PSTAT 131 HW1

Joshua Price

4/10/2022

Question 1:
Define supervised and unsupervised learning. What are the difference(s) between them? The basic difference between supervised and unsupervised learning is that in supervised learning, you have the answer to train off of.

Question 2:
Explain the difference between a regression model and a classification model, specifically in the context of machine learning. Regression models a continuous response variable while classification models categorical variables

Question 3:
Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems. Regression: price/blood pressure. Classification: Survied/died, spam/not spam.

Question 4:
As discussed, statistical models can be used for different purposes. These purposes can generally be classified into the following three categories. Provide a brief description of each.

Descriptive models: Model that best visualizes a trend in data, etc line on scatter plot

Inferential models: Asks What features are significant. Test theories, states relationship between outcome and predictors

Predictive models: Asks what combo of features fits best, to predict Y with minimum reducible error.

Question 5:
Predictive models are frequently used in machine learning, and they can usually be described as either mechanistic or empirically-driven. Answer the following questions.

Define mechanistic. Define empirically-driven. How do these model types differ? How are they similar?

In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice.

Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models.

1. Mechanistic uses theory to predict what will happen in the real world. Emperical studies real-world events to develop a theory
2. I think empirical is easier to study because we already have the data and we're just trying to make sense of it
3. Both models can be over fitted with low variance and high bias

Question 6:
A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions:

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

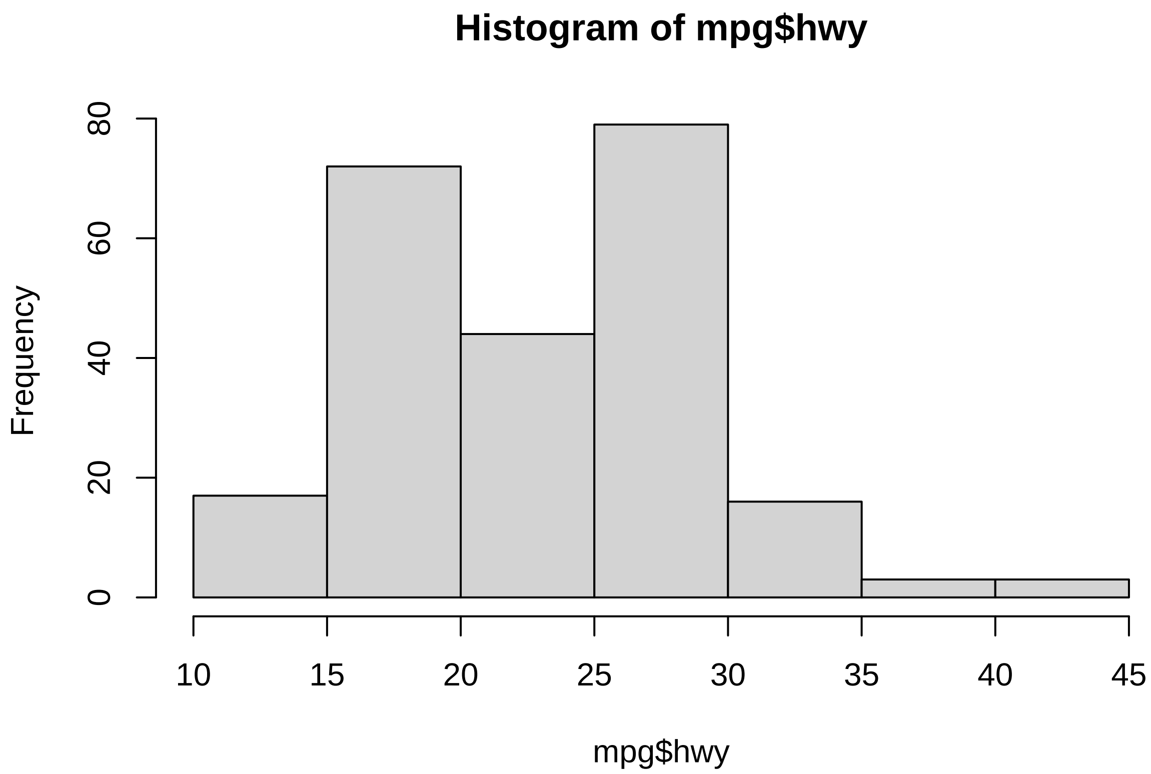
How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?

Classify each question as either predictive or inferential. Explain your reasoning for each.

