Hw1

Ying Zhou

10/1/2022

Homework 1

PSTAT 131/231

Machine Learning Main Ideas

Please answer the following questions. Be sure that your solutions are clearly marked and that your document is neatly formatted.

You don't have to rephrase everything in your own words, but if you quote directly, you should cite whatever materials you use (this can be as simple as "from the lecture/page # of book").

Question 1:

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

Answer: Supervised and unsupervised learning are both machine learning categories. Supervised learning uses labeled datasets that it accurately predict the response for future observations given predictors or understand the relationship between the response and the predictors. It separates into two types: classification and regression. The learning algorithm learns from labeled training data, helps us predict the future outcomes. Unsupervised learning algorithms Learn without a supervisor. It uses datasets with unlabeled data to discover discover hidden patterns in data.

Question 2:

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

Answer: We refer to problems with a quantitative response as regression problems and problems with a qualitative response as classification problems. A regression model uses an algorithm to understand the relationship between response and the predictors while a classification model uses an algorithm to accurately assign data into specific categories. And also, a regression model involves a continuous outcome while a classification model involves a non-continuous (categorical) outcome.

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 Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). For classification ML problems, two common metrics are accuracy and precision.

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 best visually emphasize a trend in data. For example, using a line on a scatterplot to illustrate the relationship between variables.

- Inferential models:

Used to make causal statements about the relationship between predictors and the outcome. Aim is to test theories.

- Predictive models:

Used to predict the outcome variable with minimized reducible error. It does not focused on hypothesis tests.

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 models assume a parametric form for f, the relationship between the predictors and the outcome that it won't match the true unknown f. They are more flexibility that they can add parameters. Empirically-driven models make no assumptions about the form for the relationship f. They requires a large number of observations and are much more flexible by default.

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

A mechanistic model is easier to understand since it follows certain patterns that it fits simple parametric form. It uses a theory to make a prediction, which already gives a start, while a empirically-driven model is based on experiment. We cannot predict until we have a large amount of observations.

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

Bias and variance are inversely connected. Mechanistic models tend to have higher bias and lower variance, while empirically-driven models tend to have higher variance and lower bias since they are much more flexible.

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: Predictive analysis focus on the past behavior more to predict the future, while inferential analysis extrapolates properties from tests and estimates. Thus, the first question is predictive. Because in the first question, the campaign is focus on the voters' profile that show their past behaviors and try to predict the probability of their likability of the candidate. The second question is inferential since the campaign is curious about how would the relationship change between the voters and candidate affect the outcome. They focus on the outcome after the test.

Exploratory Data Analysis

This section will ask you to complete several exercises. For this homework assignment, we'll be working with the mpg data set that is loaded when you load the tidyverse. Make sure you load the tidyverse and any other packages you need.

Exploratory data analysis (or EDA) is not based on a specific set of rules or formulas. It is more of a state of curiosity about data. It's an iterative process of:

- generating questions about data
- visualize and transform your data as necessary to get answers
- use what you learned to generate more questions

