Homework 1

Evan Ji

2022-10-02

#Homework1 ##Question 1 Define supervised and unsupervised learning. What are the difference(s) between them?

Supervised learning is machine learning in which the data has been labeled already, which means that the machine is given data with the "correct" outcome variable to guide learning. Unsupervised learning is machine learning in which you allow the model to observe features and discover its own information. According to the textbook the difference between them is that the goal of supervised learning "us to predict the value of an outcome measure based on a number of input measures; in unsupervised learning there is no outcome measure, and the goal is to describe the associations and patterns among a set of input measures.

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

The difference between a regression model and a classification model is that a regression model is used to predict a continous data set (numbers), while a classification model is used to predict discrete data sets/class labels (categories).

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

2 commonly used metrics for regression ML problems are the Test MSE (Mean Squared Error) and the Training MSE. 2 commonly used metrics for classification ML problems are Training and Test 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.

Descriptive Model: According to lecture, models are chosen to best visually emphasize a trend.

Predictive Model: According to lecture, models are chosen to best predict Y with minimum reducible error, and they are not focused on hypothesis tests.

Inference Model: According to lecture, models are chosen to best find significant features (predictors), and find causal relationships between outcome and predictor

##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?

Mechanistic models assume a parametric form for f, meaning that there are certain factors (predictors) that affect the score of f, meaning our unknown f won't be perfectly matched. Empirically-driven models make no assumptions on the form of f, and creates a model purely off the given data. These model types differ in what they try to explain. Mechanistic models try to explain why something is, while empirically drivel models explain how something is in reality. They are similar in that they may both suffer from over fitting of the training data.

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

In general I believe a mechanistic model is easier to understand. I believe this as we as humans have an intuitive sense as well as natural tendency to try and account for things to make sense of things in our mind, and it allows us to categorize these factors which satisfies our need to find WHY something is.

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

Both mechanistic and empirically-driven models are vulnerable to overfitting. This is where the model too closely follows the training data and so becomes a non-optimal predictor for the test data, this relates to the bias-variance tradeoff as we try to minimize both the bias and variance at the same time to create a model that can accurately adjust to our test data, and not be too constricted by the bias of the training data.

##Question 6 A political candidate's campaign has collected some detailed voter history data from their constituents. The campaign is interested in two questions: 1.Given a voter's profile/data, how likely is it that they will vote in favor of the candidate? 2.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.

