Homework Assignment 1

Jai Uparkar

October 02, 2022

```
## corrplot 0.92 loaded
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## 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()
```

Written Response

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

Unsupervised learning is when your model learns without a supervisor which means that the model only contains the predictor variables and no response variables. Some examples of unsupervised learning is PCA, k-means clustering, and neural networks. Supervised learning is when your model learns with a supervisor which means that the model only contains both the predictor variables and the response variables. Some examples of supervised learning is linear regression, logistic regression, k-nearest neighbors, and decision trees. In essence, the main difference between supervised and unsupervised learning is the presence of labelled training data. Supervised and unsupervised learning are types of machine learning

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

A regression model's response variable is quantitative (continuous) while a classification model's response variable is qualitative. This is how the two models differ in the context of machine learning.

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

Two common metric for regression ML problems are the training/test mean squared error (MSE) and the R-squared. Two common metric for classification ML problems are the training/test error rate and the area under the ROC curve.

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 are statistical models that best visually emphasizes a trend in the data like using a parabola in a scatter plot to demonstrate a quadratic relationship between the variables.

Predictive models are statistical models that best predicts the response variables with the least reducible errors and thus are most focused on finding the best combinations of features that fits the model the best.

Inferential models are statistical models that are used to infer and state relationships between the predictors and response variables and are most often used to test theories and causal claims.

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:

Mechanistic means that you assume a parametric form to describe the relationship between the independent and dependent variables which often doesn't match the true function f that describes their relationship. Empirically driven means that the models makes no assumption about the function f or rather the relationship between the independent and dependent variables. Empirically driven models often require a much larger number of observations compared to mechanistic models and as such as much more flexible by default. These types of models are both predictive models.

In general, mechanistic models are easier to understand because often fit simple parametric models and empirically driven models are more flexible and thus more complex and harder to interpret.

Because empirically driven models are more flexible, they do have a lower bias and a higher variance. On the other hand, mechanistic models have a higher bias and a lower variance.

6) A political candidate's campaign has collected some detailed voter history data from their constituents. 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? (ii) How would a voter's likelihood of support for the candidate change if they had personal contact with the candidate?

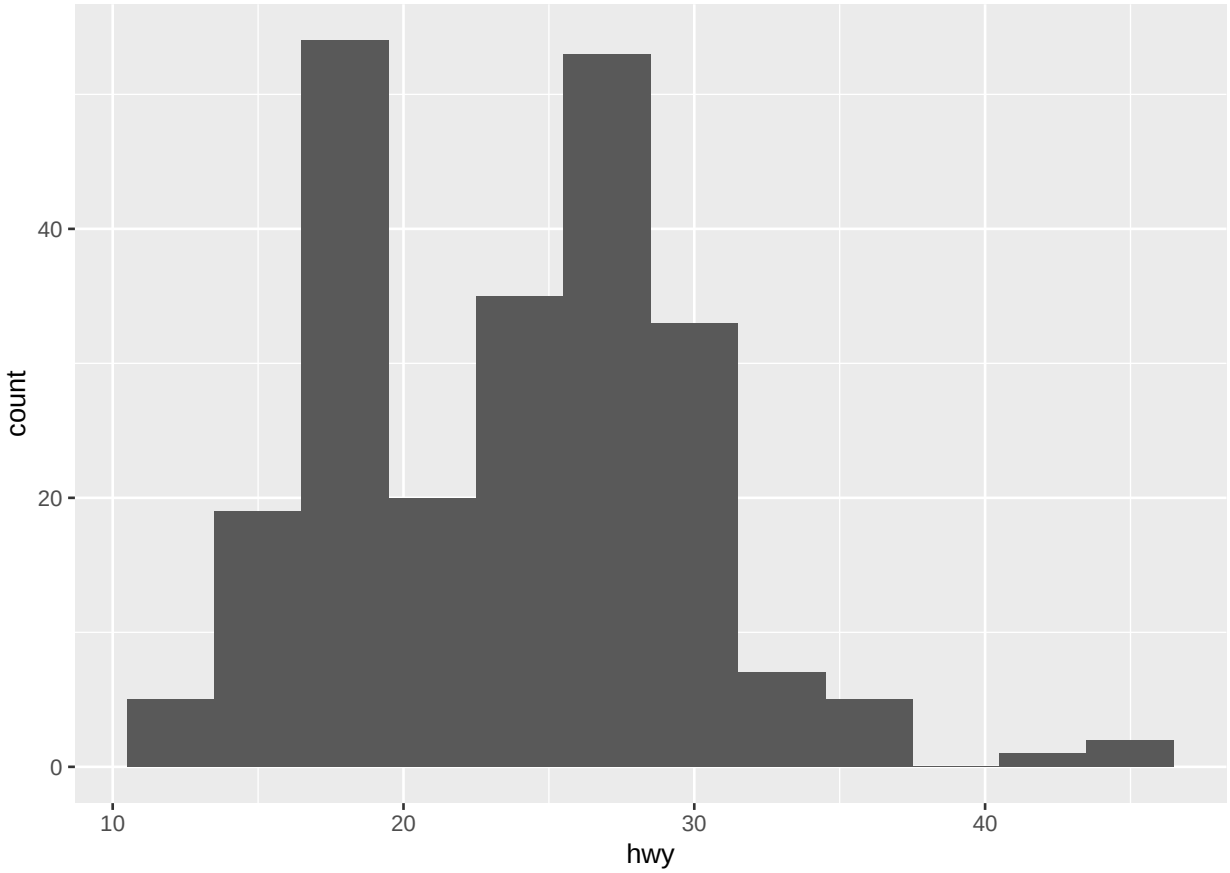
(i) This is a predictive model because the campaign is interested in the best combinations factors of voter qualities that maximizes the probability of winning for the candidate.

