# Pstat231HW1

Zihao Yang

2022-04-03

```
#install.packages("tidyverse")
#install.packages("tidymodels")
#install.packages("ISLR")

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 0.2.0 --

## v broom        0.7.12     v rsample      0.1.1
## v dials        0.1.0      v tune         0.2.0
## v infer        1.0.0      v workflows    0.2.6
## v modeldata    0.1.1      v workflowsets 0.2.1
## v parsnip      0.2.1      v yardstick    0.0.9
## v recipes      0.2.0

## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## * Learn how to get started at https://www.tidymodels.org/start/
```

```
library(ISLR)
```

Q1

Supervised learning: The actual data of response Y is known, so it can serve as the supervisor of the supervised learning. Comparing the actual data of response Y with predicted response Y, we can determine the accuary of our models. With predictors, we can accurately predict future response, understand how predictors affect response, find the "best" model for response, and assess the quality of our predictions and (or) estimation. It works with well labeled data.

Unsupervised learning: We don't know the actual data of response Y. We only know the predictors, so there is no supervisor in this type of learning. It deals with unlabeled data.

Difference: The difference between supervised and unsupervised is whether there is a supervisor or not. Here the supervisor is the actual data of response Y. If we have the actual Y, then the data is well labeled and it is supervised learning. If we don't have the actual Y, then the data is not well labeled and it should be unsupervised learning.

Q2

In the context of the machine learning, classification is the task of predicting a discrete class label(categorical or qualitative), while regression is the task of predicting a continuous quantity(quantitative).

Q3

According to the office hour, we don't need to answer Q3.

Q4

Descriptive models: Choose model to best visually emphasize a trend in data. i.e. using a line on a scatter plot.(from lecture)

Inferential models: Aim is to test theories, (Possibly) causal claims, State relationship between outcome & predictor(s). (from lecture)

Predictive models: Aim is to predict Y with minimum reducible error. Not focused on hypothesis tests. (from lecture)

Q5

Mechanistic: We assume a parametric form for function f, we can choose any suitable models according to the assumptions and estimate the coefficients. It won't match the true unknown f. It can also add parameters which means more flexibility.

Empirically-driven:There is no assumptions about function f, and it requires a large number of observations. It has much more flexibility by default.

Mechanistic needs an assumption for f, while empirically-driven doesn't require that. Empirically-driven is more flexible than mechanistic by default. But they are both over fitting.

I would say mechanistic is easier to understand, because we can use our familiar models to make assumptions. If it is not suitable, then we can make revise and select a better model. Comparing to the empirically-driven that without assumption, mechanistic is much easier to find a way to start approach. It is more controllable.

Q6

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

This should be predictive, because it aims at predicting Y, which is how likely they will vote in favor of the candidate, based on the voter's data, with minimum reducible error. And it is not focused on hypothesis tests.

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

This should be inferential, because it aims at testing a theory. We want to know about the relationship between the personal contact and the likelihood of support for the candidate.

