PSTAT 231 Homework 1

Chris Lefrak

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 involves building a model that recieves an input an gives a specific output. The model is trained with data that has an associated label, and the error between the models guess and the true label can be used to update the model.

On the other hand, unsupervised learning involves training a model where there are no specific outputs. Rather, an application of unsupervised learning could be to cluster data into "similar" groups.

Question 2:

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

Answer: A regression model is a machine learning model where the output is a quantitative variable. Most of the time, this means the model (function) is continuous: small changes in the input data should only correspond to small changes in the output.

On the other hand, a classification model is for qualitative outputs, where the output is typically (but not always) discrete. The name is kind of self explanatory, the model is trying to classify which category the input data belongs to.

Question 3:

Name two commonly used metrics for regression ML problems. Name two commonly used 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: These are models used to summarize the characteristics of the data set, e.g., by visualizing trend
- Inferential models: An inferentital model is used to test a hypothesis and see if certain characteristics that hold in your data can be generalized to a broader population.

- Predictive models: Predictive models is aimed purely at making predictions on the outcome of new data.

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?

Answer: Mechanistic essentially means parametric. There is an assumed form of the underlying funtion f, i.e., that it depends on some parameters $\beta_1, \beta_2, \ldots$. The model tries to find the best values of the parameters to fit the data.

Empirically-driven is non-parametric, i.e., no assumptions are made about the form of f. Empirically driven is more flexible that mechanistic since, if we were to assume a specific form of f, that is restricting us to a particular class of functions.

Both models are susceptible to overfitting.

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

Answer: A mechanistic model is easier to understand since we know the functional relationship. An empirically-driven approach is more like a black box.

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

Answer: We have that a mechanistic approach will tend to have a higher bias since "The bias error is an error from erroneous assumptions in the learning algorithm" (Wikipedia). In assuming a specific functional form, we are restricting ourselves to producing a function that is unlikely to match the true f.

On the other hand, an empirically-driven approach will tend to have a higher variance since "High variance may result from an algorithm modeling the random noise in the training data" (Wikipedia). This is "overfitting" and is a bigger issue with empirically-driven approaches since they require more data and are developed purely from the data itself.

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, it is trying to predict the outcome (voter preference) based on input data (voter's profile).
2) inferential, it is trying to answer a "what-if" question, extrapolating beyond the data at hand.

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.

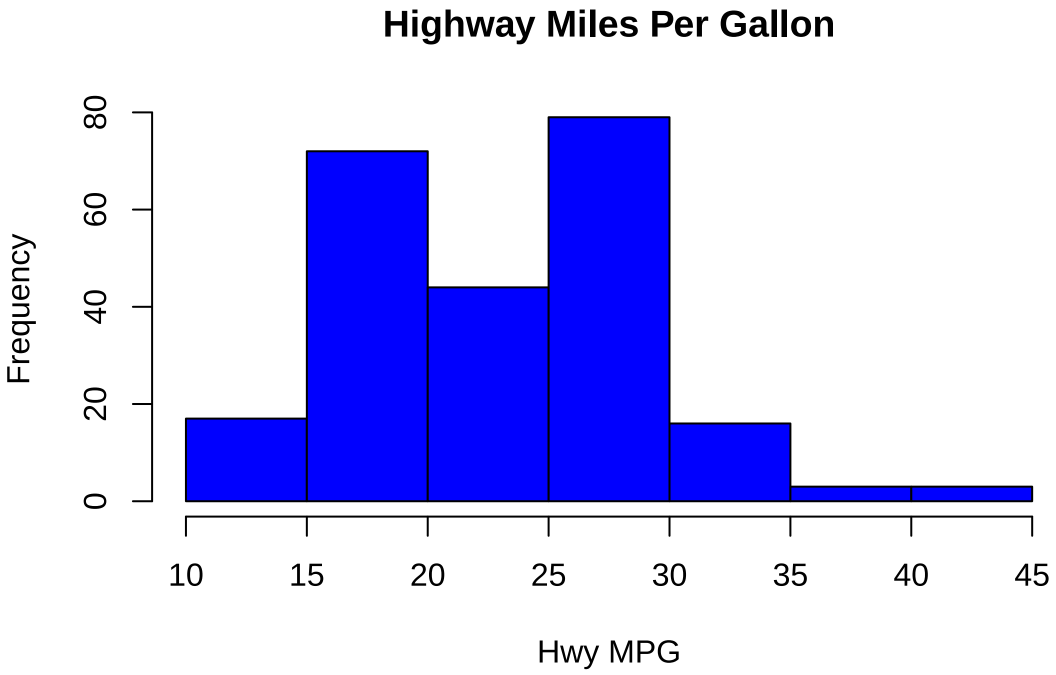
Exploratory data analysis (or EDA) is not based on a specific set of rules or formulas. It is more of a state of curiosity about data. It's an iterative process of:

- 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
hist(mpg$hwy,
     main="Highway Miles Per Gallon",
     xlab = "Hwy MPG",
     col='blue')
```

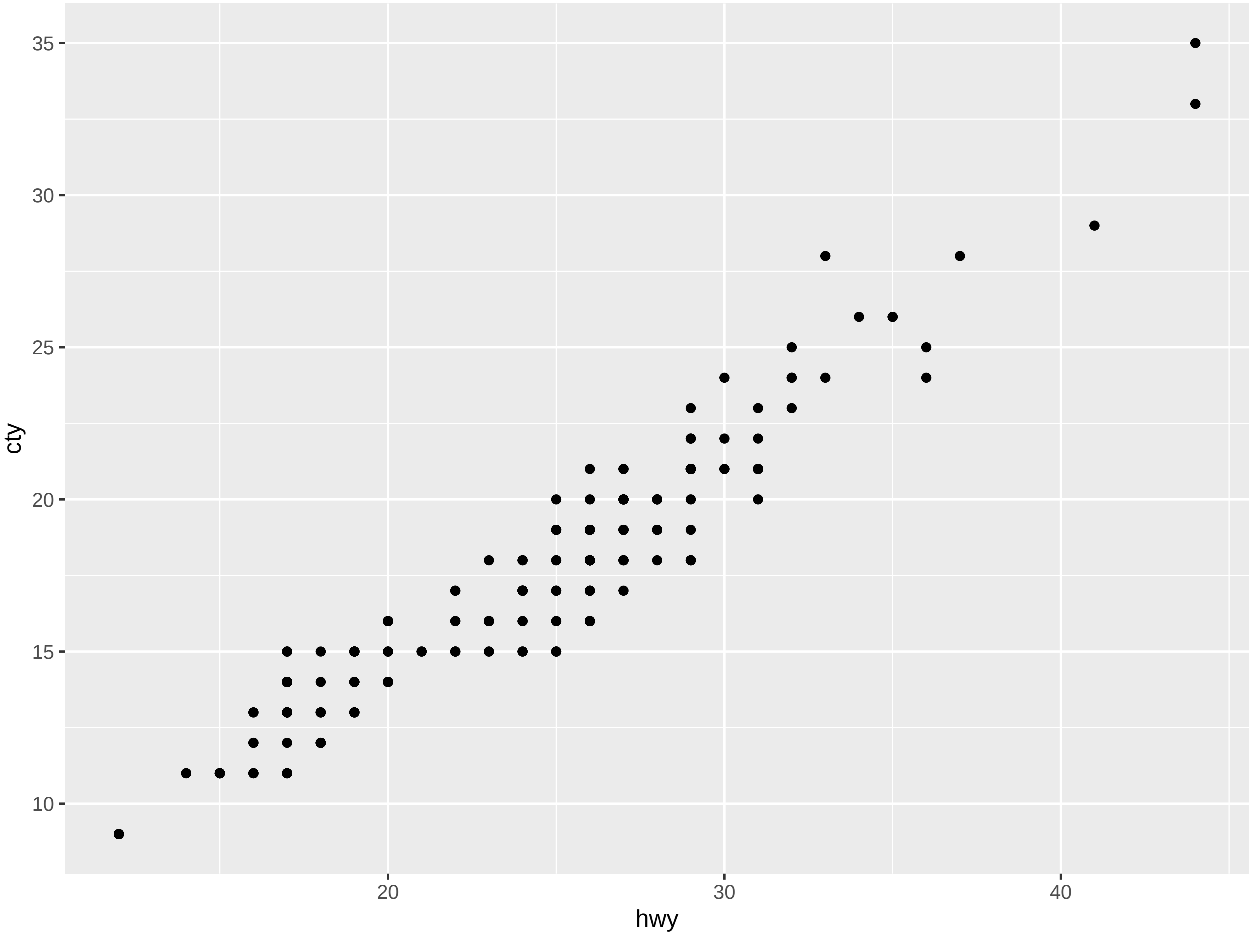


It seems that the distribution is skewed right. I have learned that my car's fuel economy is both below the mean and median, assuming this is a representative sample. :(The bin with the largest number of cars is the 25-30 bin followed by 15-20.

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?

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