(ii) This is an inferential model because the campaign is interested in testing the theory of whether or not the the voter's support for the candidate would changed based on certain conditions. In the end, the campaign wants to see if there is a relationship between the voter's support for the candidate and candidate's personal contact with the voter.

Exploratory Data Analysis

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

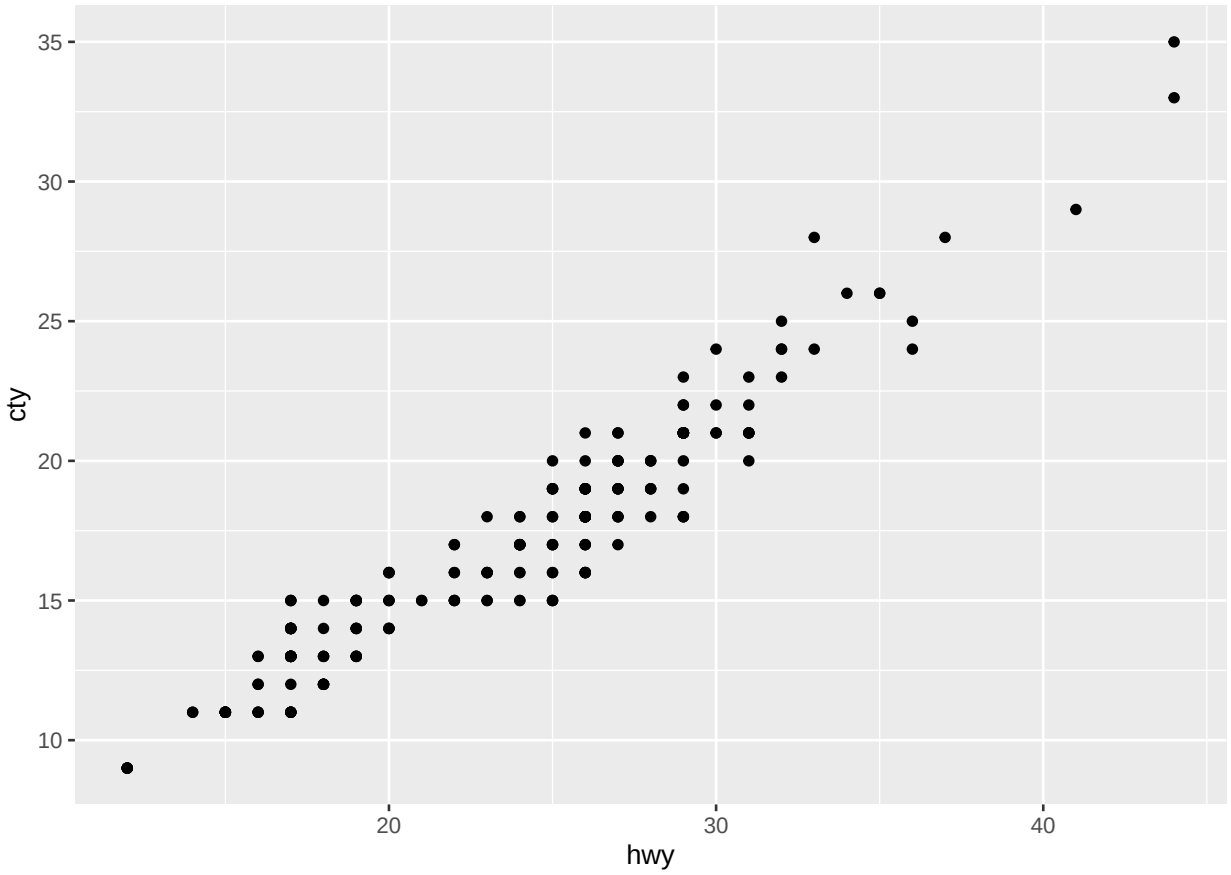
```
ggplot(mpg, aes(x=hwy)) + geom_histogram(binwidth = 3)
```



