131 Homework 1

Tonia Wu

3/29/2022

Q1. *Define supervised and unsupervised learning. What are the difference(s) between them?*

- Supervised learning involves modelling to predict an output based on input(s). This requires training data from which the model will learn.
- Unsupervised learning involves inputs and no outputs so that we can learn about the data and potentially discover patterns.
- For each input, a supervised model will have an output. An unsupervised model will lack associated responses and be unable to fit a linear regression model.

Q2. *Explain the difference between a regression model and a classification model, specifically in the context of machine learning.*

- Regression models involve continuous, quantitative output (such as expected gas mileage)
- Classification models involve qualitative output (such as a yes/no result)

Q3. *Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.*

(from the textbook)

- Regression: age, income
- Classification: marital status, brand of product purchased

Q4. *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*: describe the data of interest
- *Inferential models*: allow testing hypothesis to see if results are generalizable
- *Predictive models*: predicts the future using data collected in the past

Q5. *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?

- A mechanistic model takes a relationship or theory and imposes it on the data (from lecture)
- An empirically-driven model looks at the data and sees what best fits it (from lecture)
- Similarities: both have a tendency to overfit
- Differences (from lecture):
    - Mechanistic models assume a parametric form (though it will not match the true unknown

f)

- Empirical models make no assumptions about f and require a larger number of observations; it is also by default more flexible than the mechanistic model

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

I'm not totally sure why, but I understand the empirical model less. Perhaps that's because it is | inherently more "unknown" than the mechanistic one in that the empirical model tries to develop a theory that doesn't (yet) exist.

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

In real-life, we do not know the true f, but we must still consider the bias-variance tradeoff. Depending on the linearity of the true f, a mechanistic or empiric model may be better depending on if the model benifits from more flexibility or not.

Q6. *A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions:*

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

b. *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.

- a: as we are using past data to predict the future, this involves a predictive model.
- b: since we're interested in the relationship between the inputs and the output, this involves an inferential model.