1st question is predictive because its predicting who they'll vote for based on data
2nd question is inferential because its testing a theory

Exercise 1:
We are interested in highway miles per gallon, or the hwy variable. Create a histogram of this variable. Describe what you see/learn.

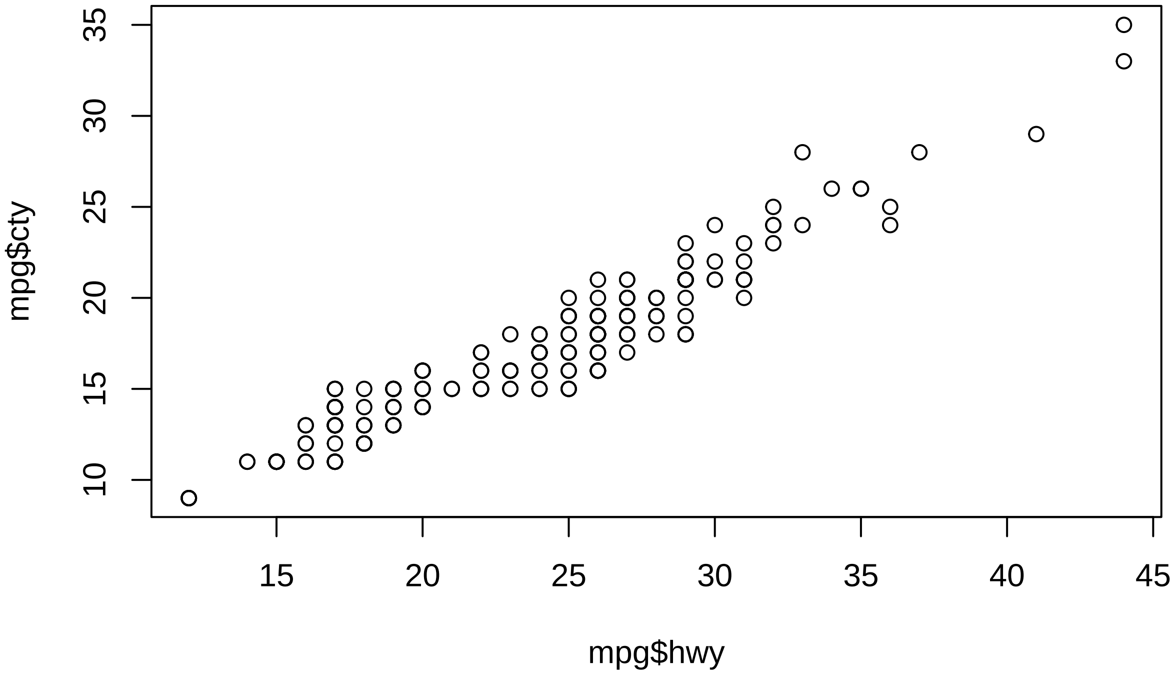
```
hist(mpg$hwy)
```



Most hwy mpg falls under 30

Exercise 2:
Create a scatterplot. Put hwy on the x-axis and cty on the y-axis. Describe what you notice. Is there a relationship between hwy and cty? What does this mean?

