PSTAT 231 Homework 1

Dylan Berneman

2022-04-01

Machine Learning Main Ideas

Question 1: Define supervised and unsupervised learning. What are the difference(s) between them?

Answer: Supervised learning is used to accurately predict future response given predictors, understand how predictors affect response, find the "best" model for response given predictors, and assess the quality of our predictions and (or) estimation. Unsupervised learning, on the other hand, does not have any measurements of an outcome.

(Lecture Slides Day 1, pgs.30, 32)

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

Answer: The main difference is that the output of a regression model is a numerical value while the output for a classification model is categorical value.

(Lecture Slides Day 1, pg.31)

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

Answer:

Regression ML: • Linear regression

- Logistic regression

Classification ML: • Principal Component Analysis (PCA)

- k-means clustering

(Lecture Slides Day 2, pg.4)

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.

Answer:

Descriptive models: • Choose model to best visually emphasize a trend in data

Inferential models: • What features are significant?

- Aim is to test theories
- State relationship between outcome & predictor(s)

Predictive models: • What combo of features fits best?

- Aim is to predict Y with minimum reducible error

(Lecture Slides Day 2, pg.7)

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.

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

Answer: Mechanistic assumes a parametric form where added parameters results in more flexibility, but too many can lead to overfitting. Empirically-driven makes no assumptions about the function, requires a larger number of observations, and is very flexible.

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

Answer: A mechanistic model is easier to understand because it provides more insight than an empirically-driven model. In addition, mechanistic models specify assumptions and attempt to incorporate known factors about the systems surrounding the data into the model, while describing the available data (Bonate, 2011).

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

Answer:

Mechanistic models: The bias-variance trade-off is the relationship between bias, variance, and test set MSE. Good test set performance of a statistical learning method requires low variance as well as low squared bias which is not easy to do because it is easy to obtain a method with extremely low bias but high variance or a method with very low variance but high bias.(ISL, pg.36)

Empirically-driven models: The bias-variance trade-off can transfer over to the classification setting with a few modifications such as using the error rate, the proportion of mistakes that are made if we apply error rate our estimate to the training observations. (ISL, pg.37)

Question 6: A political candidate's campaign has collected some detailed voter history data from their constituents. Classify each question as either predictive or inferential. Explain your reasoning for each.The campaign is interested in two questions:

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

Answer: The question is predictive because its aim is to predict the outcome, which in this case, is which candidate the voter will decide to vote for.

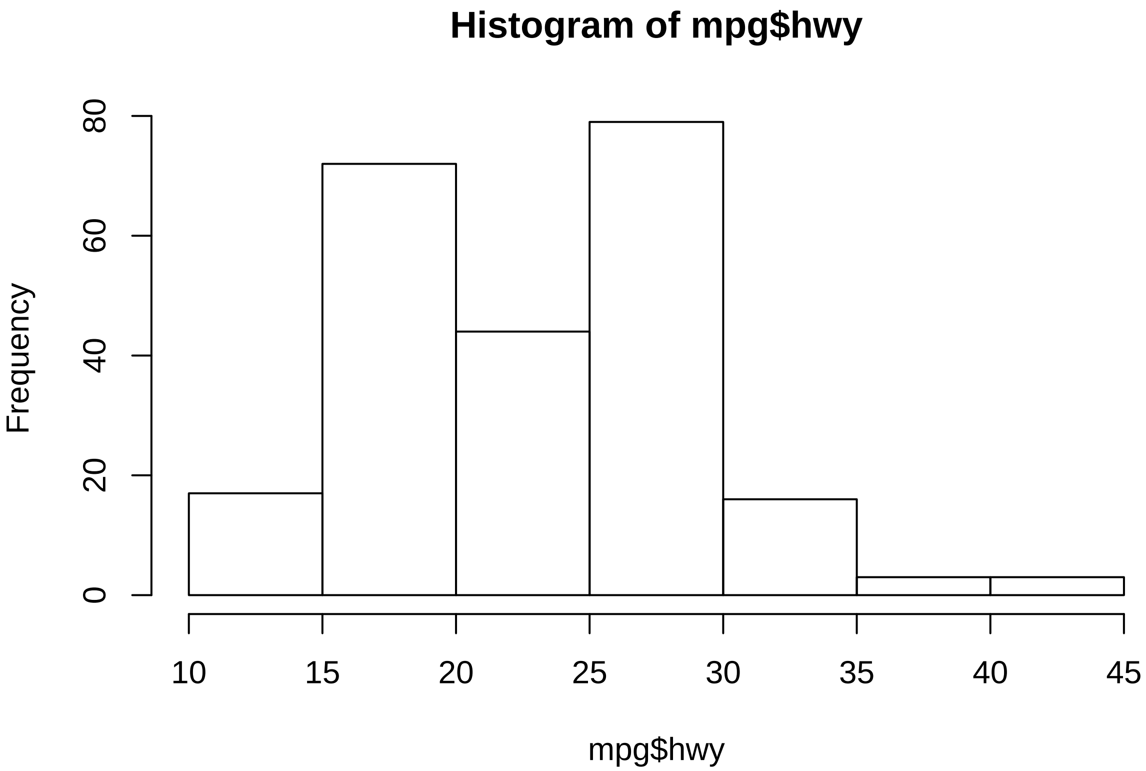
ii.) How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?