Question 1 is predictive, as we are trying to accurately predict y, the probability that a voter will vote in favor of the candidate. Question 2 is inferential, as we are testing to see if there is some sort of causal relation between meeting the candidate and voting for them.

```
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   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

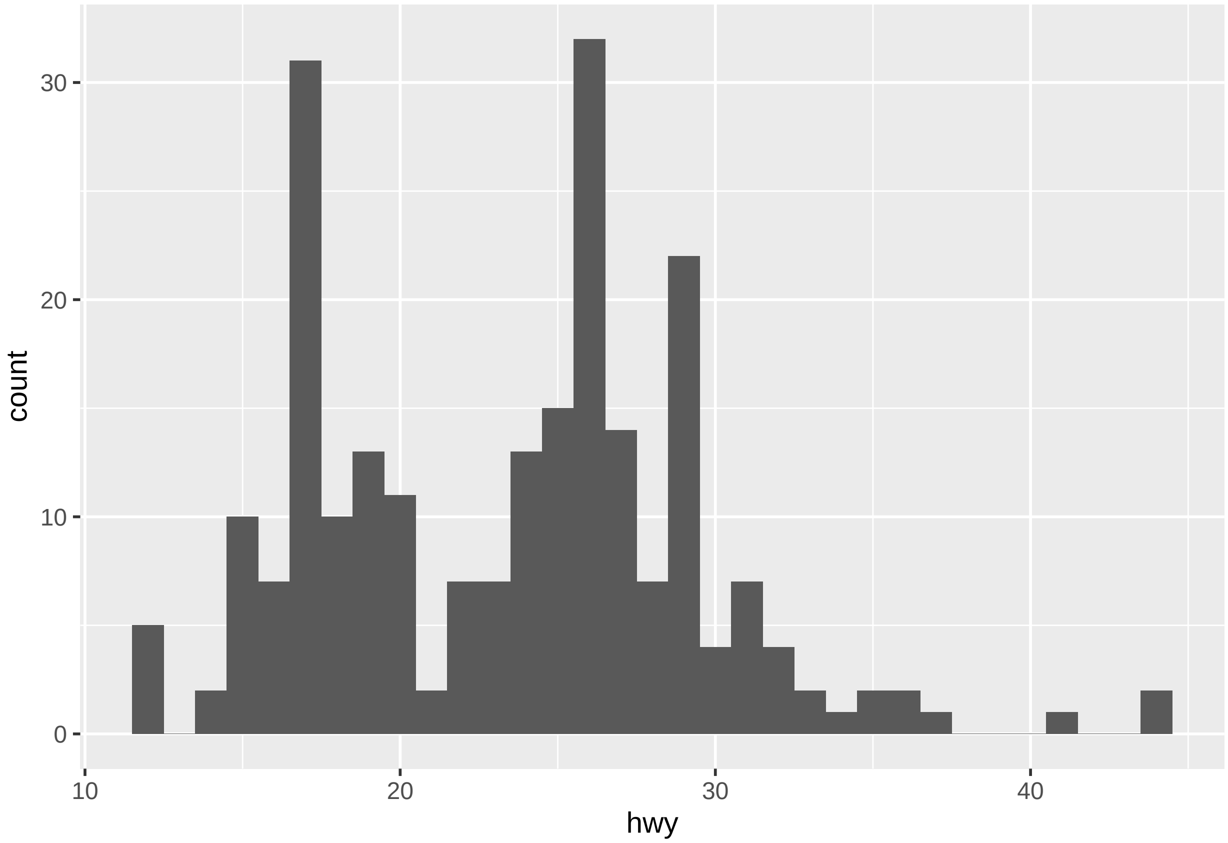
```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(ggthemes)
```

#Exploratory Data Analysis ##Exercise 1

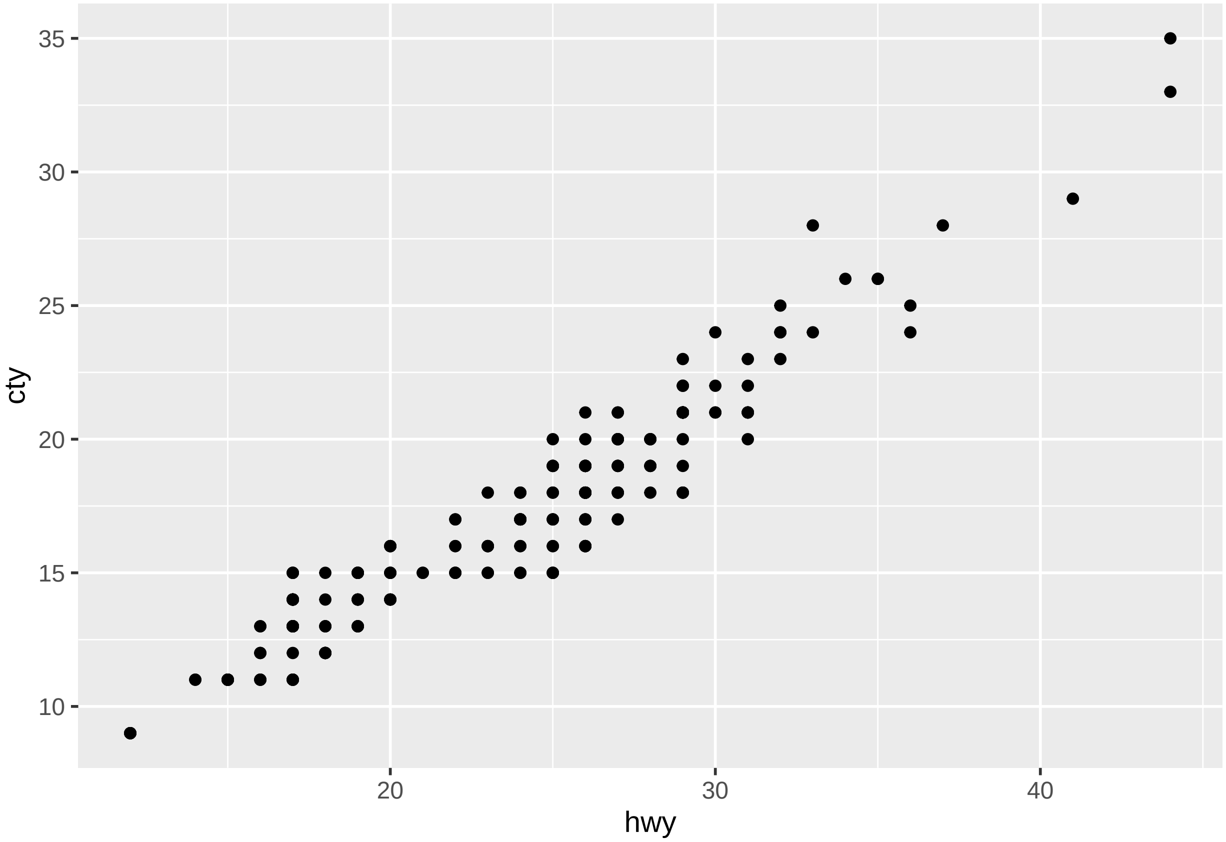
```
c <- ggplot(mpg, aes(hwy)) + geom_histogram(binwidth=1)
c
```



There seems to be a bimodal distribution with peaks at 13 mpg and 26 mpg for highway.

##Exercise 2

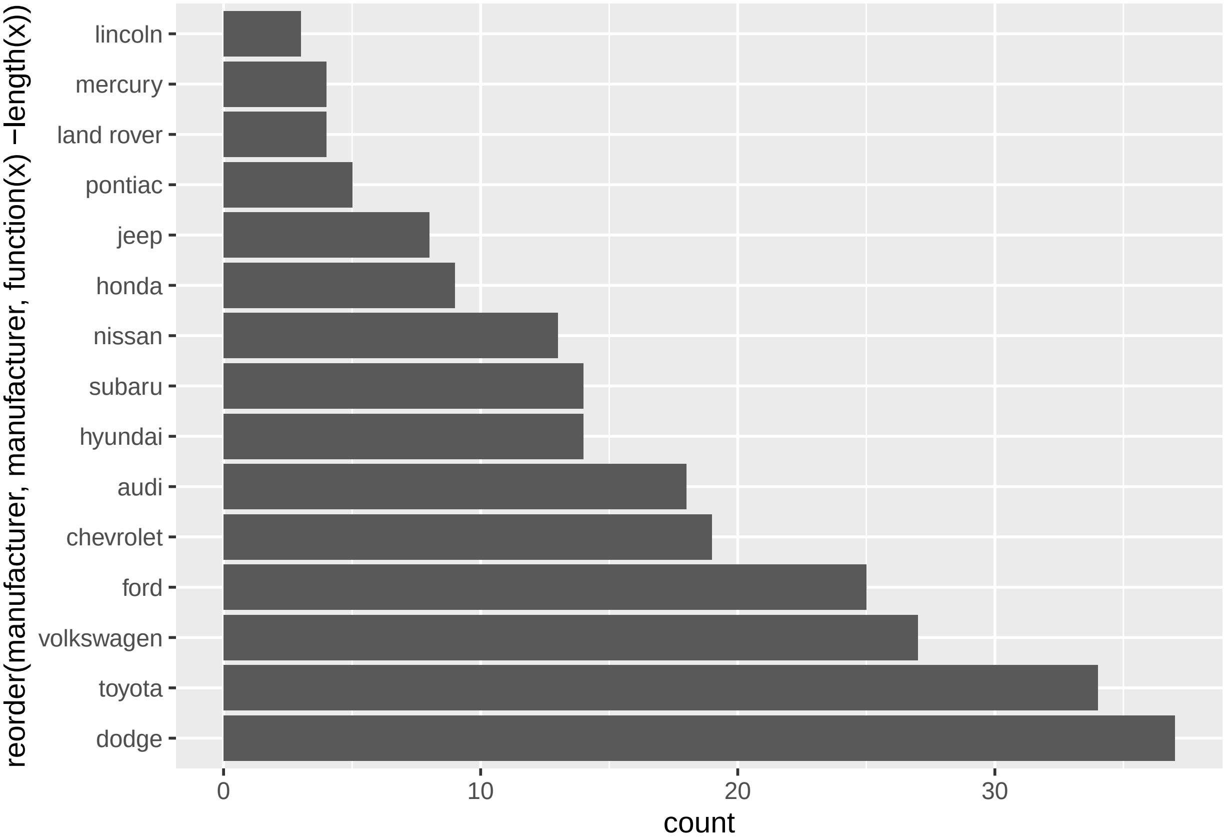