The histogram appears to be slightly bimodal, with two bumps occurring around 17.5 highway miles per gallon and 26 highway miles per gallon. The graph also appears to be slightly skewed to the right and here are not many cars with a mpg higher than 40.

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?

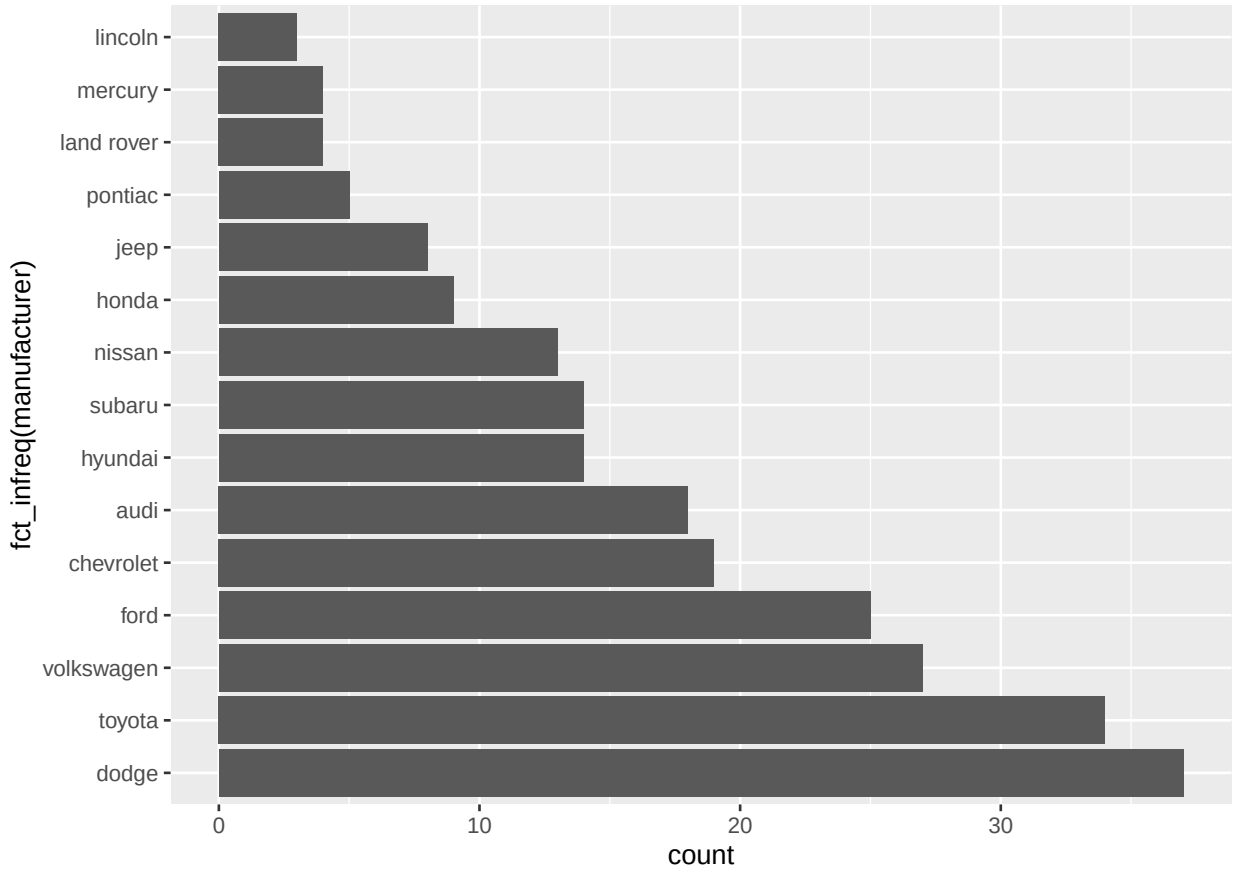
```
ggplot(mpg, aes(x=hwy, y=cty)) + geom_point()
```



