Homework 1

Lauren Huang

2022-09-24

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 learning is when a model can estimate or predict an output because the input data has the correct labels to tell how accurate the model is during training/testing.In unsupervised learning, we are not given "the correct labels", or the supervised output. We have no way to tell the numeric range of accuracy of the model, we can only infer the relationships between data.(learned from lecture 1)

Question 2: Explain the difference between a regression model and a classification model, specifically in the context of machine learning.

Answer: Regression is used when we predict outcomes that are quantitative and continuous. (ex. price, blood pressure) In contrast, classification is used to predict outcomes that are qualitative or discrete/categorical. (ex. survived/died, spam/not spam) (examples from lecture 1)

Question 3: Name two commonly used metrics for regression ML problems. Name two commonly used metrics for classification ML problems.

Answer: 2 common metrics for regression ML problems are training mean squared error and testing mean square error. 2 common metrics for classification ML problems are training error rate and testing 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.

a. Descriptive models: best for a visual representation or emphasis of a data trend. (from lecture 1)
b. Inferential models: helps us distinguish the features/predictors that are most significant to the outcome. This method tests features against one another, and also helps clarify the relationship between features.
c. Predictive models: goal is to predict outcome with the minimum reducible error (from lecture 1), often uses past behavior to help predict the future.

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.

a. Define mechanistic. Define empirically-driven. How do these model types differ? How are they similar?

Answer: Mechanistic refers to the idea that we assume data follows a certain statistical form or distribution. Empirically-driven means that we do not make any assumptions about the trends data follows. When using a mechanistic, we may never match the true unknown value of the data. With empirically-driven, we require a many more observations than we would working with a mechanistic model. An empirically-driven model is more flexible by default than a mechanistic model, though we can add parameters to a mechanistic model to help it become more flexible. One similarity between the two models is the possibility of overfitting, where the model is too sensitive and picks up on random noise that might not be consistently part of the trend. (ideas from lecture 1)

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

Answer: After learning about both types of models, I think the logic behind both are intuitive. However, from a perspective of someone with no statistical background, I would say mechanistic models are easier to understand. The reason being that if someone has no knowledge of statistical methods, we can begin learning about data analysis and manipulation by studying the most commonly used statistical distributions. After building a foundation in statistics and learning these properties, then we can move onto the more open-ended approach of handling data without making certain assumptions. (Maybe that's also why the statistics major curriculum introductory courses with mechanistic-type learning?)

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

Answer: Mean square error (MSE) is a way to evaluate the accuracy of a model, and involves the concept that variance and bias are inversely related. In more flexible models, variance increases and bias decreases. On the other hand, when the model is more restricted, variance decreases and bias increases. One shared characteristic of mechanistic and empirically-driven models is the aspect of flexibility. Empirically-driven models are generally flexible by default while mechanistic models may not be very flexible at first. But we can still add parameters to make it more flexible. The challenge that intersects between MSE and flexibility is the bias-variance trade-off. It can be easy to achieve a model that has either low variance or low bias, but our goal to find one that gives both low variance and low 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. Classify each question as either predictive or inferential. Explain your reasoning for each.

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

Answer: Predictive, because voter profile/data can be considered our input, and one characteristic of predictive models is using past knowledge or data to help predict the outcome (not focused on hypothesis tests).

b. How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?

Answer: Inferential, because this situation has not happened yet, it is hypothetical and a theory. Inferential models aim to test theories and decipher a possible relationship between predictors (ex. personal contact with candidate) and the outcome (who the voter votes for).

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.

```
library(tidyverse)
```

