Homework 1

Jules Merigot (8488256)

September 30, 2022

PSTAT 131/231 Statistical Machine Learning - Fall 2022

Machine Learning Main Ideas

Question 1

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

Supervised learning is when the user knows both the "input" and the "output" of their learning model, so the output is being "supervised", which thus allows users to find the most accurate model and determine the precision of their estimations. Unsupervised learning is when the user knows the "input", but does not know the "output" of their model, thus has "no answer key" and is "unsupervised". There is no way to check how well the model did, or how accurate it was. The user is unable to check the results of their estimations or predictions, and therefore cannot judge the accuracy of their model.
(*from 'Lecture: Course Overview and Introduction'*)
The main difference in Supervised Learning v.s. Unsupervised Learning is the that for Supervised Learning, the response is known, while for Unsupervised Learning, the response is not known. This leads to many other differences, including various regression methods that can only be used for Supervised Learned, while Unsupervised Learning uses more clustering methods.

Question 2

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

At a basic level, the difference between a regression model and a classification model lies in the predictors or values used. In the case of a regression model, a certain variable X will be quantitative, which means it will have a numerical value. On the other hand, in a classification model with a certain variable Y, Y will be qualitative, which means it will have a categorical value, such as Male/Female, True/False, etc. In the context of machine learning specifically, the difference between a regression model and a classification model is that regression algorithms are used to predict continuous numerical quantities, while classification algorithms are used to predict discrete values, such as the ones stated earlier.

Question 3

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

Two commonly used metrics for regression ML problems are Mean Squared Error (MSE) and Mean Absolute Error (RAE). Two commonly used metrics for classification ML problems are Precision and Accuracy.

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:** Used to choose a model that best visually emphasizes a trend in a set of data, such as using a line in a scatter plot.
- **Inferential models:** Used to find what combinations of model features fit best, in order to predict a response variable Y with a minimum amount of reducible error.
- **Predictive models:** Used to determine which model features are most significant, in order to test certain theories, test causal claims, and find relationships between an outcome and its predictor(s).

(*from 'Lecture: Course Overview and Introduction'*)

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?**

Mechanistic is the idea of using fundamental laws or natural theories to predict what will happen in the real world. Empirically-driven is the idea of using real-world data and evidence from events in order to develop a particular theory. These model types differ in the sense that they have opposite/reverse processes. Mechanistic models try to predict events based on theory, while empirical models try to develop theory based on real-world events. However, they are both similar since they both employ either theory or data to conclude the other.

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

In general, I believe that a empirically-driven model is easier to understand because no potentially incorrect assumptions are made at the beginning. While mechanistic models may be easier to understand for others since they are based on natural laws that may appear as common sense to many people, I have a tendency to favor trust strict, recorded data. If a model is initially made using real-world data that has been observed, we can be sure that our responses will be accurate, regardless of whether our final, developed theory, is correct or not.

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

The bias-variance tradeoff is related to the use of mechanistic or empirically-driven models in a variety of ways. In the case of a mechanistic model, since predictions are being made based on theory, there will be higher bias from the user and less variance since certain results will already be expected. In the case of an empirically-driven model, if the training data is strictly followed based on observed events, there will be high variance but much less bias. In either of these cases, the best solution is to find the optimal balance between bias and variance in order to get a good tradeoff.

Question 6

A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions:
Classify each question as either predictive or inferential. Explain your reasoning for each.

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

This question is inferential because it is attempting to infer a certain result based on known data and events.