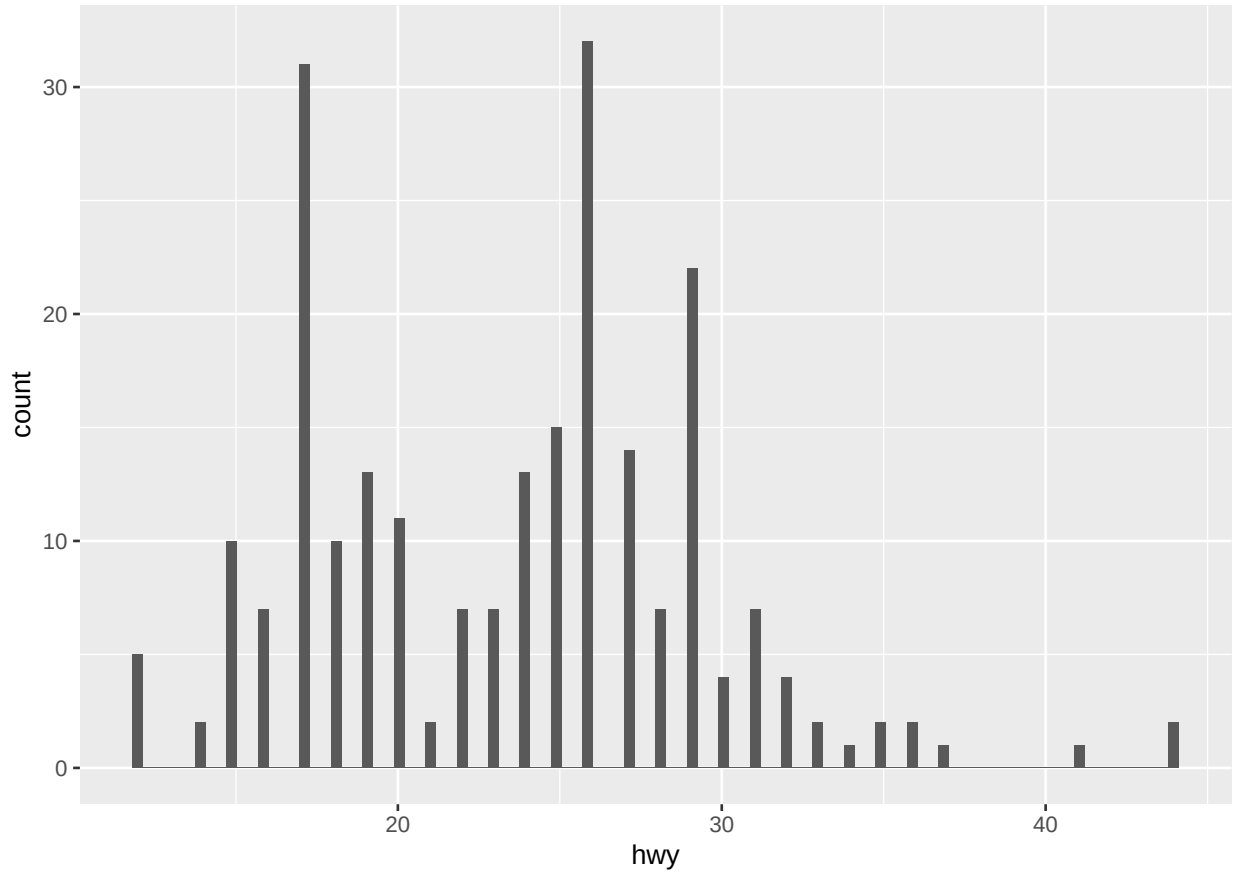
Exercise Part

```
library(ggplot2)
head(mpg)
```