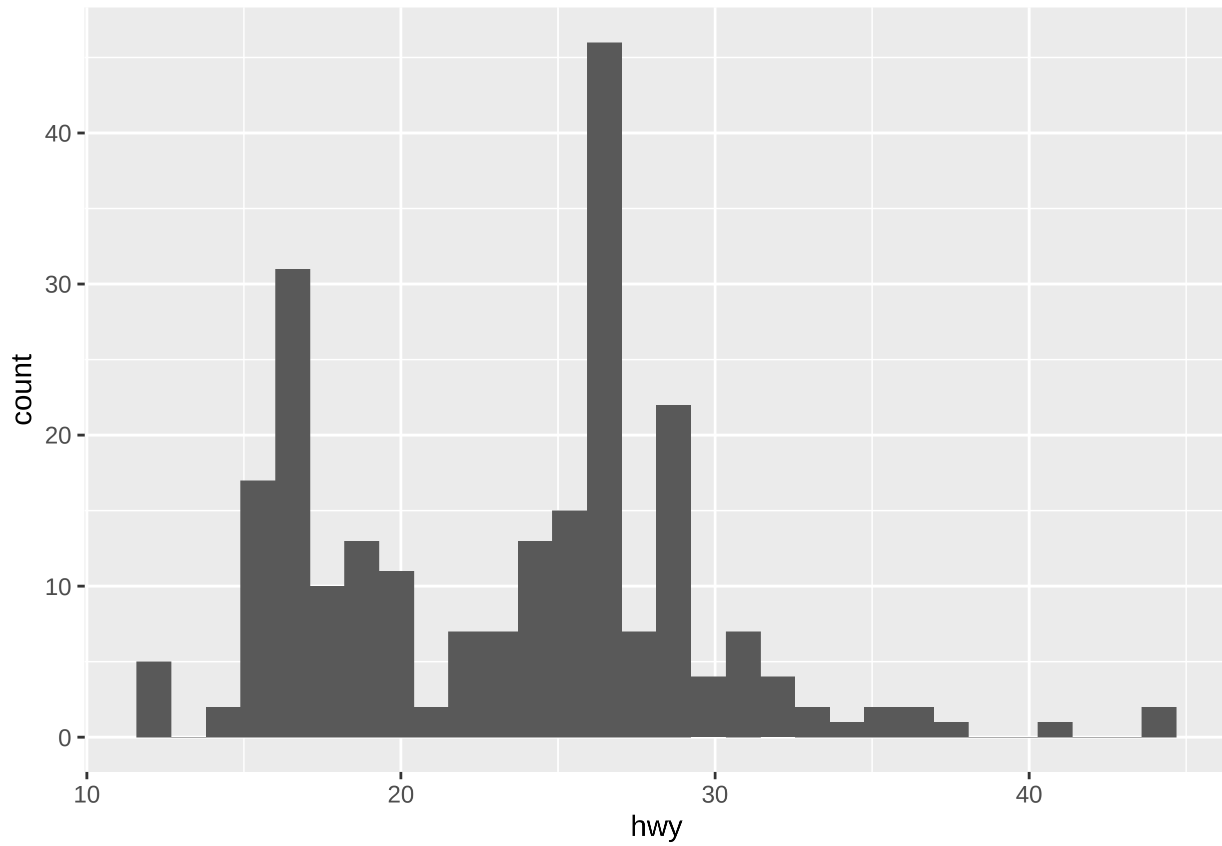
Exploratory Data Analysis: mpg

E1. *We are interested in highway miles per gallon, or the hwy variable. Create a histogram of this variable. Describe what you see/learn.*

Looks like we're dealing with nonsymmetric bimodal data:

```
ggplot(data = mpg, aes(hwy)) + geom_histogram()
```

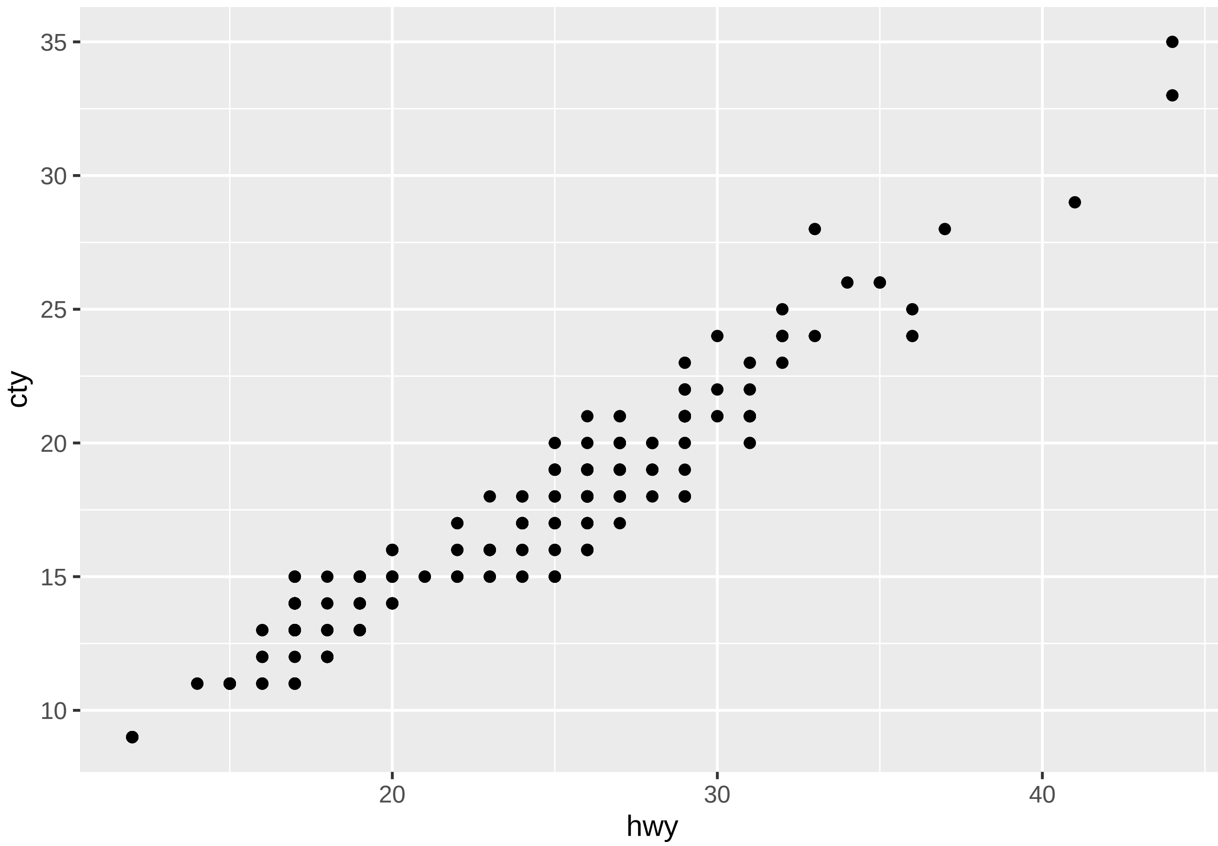
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



