HW1

Question 1

Supervised learning: consists of a target / outcome variable which is to be predicted from a given set of predictors;generates a function that map inputs to desired outputs, and training process continues until the model achieves a desired level of accuracy on the training data

Unsupervised learning:assumes that the input dataset is distributed according to some unknown underlying statistical distribution; attempts to find clusters or groups in the data without any labels for what constitutes a “cluster” or “group”

Main difference:supervised learning uses labeled input and output data, while unsupervised does not

Question 2

In the context of machine learning, classification models are used to predict a discrete class label, while regression models are tasked with predicting a continuous quantity.

Question 3

Regression: Mean Squared Error and Adjusted R-Squared Classification: Accuracy and Confusion Matrix

Question 4

Descriptive:best visually emphasizes a trend in data, such as using a line on a scatterplot

Predictive:looks for which combination of features fit best; goal to predict Y with minimum reducible error rather than focusing on hypothesis tests

Inferential: searches for significant features; test theories and possible claims, and state relationship between outcome & predictor

Question 5

Mechanistic: relating to the usage of a theory in order predict what will happen

Empirically-driven: relating to learning by experiment and observation rather than theory

The two models differ in the sense that mechanistic assumes a parametric form of f while empirically-driven does not. The latter is also more flexible by default, and requires a larger number of observations. However, they are similar in the fact that both models can lead to overfitting because of the number of parameters.

In general, I believe that neither model is more easier to understand than the other because it depends on the situation. In some cases, a mechanistic model may be easier to interpret because only a few input data

points are required for a given prediction, rather than needing a large number of observations depending on the number of variables included for empirical. However, other times, an empirical model may be easier because you just have a few concerns, and none of them require lengthy, involved computation.

Bias-variance tradeoff is related to the use of mechanistic or empirically-driven models by ensuring that the most appropriate models are selected based on the sample data. It is used to imply that a model should balance underfitting and overfitting.

Question 6

Question 1 would be considered predictive because we are estimating the value of the response variable (voting results) based on predictor or input variables (a voter's profile/data).

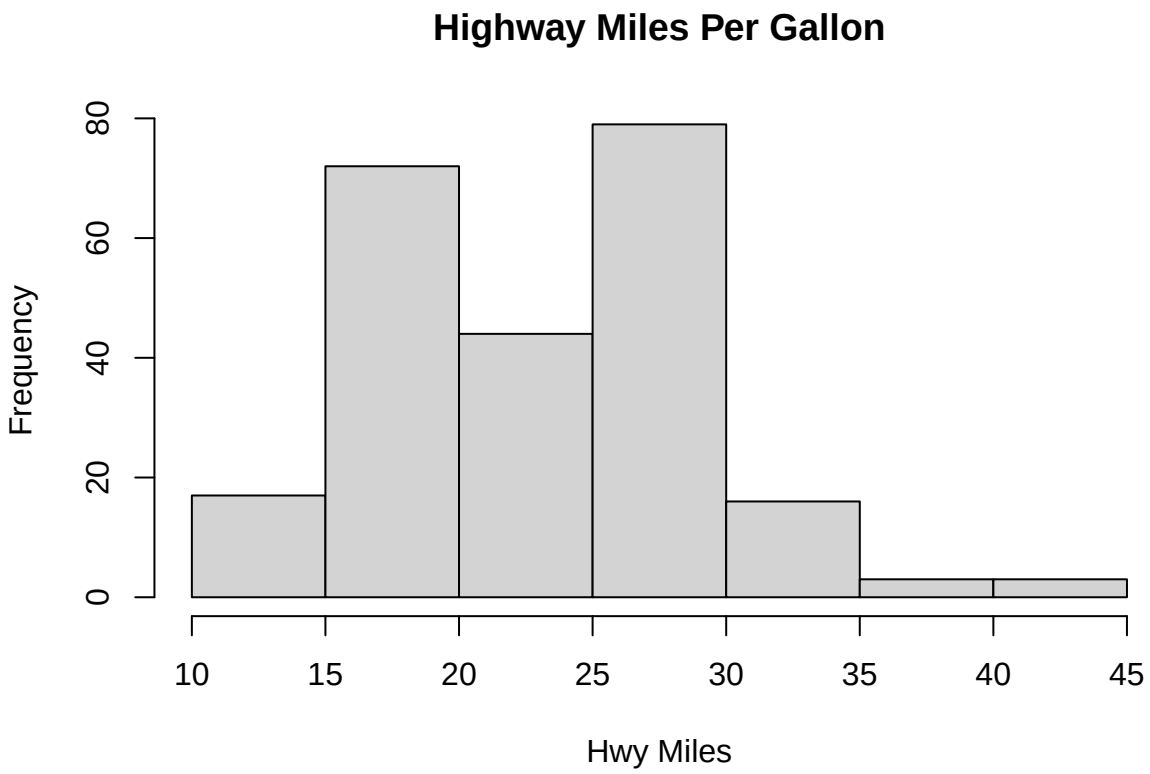
Question 2 would be considered inferential because now we are trying to understand the relationship between the response and the predictor variables. We are interested in understanding how one's personal contact with the candidate could impact their support for them.