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa~
```

Exercise 1

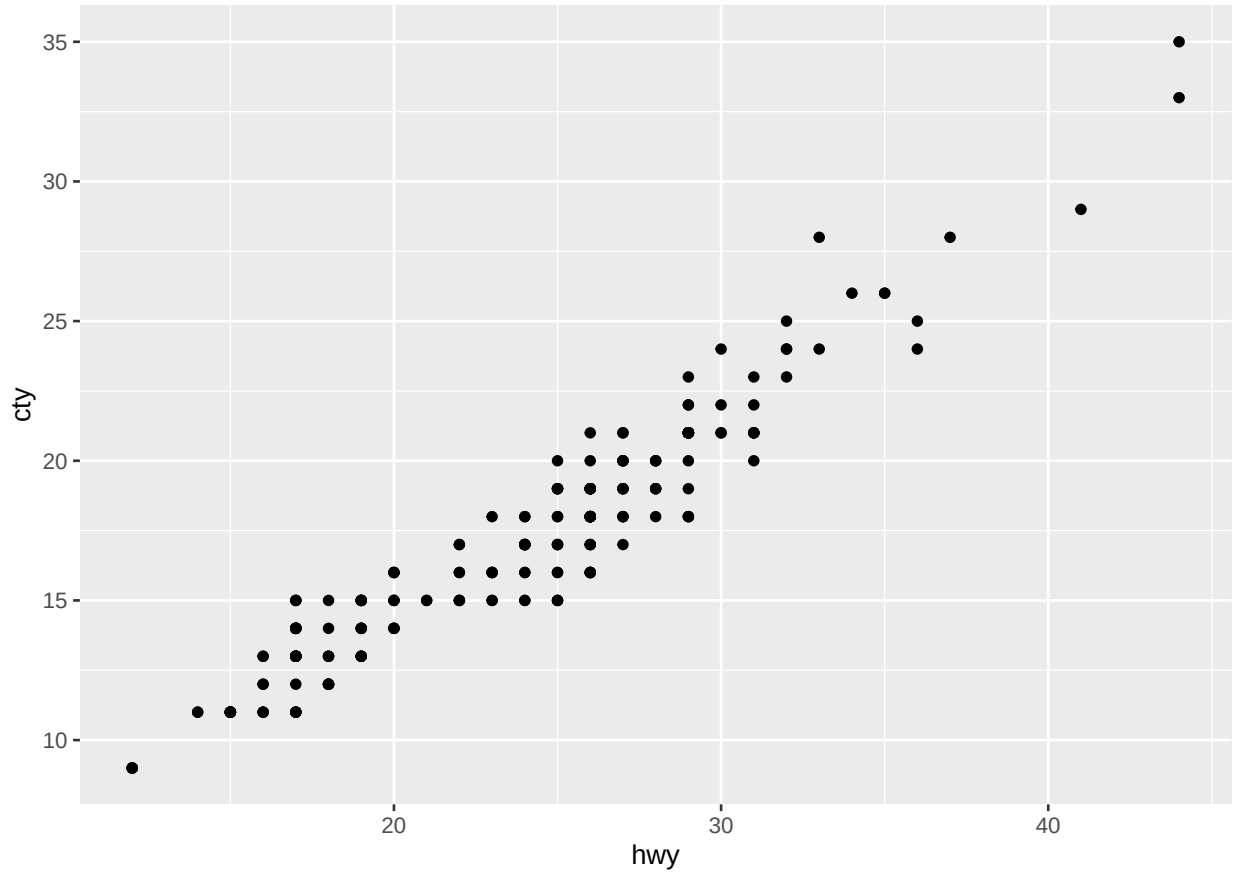
```
ggplot(mpg,aes(hwy))+geom_histogram(bins = 100)
```




According to the histogram, most cars have the highway mpg between 15-20 and 22-30. The two most frequent are at 17 and 26. The data is following bimodal pattern.

Exercise 2