```
a <-ggplot(mpg,aes(x=hwy,y=cty)) + geom_point()
a
```



There seems to be a positive correlation between hwy and city. This means that as hwy mpg increases, cty mpg also increases.

##Exercise 3

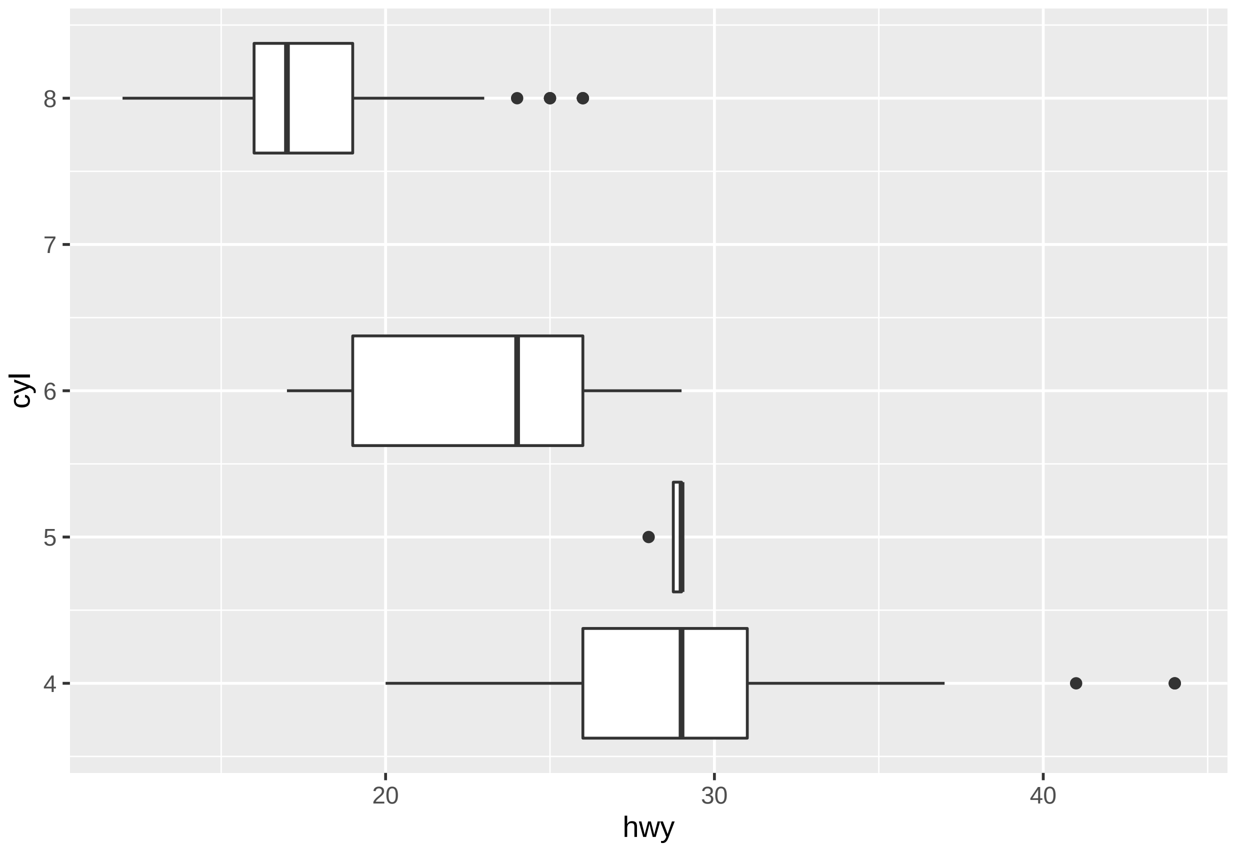
```
b <- ggplot(mpg, aes(x=reorder(manufacturer,manufacturer,function(x)-length(x)))) + geom_bar()
b +coord_flip()
```