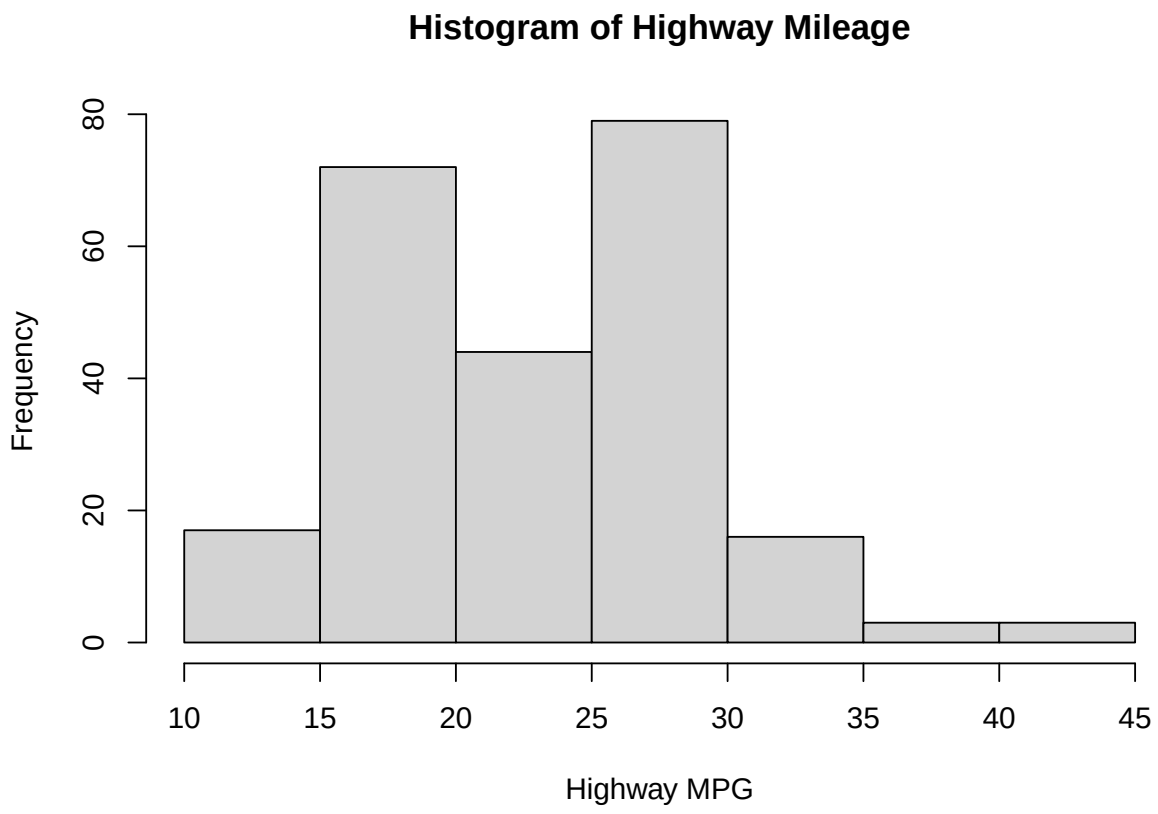
- **How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?**

This question is predictive because it is attempting to predict a future response based on past behaviors.