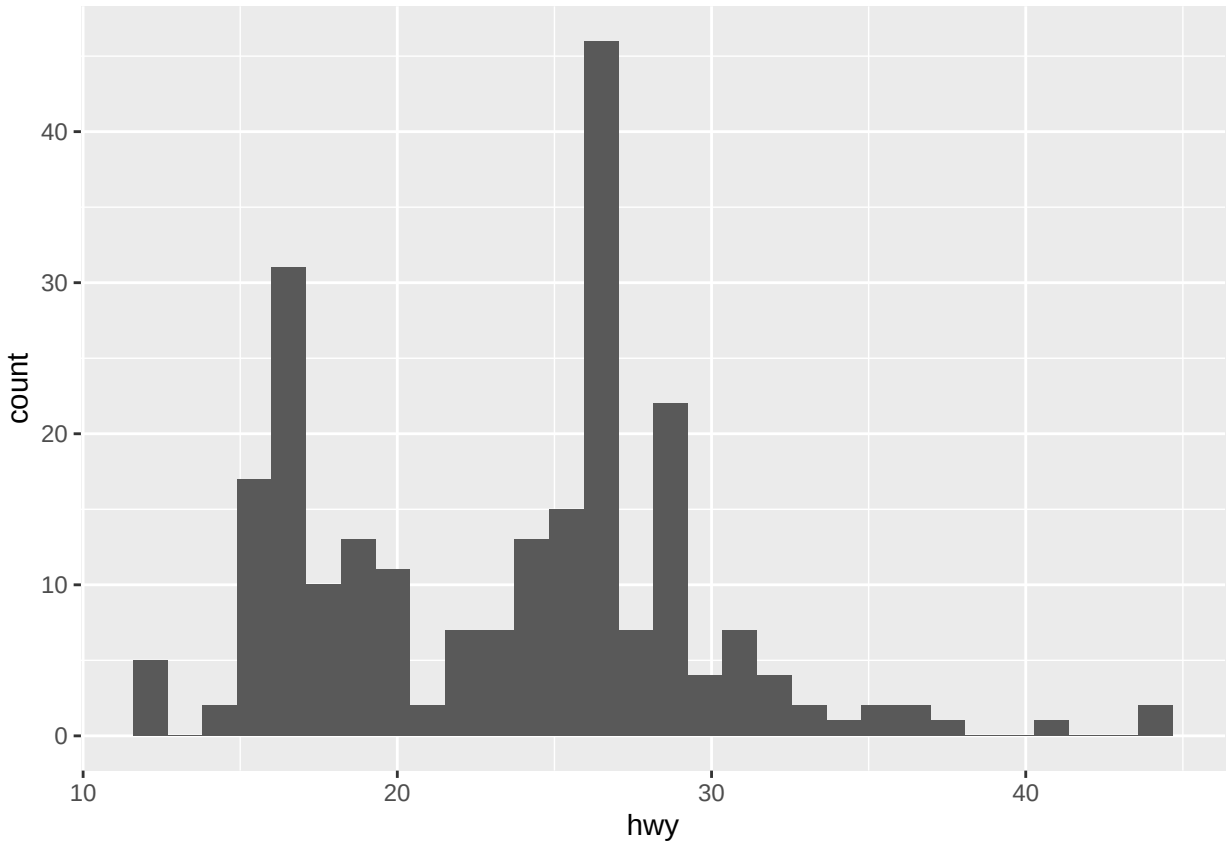
A couple questions are always useful when you start out. These are "what variation occurs within the variables," and "what covariation occurs between the variables."

You should use the tidyverse and ggplot2 for these exercises.