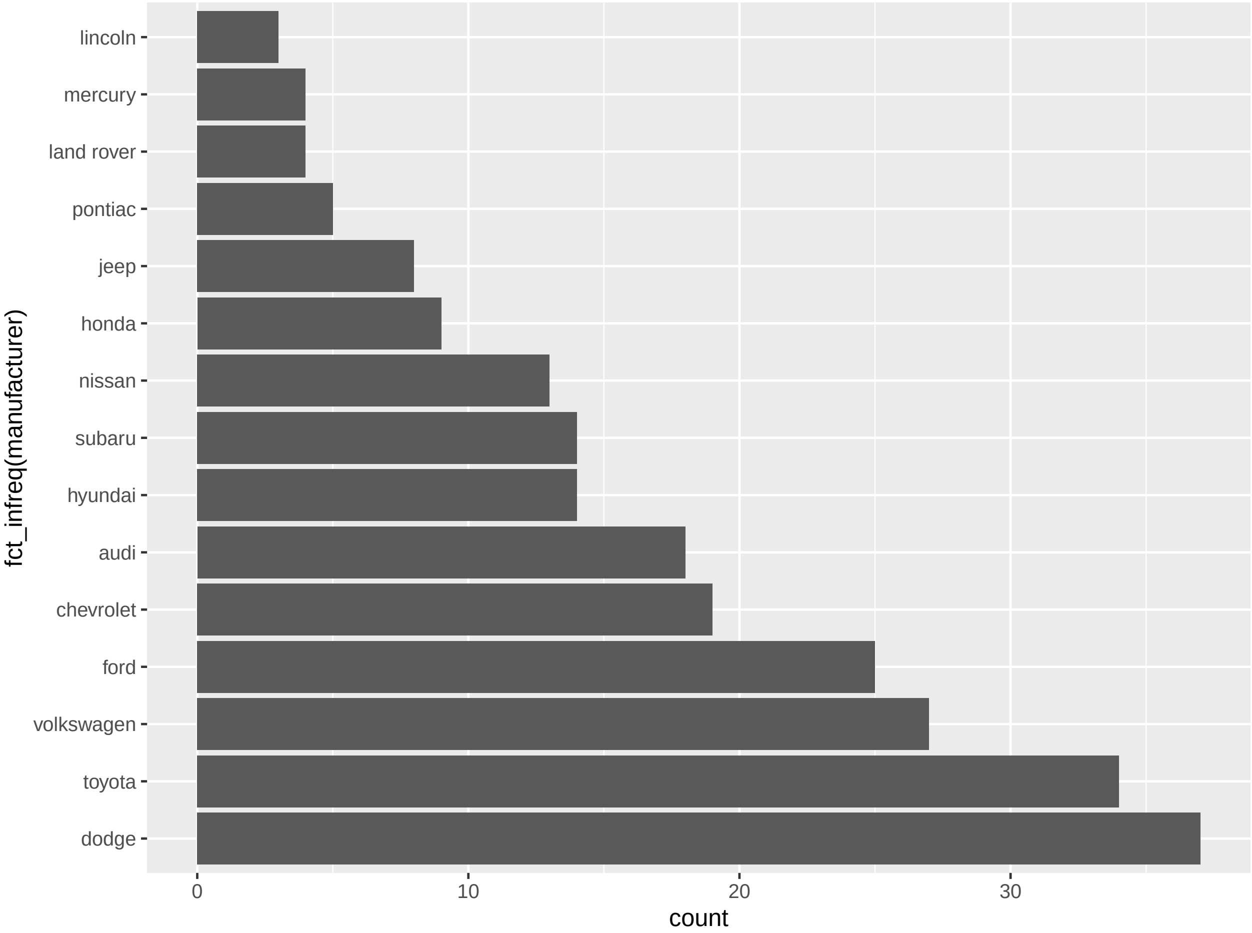
There is a very clear linear relationship between highway and city mpg, which intuitively makes sense. These two variables are highly correlated. This means two things: we could use city mpg as a strong predictor of hwy mpg (or vice versa), but it would be redundant to include both city and hwy mpg as predictors for a third variable. This is because hwy and city mpg are essentially linear combinations of each other.

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?

