Homework 1

PSTAT 131

Contents

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?

From page #26 of "An Intro to Statistical Learning", supervised learning is where for each observation of the predictor measurements (x1, x2, etc.), there exists an associated response measurement (y). Whereas in unsupervised learning, we have no response measurement, y. In this case, we are unable to fit a linear regression model as there is no response to predict unlike supervised learning.

Question 2:

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

As seen in the lecture slides, "Intro to ML", a classification model is when your outcome/result is a categorical variable (Yes/No for example). A regression model is when your outcome/result is a continuous variable (Price, Speed for example).

Question 3:

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

In the lecture slides, "Intro to ML", Mean Squared Error (MSE), and Mean Absolute Error (MAE) are two common error metrics for regression ML problems.

Accuracy, and precision are two common error metrics for classification ML problems.

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:
- Inferential models:
- Predictive models:

From the lecture “Intro to ML”, Descriptive models are when a model is chosen to best visually emphasize a data trend, like using a line on a scatterplot.

For inferential models, the aim is to test theories and possibly make casual claims. Also, to state a relationship between outcome and predictors.

In predictive models, the goal is to predict “y” with a minimum reducible error, and try to understand which combo of features fits best.

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.

From the lecture “Intro to ML”, a mechanistic model assumes a parametric form for f whereas an empirically-driven model makes no assumptions about f, and is more flexible by default. Also, extrapolation is only possible with mechanistic models. But, a mechanistic model can also gain more flexibility like an empirical model, by adding parameters. Both mechanistic and empirical models are designed to predict the output. In that sense both models are similar because both are trying to achieve the same goal.

A mechanistic model is generally easier to understand because these models consist of input and output variables in a mathematical equation which can be understood by humans. On the other hand, an empirically-driven model makes no assumptions about f, and does not represent the input-output in terms of an equation that can be understood by humans (i.e., an empirically-driven model can seem like a mysterious black-box which magically produces the correct answer for at least the training data).

The bias-variance trade-off states that if we try to reduce the bias (i.e., the error between the predicted values and the actual values) then we may increase the sensitivity of the model to the training data (i.e., the model is over-fitted to the training data and will generate a larger error with the test data). The bias-variance trade-off is related to the use of mechanistic or empirically-driven models in the sense that mechanistic models generally have a lower variance and a higher bias when compared to empirically-driven models.

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.

1). Predictive; in this question, we are trying to predict the likelihood that the voter will vote in favor of the candidate with as much certainty as possible.

2). Inferential; here, we may be trying to make casual claims, let's say that the voter did get more support if they had personal contact. Then we can make a loose claim here, or state a relationship between the two.

