PSTAT 131 Homework 1

Tammy Truong

2022-03-30

Machine Learning Main Ideas

Question 1:

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

Response From lecture 1, we know that supervised learning consists of accurately predicting future response given predictors, understanding how predictors affect response, finding the best model for response given predictors, and assessing the quality of our predictions and/or estimation. Unsupervised learning only takes the input variables with no output variables involved, concluding based on unlabeled data. While supervised learning has response variables, unsupervised does not. Supervised learning can take on the input and output variables and connect the relationship while unsupervised may only understand through variables and observations.

Question 2:

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

Response A regression model predicts quantitative variables with continuous values and a classification model predicts qualitative variables with categorical values.

Question 3:

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

Response Regression ML problems have two commonly used metrics with MSE/RMSE and R-squared while classification ML problems have error rates and F-1 score.

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.

Response Descriptive models: Models that best emphasize a trend in the data through a visual, such as a line in a scatterplot.

Inferential models: Models that test theories to understand the relationship between variables

Predictive models: Models that predicts an exact outcome with minimal 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?

Response
- Mechanistic, also known as parametric, uses a theory to estimate the outcome in the real world.
- Empirically-driven, also known as nonparametric, requires a large amount of observations and low assumptions about the data.
- These models differ through the amount of information known regarding the data, such as the assumption of f. However, empirically-driven models are more flexible than mechanistic models. These two models are similar in a way that they both have an issue with overfitting.

In general, is a mechanistic or empirically-driven model easier to understand? Explain your choice.
- Mechanistic models are easier to understand because we can have assumptions of the model so we can easily interpret the parameters.

Describe how the bias-variance tradeoff is related to the use of mechanistic or empirically-driven models.
- Since empirically-driven models have higher flexibility than mechanistic models, we know that it would have a higher variance and less bias. Mechanistic models are less flexible so it has a lower variance and higher 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.

Response The first question is predictive because we are applying the data into predicting the outcome. The second question is inferential because we are understanding the relationship between variables by adding a personal contact with the candidate option.

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:

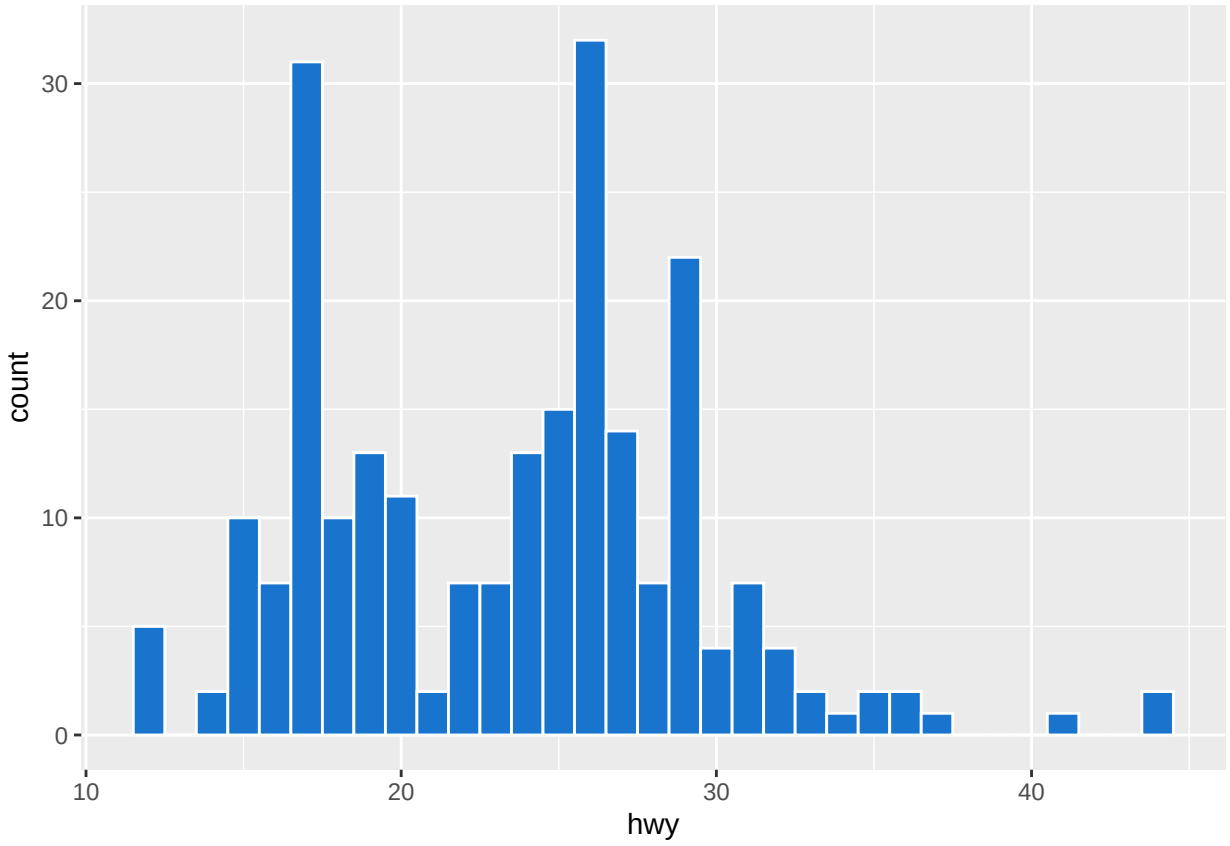
generating questions about data visualize and transform your data as necessary to get answers use what you learned to generate more questions A couple questions are always useful when you start out. These are "what variation occurs within the variables," and "what covariation occurs between the variables."