Answer: The question is inferential because its aim is to test whether a specific variable, in this case, whether the voter has had personal contact with the candidate, is significant to the outcome of which candidate the voter will vote for.

Exploratory Data Analysis

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.

```
hist(mpg$hwy)
```

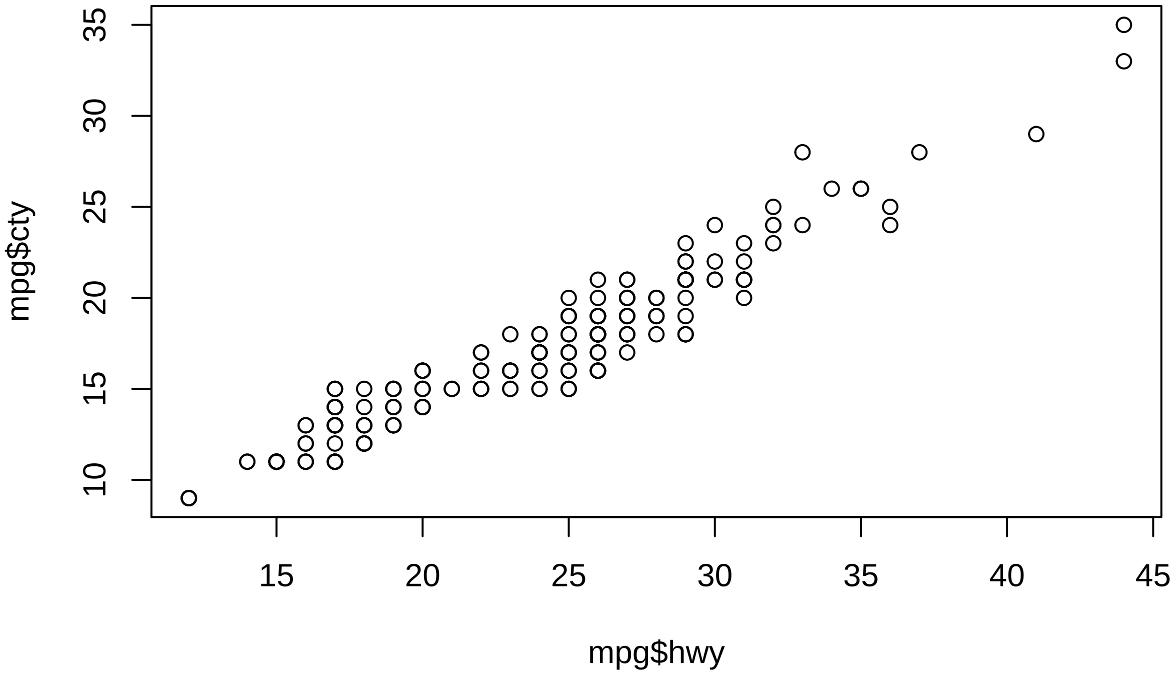


What we see is that most cars have highway miles per gallon between 15 and 30 miles. A smaller amount of cars either make trade-offs for for a slightly lower mpg between 10 and 15 miles or a slightly higher mpg between 30 to 35 miles. Finally, we see that there are a few outliers who get between 35 and 45 mpg.

