PSTAT131 HW1

Kris Li

2022-04-03

Question 1:

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

Answer: In supervised learning, we already have expected outcome from given inputs and we train the machine to learn an optimal function that approximates the relationship between the input and output. For example, regression and classification. In unsupervised learning, the data is unlabeled. It is useful to categorize data and to find underlying patterns among them.

Question 2:

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

Answer: A regression model predicts a continuous response; a classification model predicts a discrete response.

Question 3:

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

Answer: Regression - Training MSE, Test MSE; Classification - Training Error Rate, Test 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.

Descriptive models: The goal is to visually present a trend in data.

Inferential models: The goal is to test theories, such as claimed causal relationships between the predictors and response.

Predictive models: The goal is to predict response with minimum reducible error. Don't focus 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?

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.

Answer: Mechanistic models use assumed theories to predict responses. Empirically-driven models use observations to develop theories. In my opinion, mechanistic models are easier to understand because the theories are already known so there is one less step compared with empirically-driven models where you need to construct theories from observations first. A simple model has high bias and low variance, whereas a flexible model has low bias but high variance.

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.

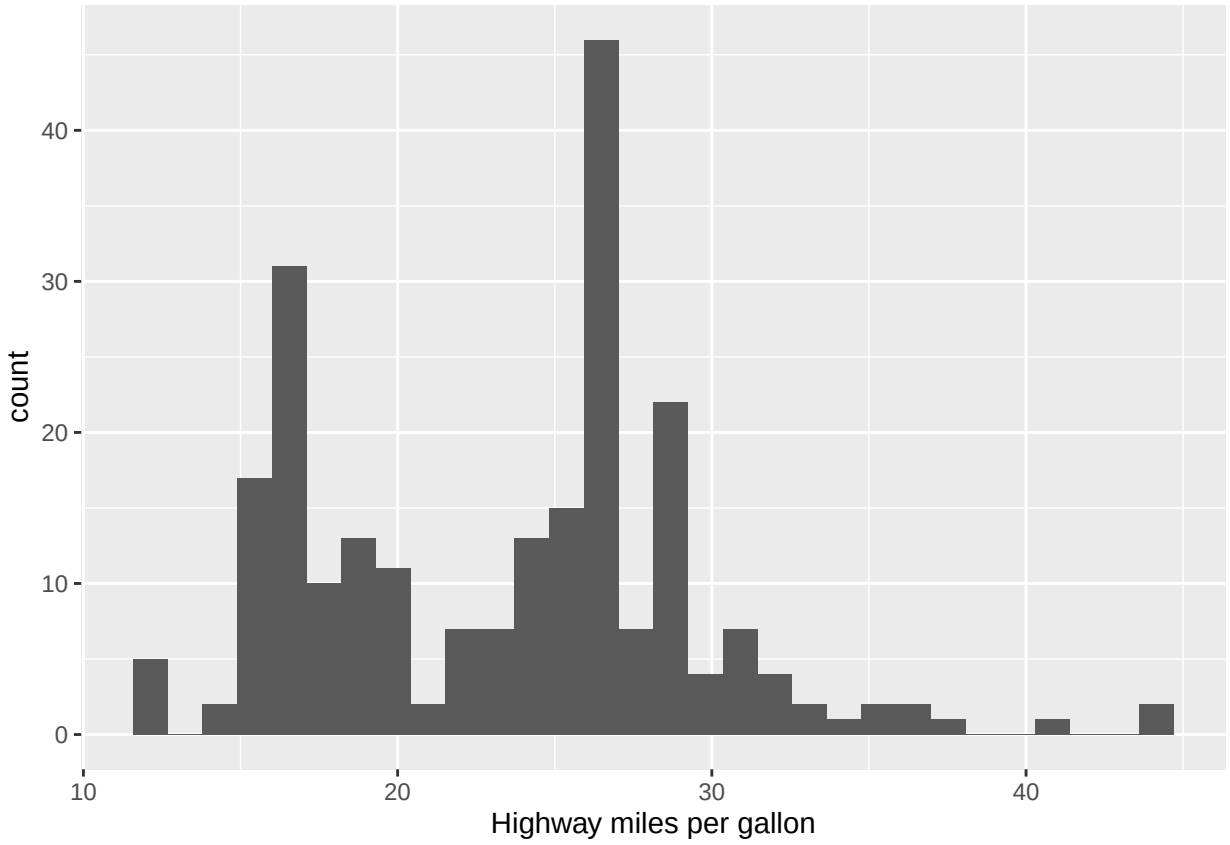
Answer: The first situation is predictive because its aim is to make a future prediction; the second situation is inferential because it aims to analyze the relationship between the predictors and the response.

