Homework 1

PSTAT 131/231

Nicole Magallanes

Contents

Main Ideas

Question 1

Define supervised and unsupervised learning. What are the difference(s) between them?

Answer: Supervised learning is a machine learning approach that is built to train or supervise algorithms to estimate or predict outcomes based on the given inputs. This approach uses labeled inputs and outputs. Unsupervised learning is an approach that takes in inputs but does not have any supervising outputs. These algorithms can help to analyze and cluster unlabeled data. The main difference between these two approaches is that unsupervised learning does not have any supervising outputs, while supervised learning has labeled inputs and outputs. Also, in supervised learning the algorithm learns from its given training set while unsupervised learning mostly works on its own to discover patterns within the unlabeled data. For the most part, supervised learning models tend to be more accurate than unsupervised models. Supervised learning can includes linear regression, logistic regression, decision trees, among others; unsupervised data includes k-means clustering, hierarchical clustering and neural networks.

Question 2

Explain the difference between a regression model and a classification model, specifically in the context of machine learning.

Answer: The difference between a classification and a regression model is that classification models in machine learning are used to assign data into specific categories while regression models are used to understand the relationships between dependent and independent variables. An example of a classification model in the real world can be spam detection in emails, an example of a regression model can be predicting the impact of test scores on college admissions. Regression is quantitative while classification in qualitative.

Question 3

Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.

Answer: For regression ML problems, two commonly used metrics are MSE and RMSE. For classification ML problems, two commonly used metrics are accuracy and error rate.

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.

Answer:

Descriptive models: These models are used when trying to visually understand trends in the data. This model is helpful to understand the given data and it relationship with the variables within it through methods such as summarizing and classifying.

Inferential models: The aim of this model is put theories to test and find the relationship between predictors and outcomes. This model allows you to identify which variables are significant within the data.

Predictive models: Predictive models aim to "predict Y with minimum reducible error" (Lecture #1 slides). This model is not focused on hypothesis testing, but rather figuring out what variables work together best to fit the data.

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?

Answer: Mechanistic models assume a parametric form for f, they can add more parameters for more flexibility and wont match true unknown f (Lecture 1 slides). Mechanistic models are helpful to provide understanding of the mechanistic functions of treatments and are useful to overcome the limitations that come with ML predictions. Empirically-driven models don't have assumptions about f, they require a larger amount of observations and are more flexible. (Lecture 1 slides).

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

Answer: In general, I would say mechanistic models are easier to understand because they use theory to understand data and outcomes. They also need few input data point to make a prediction and can make predictions outside the range of previous input values.

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

Answer: The bias-variance trade off is related to the use of the models because we always want to use an algorithm that doesn't under fit or over fit the data. The bias-variance trade off is essentially finding the right balance of values, as the increase of the variance or bias will lead to the decrease of the other, to make sure a model is accurate and decreases prediction errors.

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.

Answer: The first question is a predictive one, the question is asking to predict the percentage of likeliness that the voter would vote in favor of the candidate with their given profile. For this question we would have to use the data given to us about the voter to make this prediction. The second question is an inferential one, with this question we we are asked to find the relationships between variables within the data. In this case the variables would be the likelihood of support and that thy had personal contact with the candidate.