According to this horizontal bar plot, Dodge produced the most cars, and Lincoln produced the least.

##Exercise 4

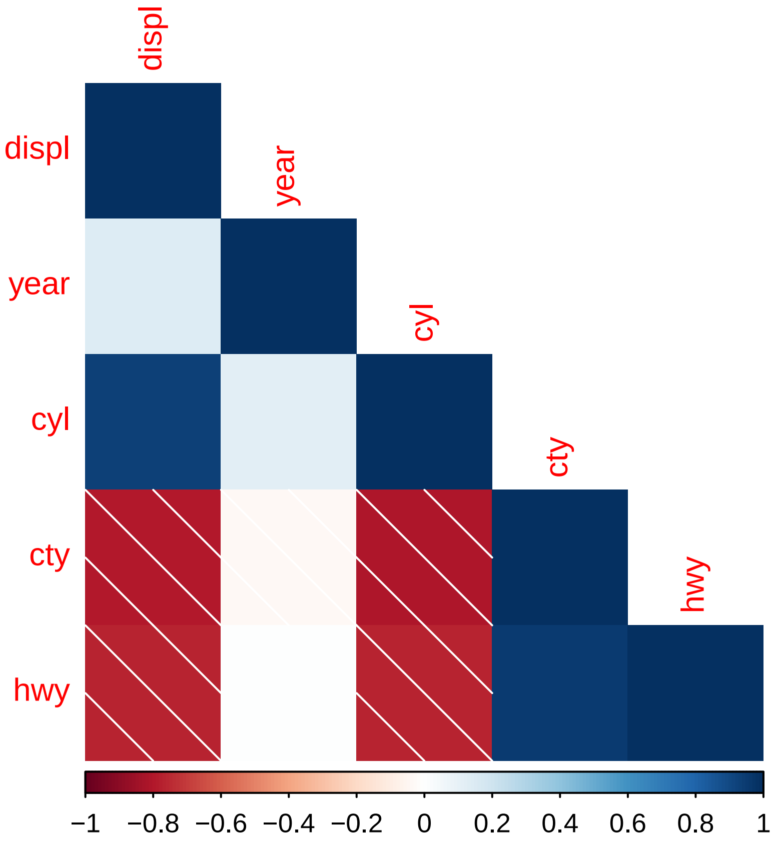
```
d <- ggplot(mpg,aes(y=cyl, x=hwy, group=cyl)) + geom_boxplot()
d
```



It seems there is is a pattern in which as the number of cylinders decrease the highway efficiency increases.

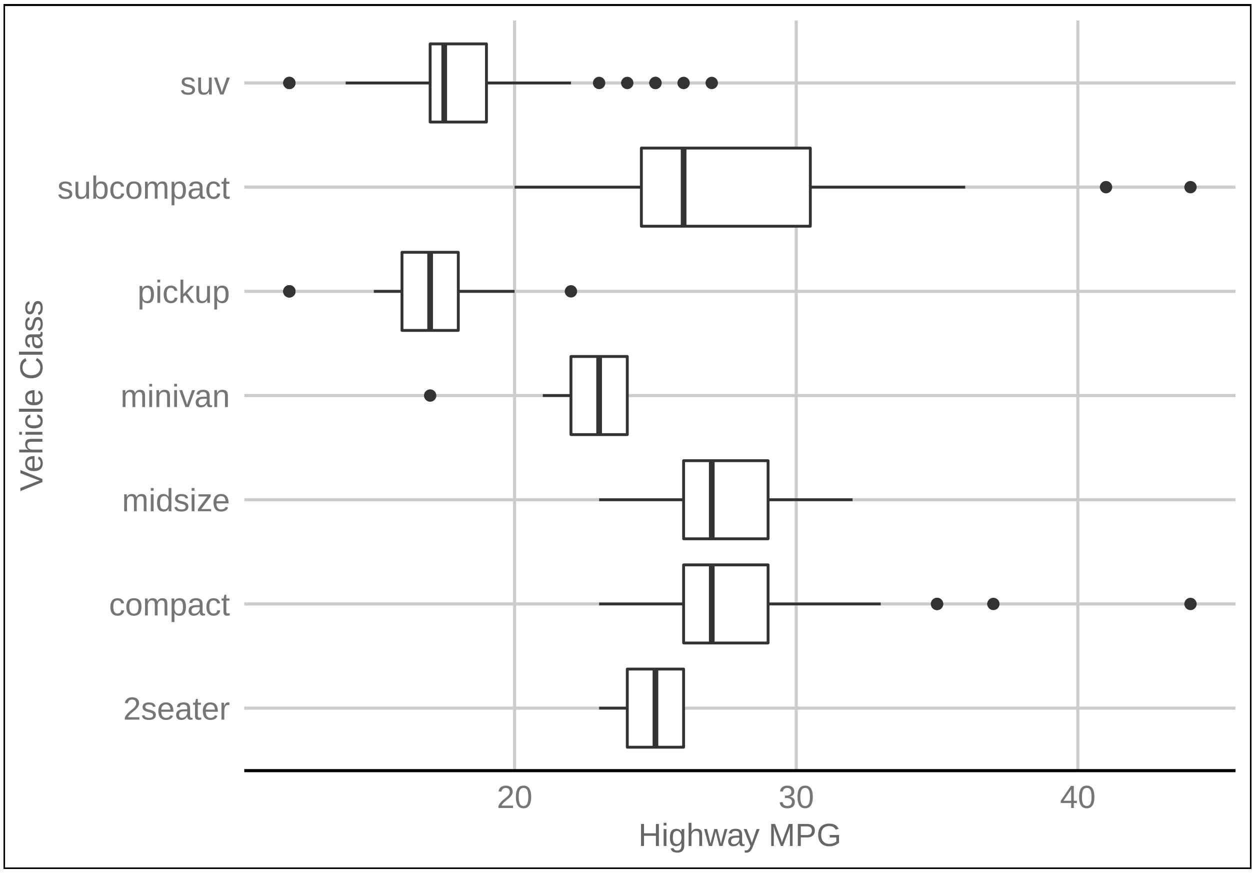
##Exercise 5

