131_HW1

Immad Ali

4/3/2022

Q1

In unsupervised learning, we do not know what the output observed by the model. In supervised learning, we do know what the output observed by the model is. The biggest difference between the 2 learning styles is the fact that we basically can double check our work in supervised learning since we have a observed output.

#Q2

Regression models can predict continuous values, such as numbers. In contrast, classification models are used to predict categorical values.

#Q3

Was not covered.

#Q4

Descriptive: Models that emphasize trends in the dataset, usually visually.

Inferential: Inferential models ask what features are significant and state the relationship between the predictors and outcomes.

Predictive: Predictive models seek to find a good combination of predictors while predicting Y with the minimum reducible errors.

#Q5

Mechanistic: We assume a parametric form for our model and then add parameters to fit our model more accurately.

Empirically-Driven: Empirically driven has no assumptions about our model, F.

Difference/Similarities: Empirically driven models require a larger number of N observations for it to work accurately, but are also more flexible by default.

Which is easier to understand: I would say mechanistic models are easier to understand because they require less N observations and can fit a simpler model.

Describe how the bias-variance trade-off is related to the use of mechanistic or empirically-driven models. Did not cover

---

#Q6

Given a voter’s profile/data, how likely is it that they will vote in favor of the candidate?:

This type of question would be predictive as we are trying to predict how likely the event is using the voters information.

How would a voter’s likelihood of support for the candidate change if they had personal contact with the candidate?

This type of question would be inferential because we are looking to see if there is a relationship/association between whether the voter personally knows the candidate and if their support for them would change.