```r
p<-ggplot(data=mpg, aes(x=fct_infreq(manufacturer))) +
  geom_bar(stat="count")

# Horizontal bar plot
p + coord_flip()
```

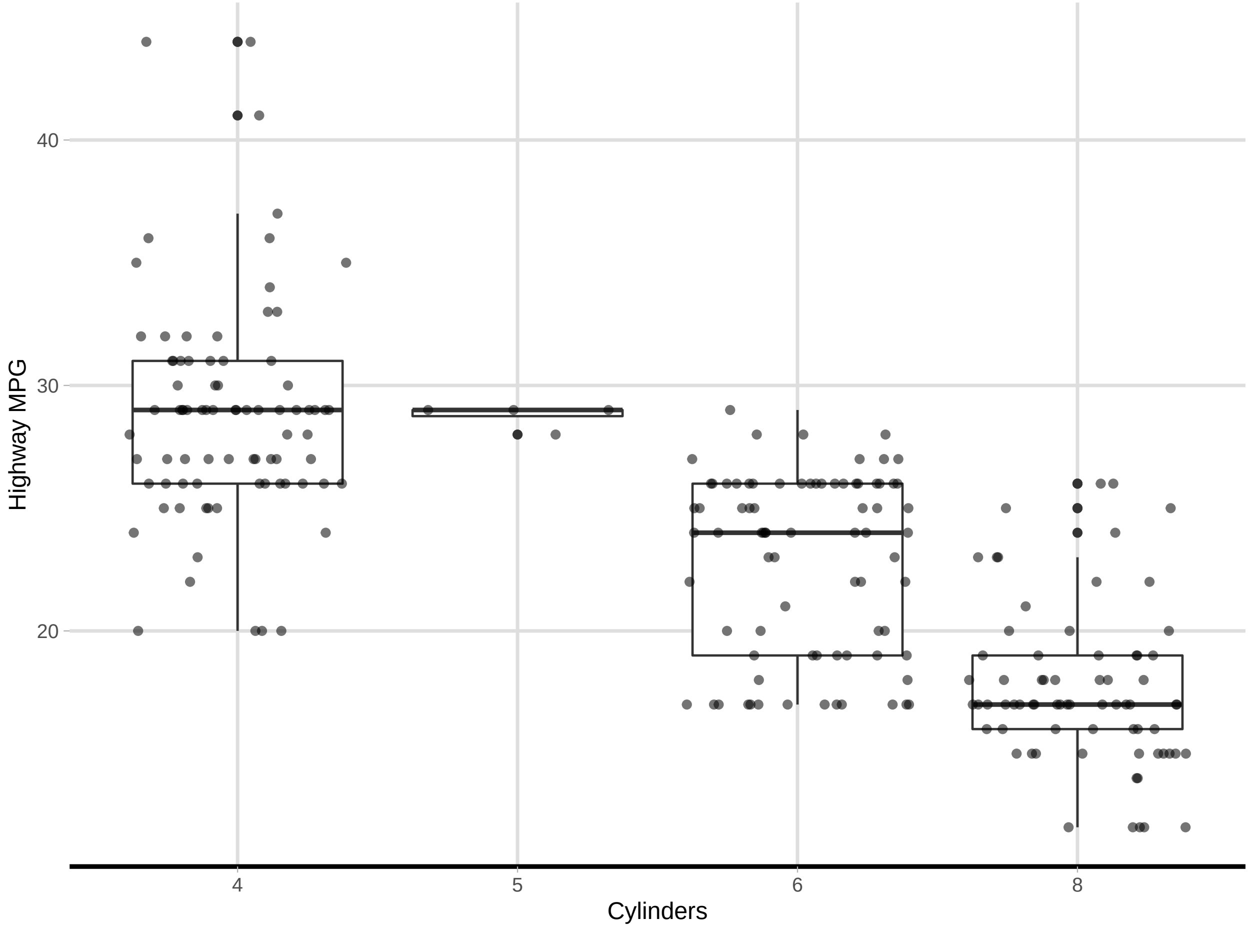


Toyota produced the most cars and Lincoln produced the least.

Exercise 4:

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