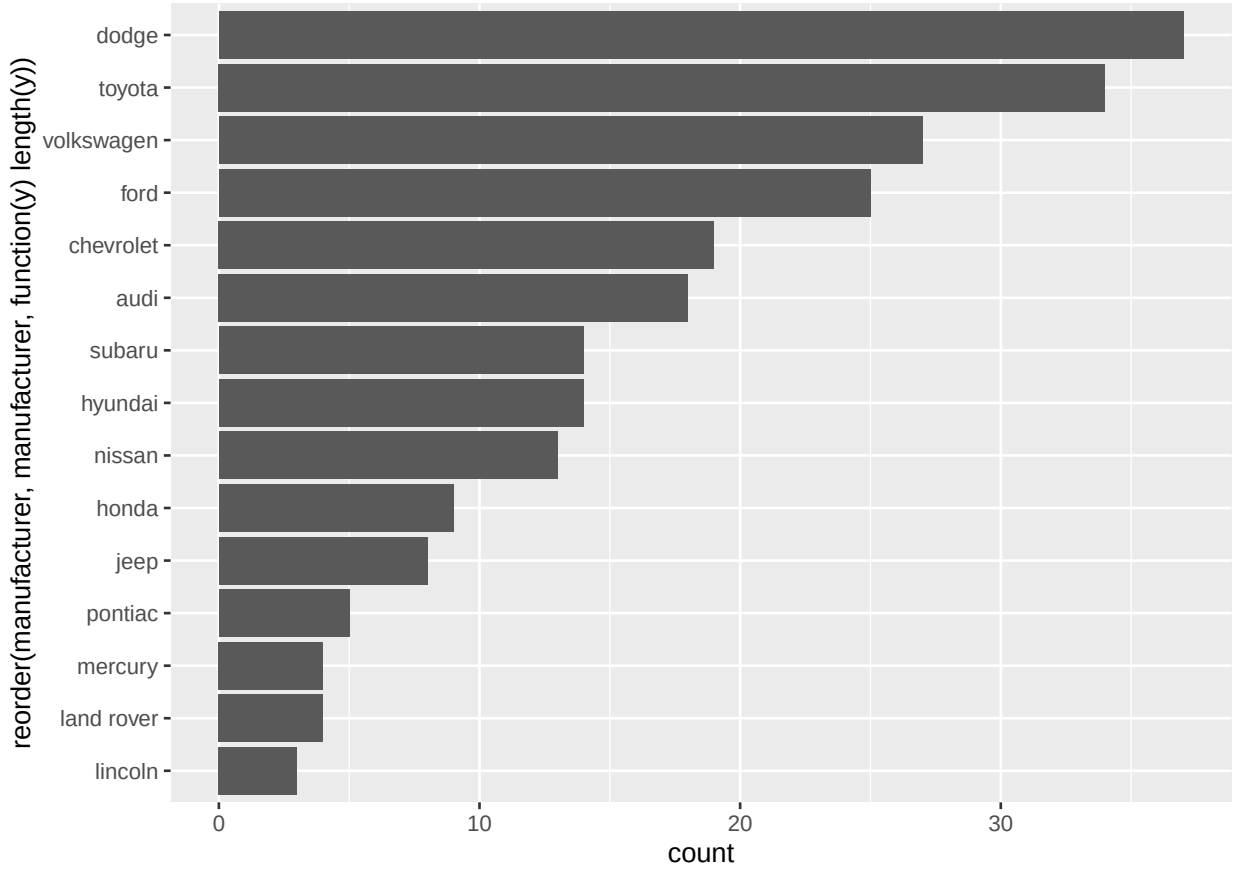
```
ggplot(mpg,aes(x=hwy,y=cty))+geom_point()
```




The scattered points locate along an increasing line, which indicates that there is a positive relationship between hwy and cty. As hwy increases, cty increases.

Exercise 3

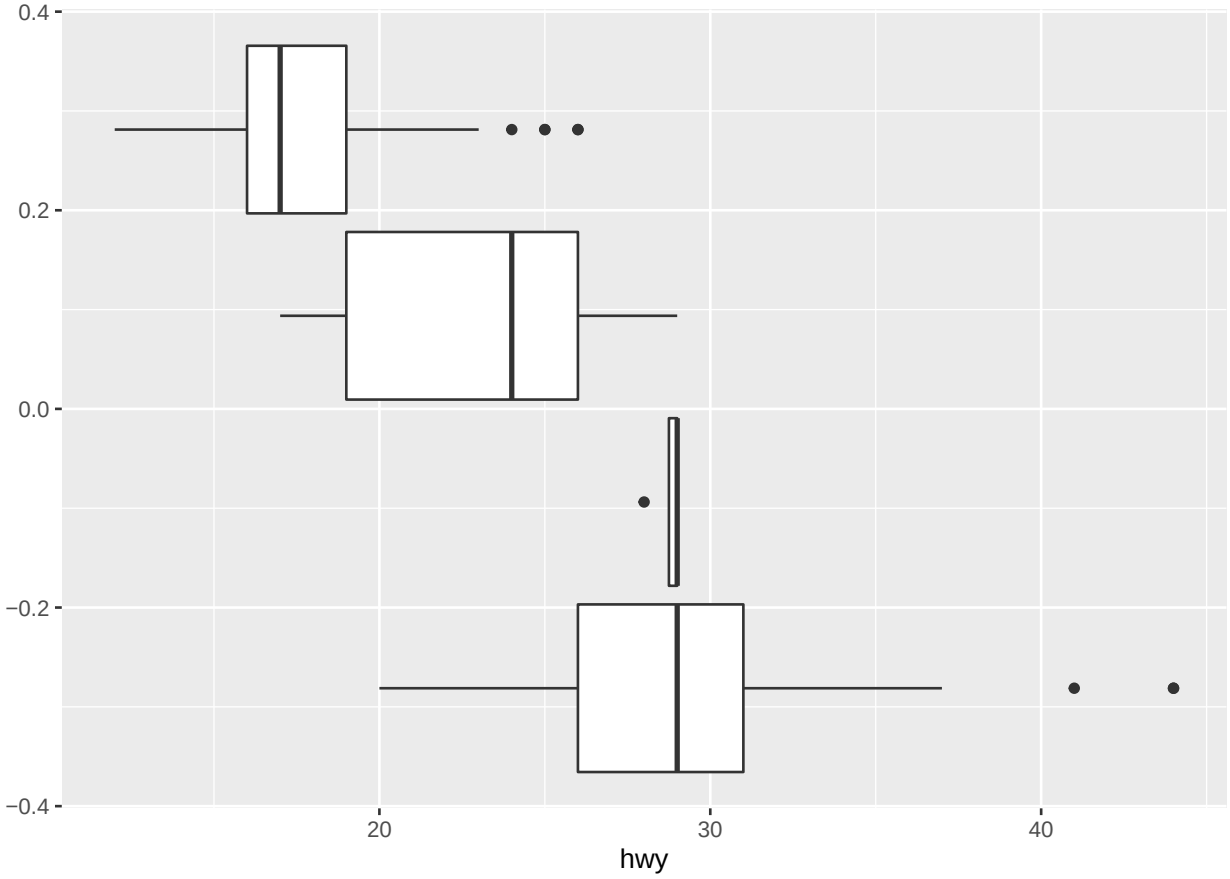
```
ggplot(mpg,aes(y=reorder(manufacturer, manufacturer, function(y) length(y))))+geom_bar(stat = "count")