E2. *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?*

There is an incredibly strong positive relationship between hwy (highway mpg) and cty (city mpg). Thus, as hwy increases, cty increases.

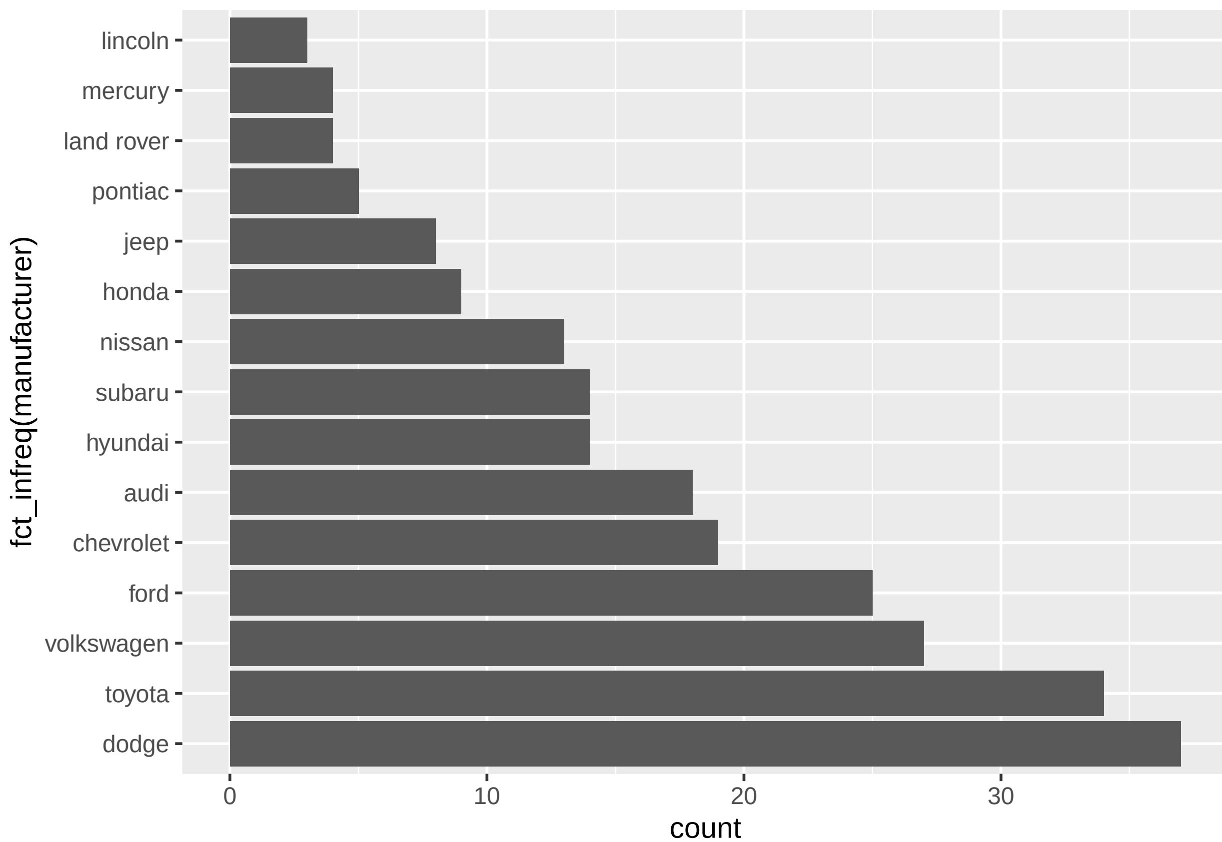
```
ggplot(mpg, aes(x = hwy, y = cty)) + geom_point()
```