From looking at the scatterplot, I can definitely tell that there is a positive relationship between the highway miles per gallon and the city miles per gallon for popular cars between 1999 and 2008. However, there appears to be an almost grid like pattern that surrounds the positive relationship between the 2 variables and there don't seem to be many points on the graph despite there being 234 observations for the data set making me believe that a lot of the points are overlapped. After observing this, I took a closer look at the data and found that both variables of interest only contain whole numbers.

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?

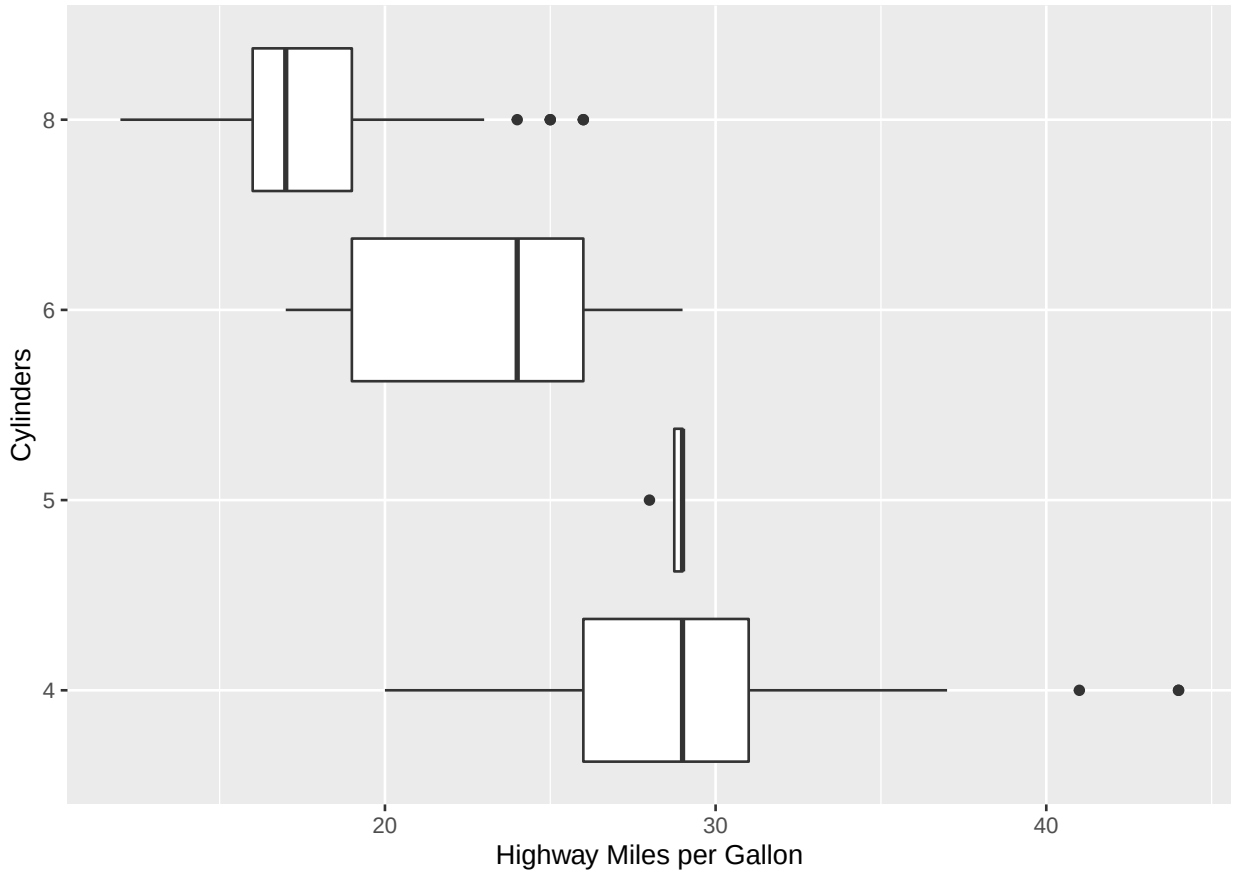
```
ggplot(mpg, aes(x=fct_infreq(manufacturer)))  + geom_bar() + coord_flip()
```



