Homework 1

Mason Wong

2022-10-01

Problem 1

Supervised learning is mainly described as giving a complete model of both input and output variables to a machine. After doing so, the machine will use these values to find how impactful each input, or predictor variable, is and assign it a weight. A variable with a higher weight would be more impactful than a lesser weighted variable if both input values increased by the same amount. In this case, the output values are considered the supervisors.

Unsupervised learning involves assigned a machine a set of ONLY input/predictor variables. Unsupervised learning finds patterns in the input variables in order to find variables that have a significant impact on the unknown outcome.

The main difference between the two is that Supervised learning has a target set of data that corresponds with the predictor variables. This means that it can be more precise in finding trends in the data. Supervised learning is capable of regression, classification and finding a "model of best fit". Unsupervised learning lacks the target data and thus is less accurate in finding predictor variables that have a significant influence over the dataset. Despite this unsupervised learning still has its uses being useful in finding patterns and groups.

Problem 2

In machine learning, regression analysis is used in the context of quantitative values, and useful in predicting such variables, Classification is more useful when dealing with qualitative values, and can help in predicting future categories that data might fall under.

Problem 3

Two machine learning metrics for regression involve mean squared error and root-mean-squared-error.

Two machine learning metrics for classification involve accuracy and error rate.

Problem 4

Descriptive model: Helps find trends in the datasets and build models to represent said trends

Inferential model: Looks to find significant predictor variables to test theories about the response variables and uncover any possible relationships that are present between the predictor and response variables

Predictive model: Intends to predict Y with the optimal amount of predictor variables such that reducible error is minimal. This is simply to get the most accurate model while also removing the most amount of unpredictable uncertainty

Problem 5

Mechanistic models primarily focus on using math and deterministic values to forecast future variables. Machanistic models also assume a parametric from the response variable.

Empirically-driven relies on observations and makes no assumptions about f

While Mechanistic is parametric and Empirically-driven is non-parametric they can both be utilized in the same study to predict future outcomes.

Mechanistic models seems more easy to understand as to me as they rely mostly on math and theories that are already present. Empirically-driven models have a sense of unpredictability and can easily be thrown off if a couple of observations result in outliers.

Simpler models tend to be high in bias, but low in variance. While empirically-driven models are flexible regardless of the research being done, flexibility in a mechanistic model is reliant on the number of paramteters included in the model. Thus the more flexible each model is, the lower the bias, but the higher the variance.

Problem 6

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

A: The above question is predictive since we are trying to see the likelihood of a future outcome using predictor variables, being the voter's profile. Analyzing their profile would show some variables being more reliable than others, reducing the risk present in the overall model.

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

A: Inferential, in this scenario we would be testing to see if meeting the candidate would cause a significant change in our prediction. Thus this would fall under the inferential model since that model is all about uncovering any possible relationships predictors may have with the response variable