You should use the tidyverse and `ggplot2` for these exercises.

Exercise 1:

Histogram of the `hwy` variable in mpg.

```
# creating histogram with ggplot
histogram <- ggplot(mpg, aes(x=hwy)) + geom_histogram(binwidth = 1, color = 'white',
                                                      fill = 'dodgerblue3')
histogram
```

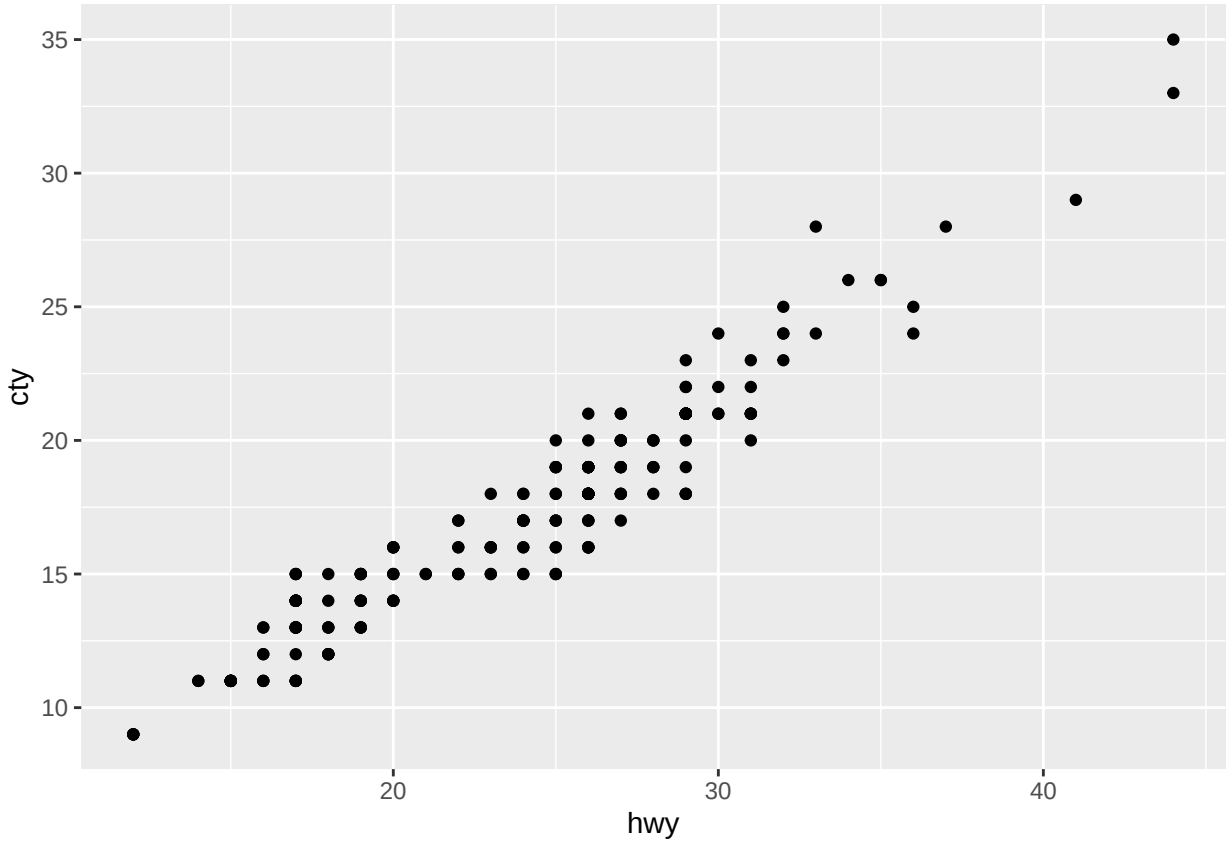


From above, the histogram shows that there are spikes around 15-20 and 25 mpg on the highway whereas there are less cars with over 30 mpg on the highway.

Exercise 2:

Creating a scatterplot with `hwy` *on the x-axis an* `cty` *on the y-axis.*

```
# creating a scatter plot with ggplot
scatter = ggplot(data = mpg, aes(x = hwy, y = cty)) + geom_point()
scatter
```