```
ggplot(mpg, aes(x = as.factor(cyl),y = hwy))+
  geom_boxplot()+
  geom_jitter(height=0,aes(alpha=0.5))+
  theme_light()+
  #coord_flip()+
  scale_y_continuous(breaks=seq(20,40,by=10),minor_breaks = NULL)+
  theme(panel.border = element_blank(),#element_rect(color = "black"),
        legend.position = "none",
        panel.grid = element_line(size = 1.5),
        axis.line.x = element_line(color="black", size = 1)
       )+
  labs(x="Cylinders", y = "Highway MPG")
```

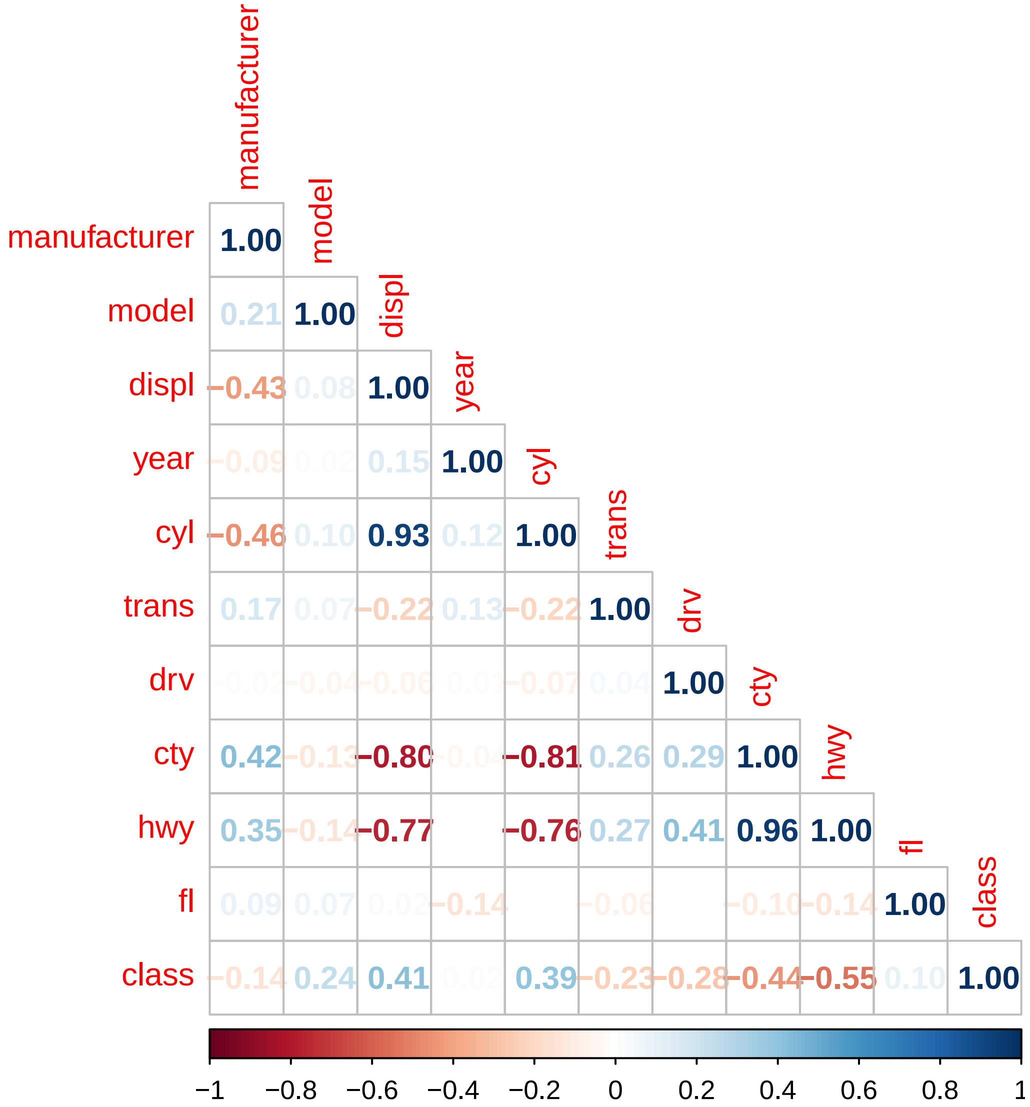


I don't really notice any patterns. What I do notice is that median fuel economy decreases with more cylinders. Also the variance of 4-cylinder cars is much greater than that of 6- or 8-cylinder.

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 on the web.)