Exploratory Data Analysis

```
library("ggplot2")
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## 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
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

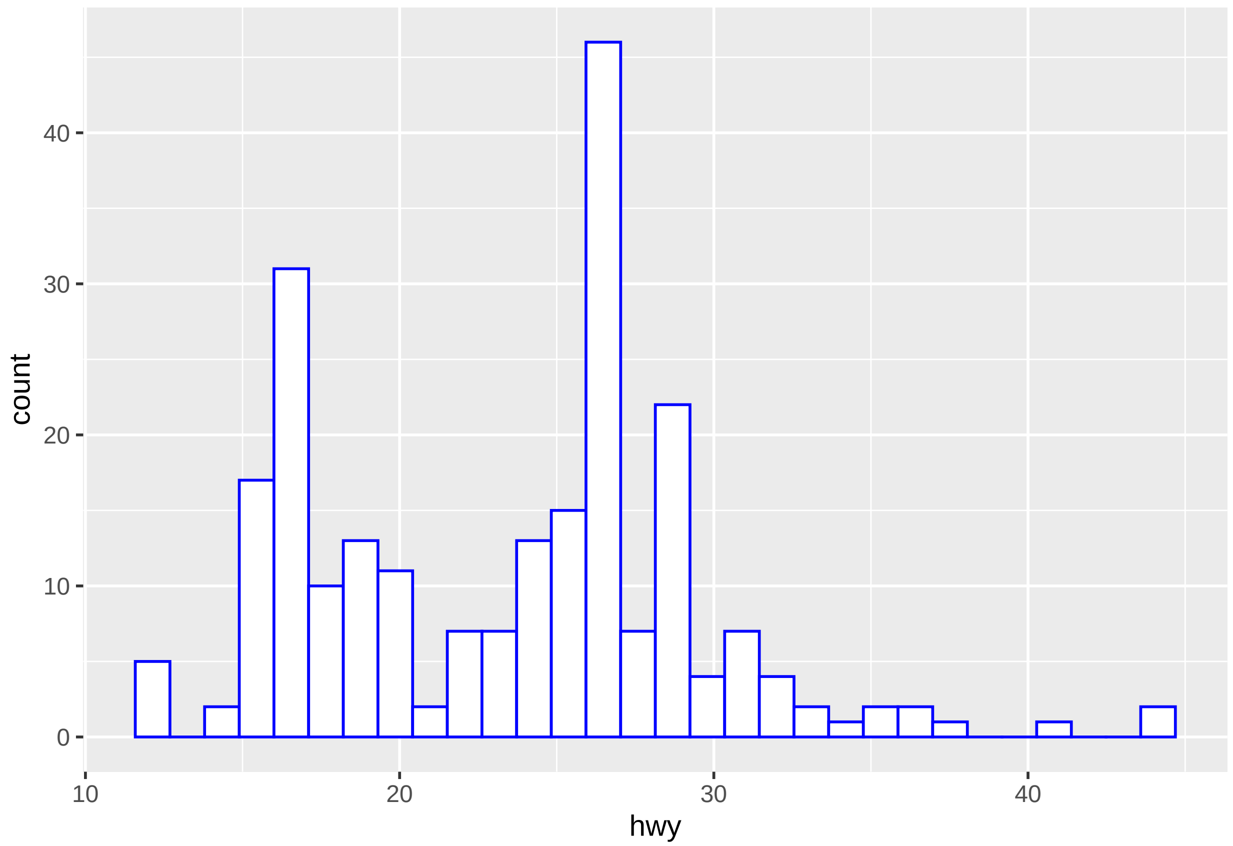
```
library("corrplot")
```