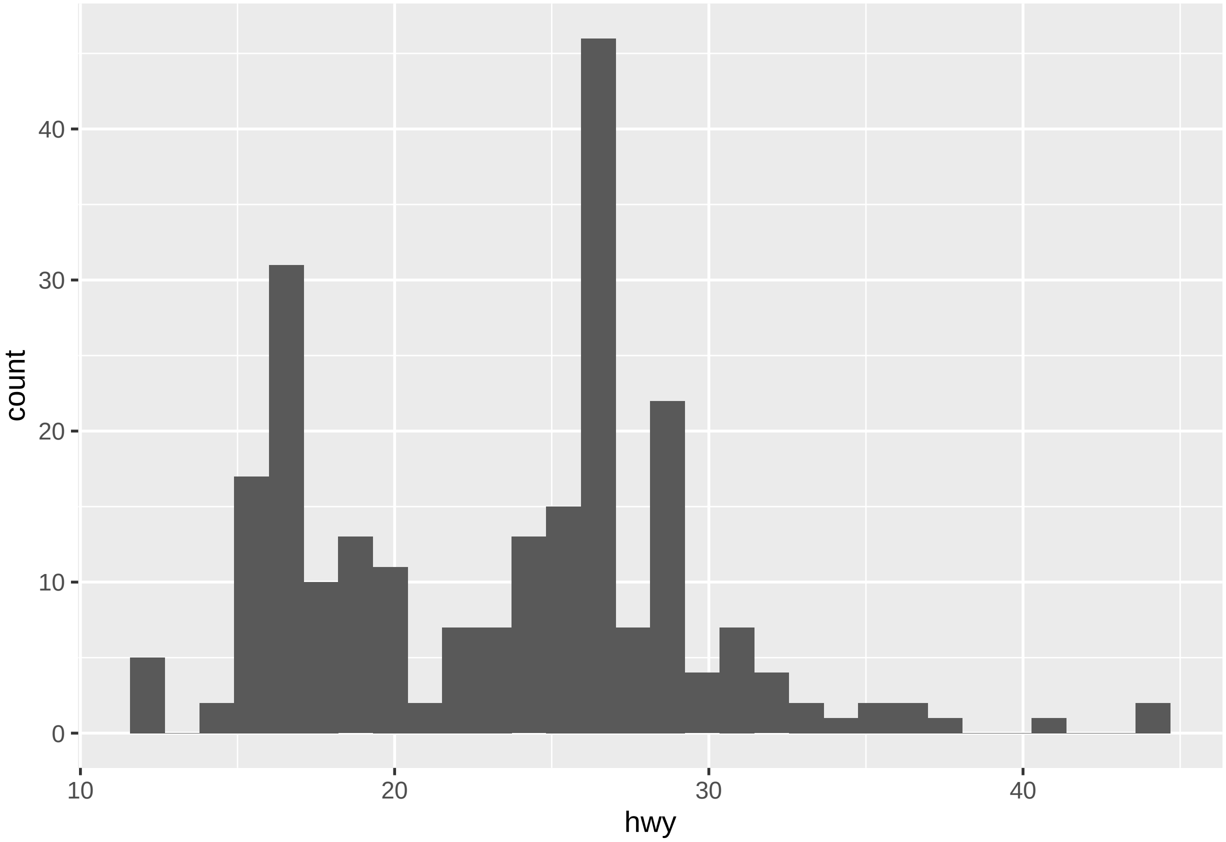
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

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.

```
# Histogram
ggplot(mpg, aes(x=hwy)) + geom_histogram()
```

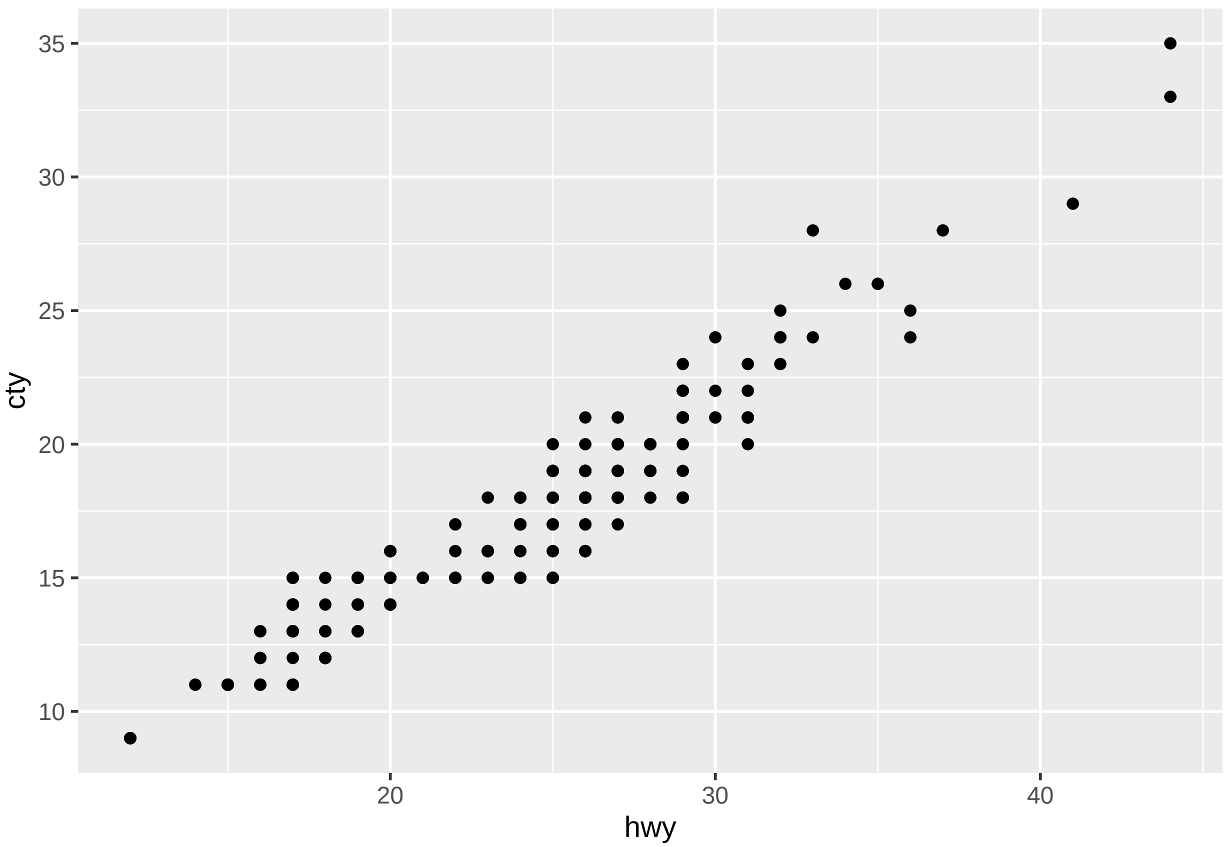
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Answer: There are a large amount of cars that can run approximately 26 highway miles per gallon. The next largest category would be cars that can run about 15 highway miles per gallon. Most cars call into the category of less than 37-38 highway miles per gallon, there are very few cars that can run over 40 highway miles per gallon. Maybe certain types of cars (sedans, SUVs, minivans, etc.) or models/brands from the same category have similar/same hwy miles per gallon? (which might help explain why some bars are so tall while others are short)