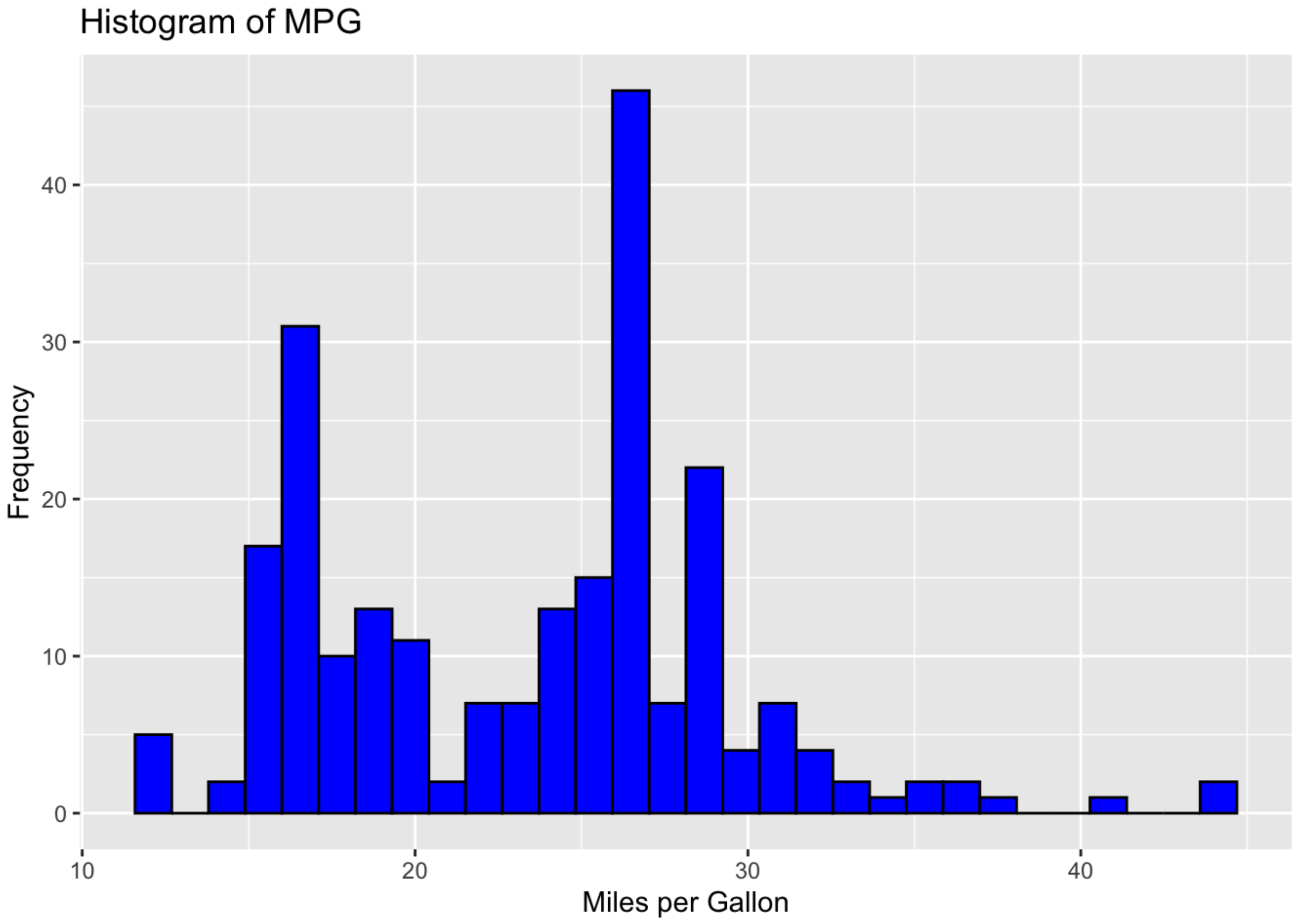
Exploratory Data Analysis

Exercise 1

```
hwyhist <- ggplot(mpg, aes(x = hwy)) + geom_histogram(fill = "blue", color = "black") +
  labs(title = "Histogram of MPG", x = "Miles per Gallon", y = "Frequency")

hwyhist
```