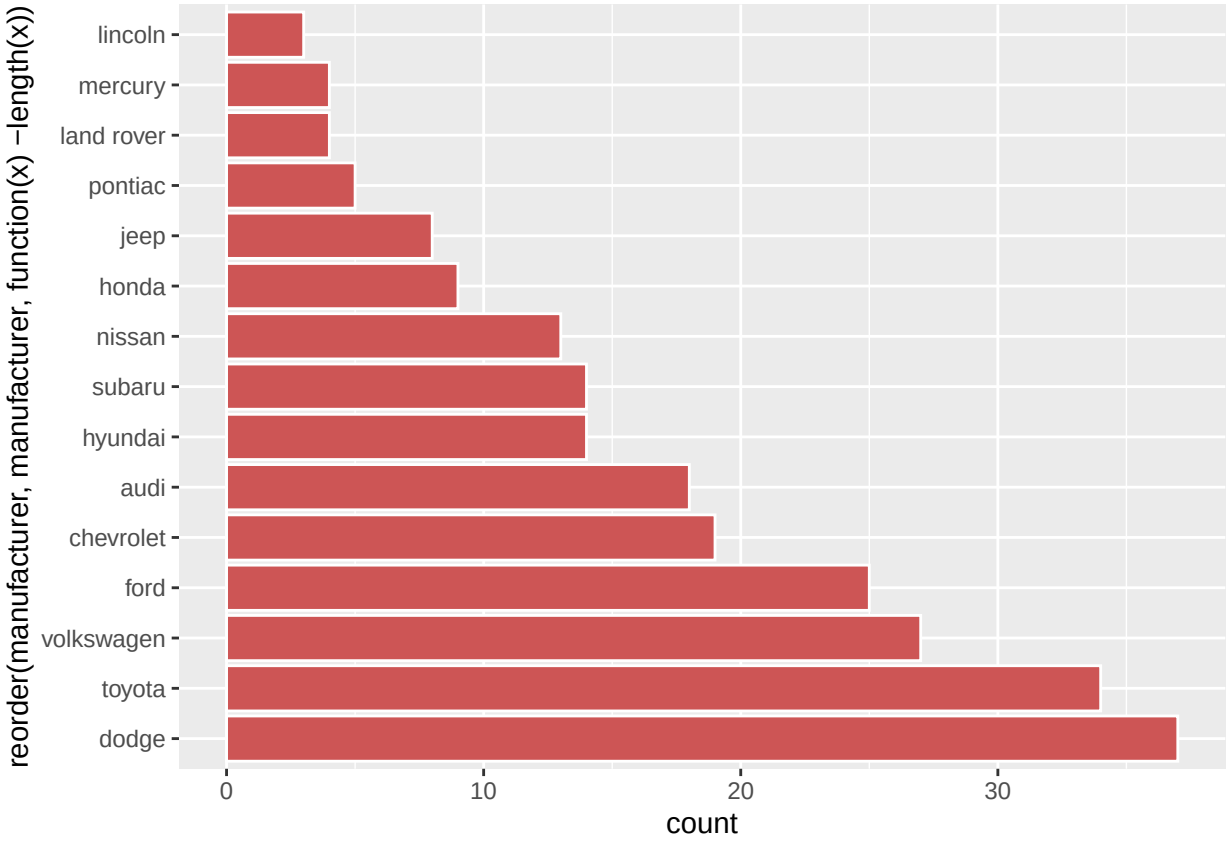


The scatter plot shows that there is a positive correlation between the variables highway and city, implying that there is a relationship where cars that drive many mpg on the highway also drive around the same amount of mpg.

Exercise 3:

Barplot of `manufacturer`

```
# reordering with manufacturer by height
ggplot(mpg, aes(x = reorder(manufacturer, manufacturer,
                            function(x) - length(x)))) +
  geom_bar(color = 'white', fill = 'indianred3') + # bar plot
  coord_flip() # flipping manufacturer as y-axis
```