Lincoln produced the least cars while Dodge produced the most.

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

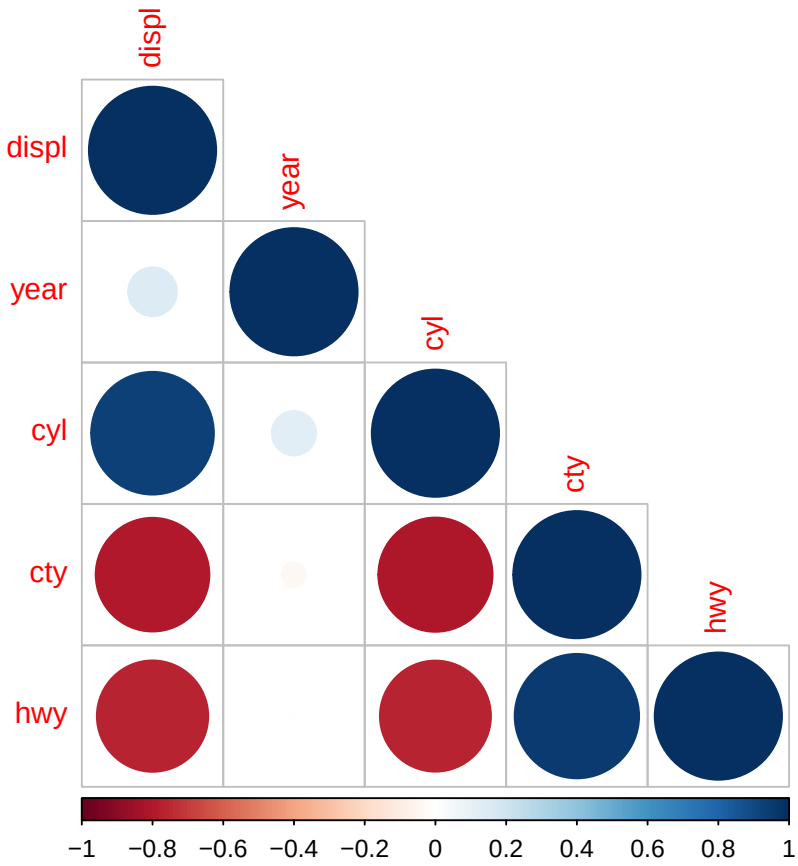
```
ggplot(mpg, aes(x=hwy, y=factor(cyl))) +
    geom_boxplot() + ylab("Cylinders") + xlab("Highway Miles per Gallon")
```



From the graph above we see that as the number of cylinders increases, the highway miles per gallon tends to decrease. There is an inverse relationship between the 2 variables. We see that cars with 8 4 cylinders have the highest highway miles per gallon values and that cars with 8 cylinders have the least mileage on average and the boxplot for cars with 5 cylinders is very small because of the small sample size.