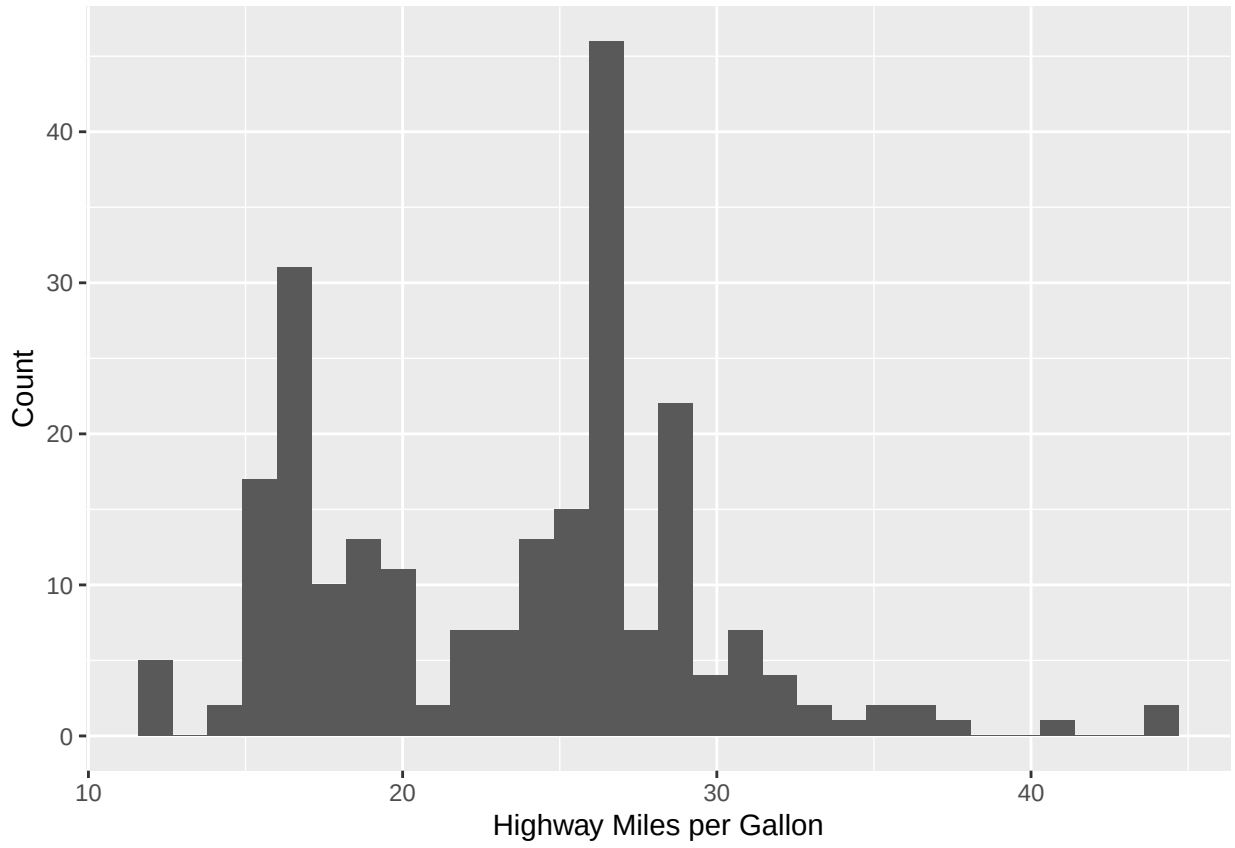
Exploratory Data Analysis

```
# install.packages("tidyverse")
# install.packages("tidymodels")
# install.packages("ISLR")

library(tidyverse)
library(tidymodels)
library(ISLR)
library(ggplot2)
library(corrplot)

view(mpg)
```

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.