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?

```
plot(mpg$hwy, mpg$cty)
```

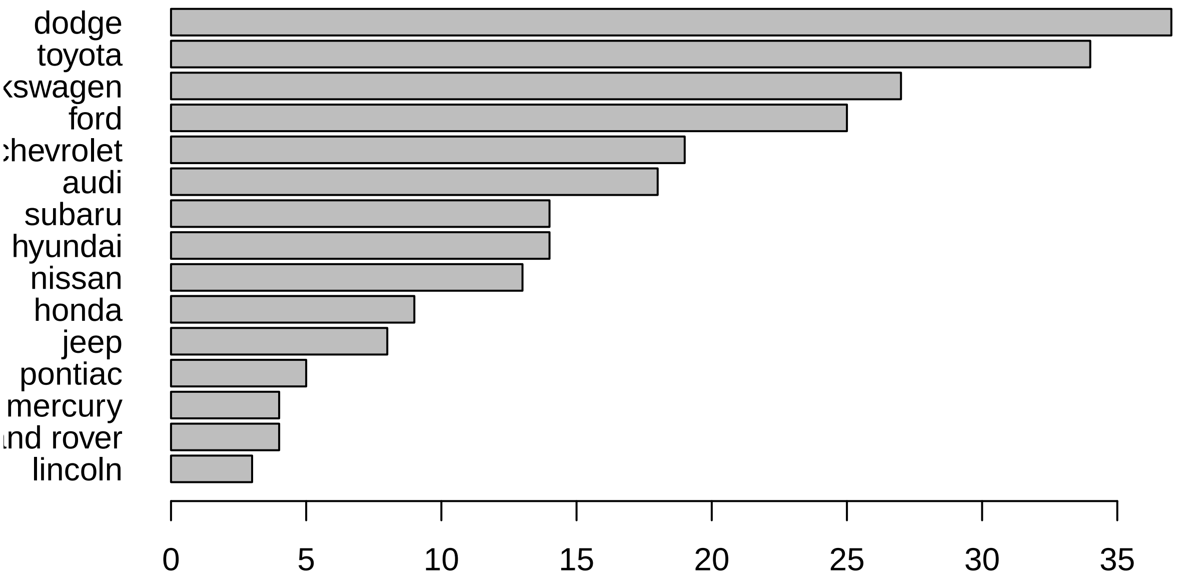


There is a positive correlation between hwy and cty. As cty increases, so does hwy. This means that the observations are fairly accurate because the ratio of hwy to cty is nearly constant which reflects on how cars will use up more gas in a city due to having to stop for traffic lights.

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?

```
manufacturer.freq <- sort(table(mpg$manufacturer))
barplot(manufacturer.freq, width = 1, horiz=TRUE, las=1)
```