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)) +
 geom_histogram(bins = 30)
```

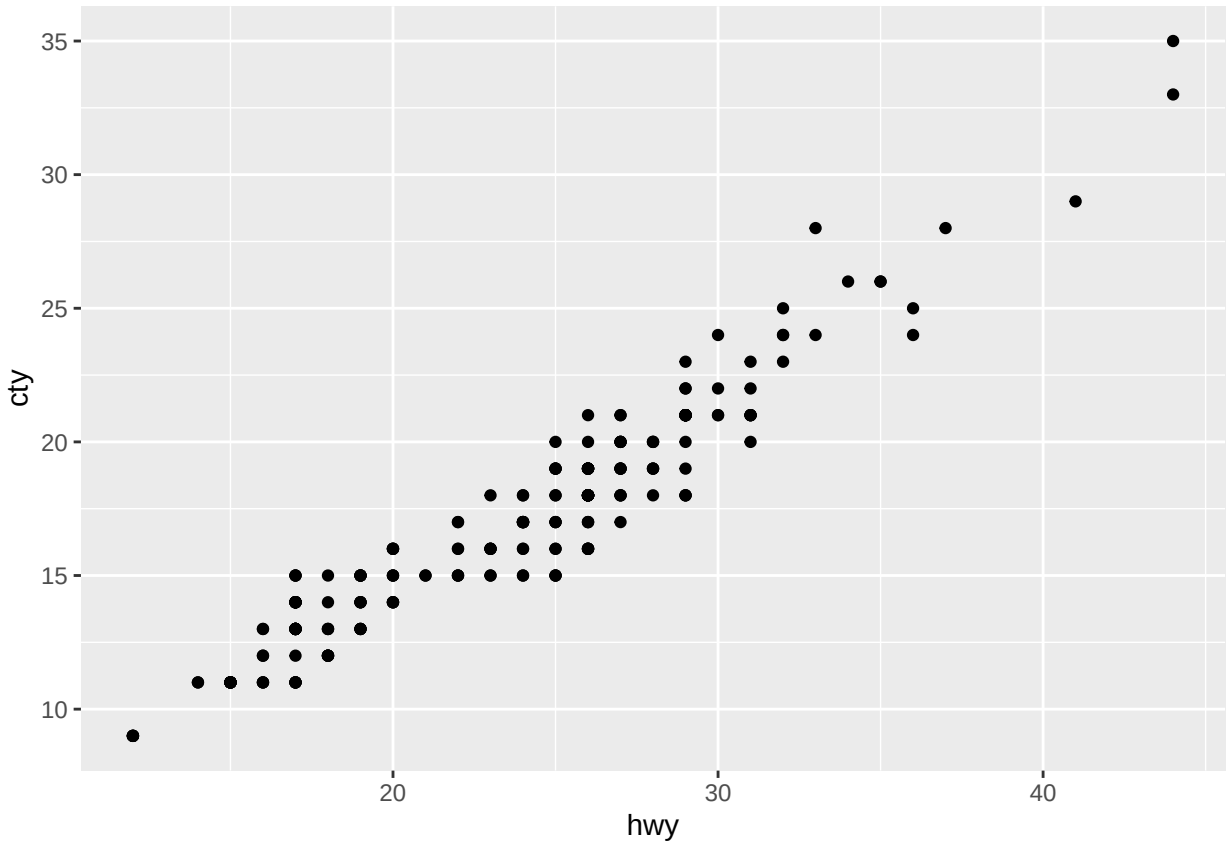


The distribution of the highway mileage seems positively skewed. There is a peak around 16-17 mpg and another one around 26-27 mpg. Few cars have above 40 mpg.

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?

```
mpg %>%
 ggplot(aes(x = hwy, y = cty)) +
 geom_point()
```

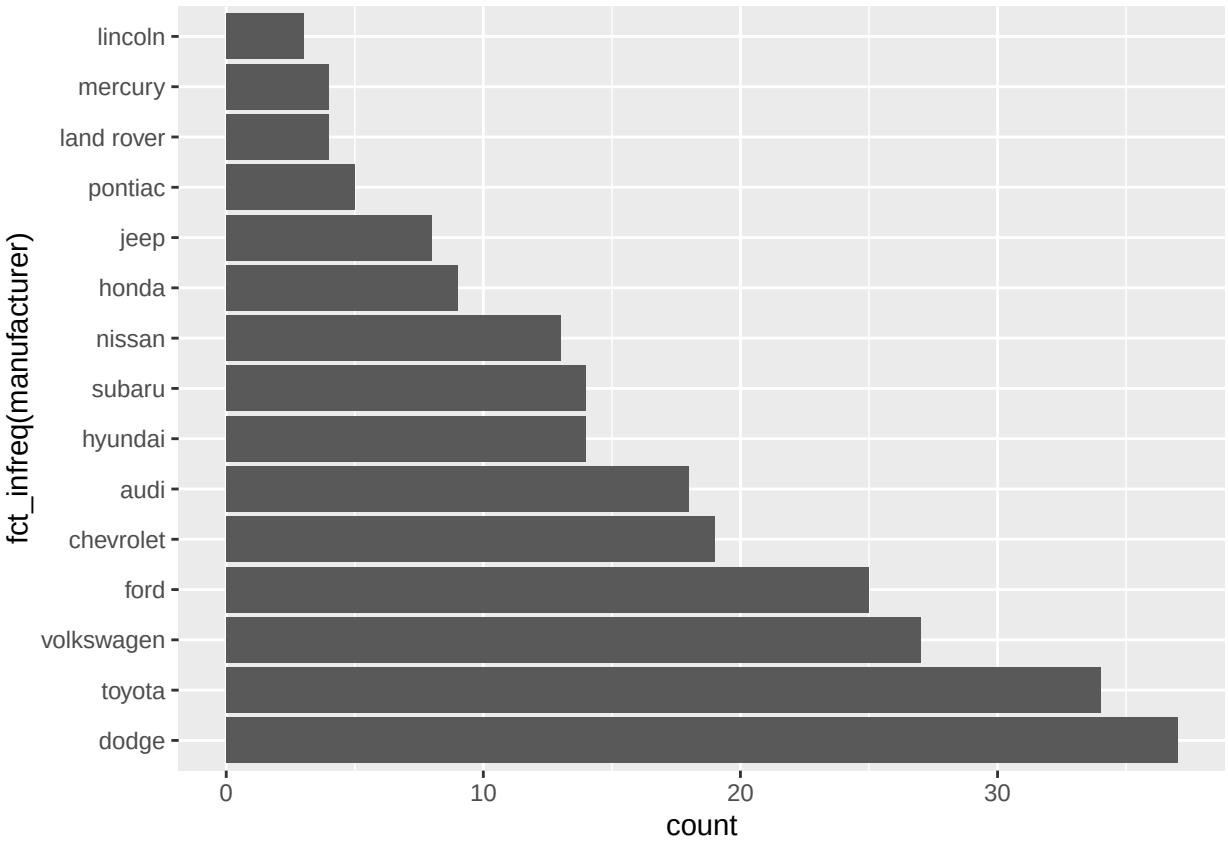


The hwy and cty seems to have a positive linear relationship. As the hwy increases, the cty also increases. The scatterplot looks not as cluttered as it generally be since the observations might be round numbers.

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?