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.)

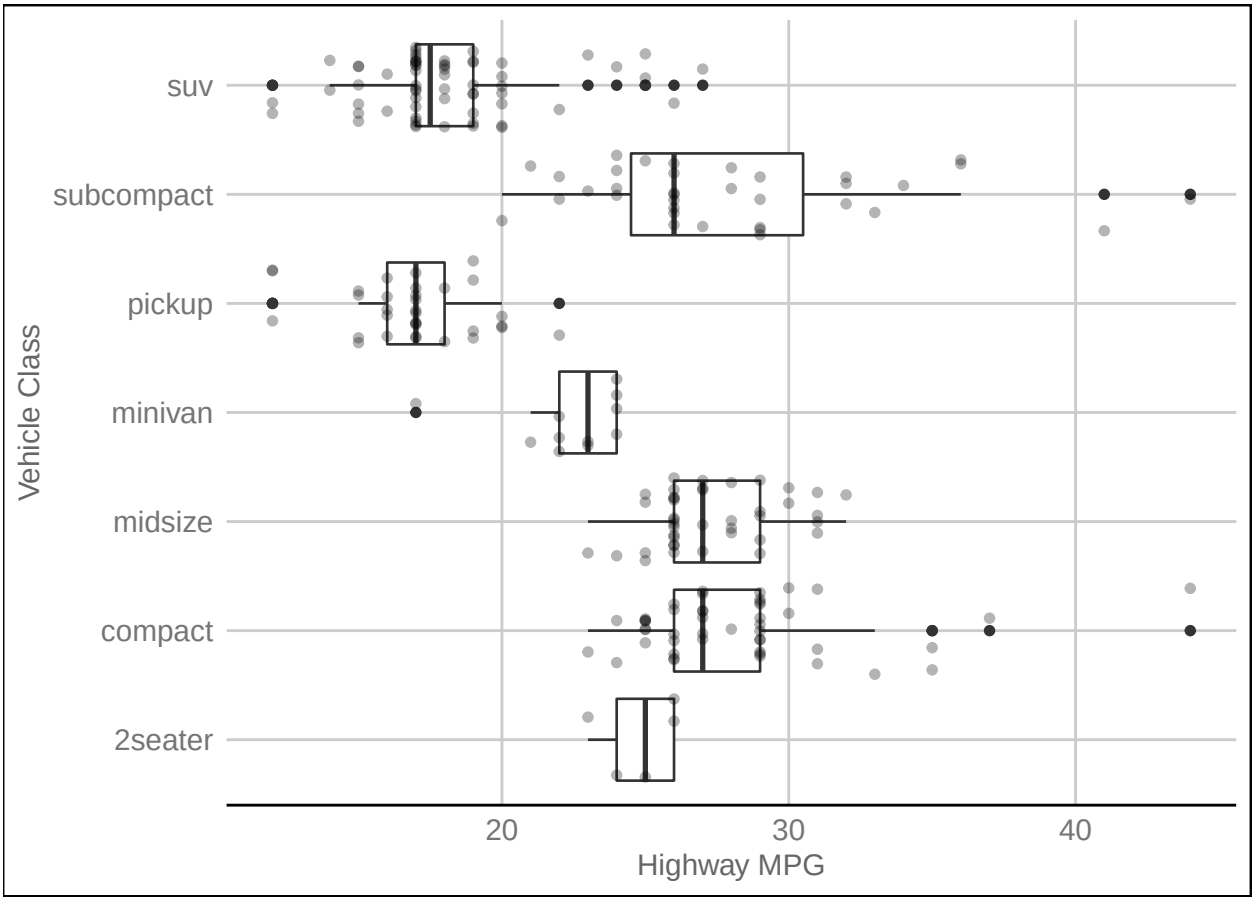
```
new_mpg<- mpg[,c(3,4,5,8,9)] # only selectint numeric variables
corrplot(cor(new_mpg), type = 'lower')
```



The variables that are positively correlated with each other are: city mileage & highway mileage and number of cylinders & engine displacement. The variables that are negatively correlated with each other are: number of cylinders & highway mileage, number of cylinders & city mileage, engine displacement & city mileage, and engine displacement & highway mileage. These relationships do make sense to me especially because engine displacement is calculated based off the volume of all the clyinders and so the negative correlation between each type of mileage and number of cylinders & engine displacement makes sense. The correlation relationship we observed were reflected in the relationships we found earlier in the homework so those findings make sense. There are no relationships which really surprise me.