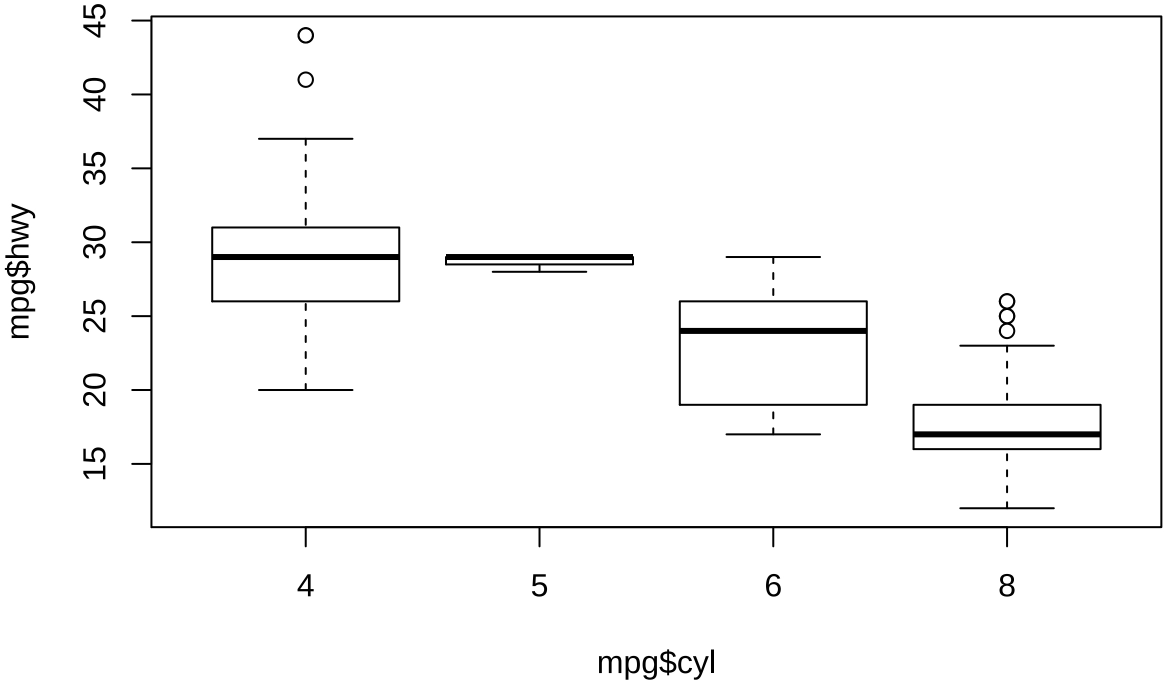


Dodge produced the most cars while Lincoln produced the least.

Exercise 4:

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

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



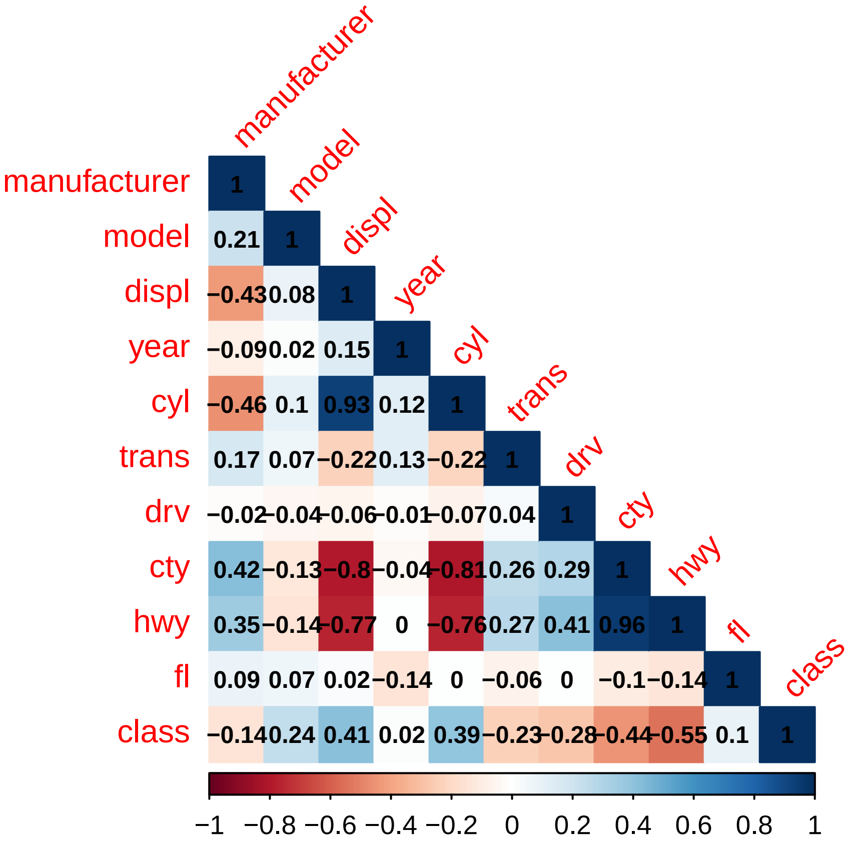
Highway miles per gallon tends to decrease as the number of cylinders increases, meaning that they have an inverse relationship.