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

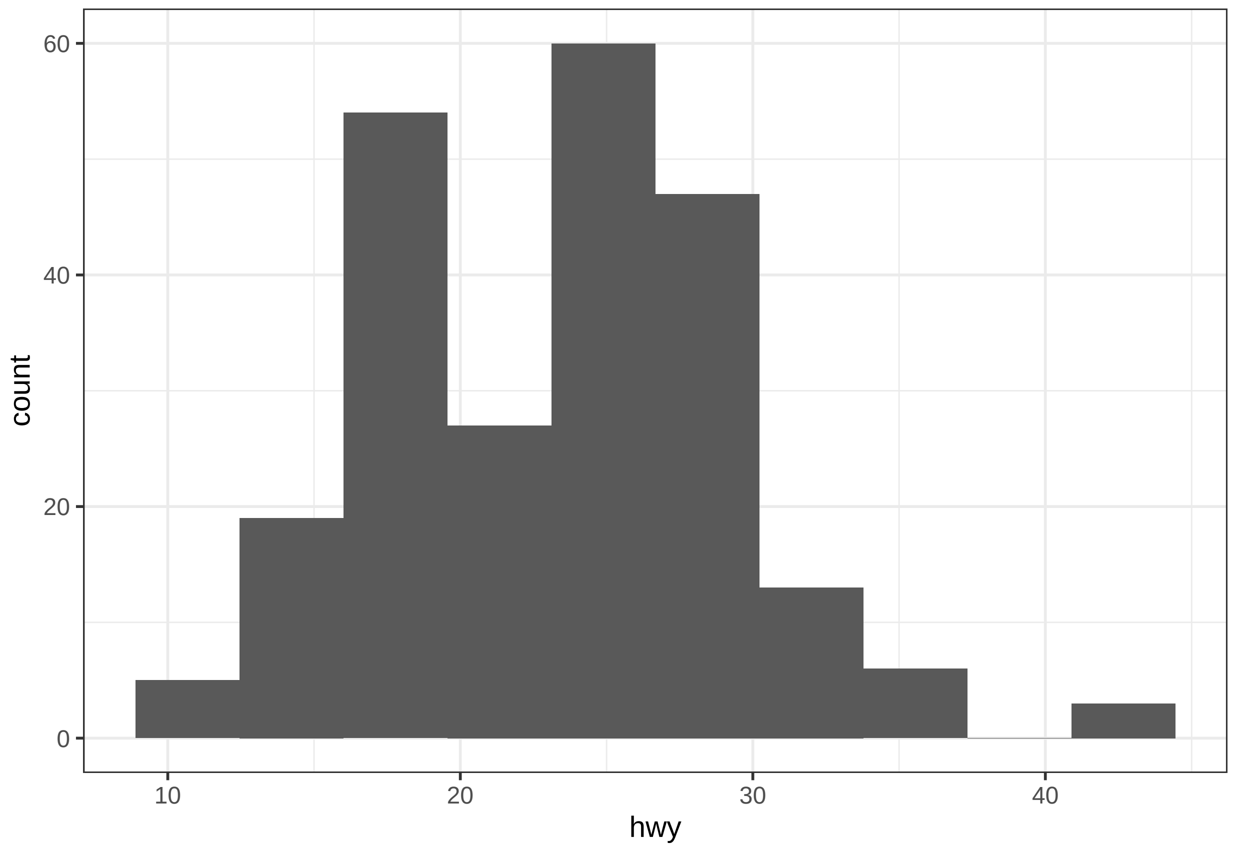
A couple questions are always useful when you start out. These are "what variation occurs within the variables," and "what covariation occurs between the variables."

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.

```r
mpg %>%
ggplot(aes(x=hwy)) +
geom_histogram(bins=10) +
theme_bw()
```