```
mpg_factor<-mpg %>%
mutate_if(sapply(mpg, is.character), as.factor)
mpg_factor%>%
mutate_if(sapply(mpg_factor, is.factor), as.numeric)%>%
cor()%>%
corrplot(method='number',type='lower')
```



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

Variables that are positively correlated include: Highway & City MPG and Cylinders & Engine Displacement.

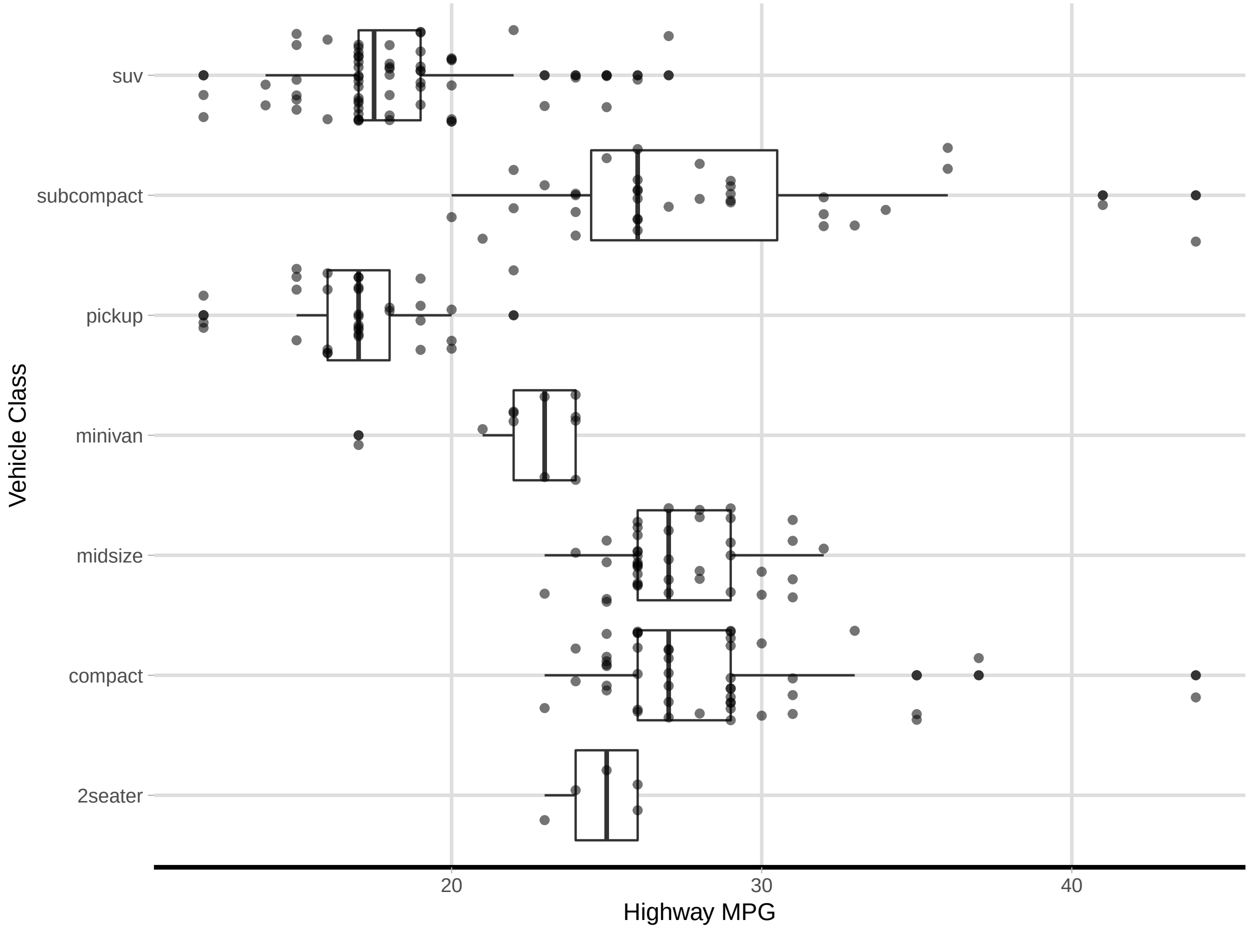