Exercise 5:

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?

```
mpg1 <- data.matrix(data.frame(unclass(data.frame(mpg))))
corrplot(cor(mpg1), type = 'lower', method = 'color', addCoef.col = "black", tl.srt = 45, tl.offset = 0
```



Positively Correlated: - Strong: (hwy, cty), (cyl, displ)

- Weak: (cty, manufacturer), (hwy, manufacturer), (class, displ), (class, cyl),

(hwy, trans), (cty, trans), (cty, drv), (hwy, drv)

Negatively Correlated: - Strong: (cty, displ), (hwy, displ), (hwy, cyl), (cty, cyl)

- Weak: (displ, manufacturer), (cyl, manufacturer), (class, drv), (class,cty),

(class, hwy)

No relationship: - All pairs with coefficients less than the absolute value of 0.25

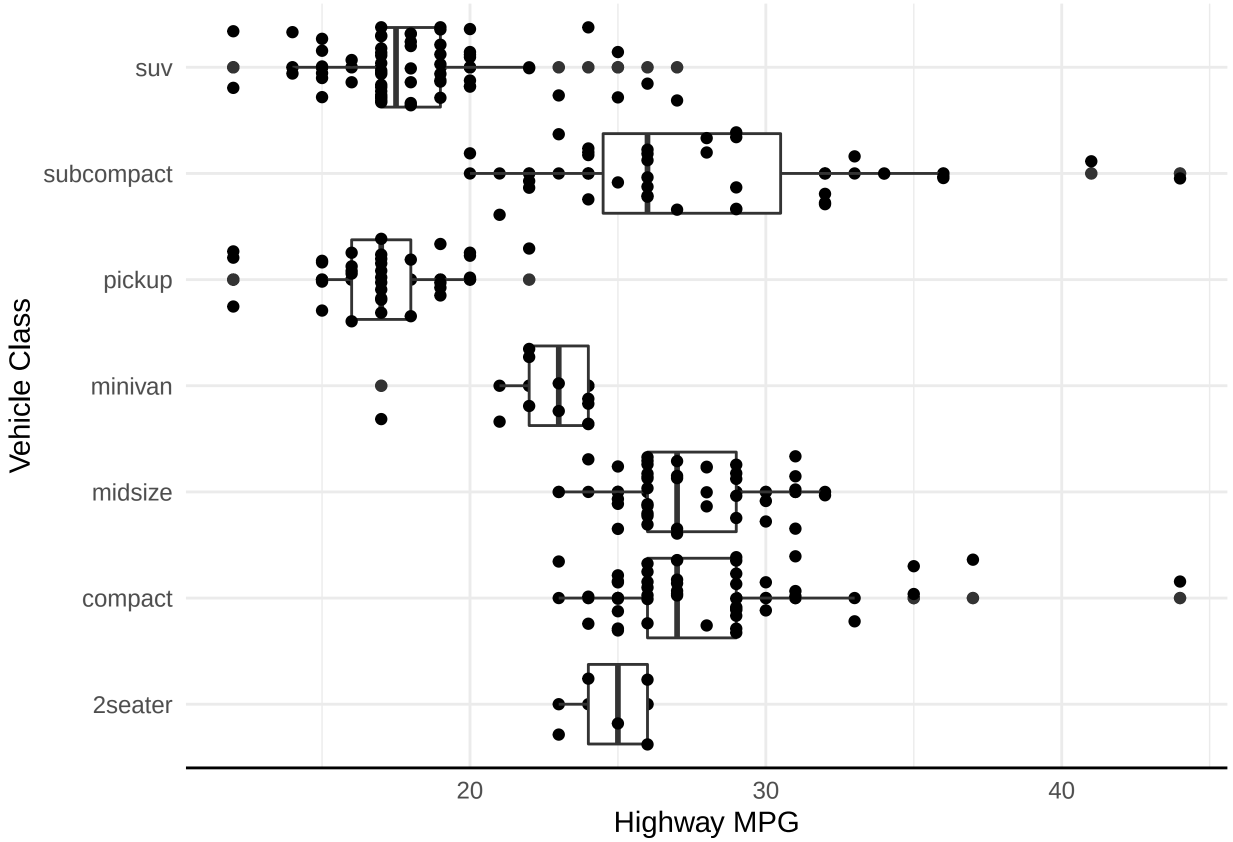
Most of the relationships make sense to me. However, I would have believed that at least a few of the variables would have a correlation with the year since technology is always advancing and the goal is to always improve products each year.