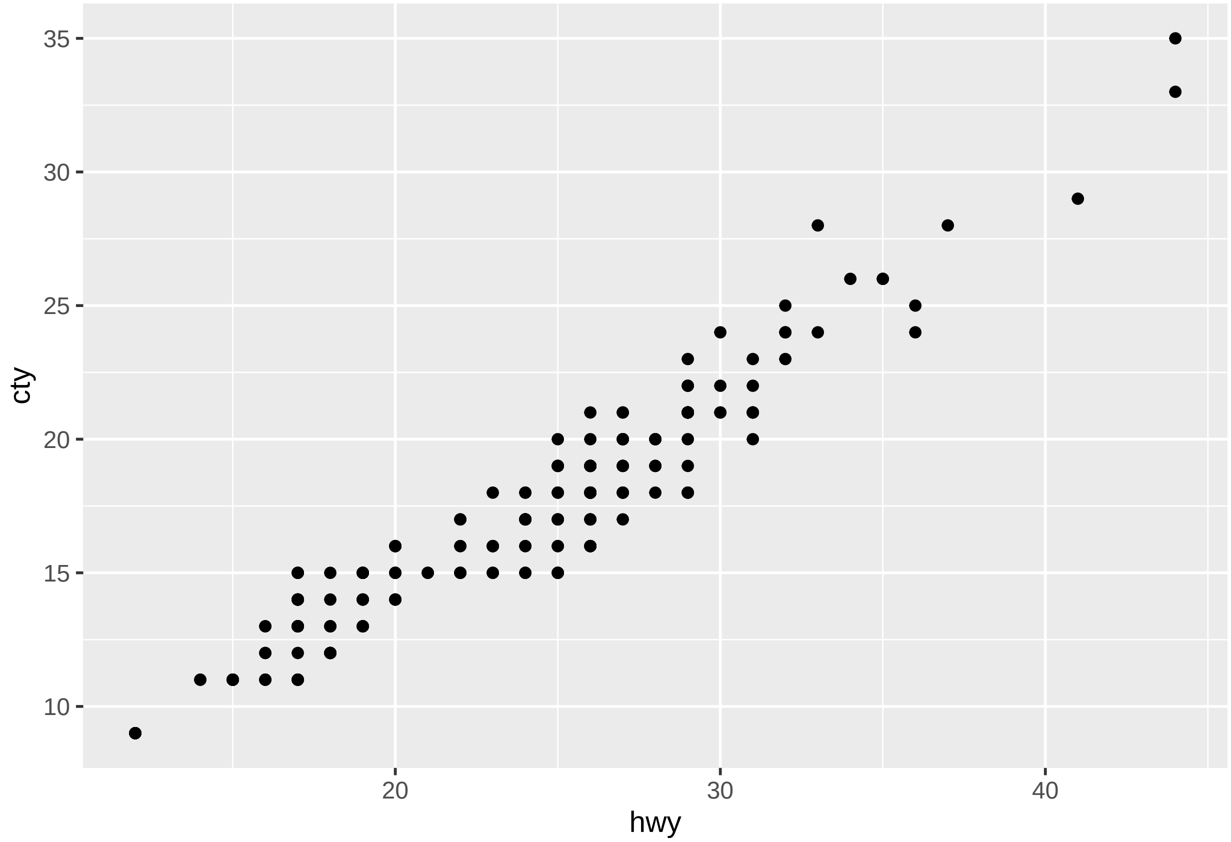


I notice that there are very few hwy miles/gallon that are over 30, and few values less than 15. Most of the hwy miles per gallon are seen to be from 16-30 in this dataset.

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()
```

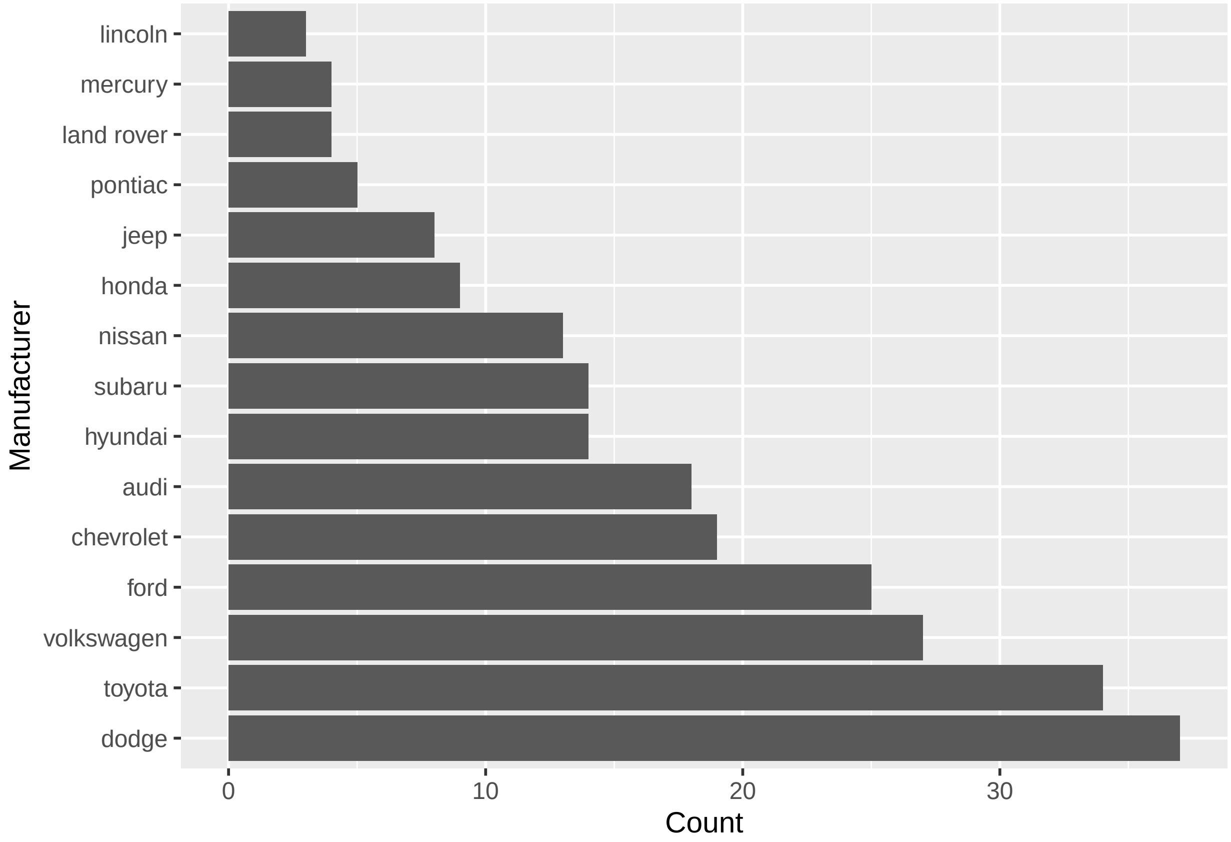


As we can see from the scatterplot, there is a positive relationship between hwy and cty. This tells us that the more miles per gallon a car drives on a highway, the more miles per gallon that car is able to drive in the city.

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?