Exercise 2: Create a scatter plot. 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?

```r
# Scatter plot
ggplot(mpg, aes(x=hwy, y=cty)) + geom_point()
```



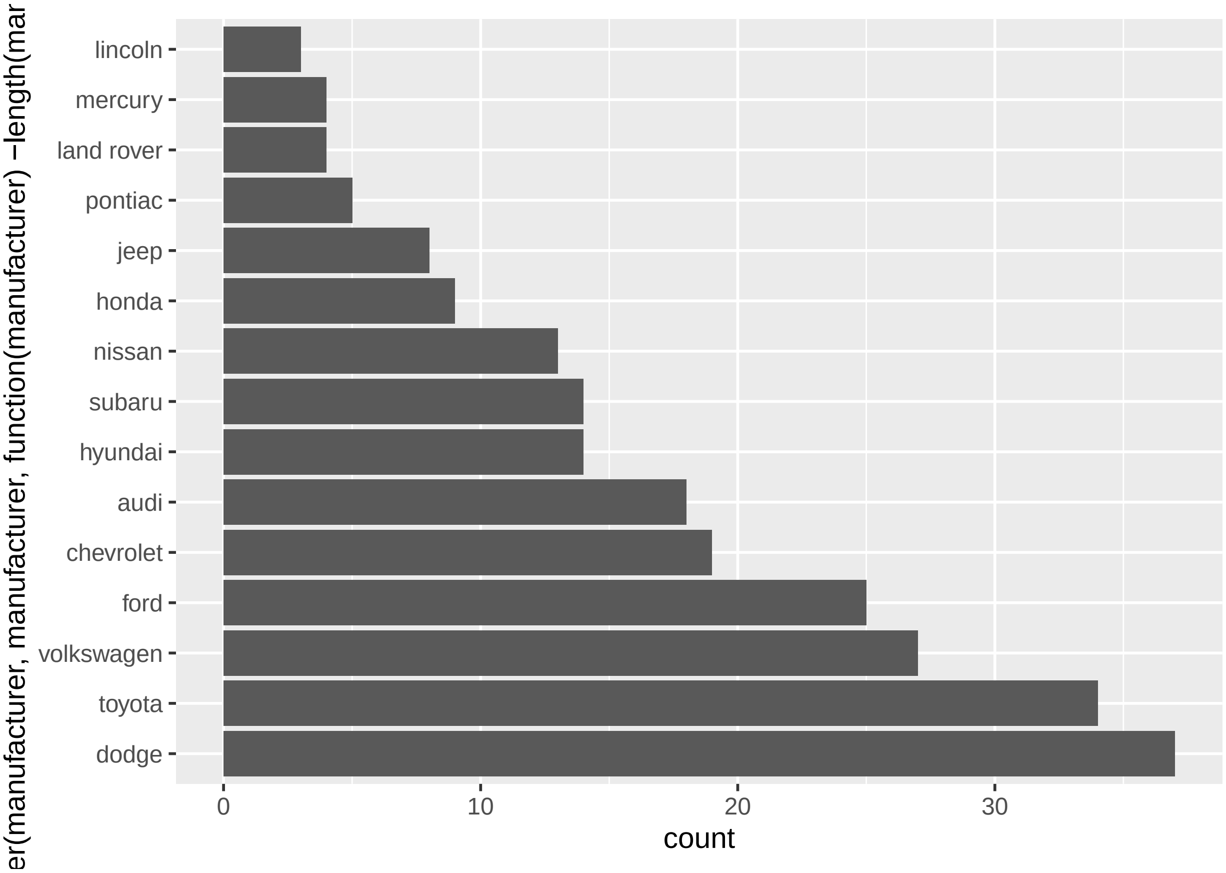
Answer: There seems to be a positive correlation between hwy and cty. This means that any changes will affect hwy and cty similarly/in the same direction. (ex. increase together, decrease together)