Exercise 1

```
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   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

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

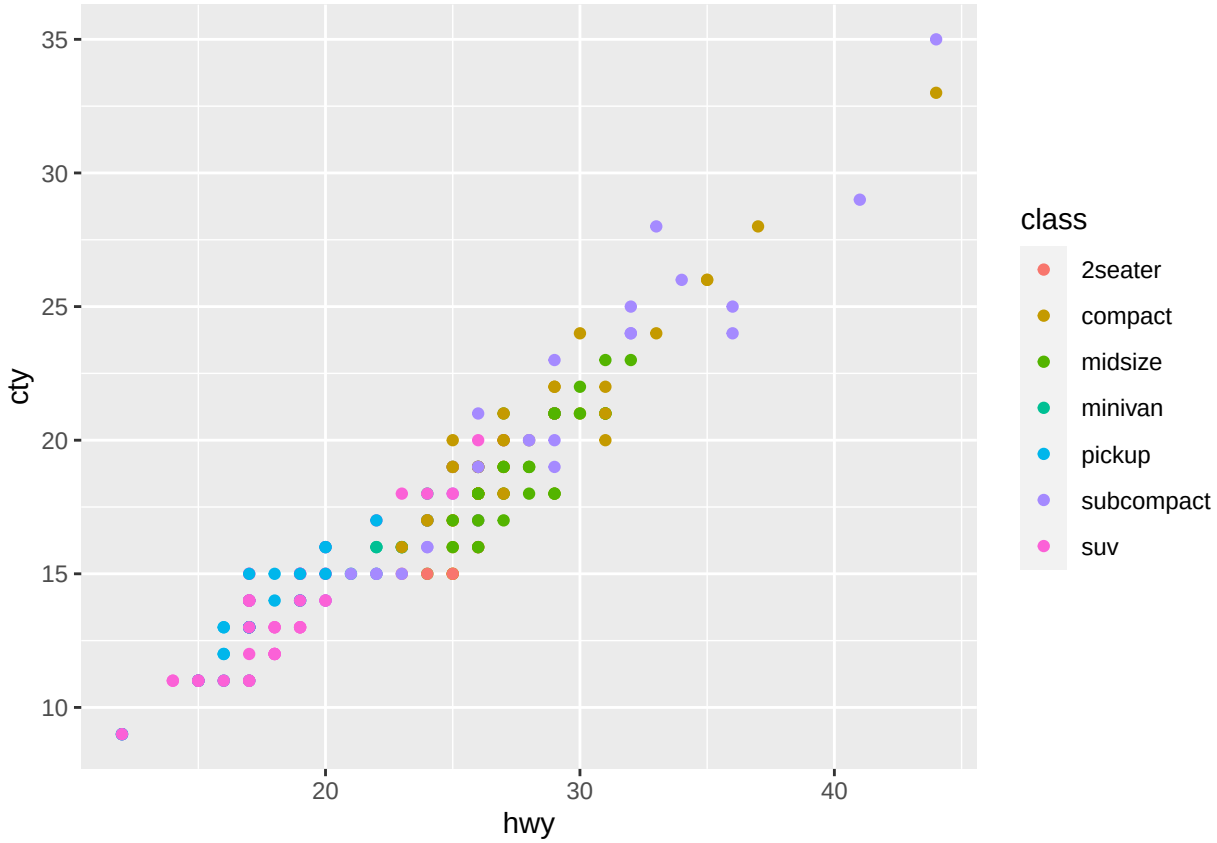
```
hist(mpg$hwy,main = "Highway Miles Per Gallon",xlab = "Hwy Miles")
```



In the histogram, we can see there is a bi-modal distribution that is slightly right-skewed. The center is also around 23, while the spread is from 10 to 45, with a range about 35.

Exercise 2

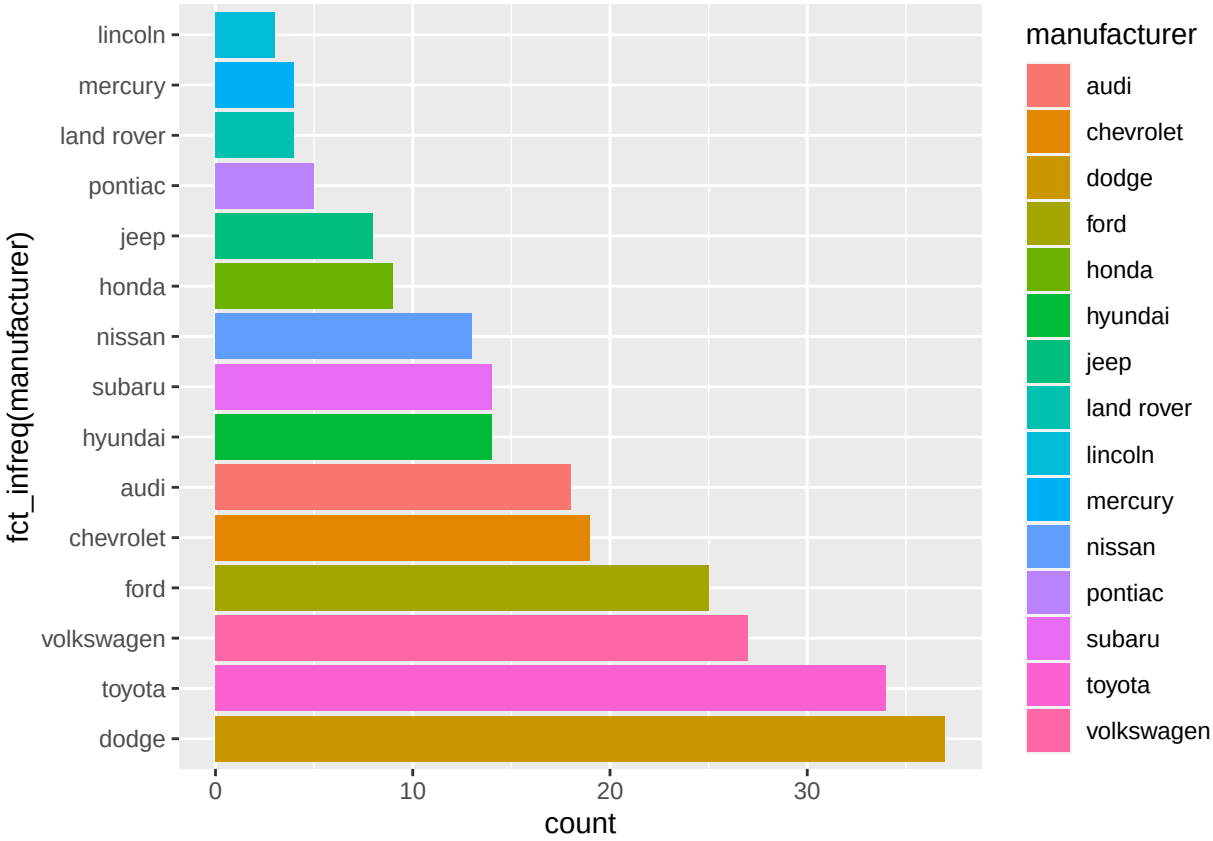
```
library(ggplot2)