```




From the bar plot above, we can see Dodge produced most cars, and Lincoln produced least cars.

Exercise 4

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



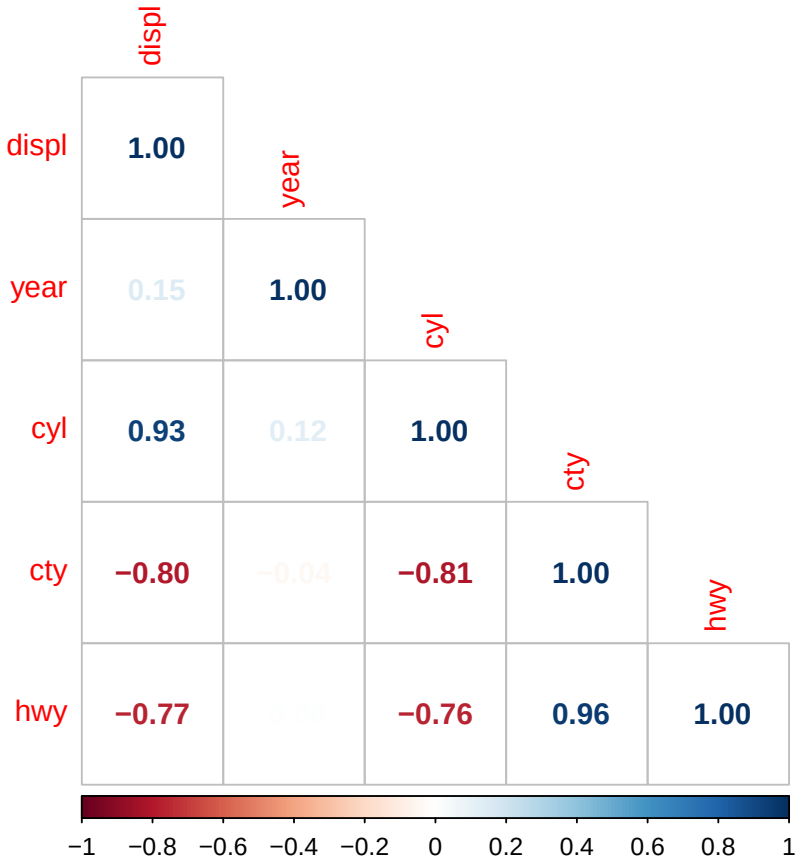

I notice that the less in cylinders,the higher values in highway mpg. It is a negative correlation between them.

Exercise 5