Exploratory Data Analysis

Question 1

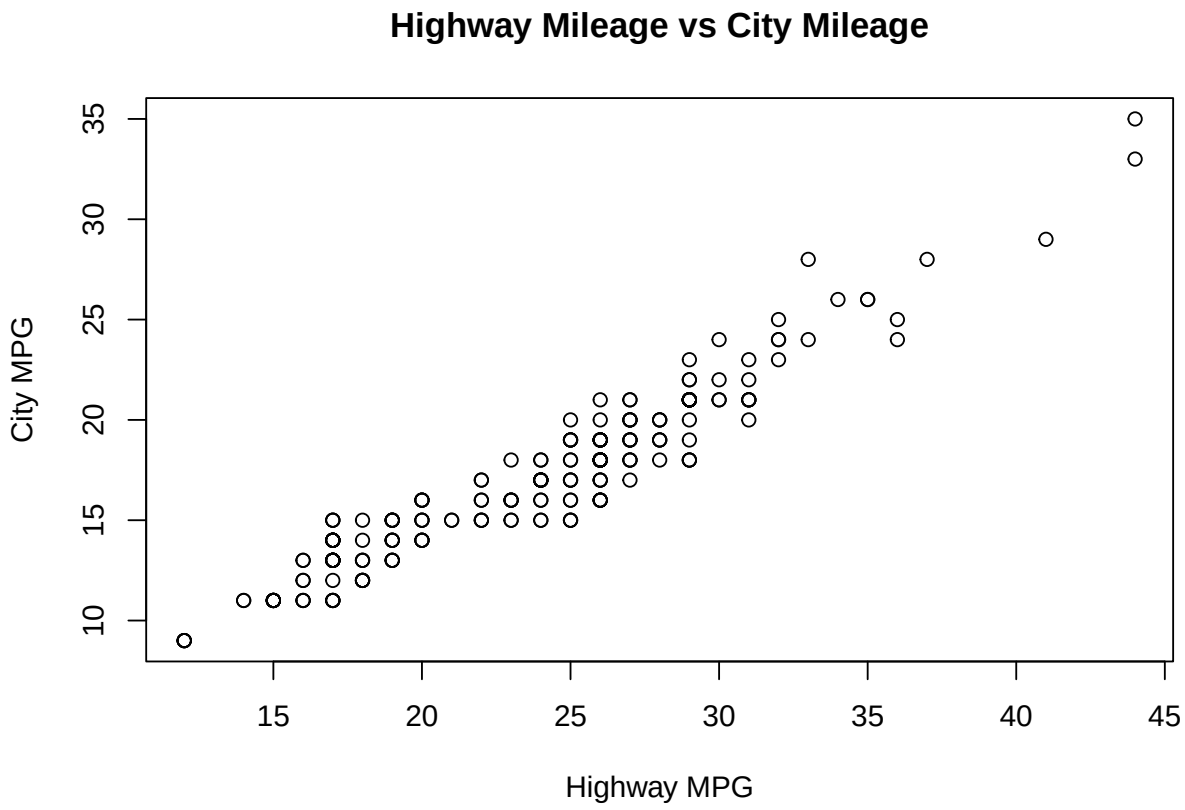
```
require(ggplot2)
hist(mpg$hwy, main="Histogram of Highway Mileage", xlab="Highway MPG")
```



The histogram of highway miles per gallon seems so be normal and slightly skewed right with a majority of cars having a highway mileage between 25 mpg and 30 mpg.

Question 2

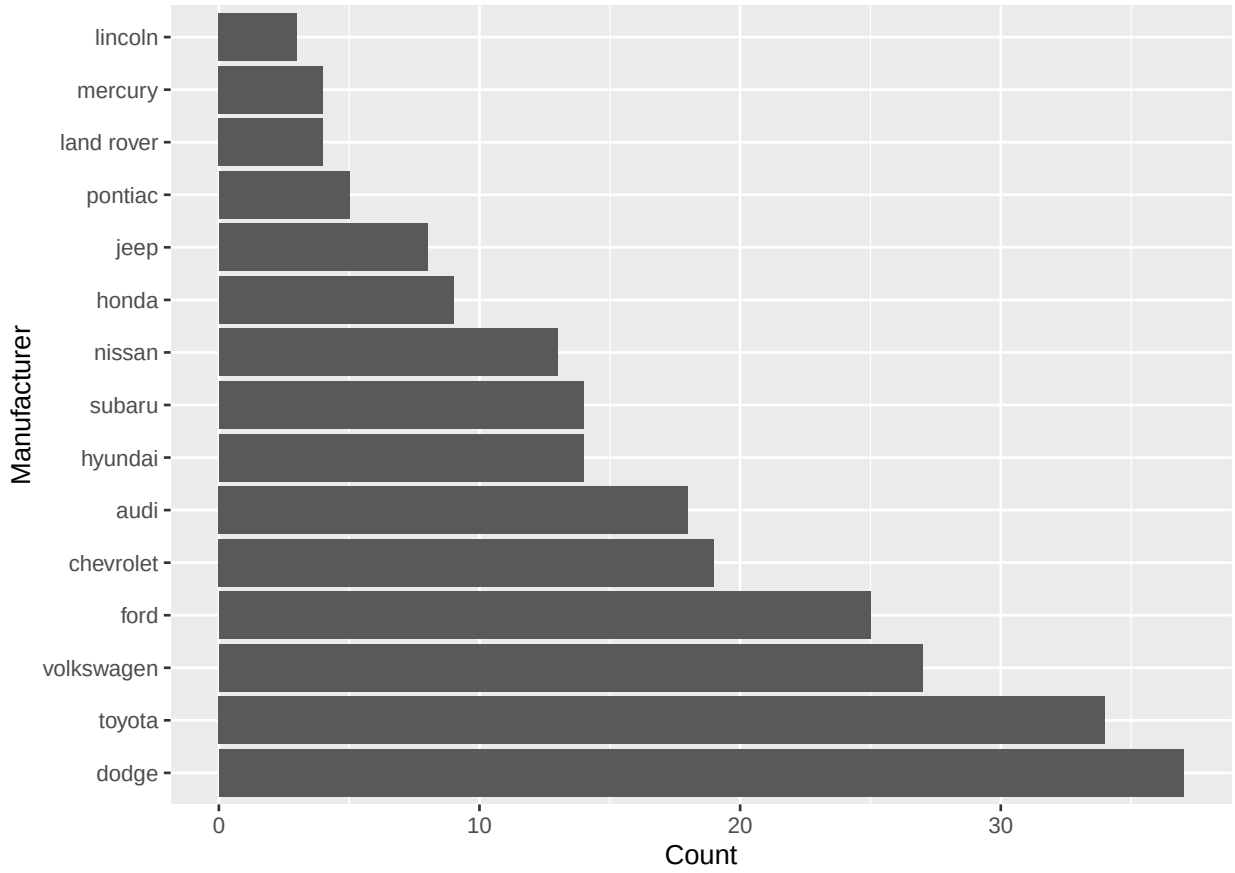
```
plot(mpg$hwy, mpg$cty, main="Highway Mileage vs City Mileage", xlab="Highway MPG",
     ylab="City MPG")
```