```
mpg %>%
 ggplot(aes(x = fct_infreq(manufacturer))) +
 geom_bar() +
 coord_flip()
```

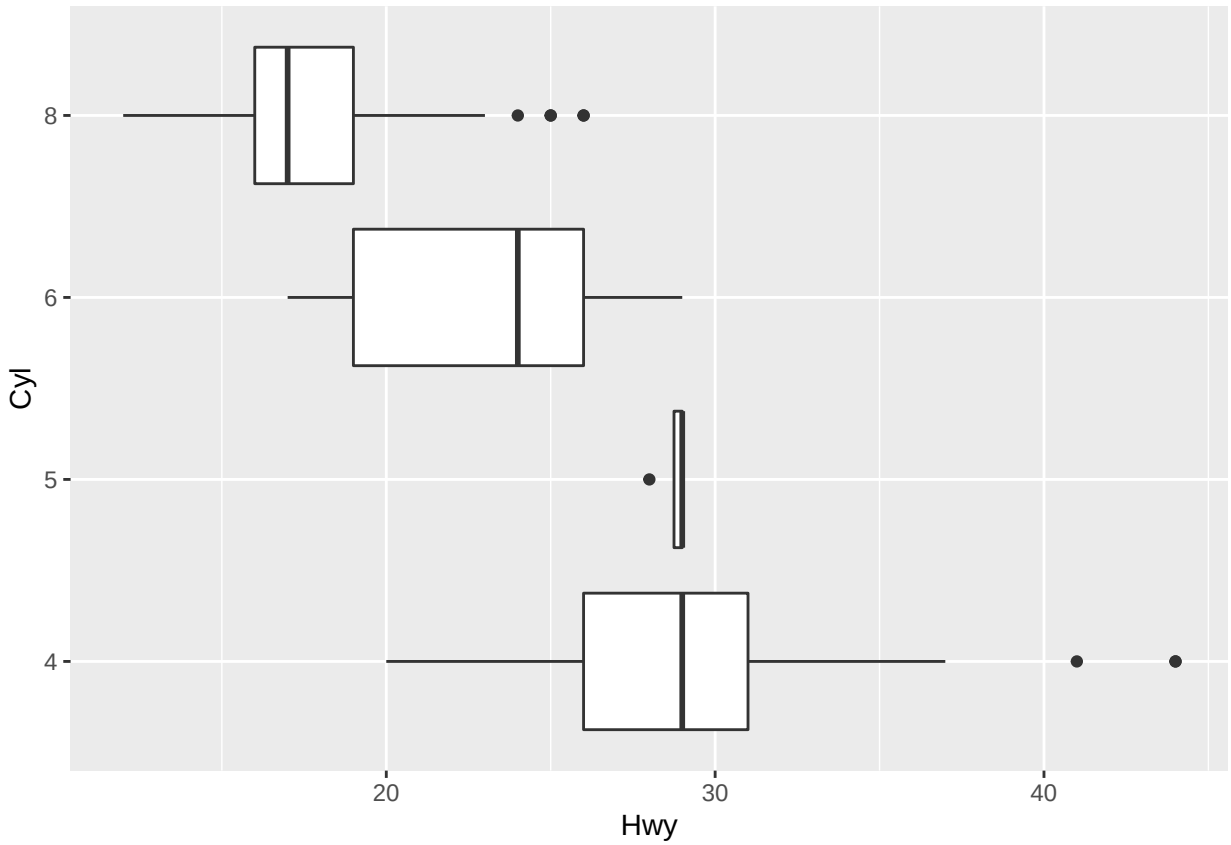


We can find that Dodge produced the most, it almost reached 40. And Lincoln produced the least that it produced below 5.

Exercise 4:

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