ggplot(mpg, aes(hwy,cty, color = class)) + geom_point()
```



Yes, there seems to be a relationship between hwy and cty as a positive correlation appears in the scatterplot. This indicates that when the highway mileage increases, so does the city mileage.

Exercise 3

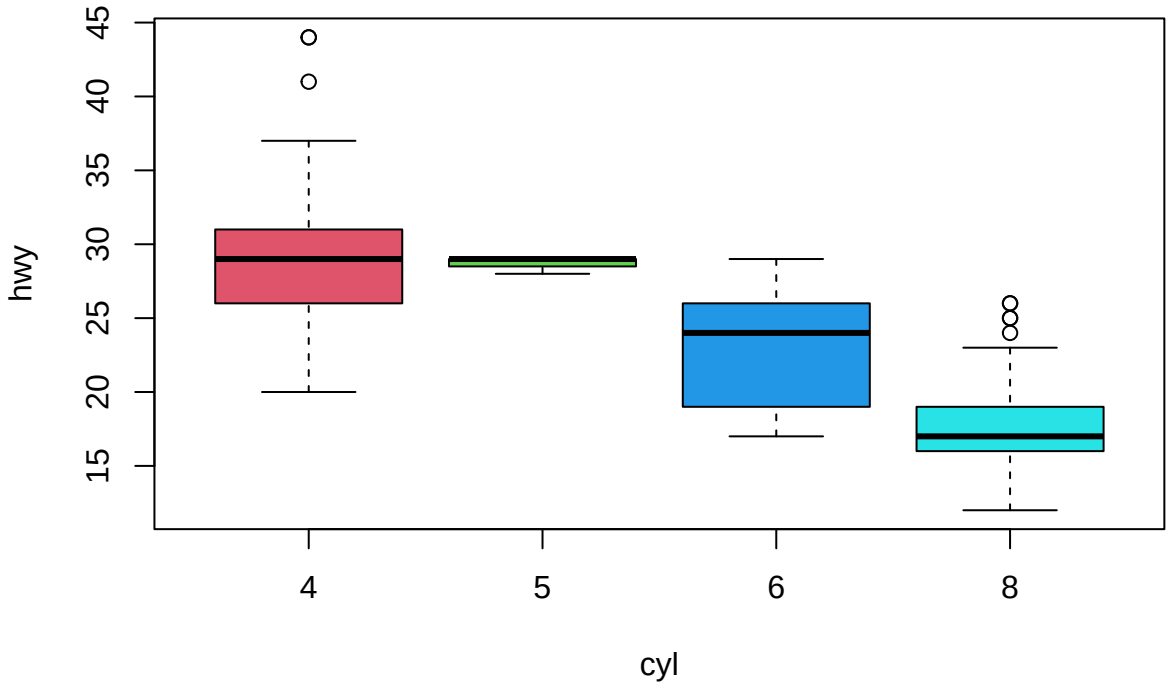
```
library(forcats)
library(ggthemes)
ggplot(mpg,aes(x = fct_infreq(manufacturer), fill = manufacturer)) +
    geom_bar(stat = 'count') +
  coord_flip()
```



The dodge produced the most cars, while the lincoln produced the least.

Exercise 4