```
library(forcats)
mpg %>%
ggplot(aes(x=fct_infreq(manufacturer))) +
labs(x='Manufacturer', y='Count') +
coord_flip() +
geom_bar()
```



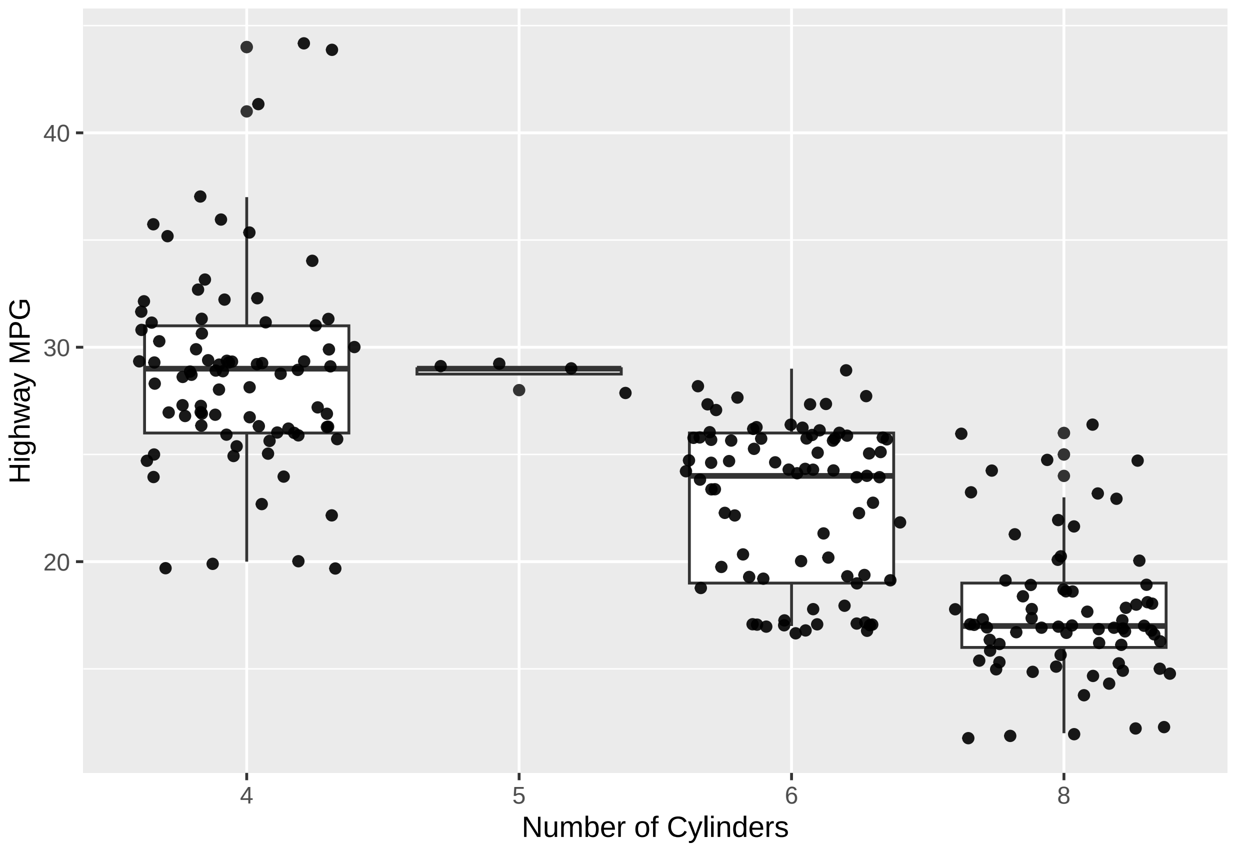
Dodge produced the most number of cars, and Lincoln produced the least number of cars.

Exercise 4:

Make a box plot of `hwy`, grouped by `cyl`. Use `geom_jitter()` and the `alpha` argument to add points to the plot.

Describe what you see. Is there a relationship between `hwy` and `cyl`? What do you notice?