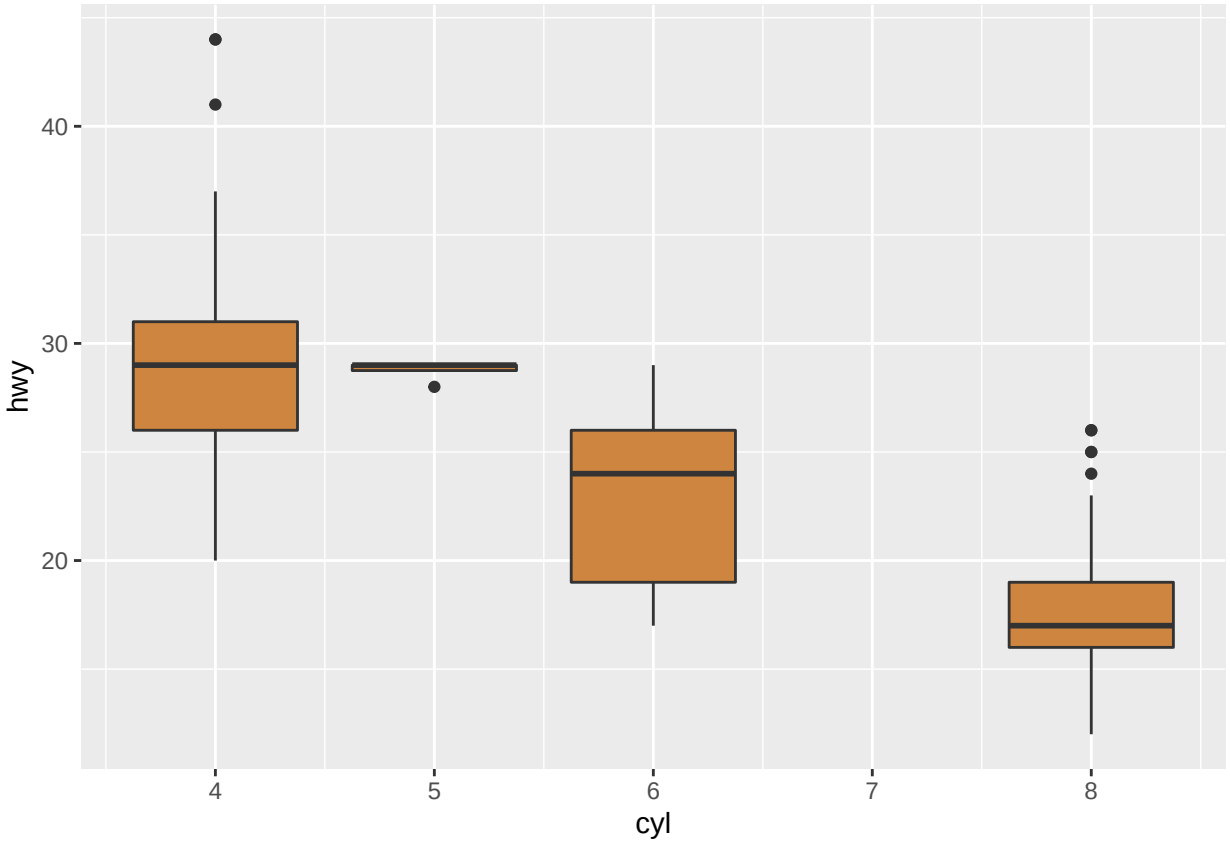


From our bar plot above, we see that Lincoln produced the least cars and Dodge produced the most cars.

Exercise 4:

Box plot of `hwy` *grouped by* `cyl`

```
# creating box plot of hwy and grouping it with cyl
box <- ggplot(data = mpg, aes(x = cyl, y = hwy)) +
  geom_boxplot(aes(group = cyl), fill = 'tan3')
box
```

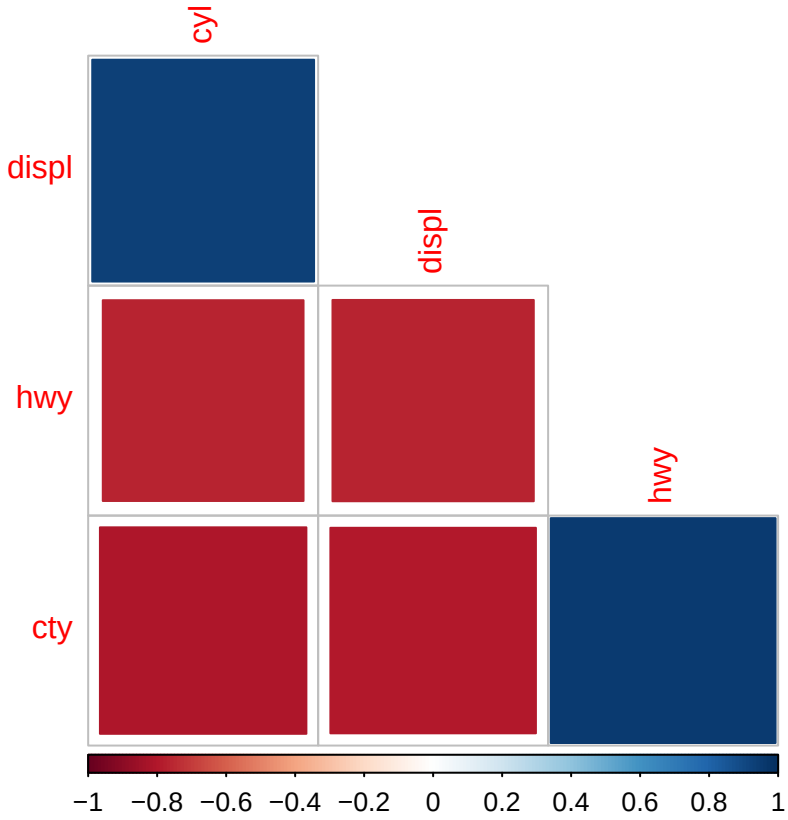


The boxplot shows a pattern of increasing mpg on highways as there are less cylinders.

Exercise 5:

Lower triangle correlation matrix of `mpg` *with numerical variables* `hwy`, `cyl`, `displ`, `cty`

```
# selecting variables without categorical values
var <- c('hwy', 'cyl', 'displ', 'cty')
M <- cor(mpg[var])
corrplot(M, method = 'square', order = 'FPC', type = 'lower',
         diag = FALSE) # creating the diagonal plot
```



The color bar shows that red implies a negative correlation and blue implies a positive correlation. It shows that `hwy` and `displ`, `hwy` and `cyl`, `cty` and `cyl`, `cty` and `displ` are negatively correlated while `cty` and `hwy`, `displ` and `cyl` are positively correlated. These relationships are expected since we saw the relationship between `hwy` and `cyl` in problem 4, so it is not surprising that the amount of cylinders would decrease mpg in the city and highway. Similarly, displacement would act the same as cylinders, causing a negative correlation between mpg in highway and city. This can be said in the other way as well.