```
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
M <- mpg %>% select_if(is.numeric) %>% cor(.)
corrplot(M, method = 'number', type = 'lower')
```



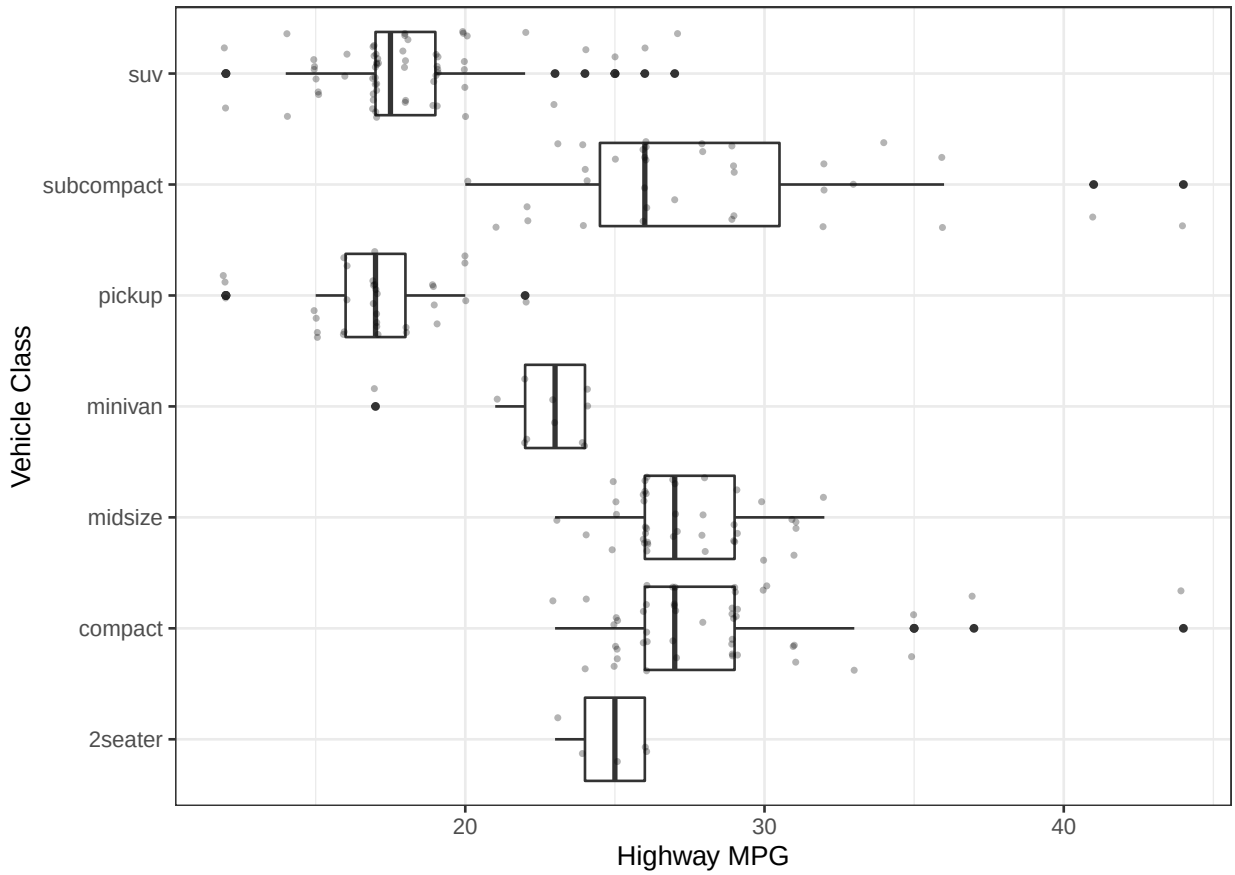

Positive correlated: Displ with cyl, Cty mpg with hwy mpg.

Negative correlated: Cty mpg with Displ, hwy mpg with Displ, cyl with cty mpg, cyl with hwy mpg.

I think they all very make sense. More cylinders lead to more engine displacement. Cty and hwy mpg are positively correlated, because the they are determined by the performance of the car. Also more cylinders lead to low mpg(as in exercise 4). And lower mpg correlate with high engine displacement.