```
mpgnum <- Filter(is.numeric,mpg)
M = cor(mpgnum)
corrplot(M, type='lower',method='shade')
```



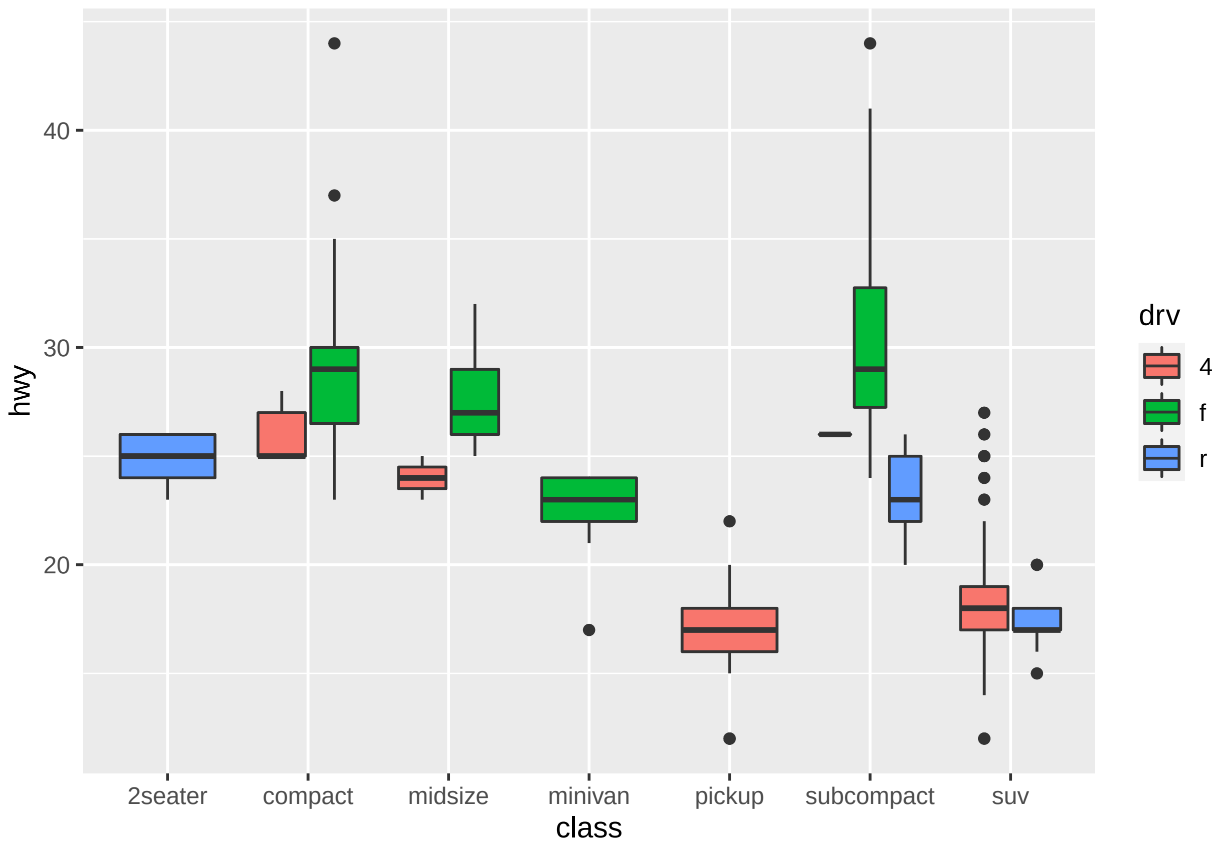
#231 Students ##Exercise 6