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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.

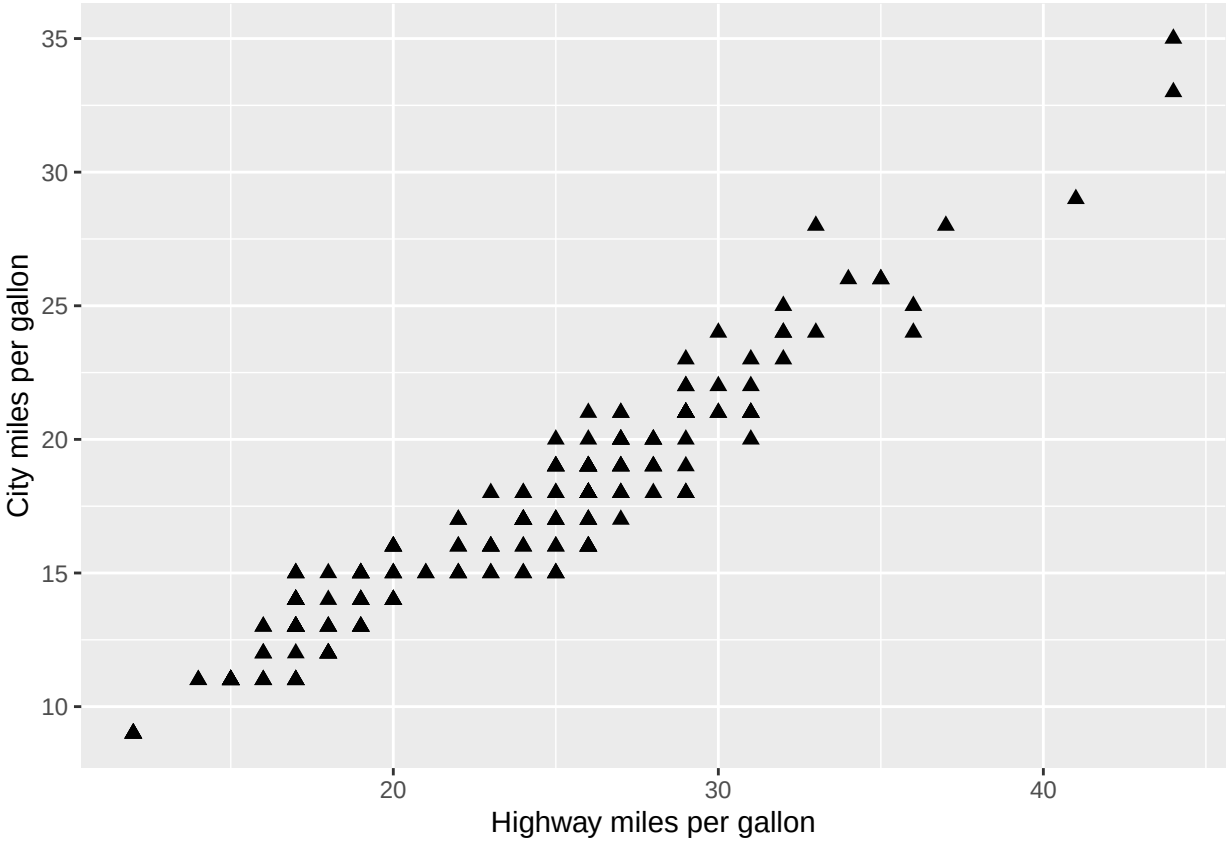
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



From the histogram, we can see a bimodal distribution that has a right skewed shape.

Exercise 2:

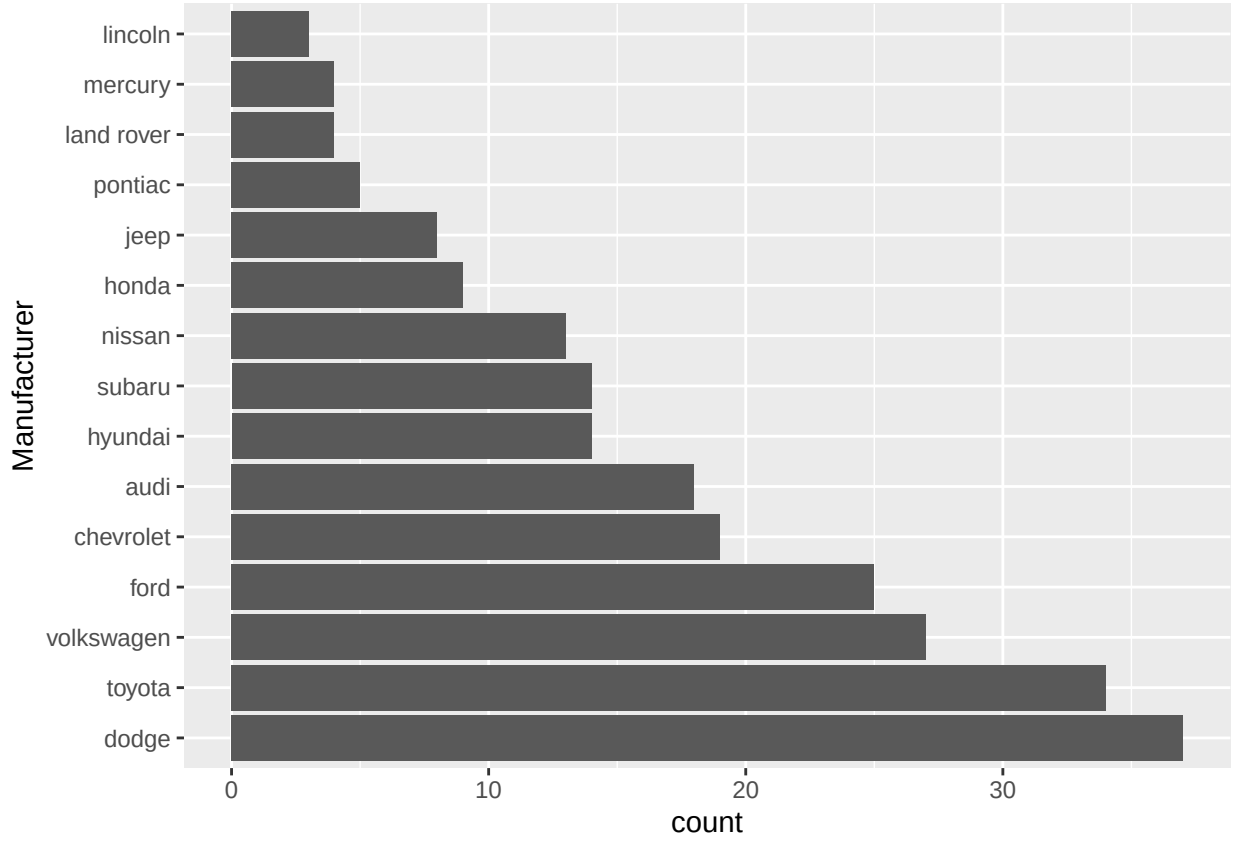
Create a scatterplot. Put hwy on the x-axis and cty on the y-axis. Describe what you notice. Is there a relationship between hwy and cty? What does this mean?



From the scatterplot, we can observe a strong positive linear correlation between hwy and cty.

Exercise 3:

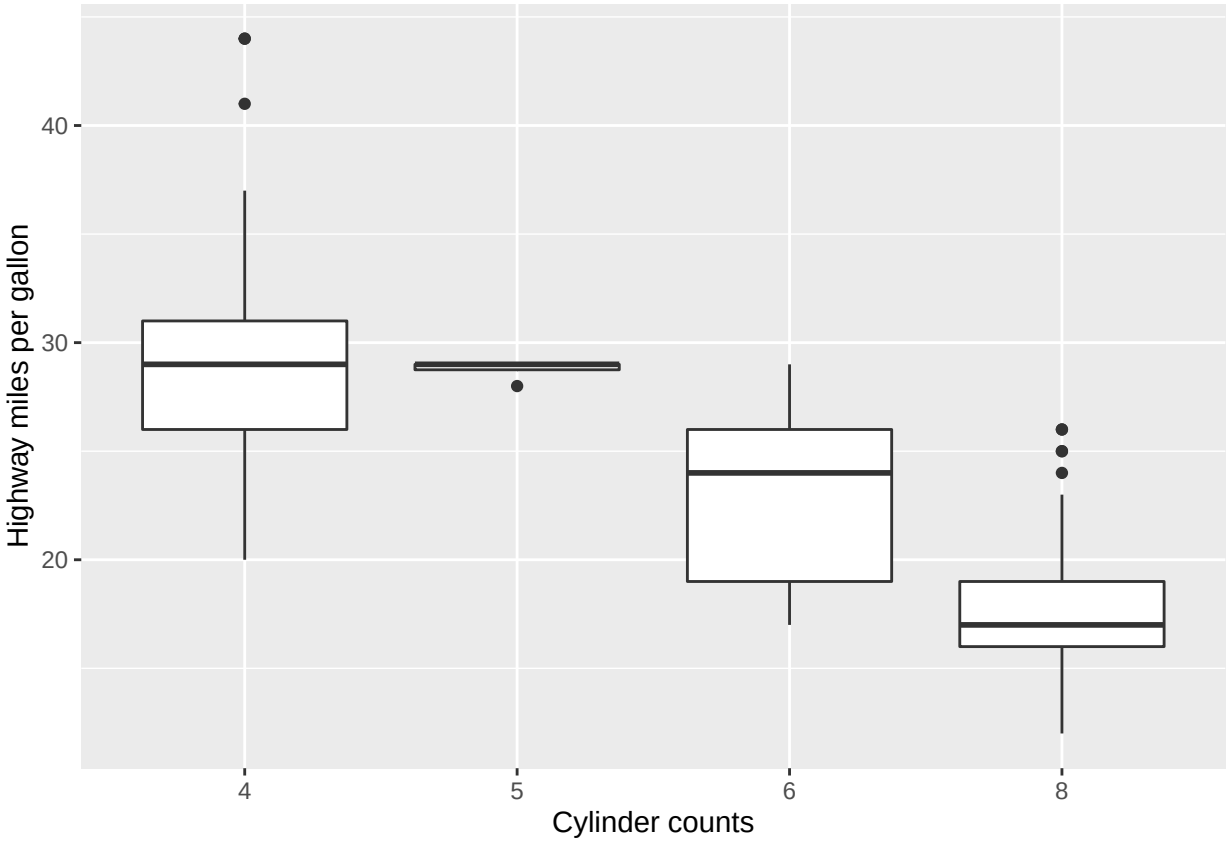
Make a bar plot of manufacturer. Flip it so that the manufacturers are on the y-axis. Order the bars by height. Which manufacturer produced the most cars? Which produced the least?