E3. *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?*

Lincoln produced the least cars and Dodge produced the most:

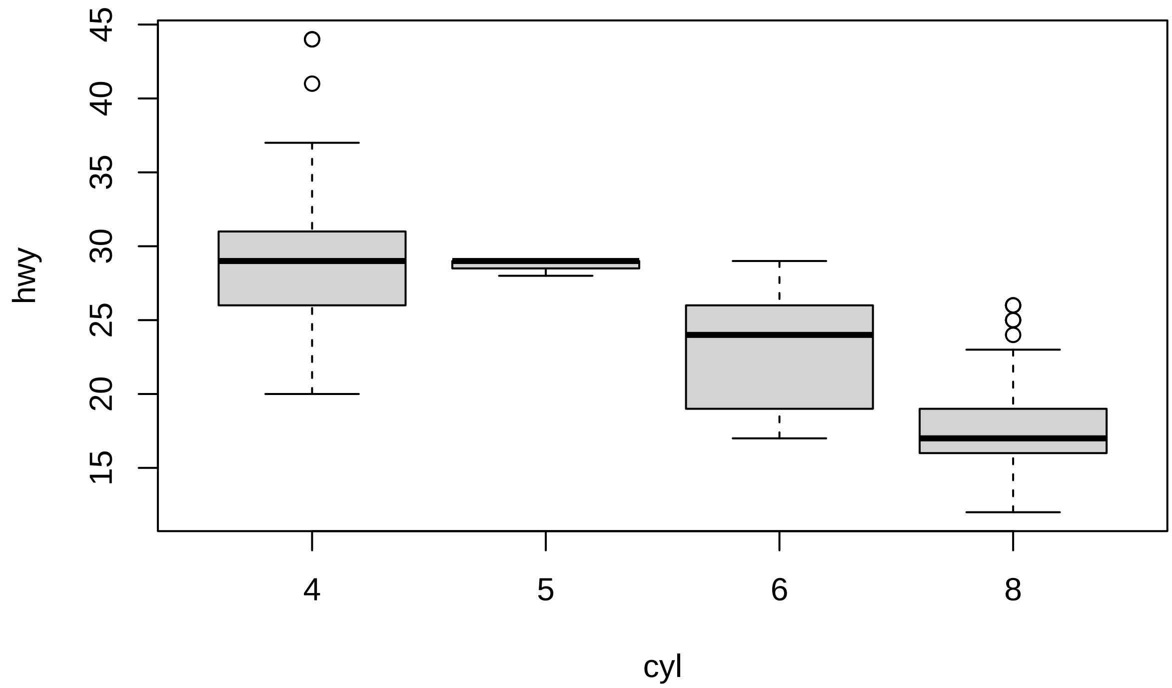
```
ggplot(data = mpg, aes(x = fct_infreq( manufacturer))) +
  geom_bar(stat = 'count') +
  coord_flip()
```



E4. *Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?*

Looks like highway mpg decreases as the number of cylinders increases.

```
hwy <- mpg$hwy
cyl <- mpg$cyl
boxplot(hwy ~ cyl)
```



E5. *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?

