# hw-1-code

Shiloh Rivers

4/2/2022

## Question 1

The difference between supervised and unsupervised data is that all data sets are labeled in supervised data while this is not the case with unsupervised data. Supervised data uses classification and regression while unsupervised uses clustering, association, and dimensionality reduction.

## Question 2

Classification has an output that is categorical while regression has an output that is quantitative.

## Question 3

2 metrics commonly used for classification models are accuracy and precision. 2 metrics commonly used for regression models are mean squared error and mean absolute error.

## Question 4

Descriptive models: Create a visual emphasis of patterns in data. Inferential models: Test a theory and determine cause and effect between outcome and predictors. Predictive models: Predict new data using the patterns found in the old data.

## Question 5

Mechanistic models use a provided theory and fit the data using the model we already have in mind. We assume a function f. Requires less data. Empirical models take the data and allow the computer to create/find a model based on the data. We don't assume a function f. Requires more data. A mechanistic model is easier to understand because we have created a theory and we can explain the theory we created to another person. Empirical models were not created by a "black box". Empirical models will tend to overfit data thereby having a low bias but high variance: not necessarily being able to predict future data well. Mechanical models are more controllable in this manner. We can use less parameters to increase bias and lower variance.

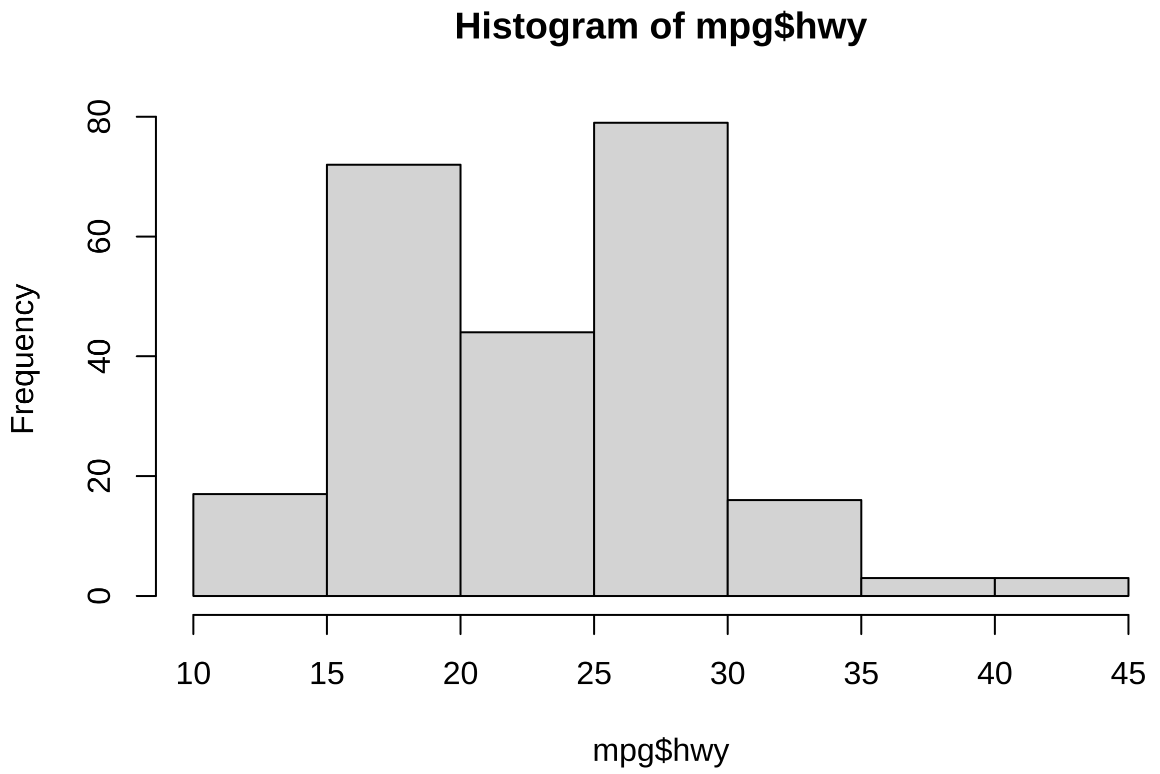
## Question 6

Given a voter's profile/data, how likely is it that they will vote in favor of the candidate? Inferential. We are determining if our theory about a connection between predictors and outcomes is plausible.

How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate? Predictive. We are predicting a future event.