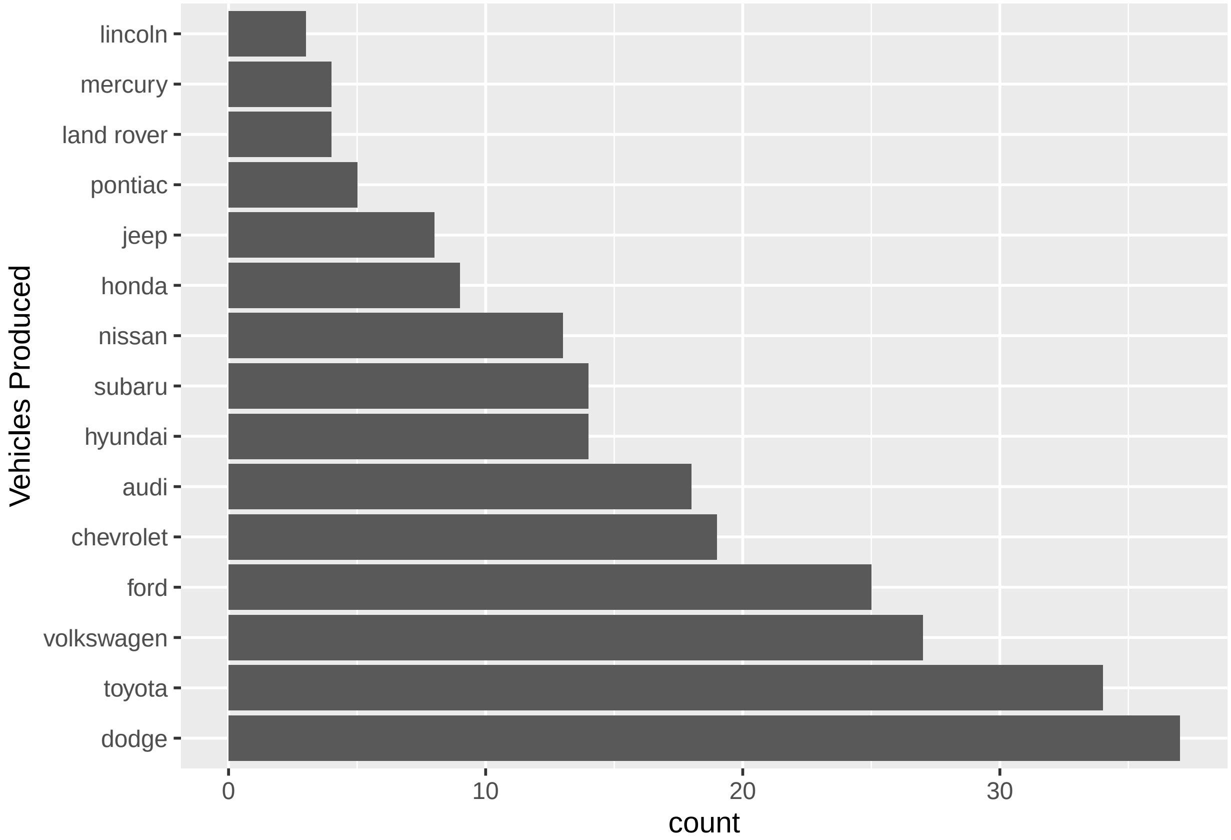
```
plot(mpg$hwy, mpg$cty)
```



It looks like there is positive linear correlation between cty mpg and hwy mpg. As highway mpg increases, city mpg increases

Exercise 3:
Make a bar plot of manufacturer. Flip it so that the manufacturers are on the y-axis. Order the bars by height. Which manufacturer produced the most cars? Which produced the least?

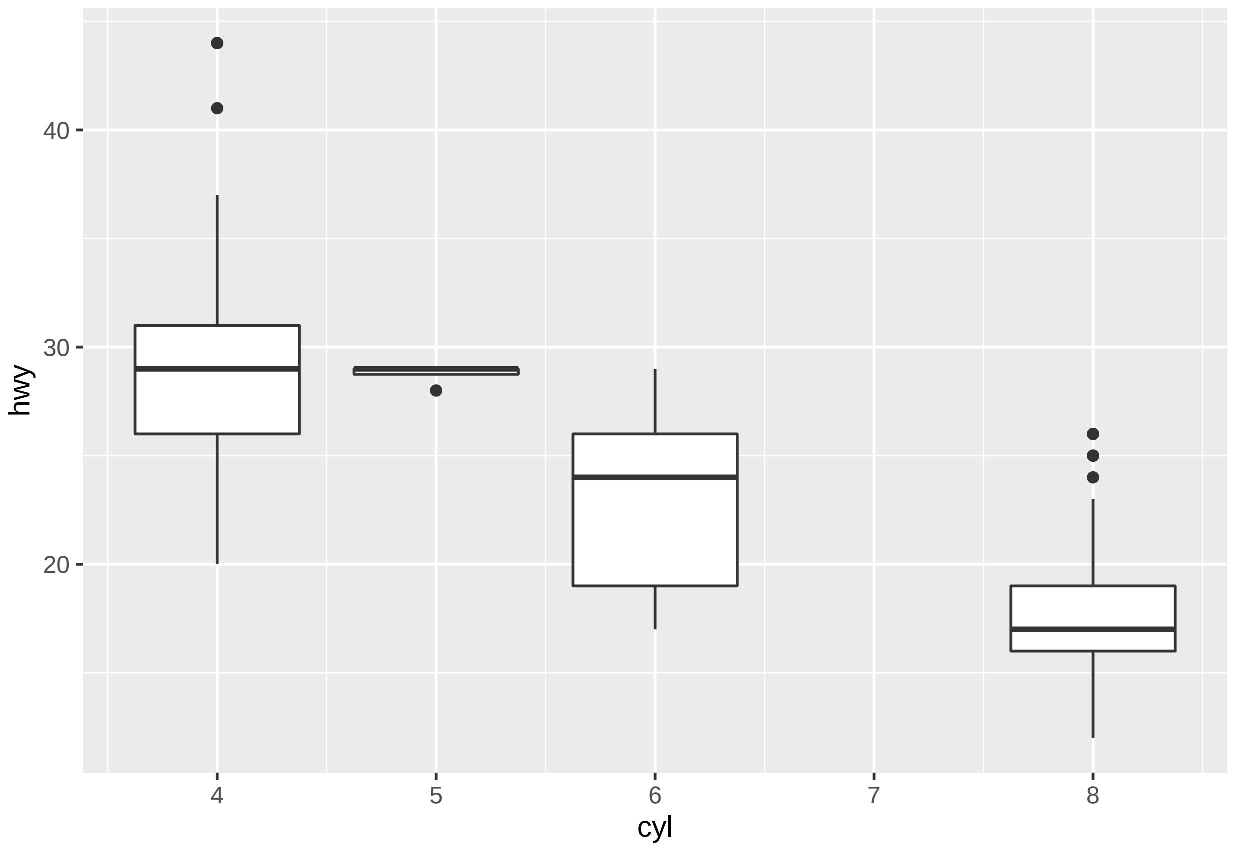
```
ggplot(mpg,aes(x=reorder(manufacturer, manufacturer,function(x)-length(x)),),horizontal =T) + geom_bar(
```