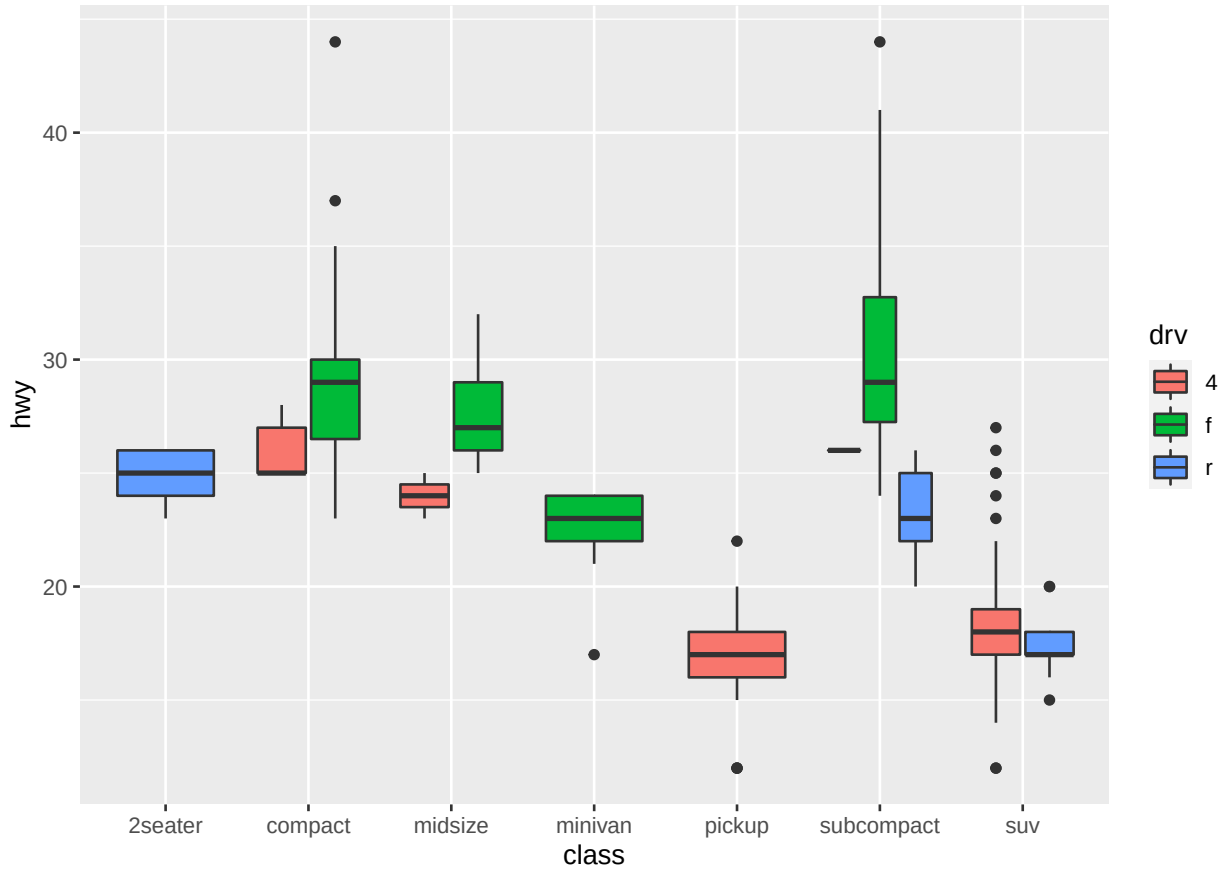
6) Recreate the following graphic, as closely as you can. Hint: Use the ggthemes package.

```
ggplot(mpg, aes(x=hwy, y=factor(class))) +
  geom_boxplot() + ylab("Vehicle Class") + xlab("Highway MPG") +
  geom_jitter(alpha= 0.3, width = 0) + theme_gdocs()
```



7) Recreate the following graphic.