There is a strong, positive, linear association/relationship between the cars' highway mileage and city mileage. This means that as a car's highway mileage increases, so does its city mileage, and vice versa.

Question 3

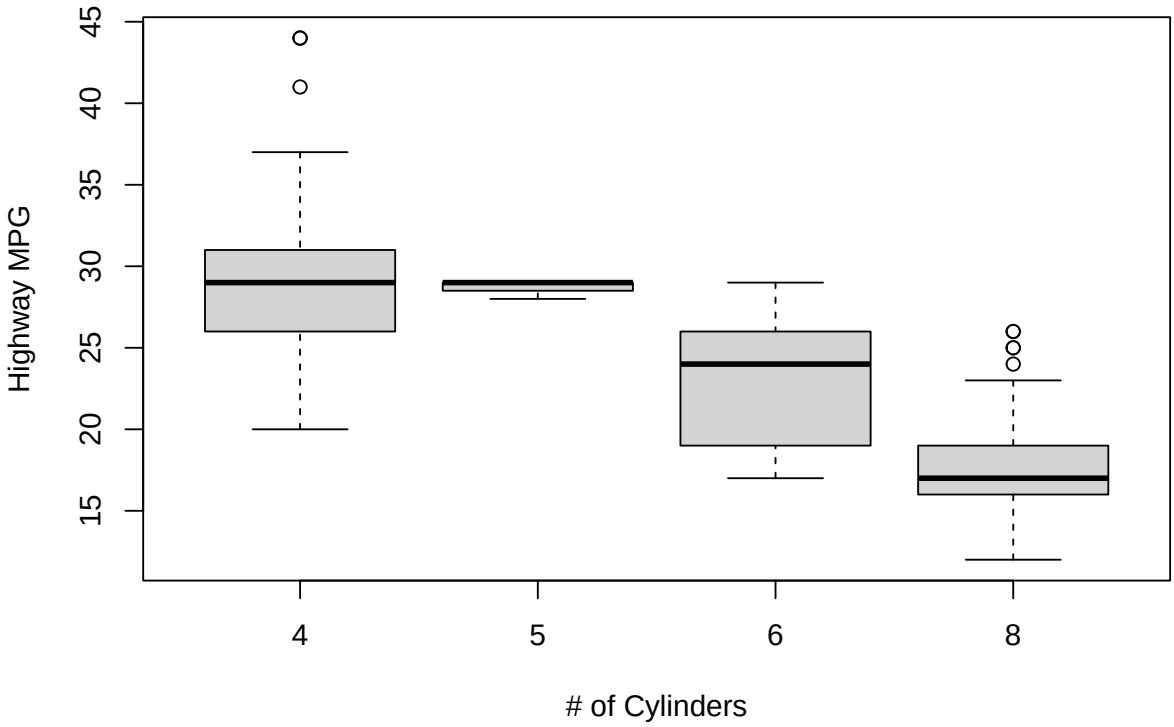
```
ggplot(data=mpg, aes(y=reorder(manufacturer, manufacturer, function(y)-length(y)))) +
  geom_bar(stat="Count") +
  labs(x="Count", y="Manufacturer")
```