```
e <- ggplot(mpg,aes(y=class, x=hwy, group=class)) + geom_boxplot()+ theme_gdocs()+ labs(y= "Vehicle Cla
e
```



##Exercise 7

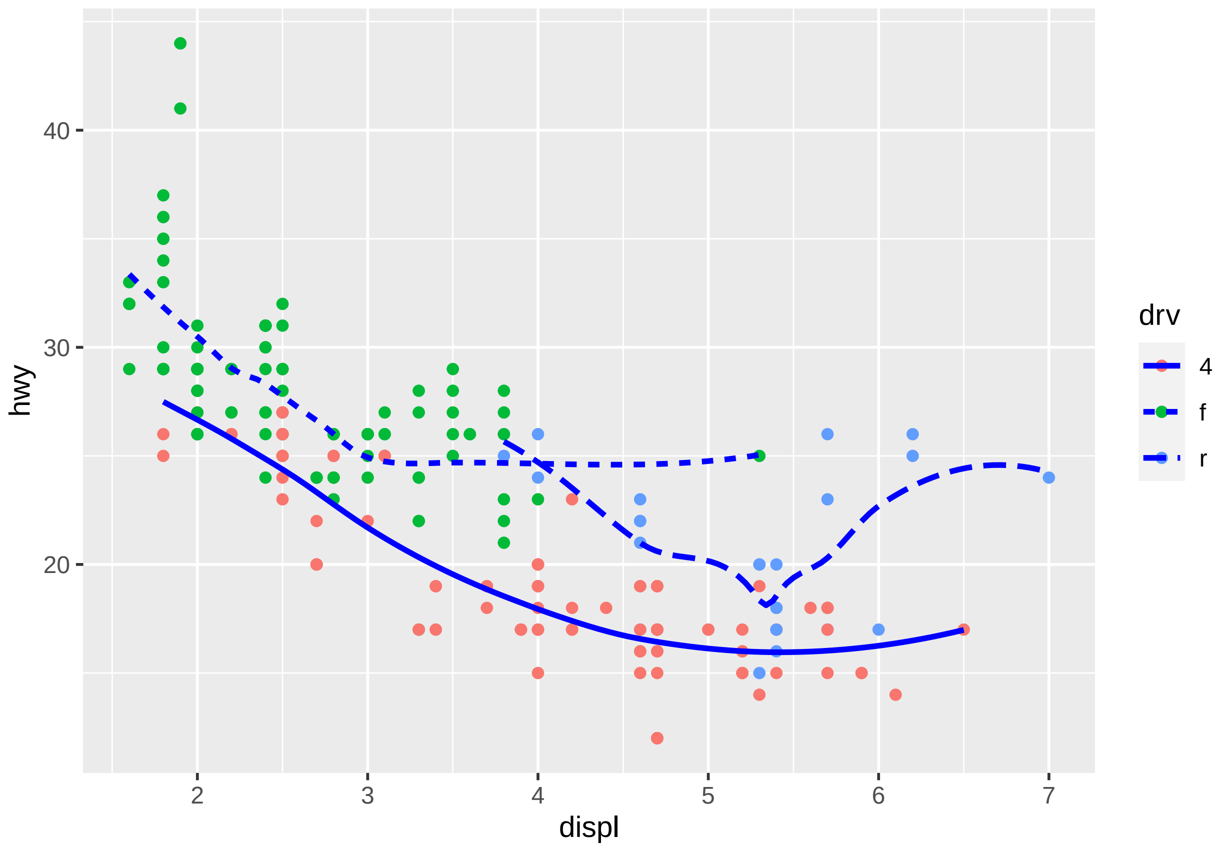
```
f <- ggplot(mpg,aes(y=hwy, x=class,fill=drv)) + geom_boxplot()
f
```



##Exercise 8

```
g <- ggplot(mpg,aes(x=displ,y=hwy,fill=drv,color=drv)) +geom_point()+geom_smooth(se=FALSE,color='blue',
g
```

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



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