Dodge produced the most cars while lincoln produced the least

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

```
ggplot(mpg, aes(group=cyl,x=cyl,y=hwy)) + geom_boxplot()
```

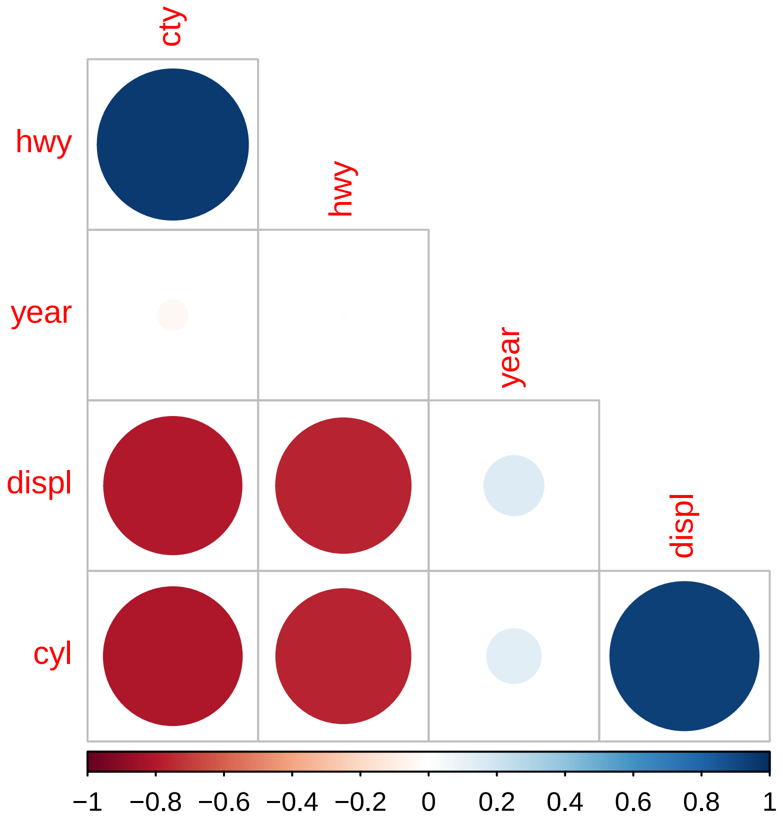


There is negative correlation between cylinders and hwy mpg. The more cylinders, the less hwy mpg

Exercise 5:
Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset. (Hint: You can find information on the package here.)

Which variables are positively or negatively correlated with which others? Do these relationships make sense to you? Are there any that surprise you?

```
mpgNums <- mpg %>% select(displ, year, cyl, cty, hwy)
corM = cor(mpgNums)
corrplot(corM,type='lower',diag =F,order='FPC')
```



Mpg in city is positively correlated with hwy and negatively correlated with displacement and cylinders with small negative correlation with year. hwy mpg is negatively correlated with displacement and cylinders, year has small positive correlation with displacement and cylinders. And displacement has high positive correlation with cylinders
I'm not surprised but I'm also not a car expert