems 1.Confusion Matrix 2.Precision

Question 4:
Descriptive Model:A model that best visually reports the trend of the data set. Inferential Model:A model used to "asses the quality of our predictions and (or) estimation" (slide 30, lecture_day1). It is a model that "aims to test theories" and "state the relationship between outcome & predictors" (slide 7 lecture_day2) Prediction model: a model that uses past data on predictors to "accurately predict future response" (slide 30, day_1_131_231). It "aims to predict Y with minimum reducible error" (slide 7, lecture_day2).

Question 5:
Mechanistic means that we "assume a parametric form of the function, f. Empirically-driven means that we do not assume anything about f. Mechanistic model can be made more flexible by adding parameters, however empirically-driven models are flexibly by default due to the fact that the actual data points determine the function f (in an empirically driven model). That said, Empirically-driven models tend to require more observations. Both models are at risk of over fitting. Both models are used to made predictions; they are predictive models.

Empirically-driven models are easier to understand. This is because it makes sense to form a model based on the actual observations, rather than use theory to form a model (in the case of mechanistic models).

Bias-variance trade off occurs when we want to decrease the variance of the model or decrease the bias of a model. Whenever we attempt to lower one of these characteristics, the other will increase. In general, we see that a simple model will have high bias and low variance, and a flexible model will have low bias and high variance. Thus under prediction models, mechanistic or empirically-driven, this bias-variance trade-off comes into play. For mechanistic models, when we add variable to increase model flexibility, the variance will increase as we reduce the bias. Similarly, empirically-driven models may seem easier to understand because they are based off of observation however as a more flexible model, a higher variance will occur in result of the low-bias nature of the model.

Question 6:
The first question is predictive. We are given a voter profile, and we want to predict the outcome of their vote.
The second question is inferential. This question tries to investigate the relationship between voter and their support on a candidate.There is no implication of outcome, rather this question looks into what about a voter's characteristics is important to their candidate support.

Exploratory Data Analysis

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'tidyr' was built under R version 4.0.5