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?

```
# Check the numbers of each type of car from manufacturers
#sum(mpg$manufacturer == "dodge")
#sum(mpg$manufacturer == "toyota")
#sum(mpg$manufacturer == "land rover")
#sum(mpg$manufacturer == "mercury")
#sum(mpg$manufacturer == "lincoln")

# Bar plot
bar_plot <- ggplot(data=mpg, aes(x=reorder(manufacturer, manufacturer, function(manufacturer)-length(man
  geom_bar()

# Flip the Bar plot so manufacturer is on the y-axis
bar_plot + coord_flip()
```

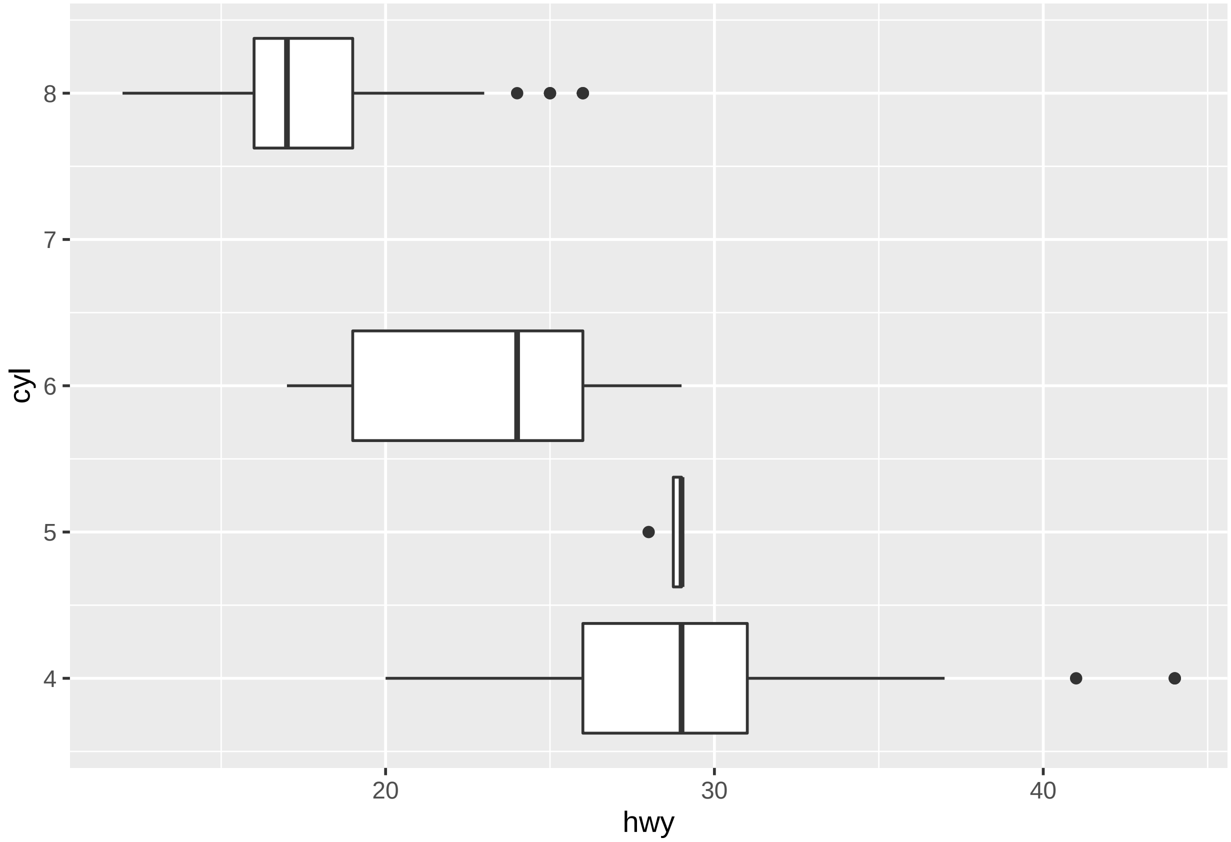


Answer: From the bar chart, Dodge produced the most cars and Lincoln produced the least cars.