For Exercises 6 - 8:

Recreate the following graphics, as closely as you can. Hint: Use the ggthemes package.

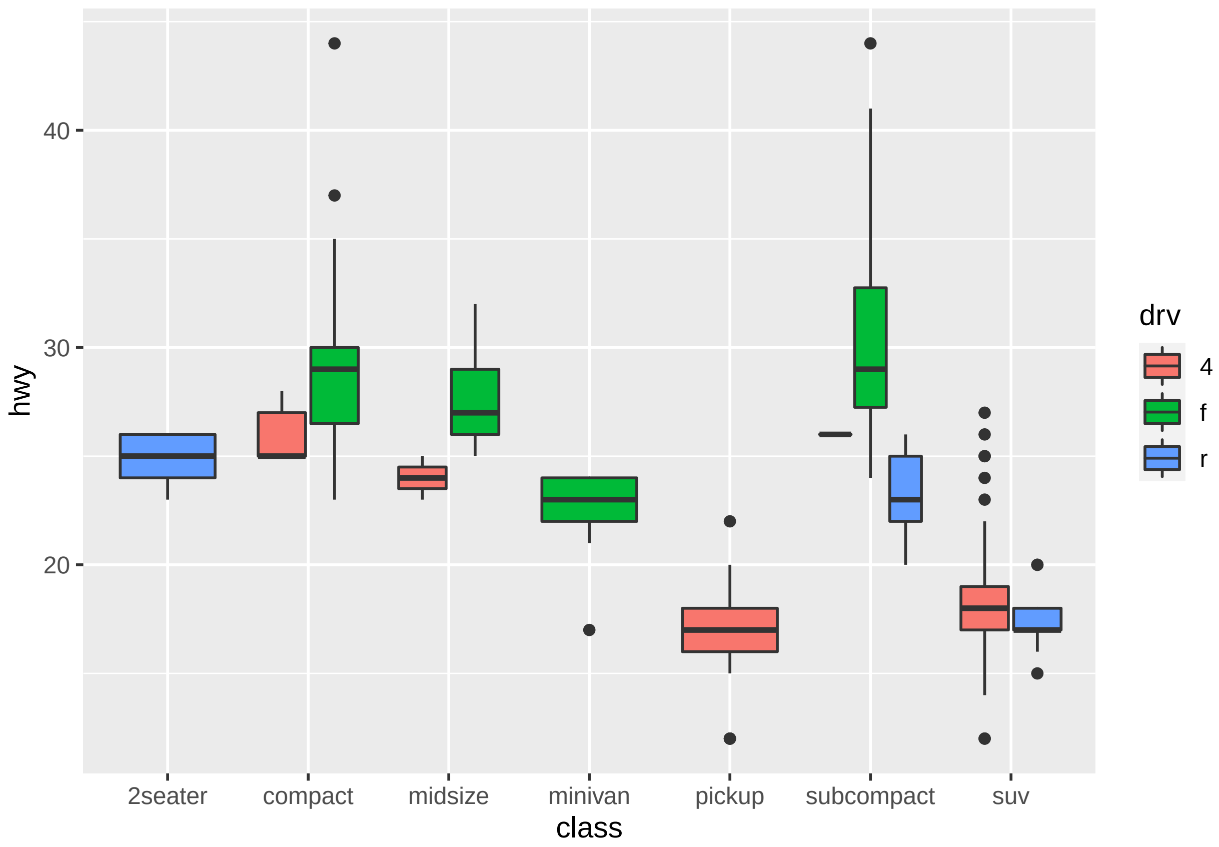
Exercise 6:

```
ggplot(mpg, aes(x = hwy, y = class)) +
  geom_point(shape = 19) + geom_boxplot() + xlab("Highway MPG") + ylab("Vehicle Class") + geom_jitter(w
axis.line.x = element_line(size = 0.5, linetype = "solid", colour = "black"))
```