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

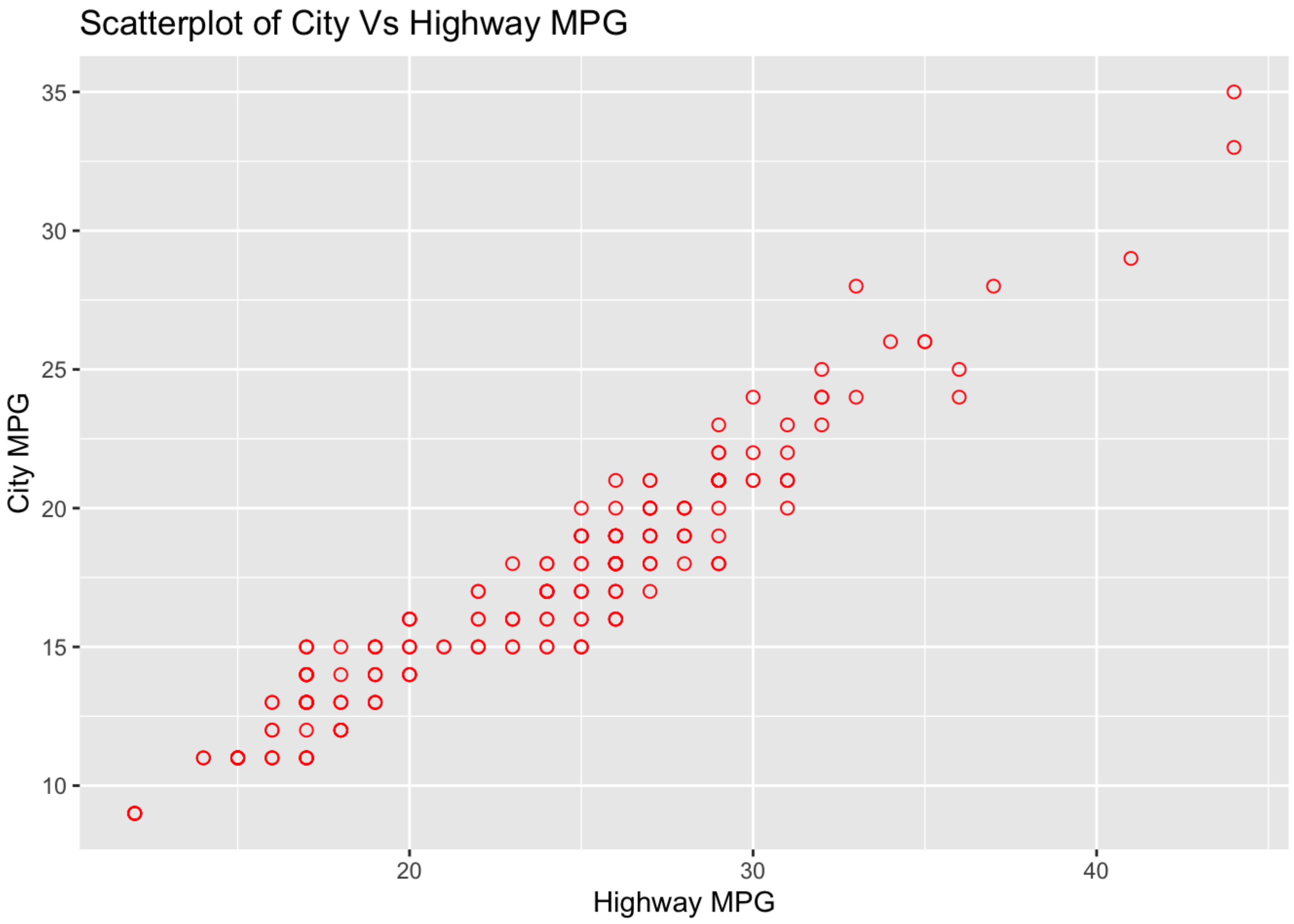


We can see the histogram is skewed slight to the left and that the maximum frequency of similar MPG reportings is around 26/27 MPG.

Exercise 2

```
hwyscatter <- ggplot(mpg, aes(x = hwy, y = cty)) + geom_point(size = 2, shape = 1, color = "red") +
  labs(title = "Scatterplot of City Vs Highway MPG", x = "Highway MPG", y = "City MPG")

hwyscatter
```



We can observe a positive correlation between highway and city MPG of cars. We interpret this as the higher a cars highway MPG is, the more likely it is the car's city MPG will also be higher.