According to the manufacturer bar plot, Dodge produced the most cars and Lincoln produced the least cars.

Question 4

```
boxplot(formula=hwy ~ cyl, data=mpg, xlab="# of Cylinders", ylab="Highway MPG")
```



There is a clear pattern that as the number of cylinders that cars have increases, the highway mileage of the cars will decrease.

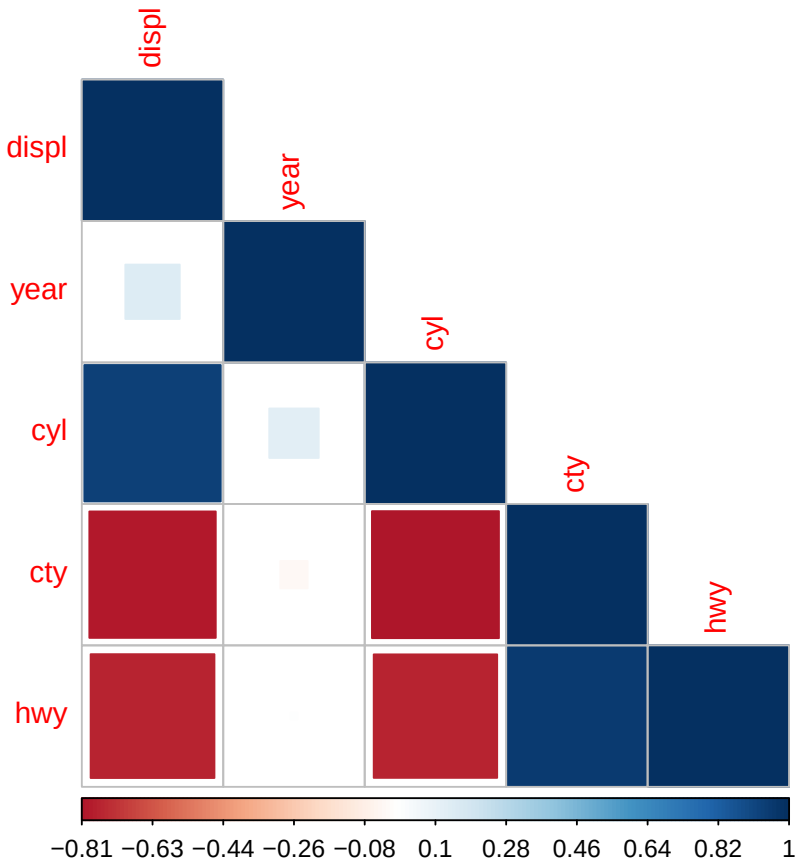
Question 5

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.1.3
```

```
## corrplot 0.92 loaded
```

```
df <- mpg[,c(-1,-2,-6,-7,-10,-11)]
M = cor(df)
corrplot(M, method="square", type='lower', is.corr=FALSE)
```



Variables that are positively correlated with each other include *cty* and *hwy*, as well as *displ* and *cyl*. While variables that are negatively correlated include *displ* and *cty*, *displ* and *hwy*, *cyl* and *cty*, as well as *cyl* and *hwy*. For variables *displ* and *year*, *year* and *cty*, *year* and *hwy*, as well as *year* and *cyl*, there seems to be no strong correlation. These relationships all make sense in in context, since city mileage and highway mileage will obviously be positively correlated, while cylinders and highway mileage will be negatively correlated as shown in question 4. I am surprised by the fact that the years of the cars has no correlation to the other variables considering that newer cars tend to be more fuel efficient and would therefore have a high city and highway mileage.