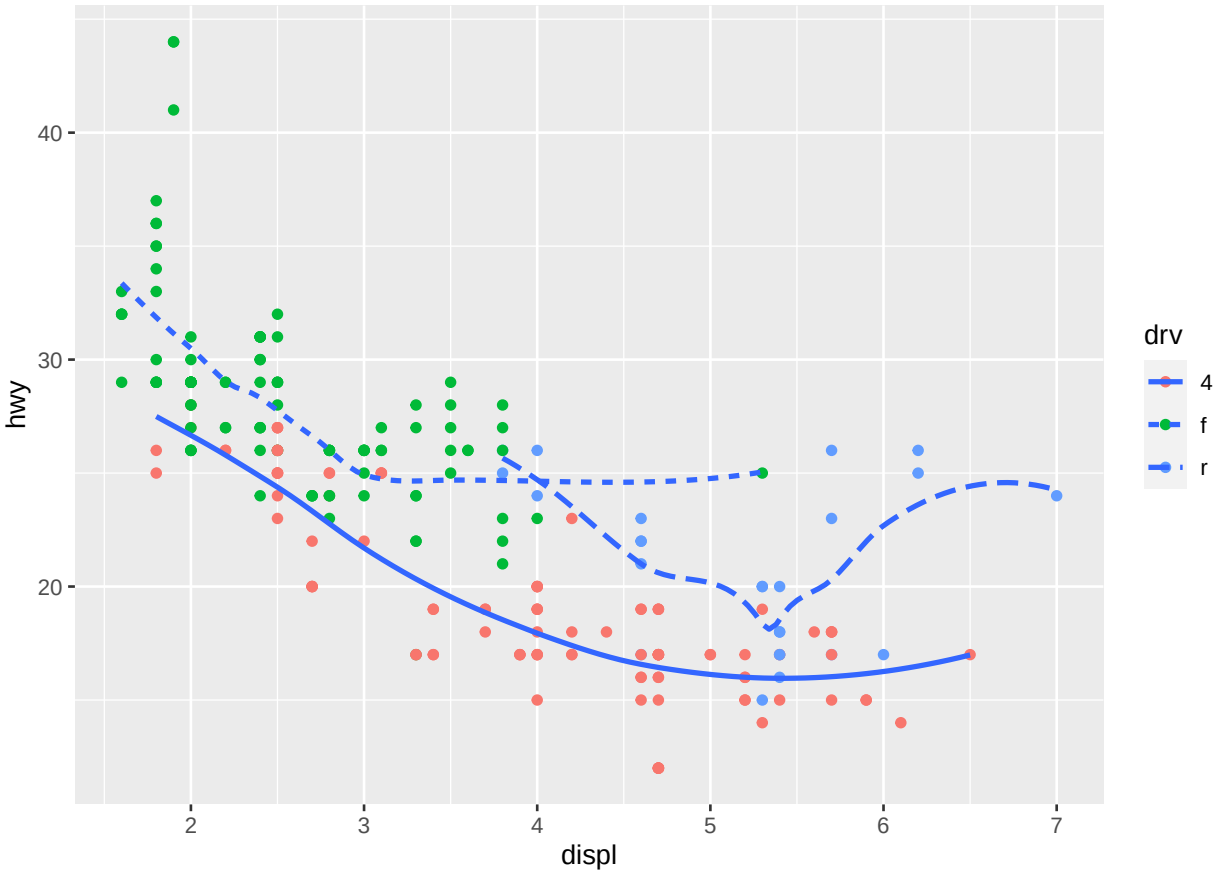
```
ggplot(mpg, aes(x=hwy, y=class, fill = drv)) +
  geom_boxplot() + ylab("class") + xlab("hwy") + coord_flip()
```



8) Recreate the following graphic.

```
ggplot(mpg, aes(x=displ, y=hwy, fill = drv)) + ylab("hwy") + xlab("displ") +
  geom_point(aes(color = drv)) + geom_smooth(aes(linetype = drv), se = FALSE)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```


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