Exercise 4: Make a box plot of hwy, grouped by cyl. Do you see a pattern? If so, what?

```
# Box plot
box_plot <- ggplot(mpg, aes(group=cyl, x=hwy, y=cyl)) +
  geom_boxplot()

box_plot
```



Answer: The general pattern I see is that a cyl and hwy are negatively correlated (lower number of cylinders corresponds to greater number of highway miles per gallon). Something interesting I see is that the median number of highway miles per gallon for a car with 5 cylinders is the same (from eyeballing) as the median number of highway miles per gallon for a car with 6 cylinders. I wonder why that is?

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?

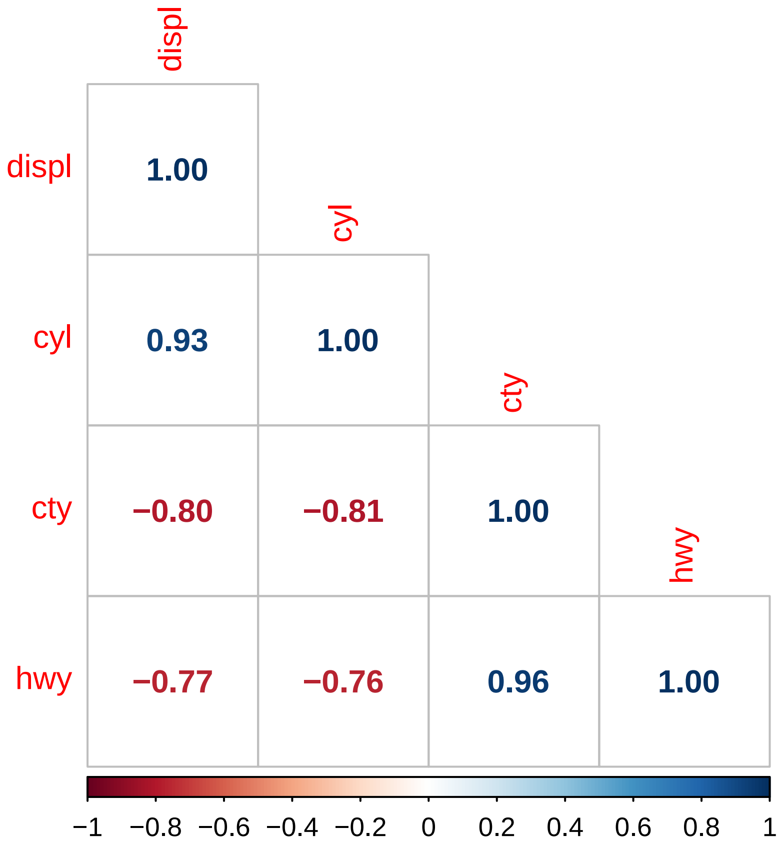
```
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
# mpg dataset, select numeric variables
cars <- mpg %>% select(displ, cyl, cty, hwy)
# put through correlation function
car_corr <- cor(cars)
head(car_corr)
```

```
##            displ        cyl        cty        hwy
## displ  1.0000000  0.9302271 -0.7985240 -0.7660200
## cyl    0.9302271  1.0000000 -0.8057714 -0.7619124
## cty   -0.7985240 -0.8057714  1.0000000  0.9559159
## hwy   -0.7660200 -0.7619124  0.9559159  1.0000000
```

```
# plot lower triangular correlation matrix
corrplot(car_corr, type = "lower", method = 'number')
```



Answer: *note: I only used the numeric variables to make the correlation matrix because I wasn't sure how to use non-numeric variables with the corrplot() function. I can see that cty and hwy are positively correlated, which matches our result from Exercise 2. hwy and cyl are negatively correlated, which matches results from Exercise 4. I haven't graphed/tested the other correlations, but looking at the relationships, most seem to make sense. For example, cty and cyl have negative correlation, cty and displ have negative correlation, so it makes sense that cyl and displ have positive correlation. (like in math, negative* negative = positive kind of thought process?) Working backwards with the same though process, since hwy and cty are positive, it makes sense that hwy and cyl are negative and cty and cyl are also negative. The one relationship that isn't super clear to me is hwy and displ being negatively correlated.