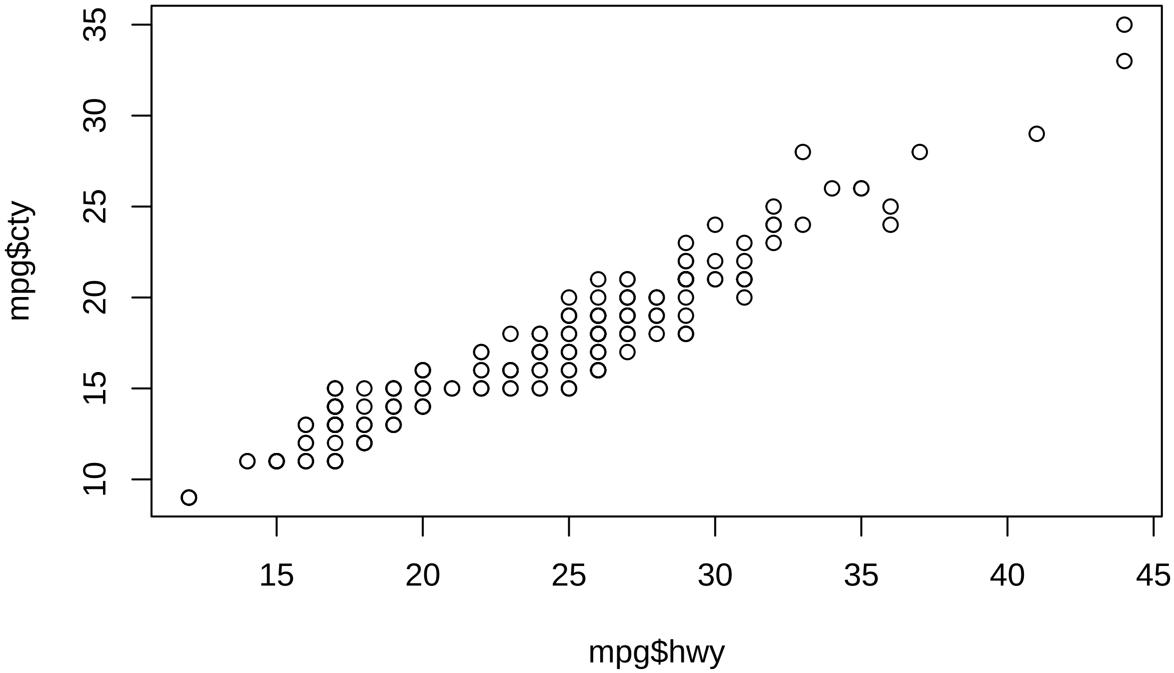
### Excercise 1

```
## -- 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   2.1.2     v forcats 0.5.1

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



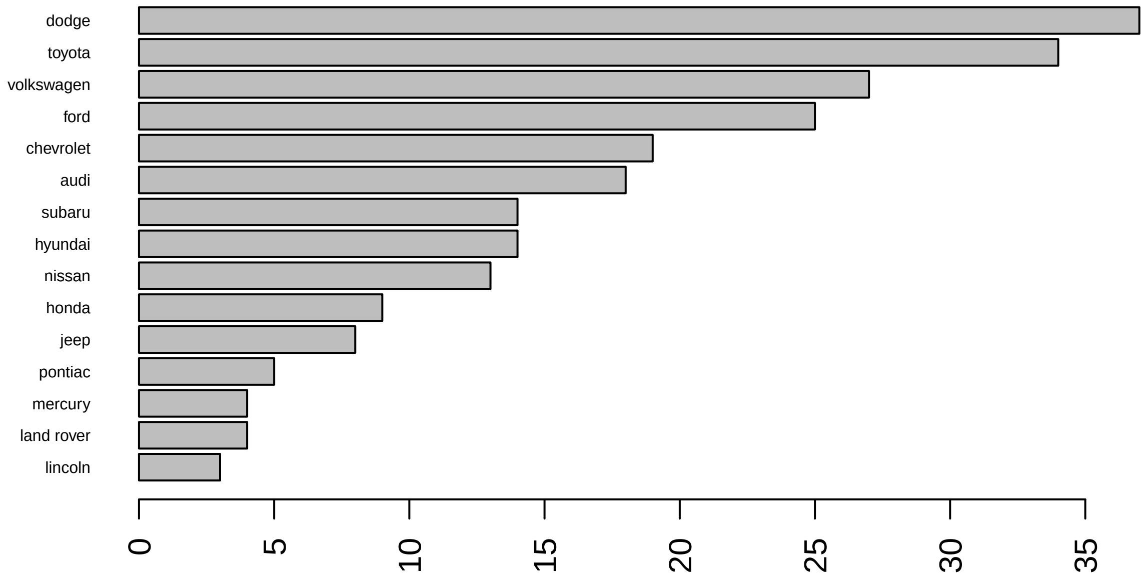
I can see that highway mileage seems to be bimodal with peaks between 15-20 and 25-30 with a long tail up to 45 mpg.

**Excercise 2**



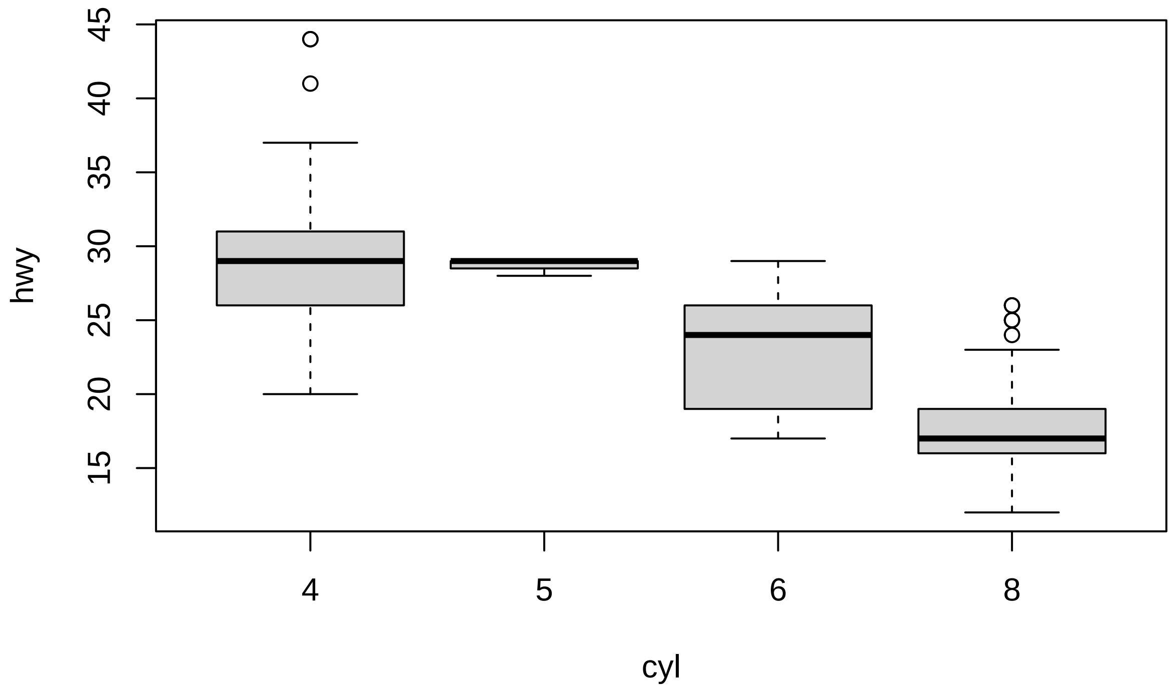
There is a positive trend between highway mpg and city mpg which makes sense. If a motor is efficient on city streets it will also tend to be efficient on the freeway.

## Excercise 3



Dodge produced the most cars while lincoln produced the least.