```
## corrplot 0.92 loaded
```

Exercise 1

```
ggplot(mpg, aes(x=hwy)) + geom_histogram(color = "blue", fill = "white")
```

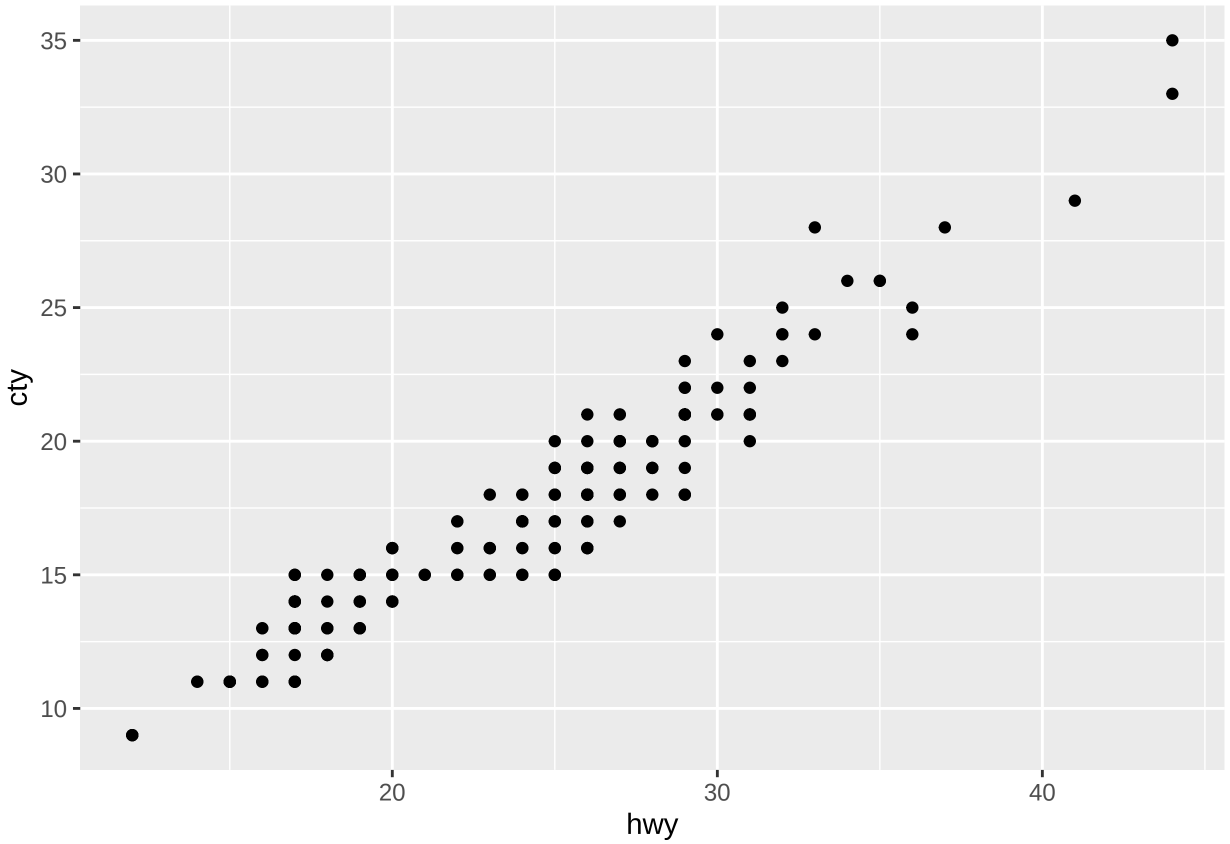
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In the histogram i can see that the mpg that appears the most often is about 27 mpg with 15 being a close