```
boxplot(hwy ~ cyl, data = mpg, col = c("2", "3", "4","5"))
```



The higher the value of cyl, the lower the value of hwy. The lower the value of cyl, the higher the value of hwy. The lowest and highest values of cyl have outliers.

Exercise 5

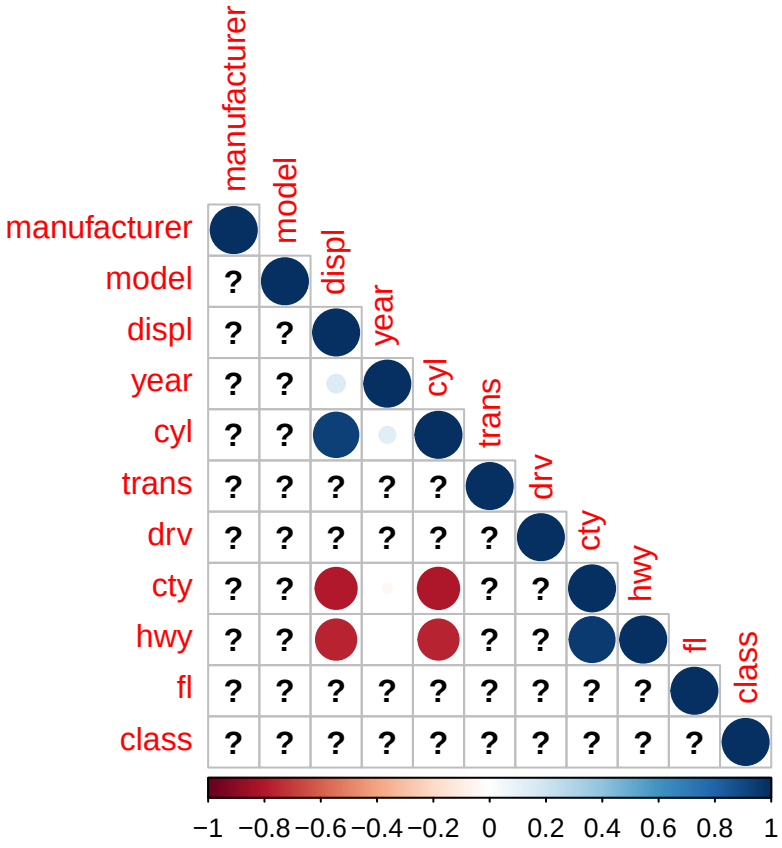
```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
mpg[] <- lapply(mpg, as.numeric)
```

```
## Warning in lapply(mpg, as.numeric): NAs introduced by coercion

## Warning in lapply(mpg, as.numeric): NAs introduced by coercion

## Warning in lapply(mpg, as.numeric): NAs introduced by coercion

## Warning in lapply(mpg, as.numeric): NAs introduced by coercion

## Warning in lapply(mpg, as.numeric): NAs introduced by coercion

## Warning in lapply(mpg, as.numeric): NAs introduced by coercion
```

```
M = cor(mpg)
corrplot(M, type = "lower")
```



Strong negative correlation: Cty and displ, cty and cyl, hwy and displ, hwy and cyl

Strong positive correlation: cyl and displ, hwy and cty

Weak positive correlation: year and displ, cyl and year

The positive correlation pairs make sense to me, but the negative correlations are rather surprising to me. However, I personally do not know much about cars, so my interpretation of the roles of these variables in cars may be completely wrong.