Dodge produced the most cars; Lincoln produced the least.

Exercise 4:

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



There is a pattern where the cylinder counts and highway miles per gallon have an inverse relationship.

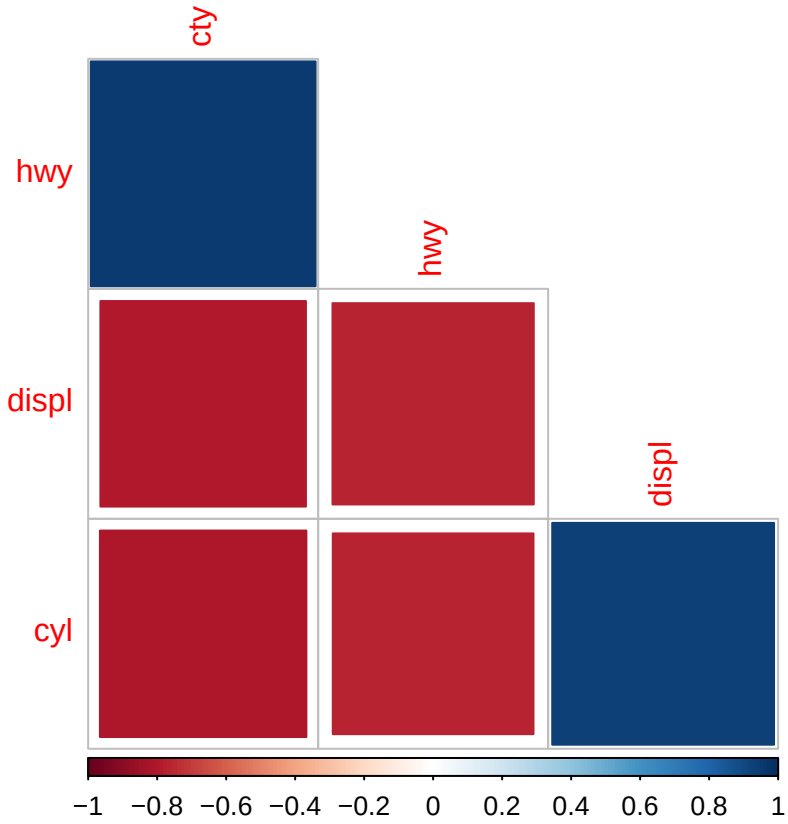
Exercise 6:

Use the corrplot package to make a lower triangle correlation matrix of the mpg dataset.

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

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

```
## corrplot 0.92 loaded
```



Highway miles per gallon and city miles per gallon are positively correlated, which makes sense because cars that burn more gas in the city would often burn more gas on highway as well. Also, number of cylinders and engine displacement are positively correlated, which is also expected because more cylinders lead to more volume. All other pairs: engine displacement & highway miles per gallon, engine displacement & city miles per gallon, number of cylinders & city miles per gallon, number of cylinders & highway miles per gallon, are negatively correlated. These are all surprising to me because I always thought that cars that have more cylinders or larger engine displacements would burn more gas faster.