Variables that are negatively correlated include Highway/City Mpg & Cylinders/Engine Displacement.

These make sense to me. I don't see anything that surprises me.

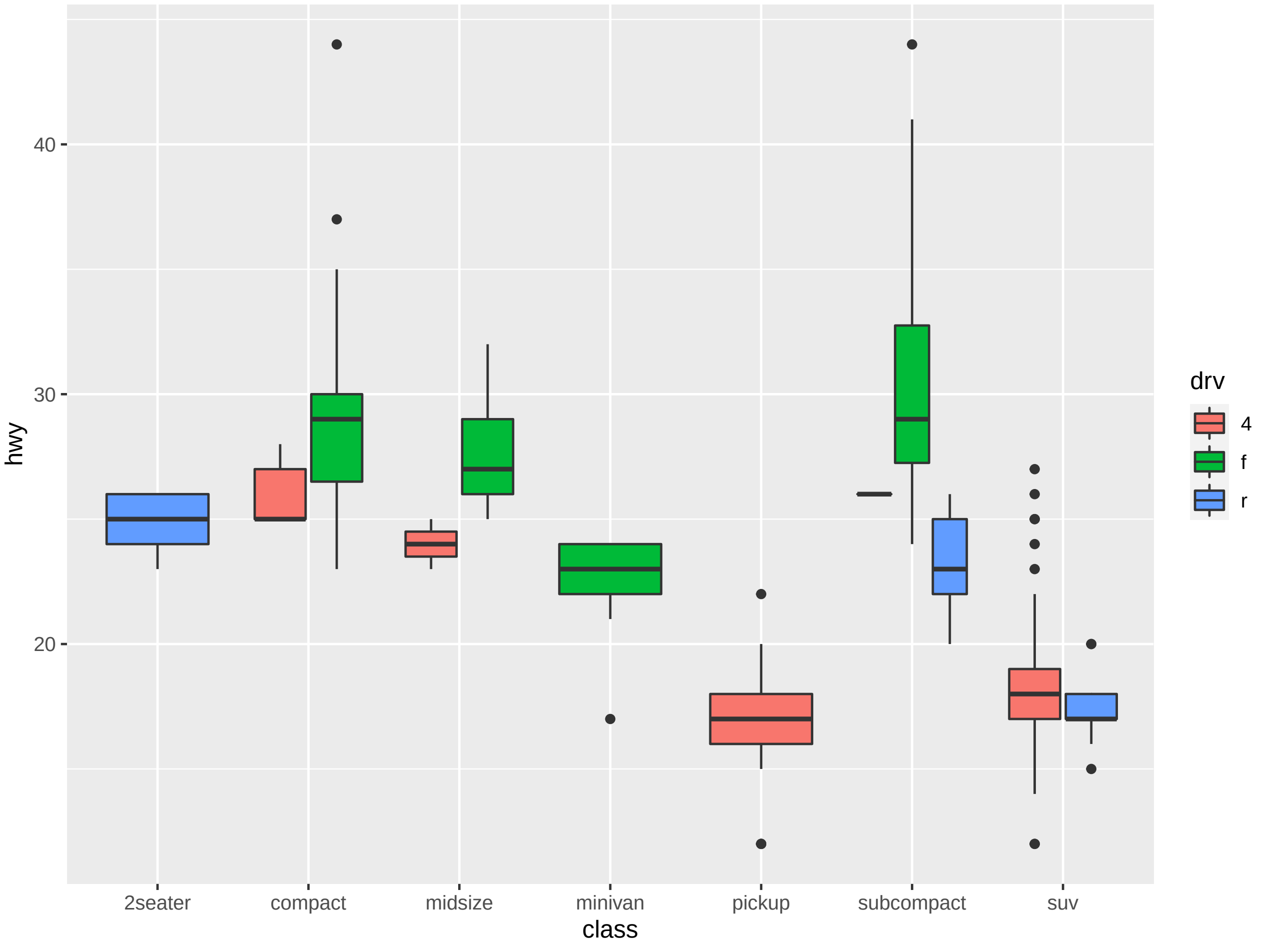
For 231 Students

Reconstructing Figures

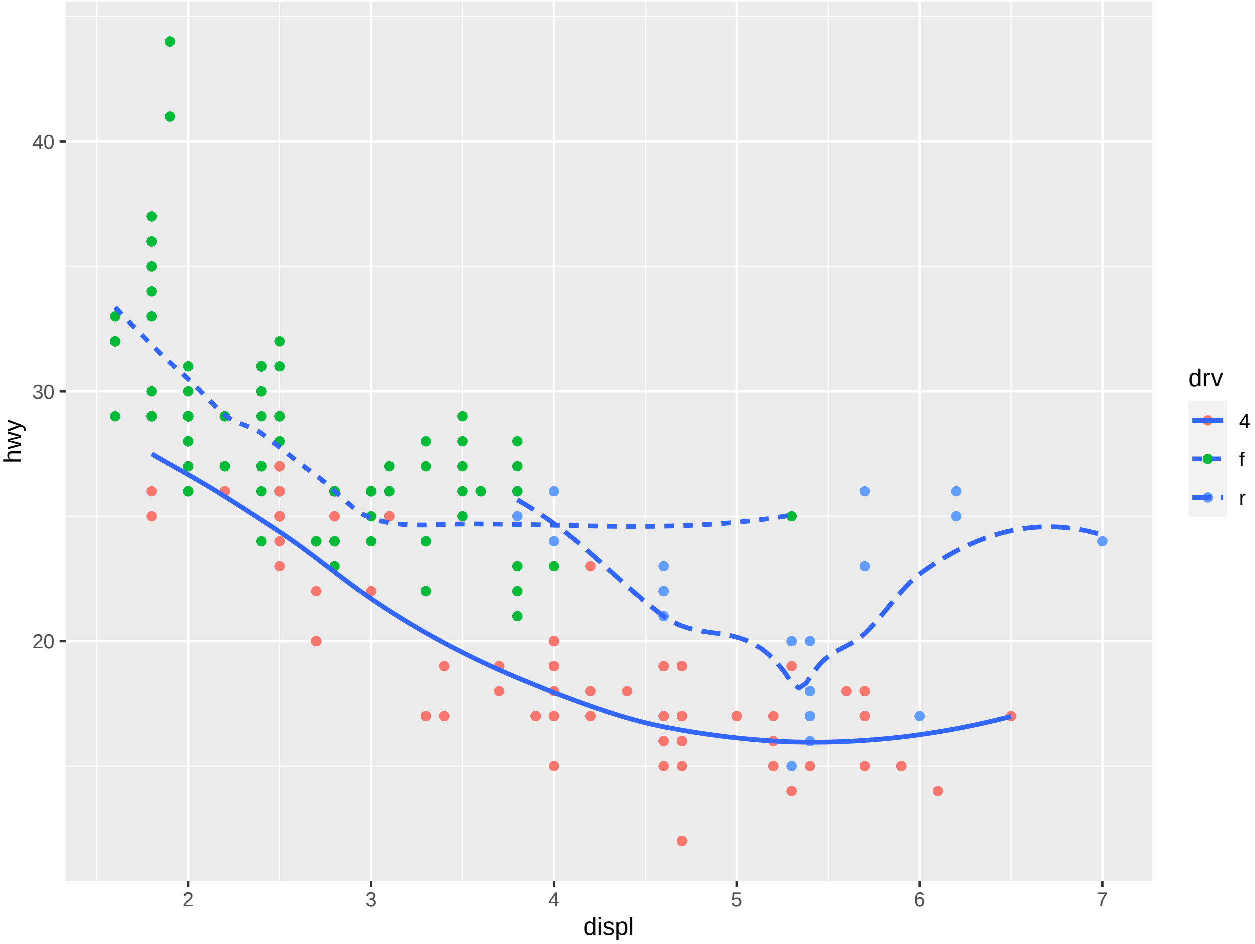
```
ggplot(mpg, aes(x = hwy,y = class))+
  geom_boxplot()+
  geom_jitter(width=0,aes(alpha=0.5))+
  theme_light()+
  labs(x="Highway MPG",y="Vehicle Class")+
  scale_x_continuous(breaks=seq(20,40,by=10),minor_breaks = NULL)+
  theme(panel.border = element_blank(),
        legend.position = "none",
        panel.grid = element_line(size = 1.5),
        axis.line.x = element_line(color="black", size = 1)
       )
```



```
ggplot(mpg, aes(x = class,y = hwy,fill = drv))+
  geom_boxplot()
```



```
ggplot(mpg, aes(x = displ,y = hwy))+
  geom_point(aes(col=drv))+
  geom_smooth(se=FALSE,aes(linetype=drv))
```


hwy
displ
drv
f
r