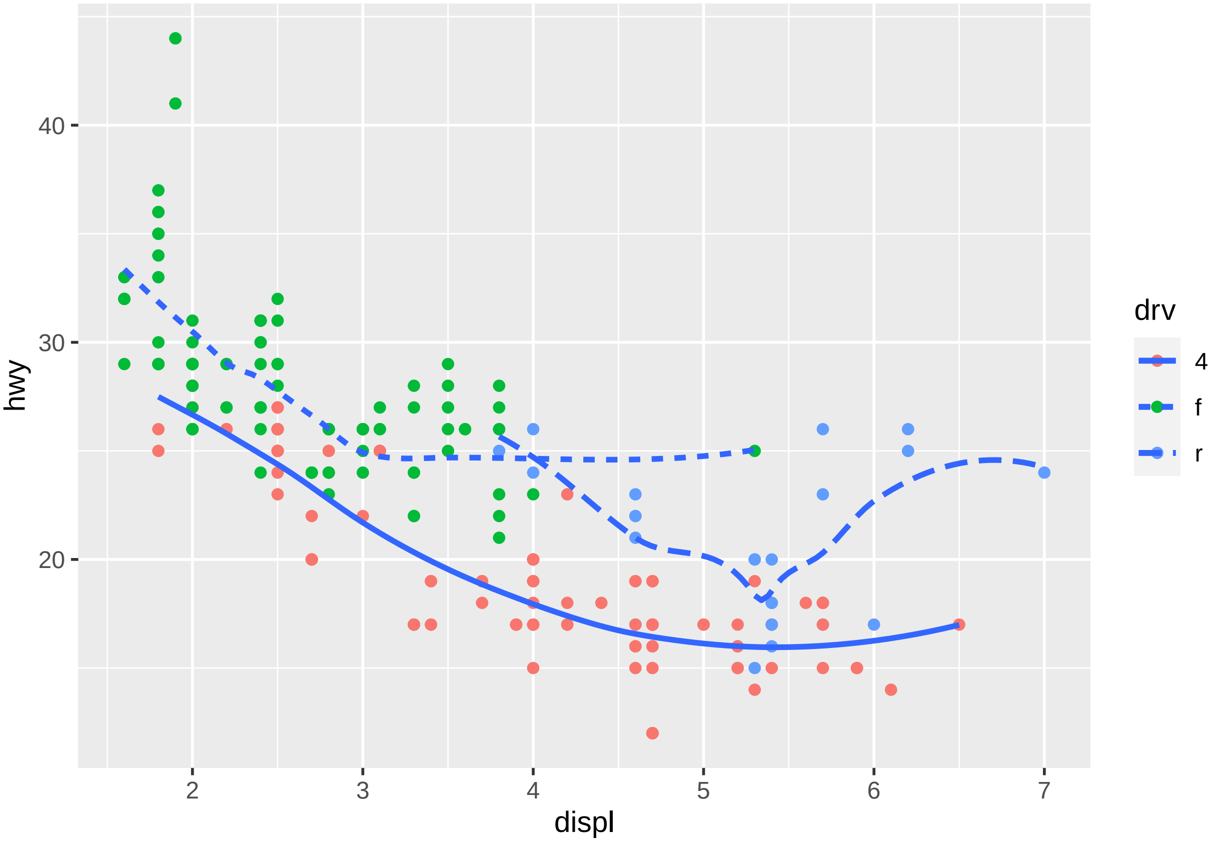
Exercise 7:

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



Exercise 8:

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy, linetype = drv)) +
  geom_point(mapping = aes(color = drv)) + geom_smooth(formula = y ~ x, method = "loess", data = filter
```


40
30
20
hwy
2
3
4
5
6
7
displ
drv
4
f
r