## Warning: package 'dplyr' was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

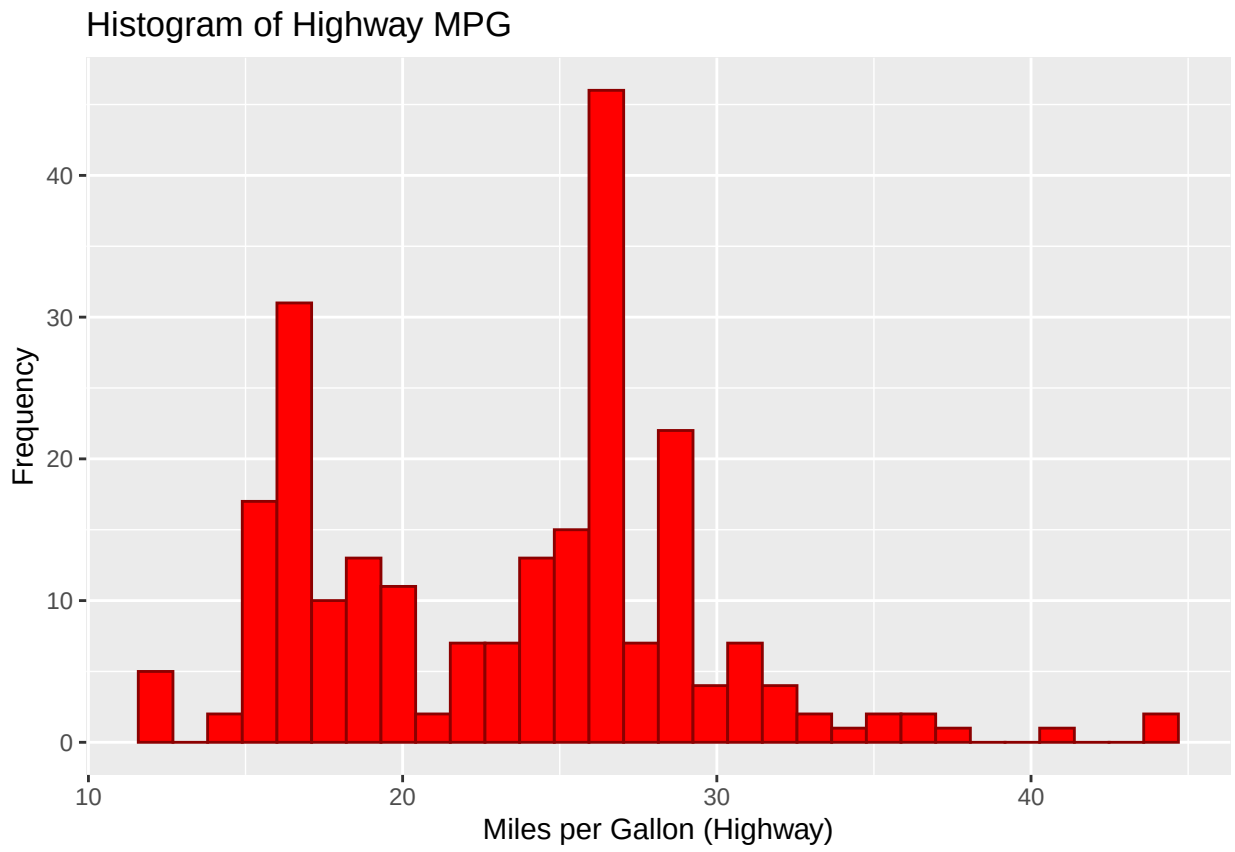
```
library(ggplot2)
head(mpg)
```

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa~
```

Exercise 1:

```
ggplot(mpg, aes(x=hwy))+
  geom_histogram(color="darkred", fill="red")+
  labs(title="Histogram of Highway MPG", x="Miles per Gallon (Highway)",
       y="Frequency")
```

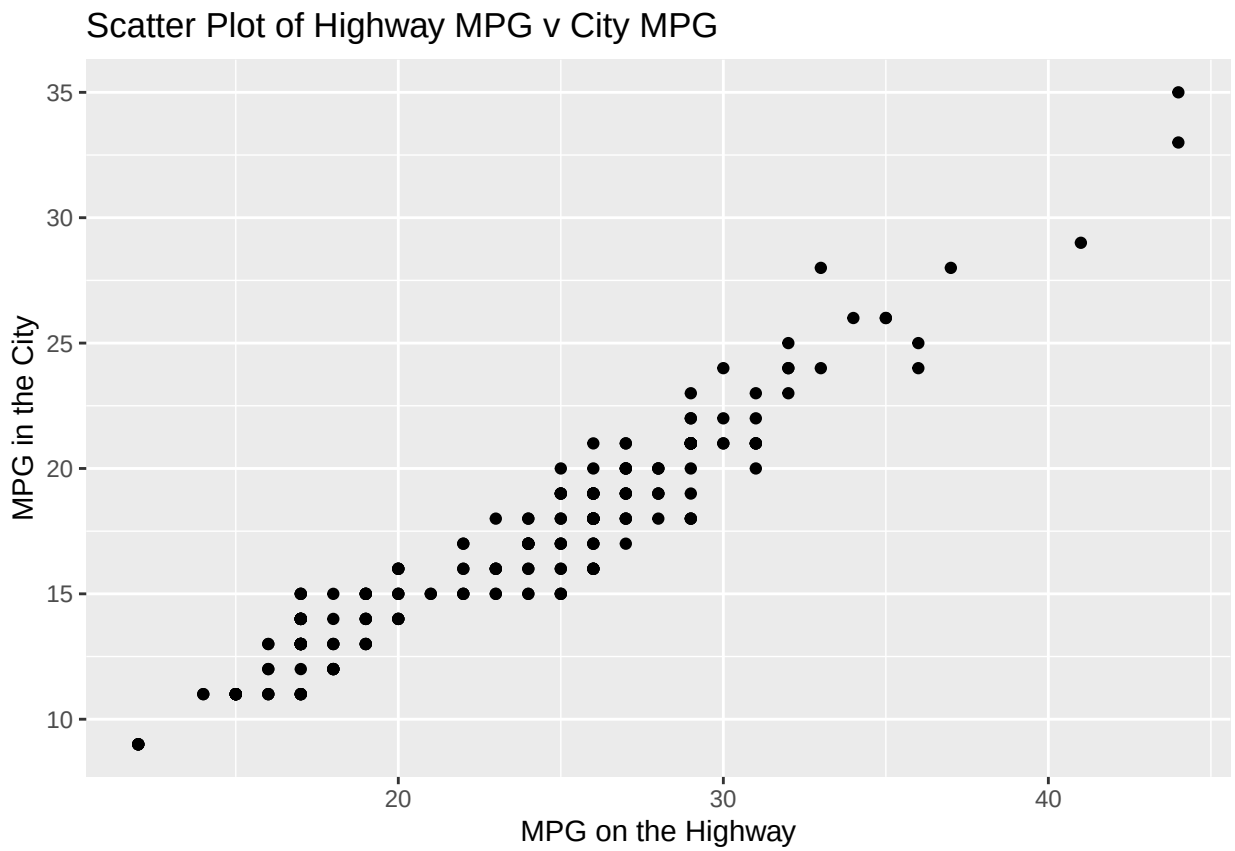
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This spread is somewhat bi-modal. There are peaks at 17 and 27. This implies that there are many car models that tend to have 17 mpg and 27 mpg on the highway. This bi-modal nature may reflect the different sizes of cars like sedans v suvs.

Exercise 2:

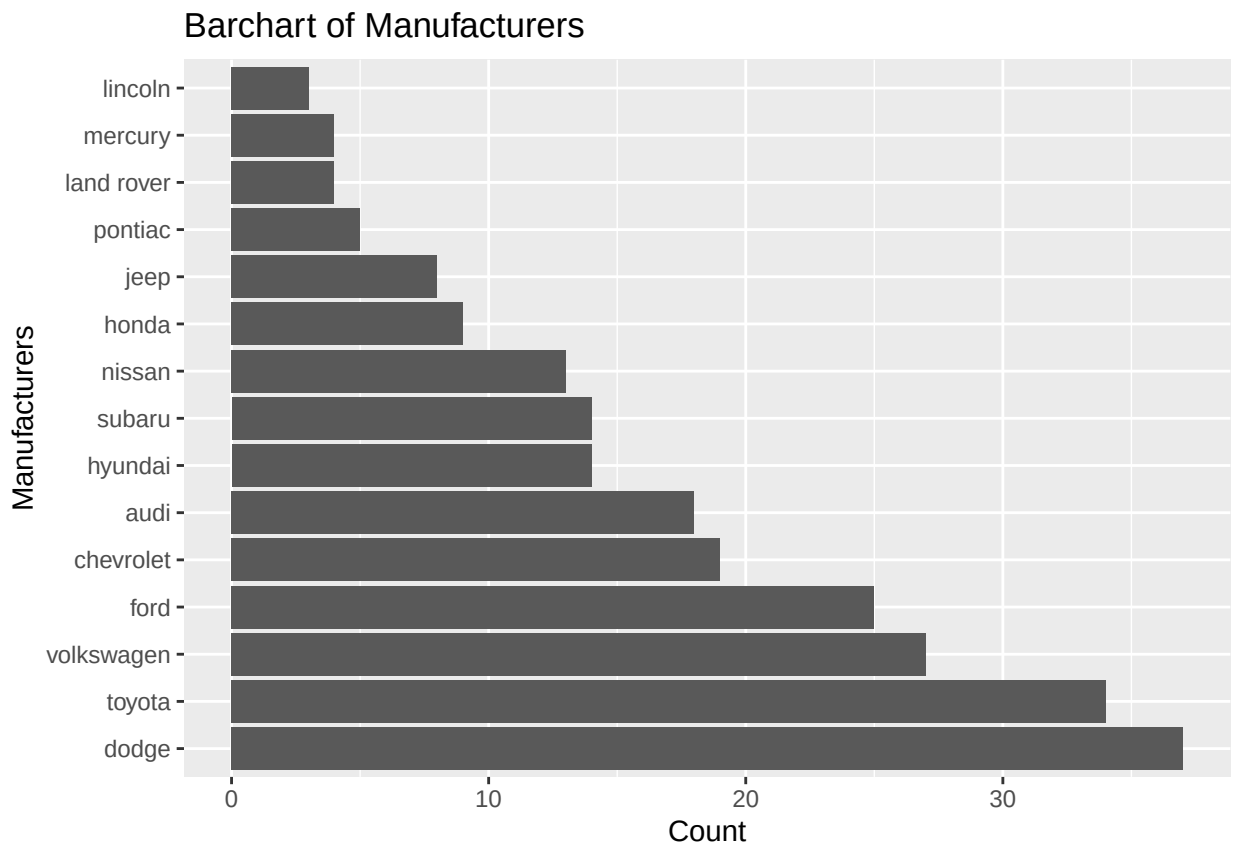
```
ggplot(mpg, aes(x=hwy,y=cty))+
  geom_point()+
  labs(title="Scatter Plot of Highway MPG v City MPG", x="MPG on the Highway",