Strong positive: displacement & cylinders, highway mpg & city mpg

Neither of these are surprising since I'd expect highway and city mpg to be quite similar, and engine displacement depends on cylinders in the first place.

Strong negative: highway mpg & displacement, city mpg & displacement, city mpg & cylinder, highway mpg & cylinder

For these, I tried Googling what effect cylinders have on gas usage, but came away with no clear answer. So I am surprised by this result, but mainly because I don't understand cars.

Weak positive correlation: year & displacement, cylinder & year

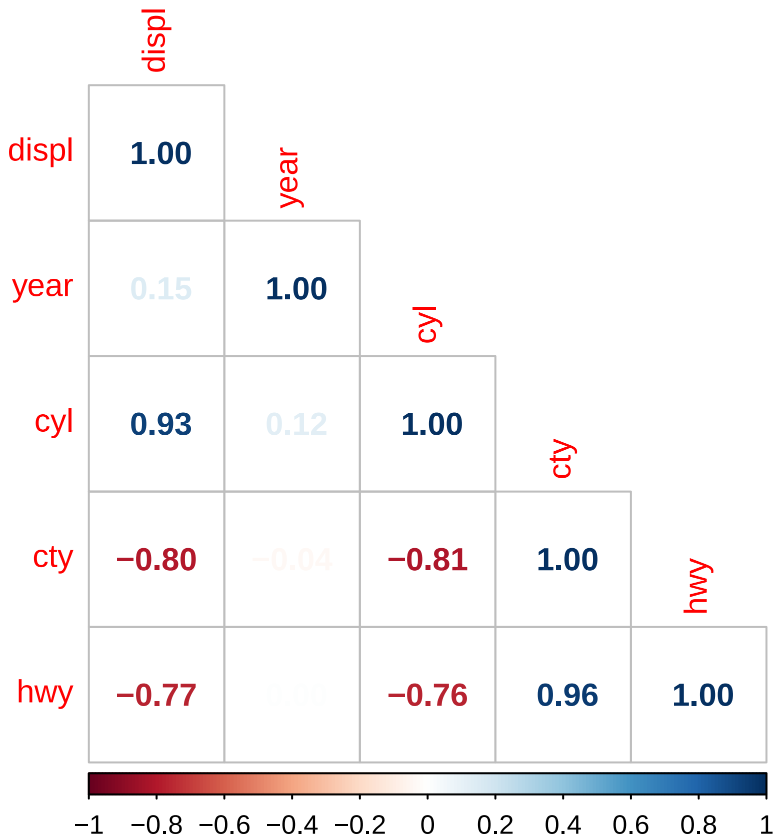
Little to no correlation: city & year, highway mpg & year

The lack of correlation between the mpgs and year was surprising to me because I thought mpg would improve over time.

```
# select only numeric variables
# source: statistics globe
mpg_2 <- select_if(mpg, is.numeric)
mpg_2
```

```
## # A tibble: 234 x 5
##    displ  year   cyl   cty   hwy
##    <dbl> <int> <int> <int> <int>
##  1   1.8  1999     4    18    29
##  2   1.8  1999     4    21    29
##  3   2    2008     4    20    31
##  4   2    2008     4    21    30
##  5   2.8  1999     6    16    26
##  6   2.8  1999     6    18    26
##  7   3.1  2008     6    18    27
##  8   1.8  1999     4    18    26
##  9   1.8  1999     4    16    25
## 10   2    2008     4    20    28
## # ... with 224 more rows
```

```
M <- cor(mpg_2)
corrplot(M, method = 'number', type = 'lower')
```


displ
year
cyl
cty
hwy
1.00
0.15
1.00
0.93
0.12
1.00
−0.80
−0.04
−0.81
1.00
−0.77
−0.76
0.96
1.00
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1