runner up. The highest value is about 45 with the lowest hwy mpg being about 2. The following data has a right skew.

Exercise 2

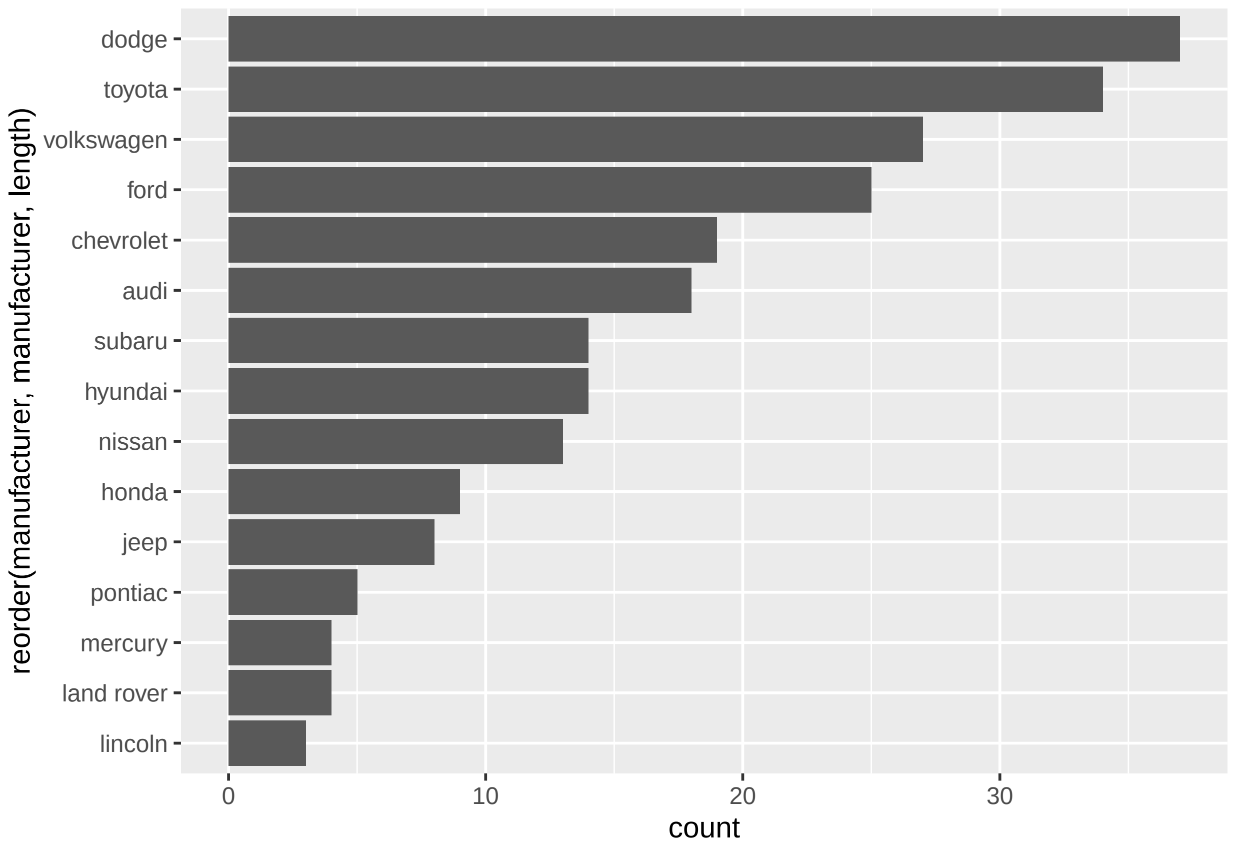
```
ggplot(mpg, aes(hwy,cty)) + geom_point()
```



The two variables seem to have a linear relationship with positive correlation, meaning that a car with high