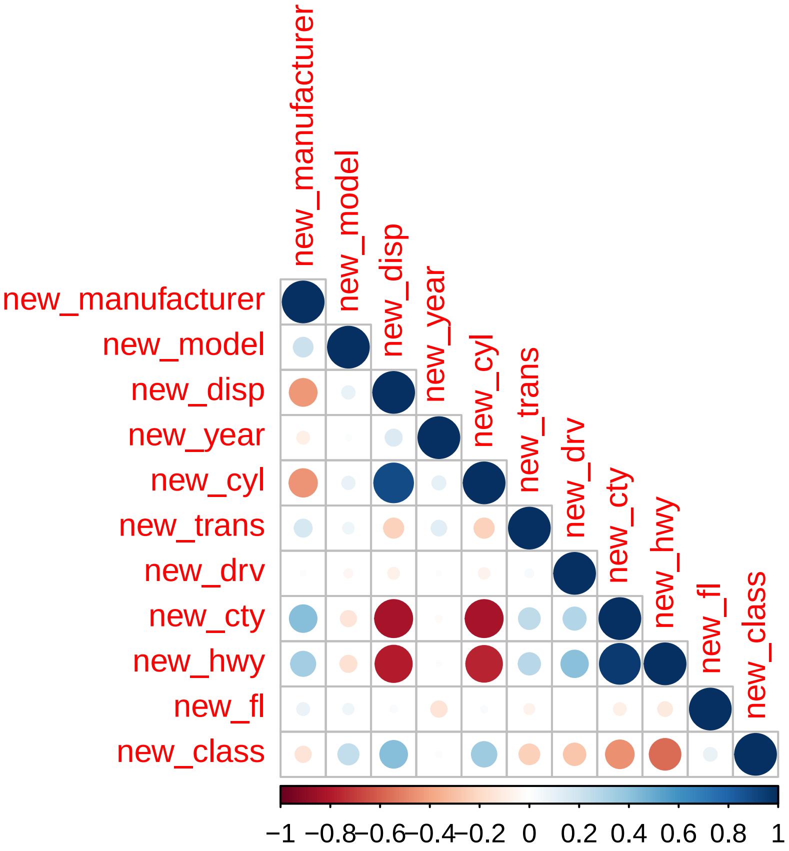
**Excercise 4**



I see that the more cylinders a car has it will generally have less highway miles per gallon.

### Excercise 5

```
## corrplot 0.92 loaded
```



City and highway mileage are strongly uncorrelated with engine displacement and number of cylinders. Highway mileage and City mileage are strongly correlated. Number of cylinders and displacement are correlated. Class of car and highway mileage are negatively correlated which may be because of my method of cleaning the data. However, it does make sense that there would be a pattern between class of car and mileage. Same with manufacturer and (mileage(both cty and hwy), disp, and cyl.)