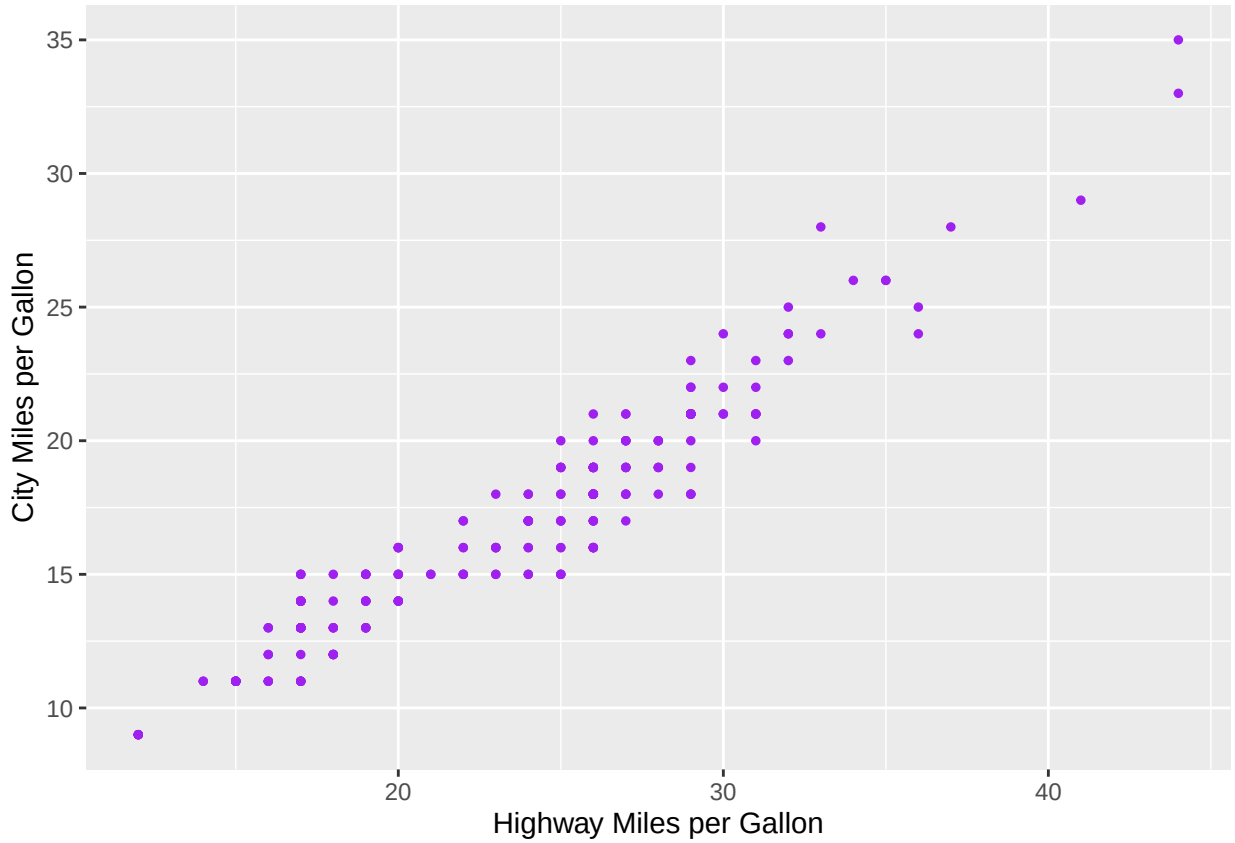
```
mpg %>%
  ggplot(aes(x=hwy)) +
  labs(x="Highway Miles per Gallon", y="Count") +
  geom_histogram(bins=30)
```



We can see that the distribution for hwy is slightly skewed to the right, meaning that there is a good amoount of values that are at the lower end of the graph and it has a tail to the right.

Exercise 2

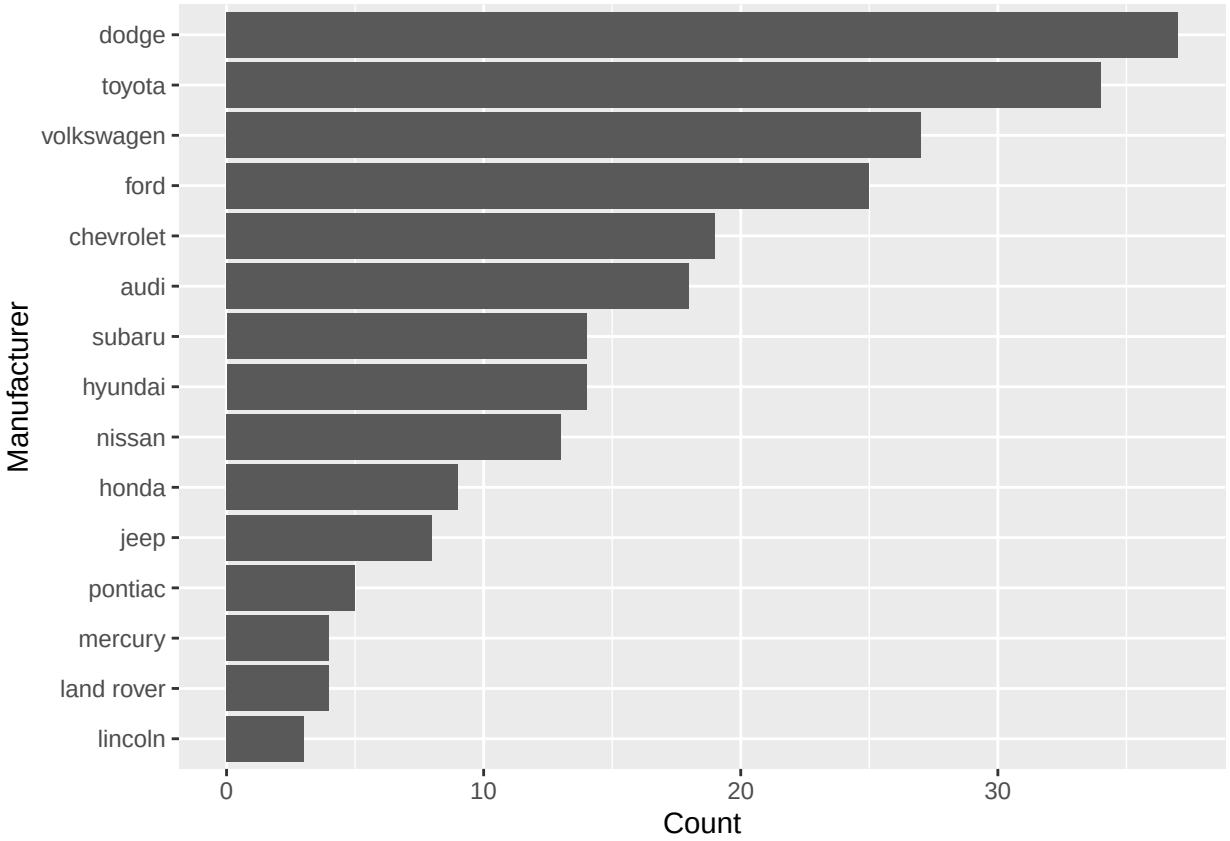
```
mpg %>%
  ggplot(aes(x=hwy, y=cty)) +
  labs(x="Highway Miles per Gallon", y="City Miles per Gallon") +
  geom_point(color='purple', size=1)
```