mpg in the city is almost guaranteed to have a higher mpg on the hwy when compared to other cars with lower mpg in the city.

Exercise 3

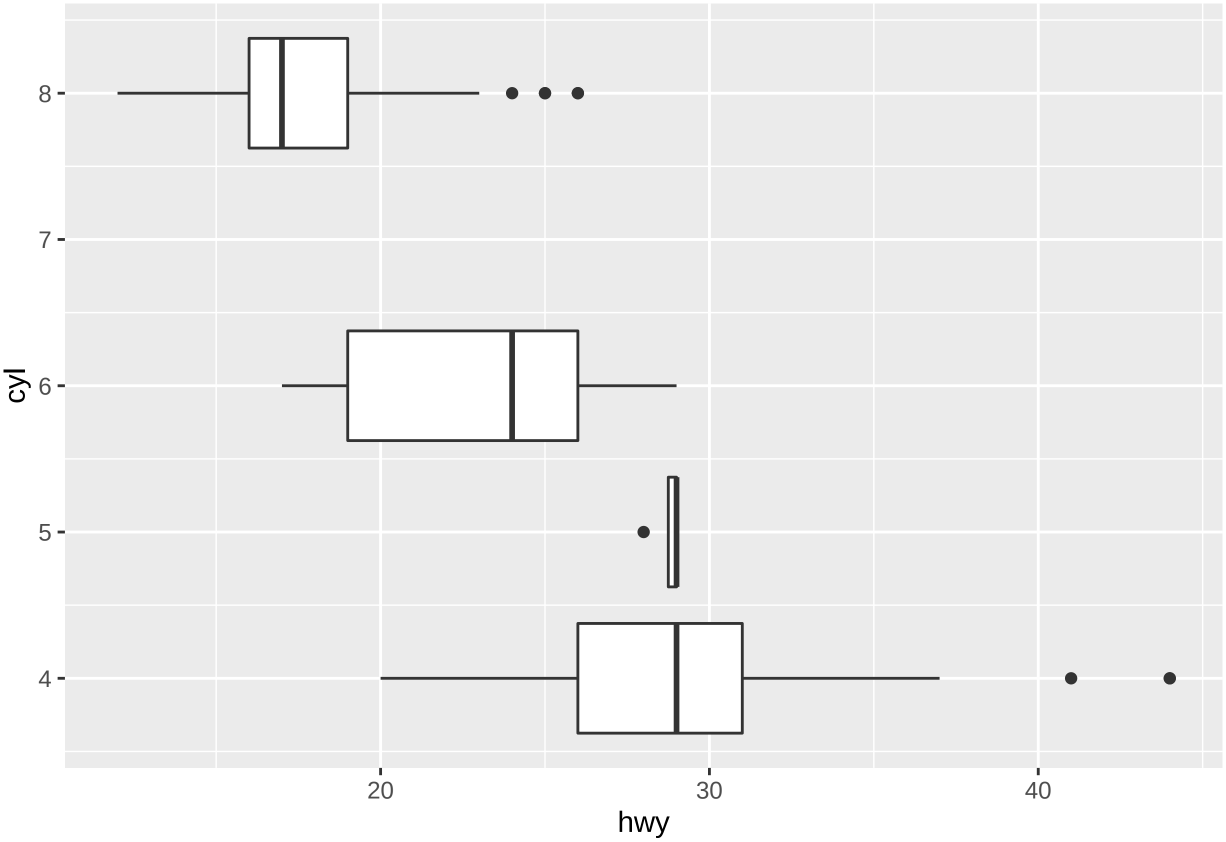
```
ggplot(mpg, aes(y = reorder(manufacturer,manufacturer,length))) + geom_bar()
```



Dodge produced the most cars, at around 38, and Lincoln produced the least cars at roughly 3