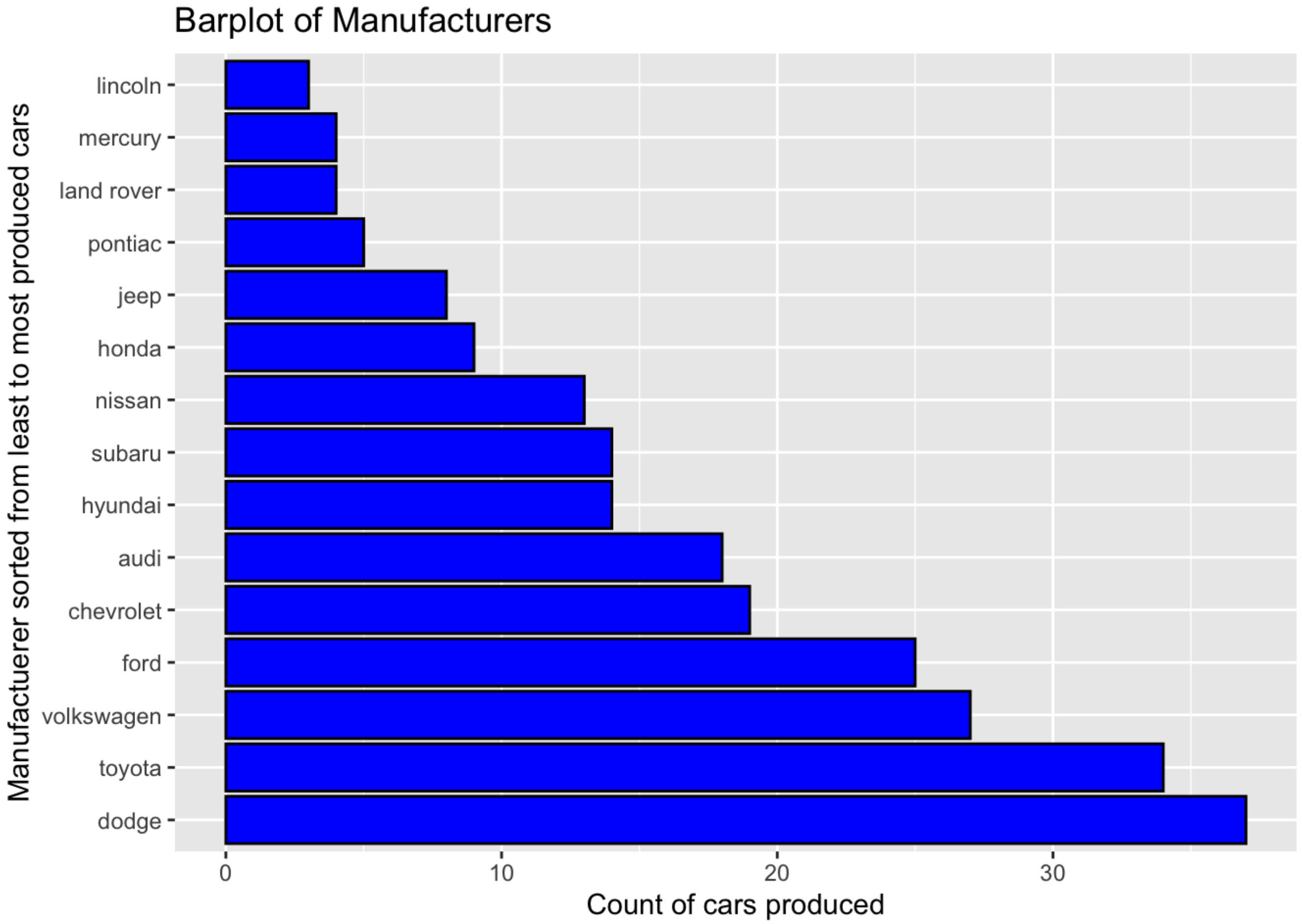
Exercise 3

```
#In order to sort the manufacturers, I looked online and found multiple sites with information, however I primarily used geeksforgeeks for the code below.

manufacter_bar <- ggplot(mpg, aes(y = reorder(manufacturer, manufacturer, function(x)-length(x)))) + geom_bar(color = "black", fill = "blue", state = "Count") +
  labs(title = "Barplot of Manufacturers", x = "Count of cars produced", y = "Manufactuerer sorted from least to most produced cars")
```

```
## Warning: Ignoring unknown parameters: state
```

```
manufacter_bar
```