```
mpg %>%
ggplot(aes(x=factor(cyl), y=hwy)) +
labs(x='Number of Cylinders', y='Highway MPG') +
geom_boxplot() +
geom_jitter(alpha=0.9)
```



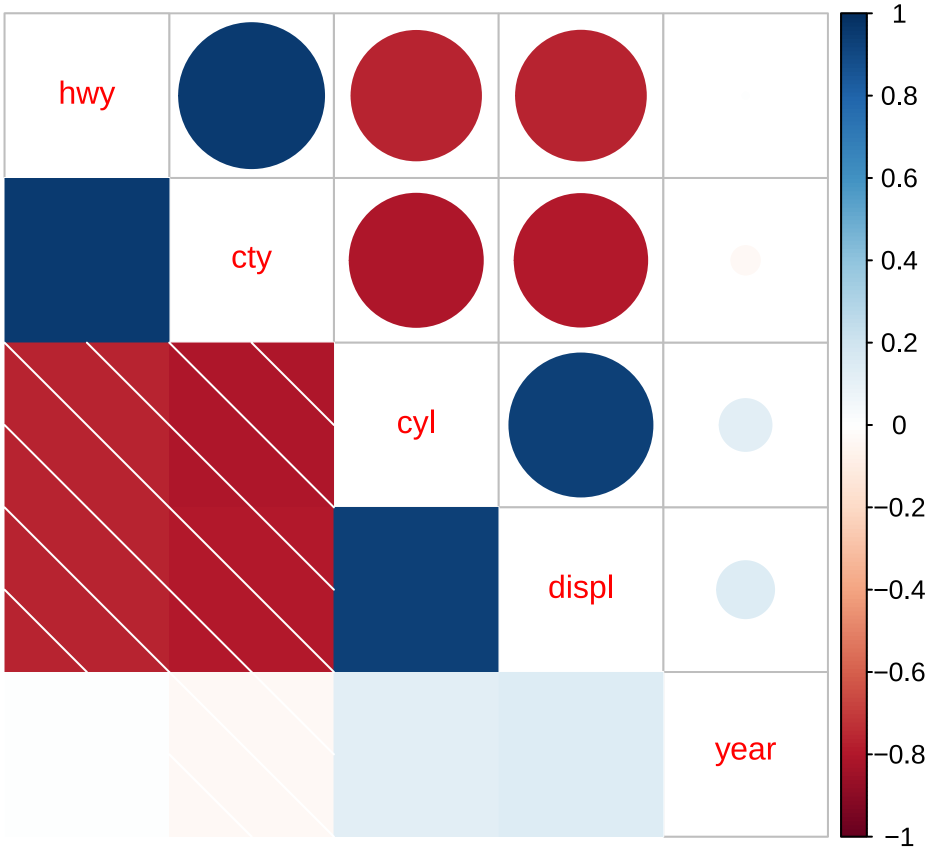
There is a negative relationship between hwy and cyl. That is, as the number of cylinders in the car increases, the highway MPG decreases. We also see that there are very few cars with 5 cylinders. The most common number of cylinders is 6, followed by 4, and then 8. There are no cars with 7 cylinders, and 6 cylinder cars have the most similar hwy (mpg) values. We see that there are 5 high outliers for 4 cylinder cars that have a hwy over 40. These cars could be hybrid cars, like a Honda Civic (42 hwy/mpg).

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.mixed(order='AOE', lower='shade')
```



As expected, hwy and cty were very positively correlated. This is because the more miles a car is able to drive per gallon on the highway should be positively correlated with the miles a car is able to drive per gallon in the city. In addition, cyl and displ were also very positively correlated. After understanding the variables, this makes sense because the more cylinders a car has, the more engine displacement (in litres), or a higher volume of engine a car would have. The negative correlation between cty and cyl also makes sense because heavier vehicles usually aren't able to drive as many miles/gallon. Same goes for hwy and cyl. Also there was no correlation between year and highway, which did surprise me a little as I expected a slight positive correlation. This means that cars didn't really become more efficient from 1999-2008 in terms of more miles driven per gallon.