Exercise 4

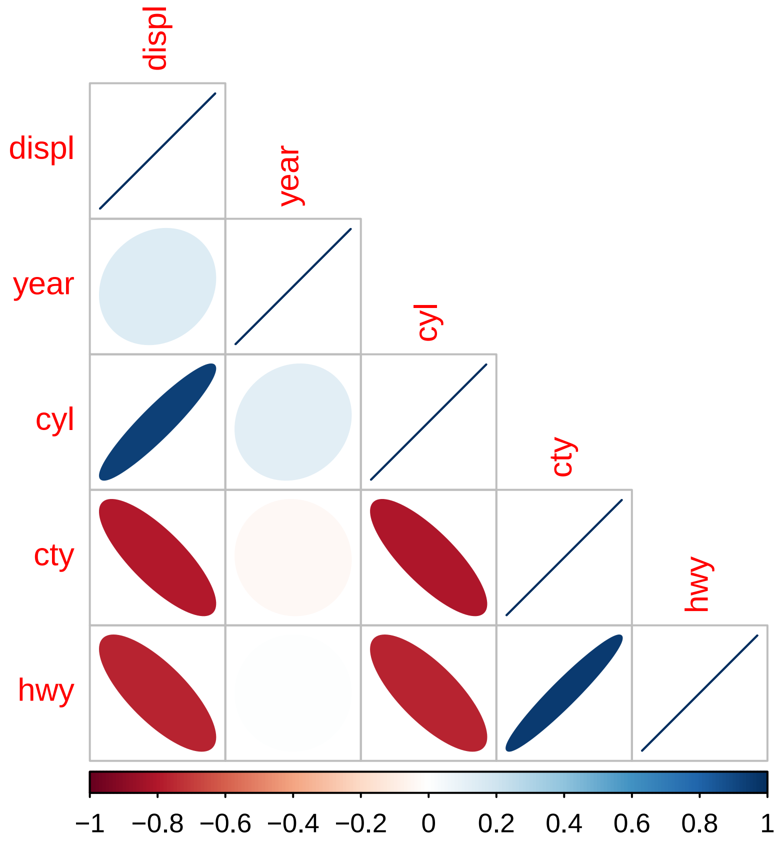
```
ggplot(mpg, aes(hwy, cyl)) + geom_boxplot(aes(group = cyl))
```



On average, cars with fewer cylinders have a higher mpg on the highway.

Exercise 5

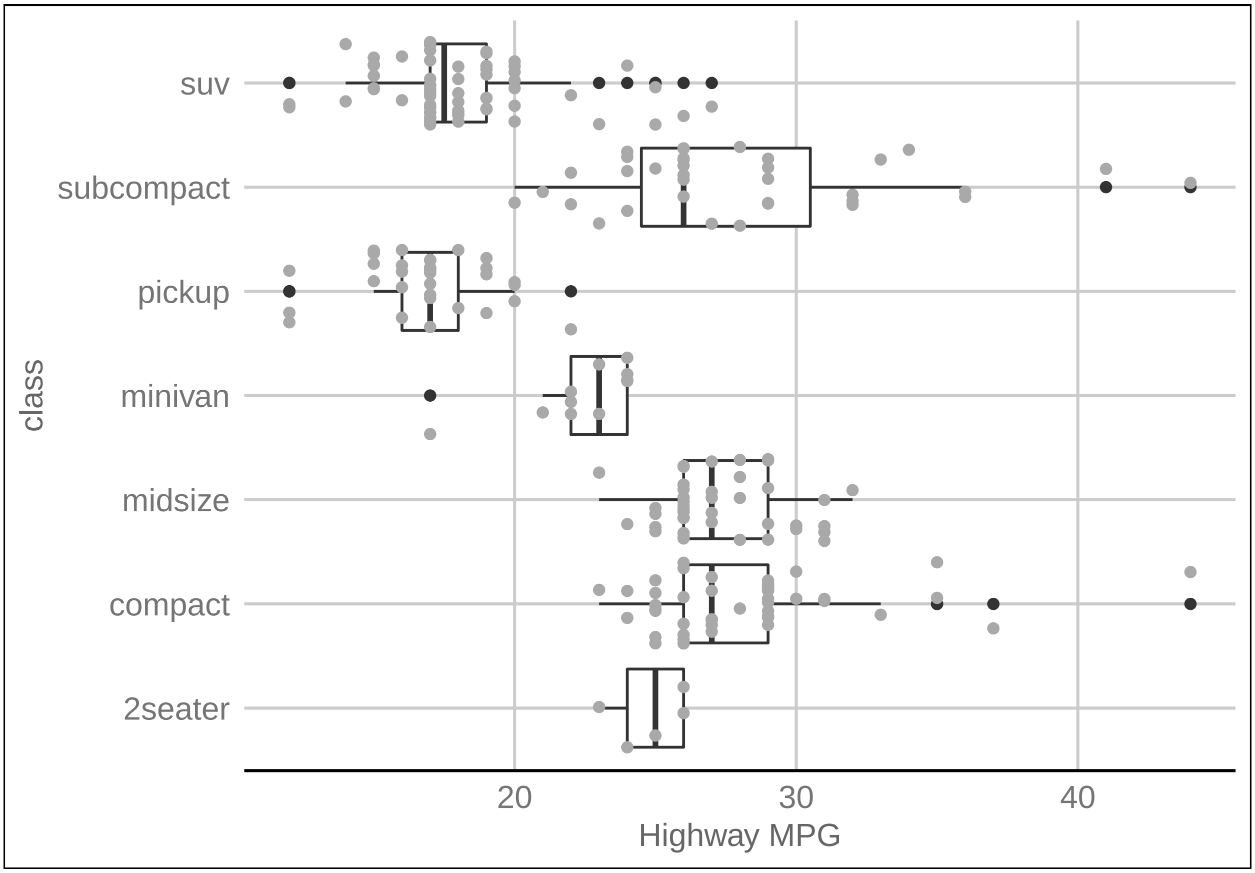
```
mpg_small = select_if(mpg, is.numeric)