      y= "MPG in the City")
```



There is a clear positive relationship between Highway MPG and City MPG. Cars that tend to have better highway MPG also have better city MPG. This makes sense because a better car, in general, will have better milleage on both the highway and in the city.

Exercise 3: https://stackoverflow.com/a/9231857 (for re-ordering bars)

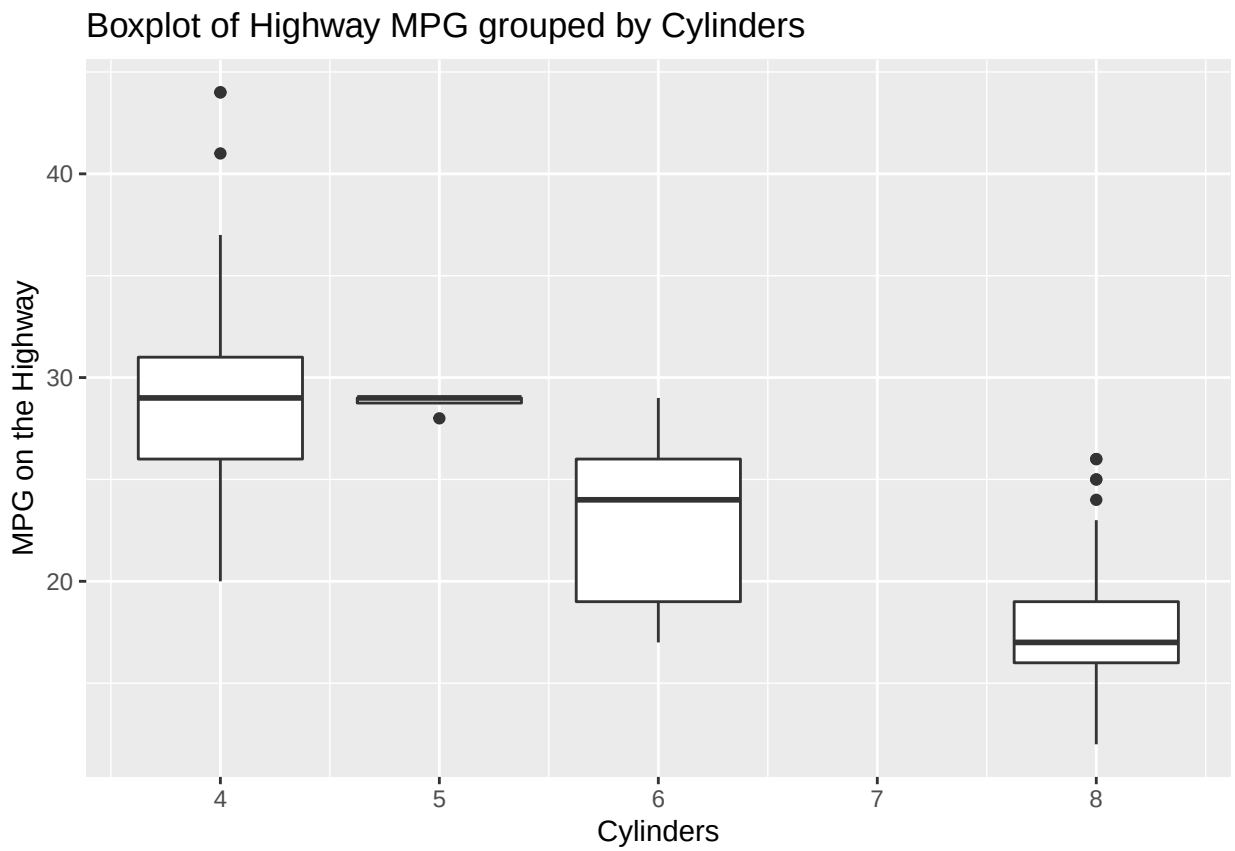
```
ggplot(mpg, aes(x=reorder(manufacturer, manufacturer,
                          function(x)-length(x))))+
  geom_bar()+
  coord_flip()+
  labs(title="Barchart of Manufacturers", x="Manufacturers",
         y="Count")
```



Dodge produced the most.
Lincoln produced the least.

Exercise 4:

```
ggplot(mpg, aes(group= cyl, x=cyl, y=hwy))+
  geom_boxplot()+
  labs(title="Boxplot of Highway MPG grouped by Cylinders", x="Cylinders",
       y="MPG on the Highway")
```



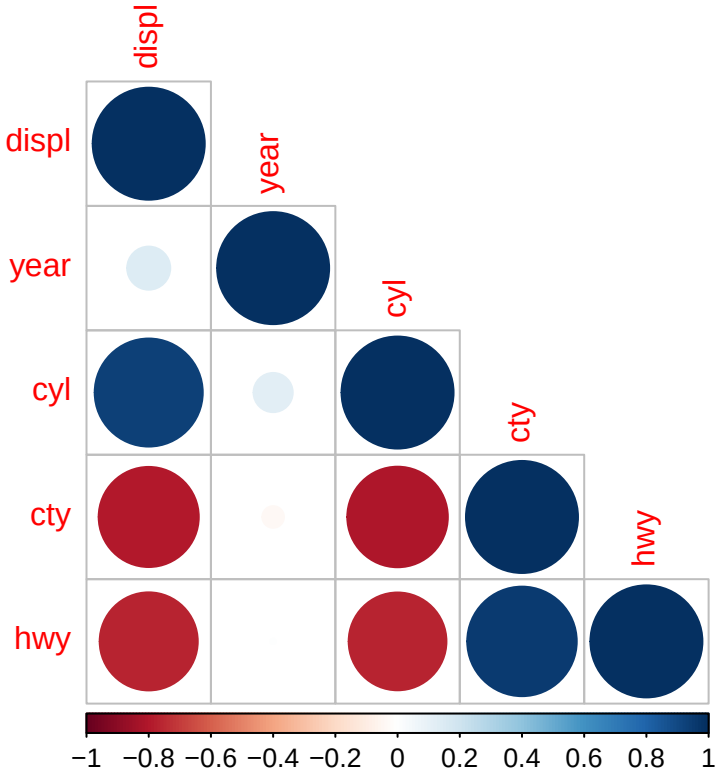
We can see from the boxplots, that as the number of cylinders increase, the average highway MPG tends to decreases. Which makes sense: Generally more cylinders means more power but less fuel efficiency.

Exercise 5:

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
keeps<-c("displ", "year", "cyl", "cty", "hwy")
mpg_num <-mpg[keeps]
M<-cor(mpg_num)
corrplot(M, type ="lower")
```



The positively correlated pairs are year and displacement, number of cylinder and displacement, number of cylinders and year, city mpg and highway mpg. Some of these make sense. It makes sense that engine displacement will improve over time as innovation occurs. The positive correlation between cylinder count and displacement makes sense because both contribute the the overall power of the engine. And as mentioned in the previous question, it makes sense for highway mpg and city mpg to be positively correlated because a better car will have better mileage on both the highway and in the city overall (and vice versa). It surprises me that the number of cylinders and year is positively correlated. To my limited car knowledge cylinder count is vehicle-model specific. That said, in any given year the number of cylinders var based on the car and have nothing to do with what year the car is made in (especially considering the time frame of this data). The negatively correlated pairs are city mpg and displacement, highway mpg and displacement, city mpg and number of cylinders, highway mpg and displacement, and city mpg and year. the negative correlations between the mileage variables and cylinder variable makes sense. A vehicle with less cylinders gets better fuel efficiency. Similarly, the negative correlation between the mileage variables and displacement make sense because lower displacement uses less power and therefore is more fuel efficient. The negative correlation between city mpg and year surprises me. I would suspect a positive correlation between these two variables. It would seem that mpg in the city would improve overtime with innovation. But this data indicates the opposite.