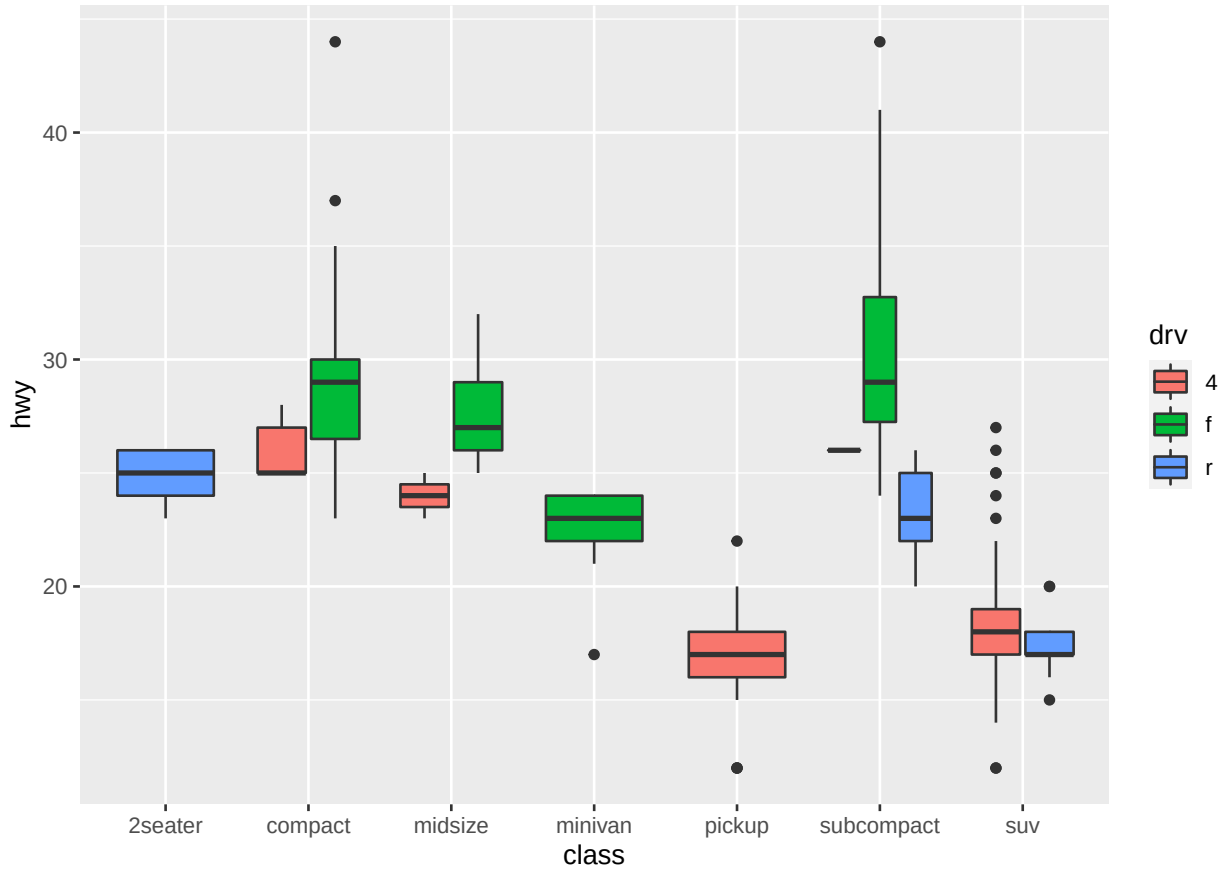
Exercise 6

```
#install.packages("ggthemes")
library(ggthemes)
ggplot(mpg, aes(hwy,class))+geom_boxplot(outlier.size = 1)+
  geom_jitter(shape = 16, position = position_jitter(width = 0.1), alpha = 0.3,size=1)+
  ylab("Vehicle Class")+xlab("Highway MPG")+theme_bw()
```



Exercise 7

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



Exercise 8

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

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