In this scatter plot we can see that there is a linear relationship between hwy and cty. As the values for hwy increase, the values for cty also increase. This relationship can tell us that as the value is high for the city, values for the highway will also be high, and it would be the same the other way around.

Exercise 3

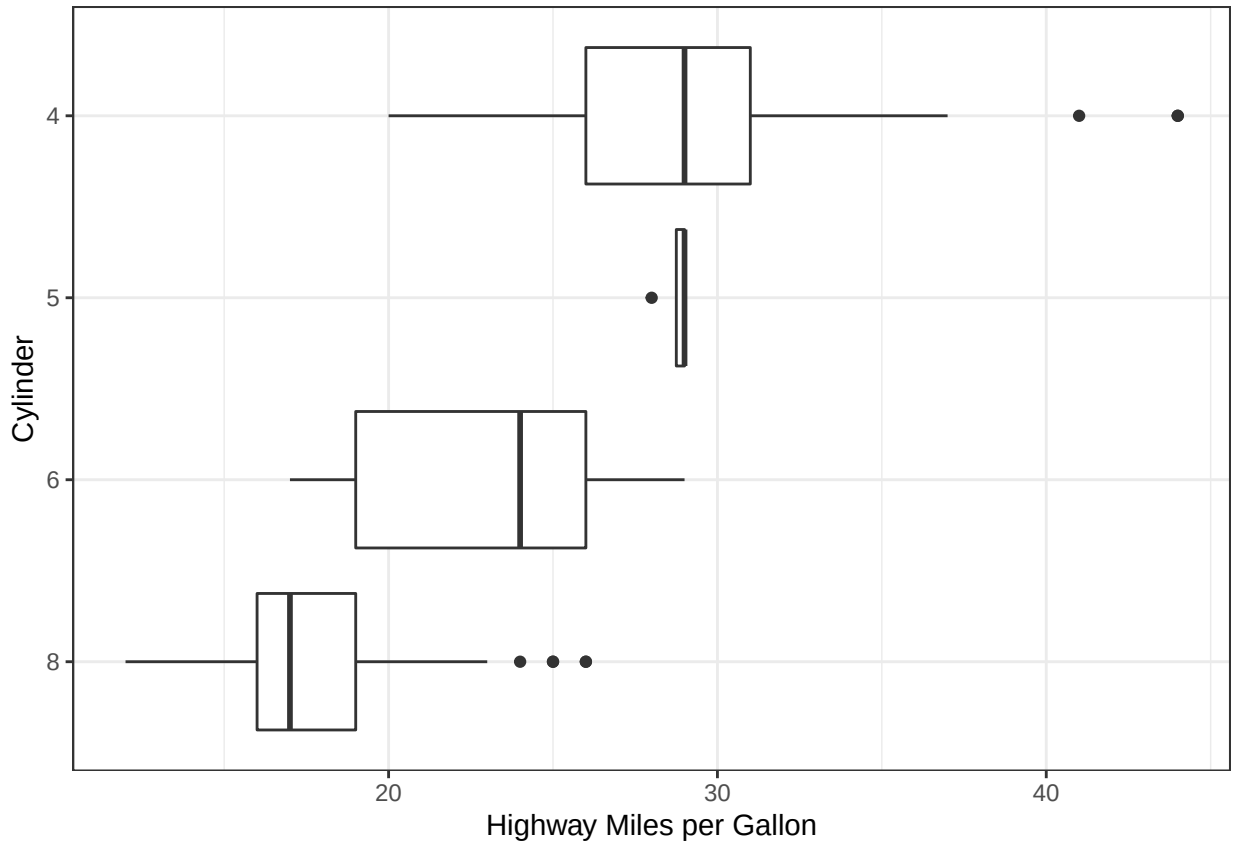
```
barplot <- ggplot(mpg, aes(x=reorder(manufacturer, manufacturer, length))) +
  labs(x="Manufacturer", y="Count") +
  geom_bar()
barplot + coord_flip()
```