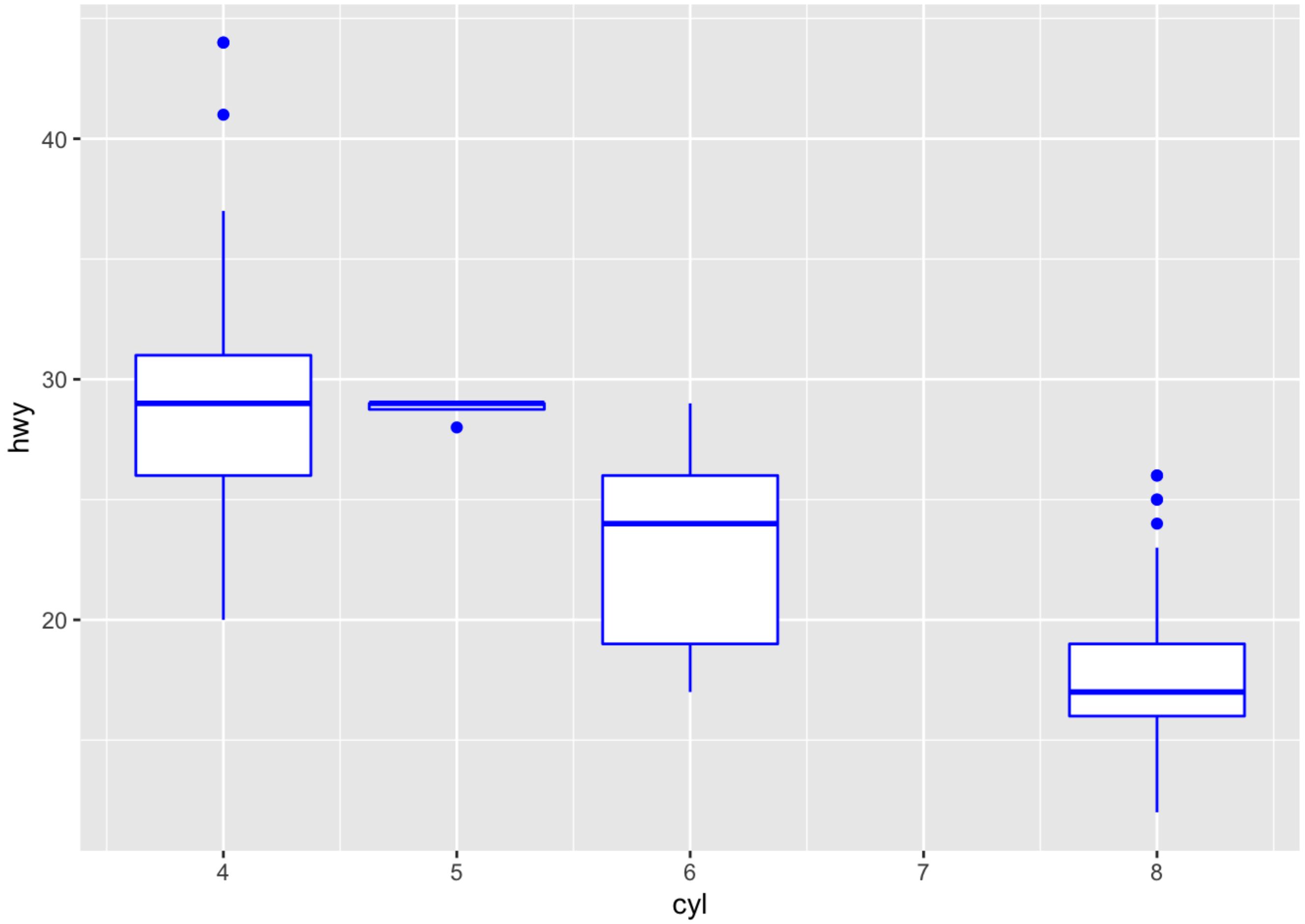


We can observe from the bar graph above that the amount of cars manufactured by company is ordered from least to greatest. Lincoln holds the title for least cars manufactured whereas Dodge hold the title for most cars manufactured.

Exercise 4

```
box_hwycyl <- ggplot(mpg, aes(x = cyl, y = hwy)) + geom_boxplot(aes(group = cyl), color
= "Blue")

box_hwycyl
```



From the box plot we can infer that the vehicles with more cylinders tend to have less spread in highway MPG as well as have lower MPG values, compared to vehicles with 4 cylinders.

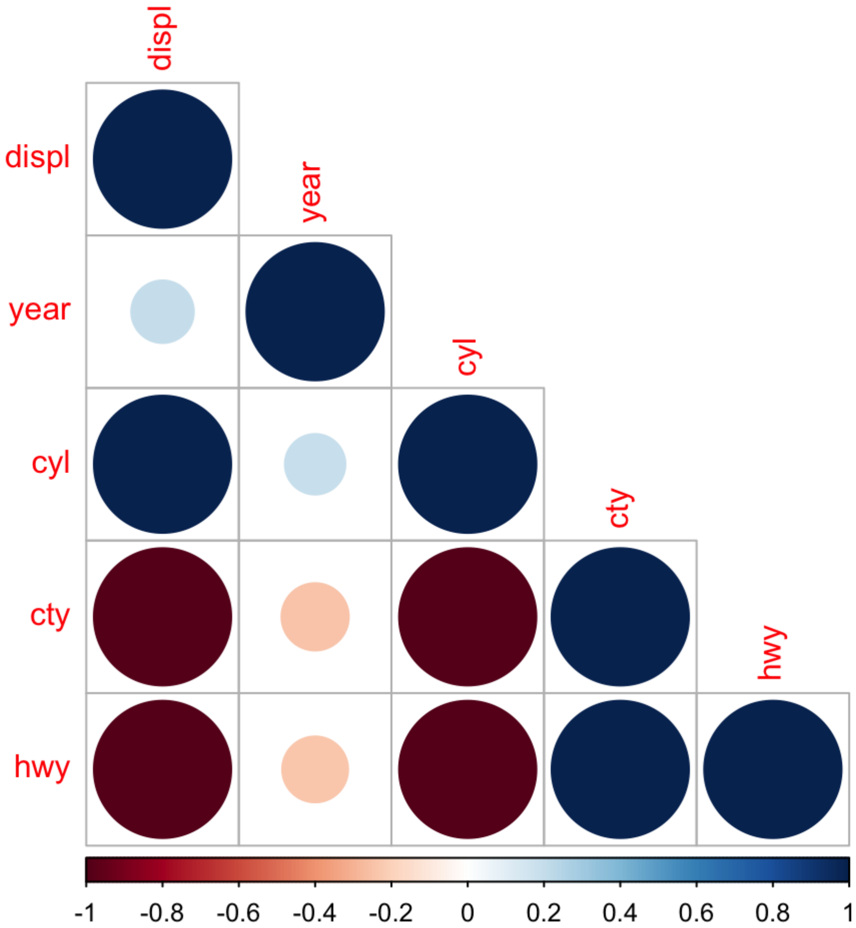
Exercise 5

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
#select quantitative vars
tidy_mpg <- cor(mpg[, c('displ', 'year', 'cyl', 'cty', 'hwy')])

#Based on the corrplot documentation, I used the command below to make the corrplat.
corrplot(cor(tidy_mpg), type = "lower")
```



The following variables have a negative correlation with each other:

Highway and displacement, Displacement and City MPG, Highway and Cylinders, City and Cylinders

The following variables have a positive correlation with each other:

Displacement and Cylinders, City and Highway

The following have very little (positive or negative) correlation:

Displacement and year, Cylinder and Year, City and Year

These correlations makes sense to me. We have already explored the correlation between certain variables in previous problems.