```
mpg %>%
 ggplot(aes(x = hwy, y = factor(cyl))) +
 geom_boxplot() +
 xlab("Hwy") +
 ylab("Cyl")
```



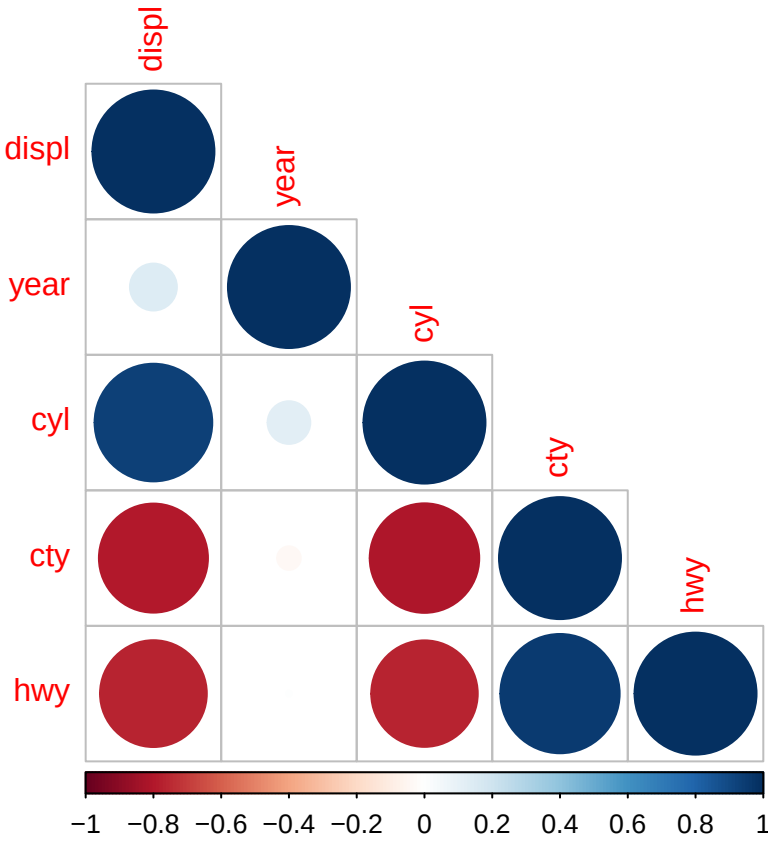
From the box plot, we can observe that with the increasing of the number of cylinders, the mileage tends to decrease. There are few five cylinders cars in the data and cars with four cylinders have the highest average mileage.

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?

```
mpg %>%
 select(is.numeric) %>%
 cor() %>%
 corrplot(type = "lower")
```



Number of cylinders and displacement are positively correlated with each other. Hwy and cty are also positively correlated. Displacement is negatively correlated with both the mileage variables, hwy and cty, so as number of cylinders, also negatively correlated with them. I was surprised by the negative correlation of displacement and mileage. I originally thought that the more mileage, the greater the displacement. Now I learn to analysis the correlation by making plot.