According to this bar plot, the manufacturer that produced the most cars is Dodge, while the manufacturer that produced the least cars is Lincoln.

Exercise 4

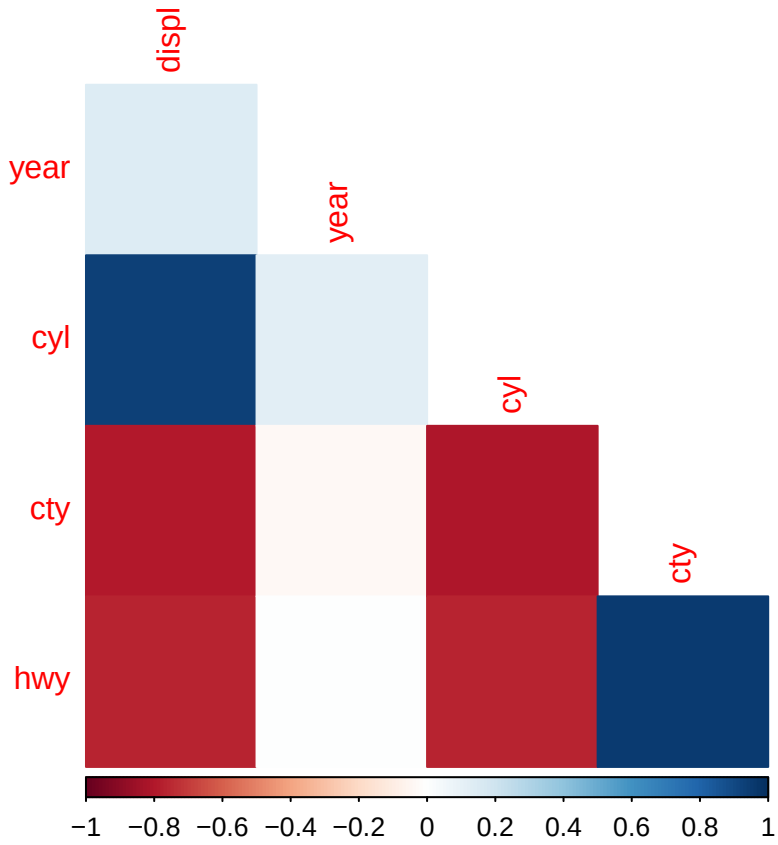
```
mpg %>%
  ggplot(aes(x = hwy, y = reorder(cyl, hwy))) +
  labs(x="Highway Miles per Gallon", y="Cylinder") +
  geom_boxplot() +
  theme_bw()
```



Some of the patterns I can see through this box plot is that the smaller cyl is, the higher the values are for hwy. The same is for the other way around, the bigger the value is for cyl, the smaller it is for hwy. I can also see that the the highest and lowest values of cyl have outliers.

Exercise 5

```
mpg %>%
  select(is.numeric) %>%
  cor() %>%
  corrplot(type = 'lower', diag = FALSE,
           method = 'color')
```



The values that are positively correlated are hwy and cty, cyl and displ, year and displ, cyl and year. The values that are negatively correlated are hwy and displ, cty and displ, hwy and cyl, and lastly cty and cyl. Most of these relationships make sense to me, one that did surprise me that has a strong negative correlation is displ and cty.