data = cor(mpg_small)

corrplot(data, method = 'ellipse', type = 'lower')
```



cty and hwy have a positive correlation, as well as cyl and disp. On the other side, cty and displ, hwy and displ, city and cyl, and hwy and cyl have negative correlations. These make sense to me as the relationship between cyl, hwy, and cty can be inferred from the graphs above. Although we never graphed displ, since it has a positive correlation with cyl, it only makes sense that it would have a negative correlation with cty and hwy. Furthermore, it seems that year has no real impact on the dataset and only has a correlation value of 0.1 in some cases. Overall no surprises.

Exercise 6

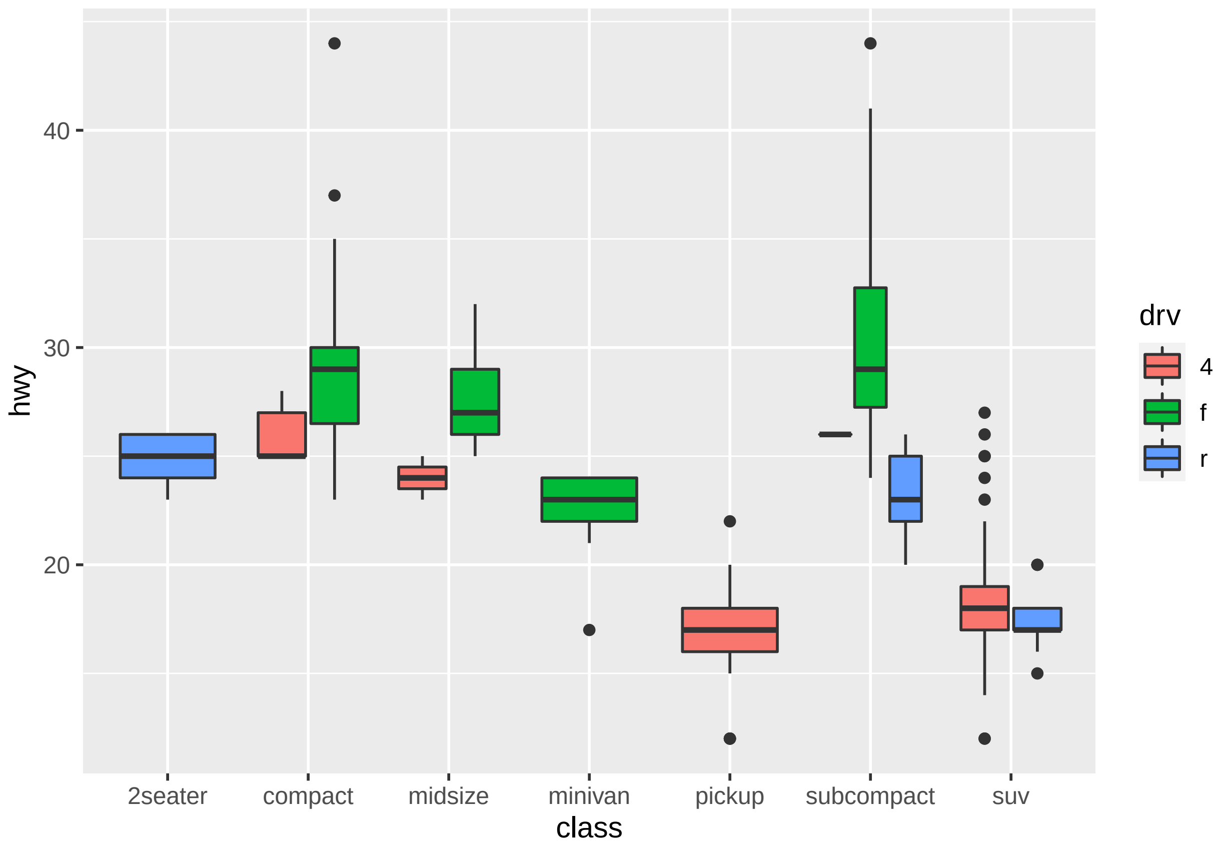
```
library(ggthemes)
ggplot(mpg, aes(hwy, class)) + geom_boxplot() + geom_jitter(width = 0, height = 0.4, color = "darkgrey")
```



From my understanding, these points move every time I rerun the code so this is the closest I could get

Exercise 7

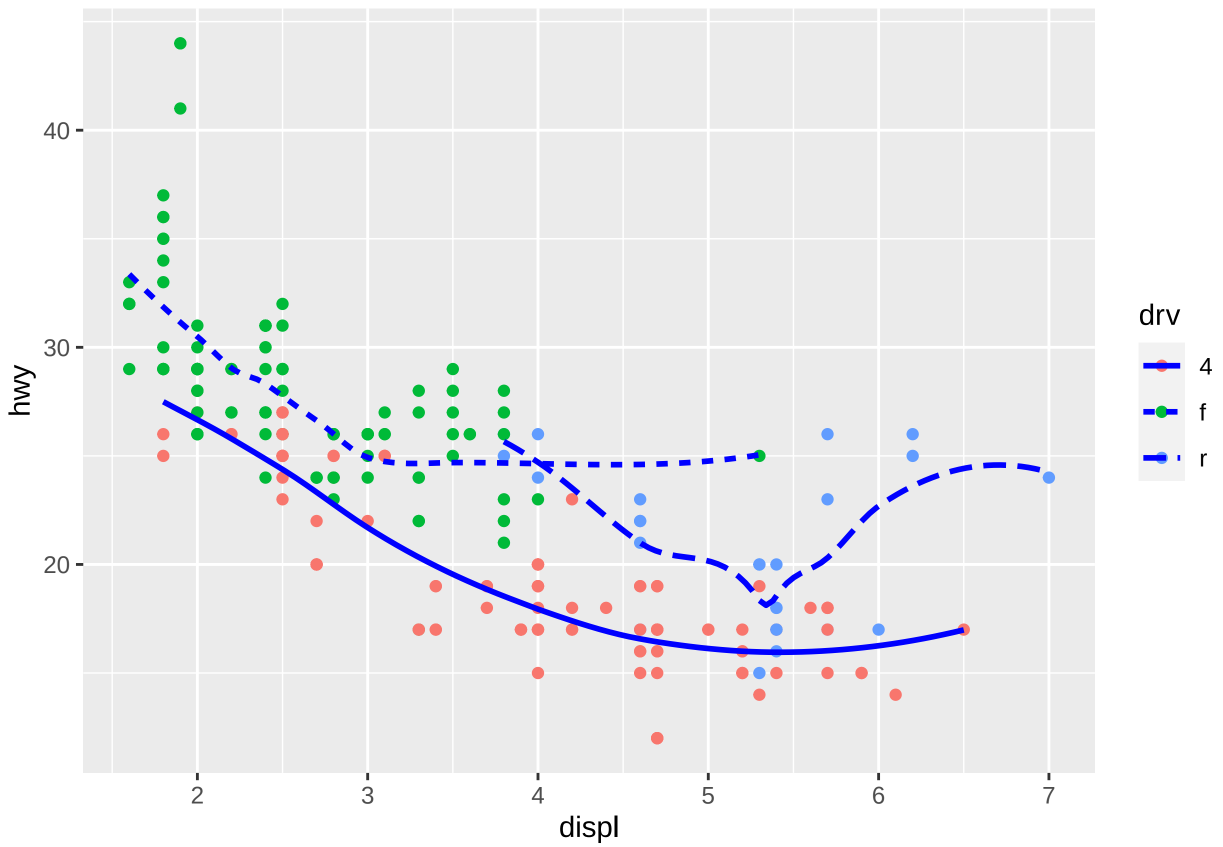
```
ggplot(mpg, aes(class,hwy, fill = drv)) + geom_boxplot() + scale_x_discrete(expand = )
```



Exercise 8

```
ggplot(mpg, aes(displ, hwy, color = drv, linetype = drv)) + geom_point() + geom_smooth(se = F, method =
```

```
## 'geom_smooth()' using formula 'y ~ x'
```


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