PSTAT231 hw1

Olivia Dong

2022/3/30

Main Ideas

1. According to the lecture, supervised learning is that the machine learns the rule from observations x_i with its corresponding response y_i data(observed outcome). The actual response data or the outcome Y is the supervisor. Different from supervised learning, there is no supervisor in unsupervised learning, which means no available true response data.

2. The difference between regression model and classification model lies in the categories of their outcome. According to the lecture, the former has quantitative response Y; the latter has qualitative response Y.

3. Regression: Root Mean Squared Error(RMSE)/MSE and Mean Absolute Error (MAE)
   classification: rate error

4. According to the lecture,
   Descriptive model: model that could represent the trend of data.
   inferential model: used to test significant features of data. To explore the relationship between outcomes and predictors
   Predictive model: find the combo of features that fit the data best. It is used to predict Y with the minimum reducible error.

5.

a. Mechanistic model have an assumption of parametric form of function f($\beta_1 + \beta_2 + \ldots$), which will not match the true f. Empirically-driven model does not have assumption about f. Instead, according to the book, it seeks an estimate of f to fit the data points as close as possible.
   difference: mechanistic has an assumption of functional form of f, whereas empirically-driven does not
   similarity: both of them could be overfitting.

b)The mechanistic model is easier to understand since we have an given underlying form f. According to page 22 of ISLR , finding a series parameters is easier than finding an entirely arbitrary function f.

c. The mechanistic model can become more flexible by adding more parameters in f. However, when too many parameters are added, the model might be overfitting. The variance is large and the bias is small. However, when the model is simple and with only few parameters, the variance is small but the bias will be large. The empirically-driven model is by default flexible, which means that it tend to has large variance and small bias.

6. First one is predictive. We are trying to predict the outcome of votes by the given voter's data. Second one is inferential because it explores the relationship between the voter's likelihood of support and whether they have personal contact with candidates.

EDA

1.

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

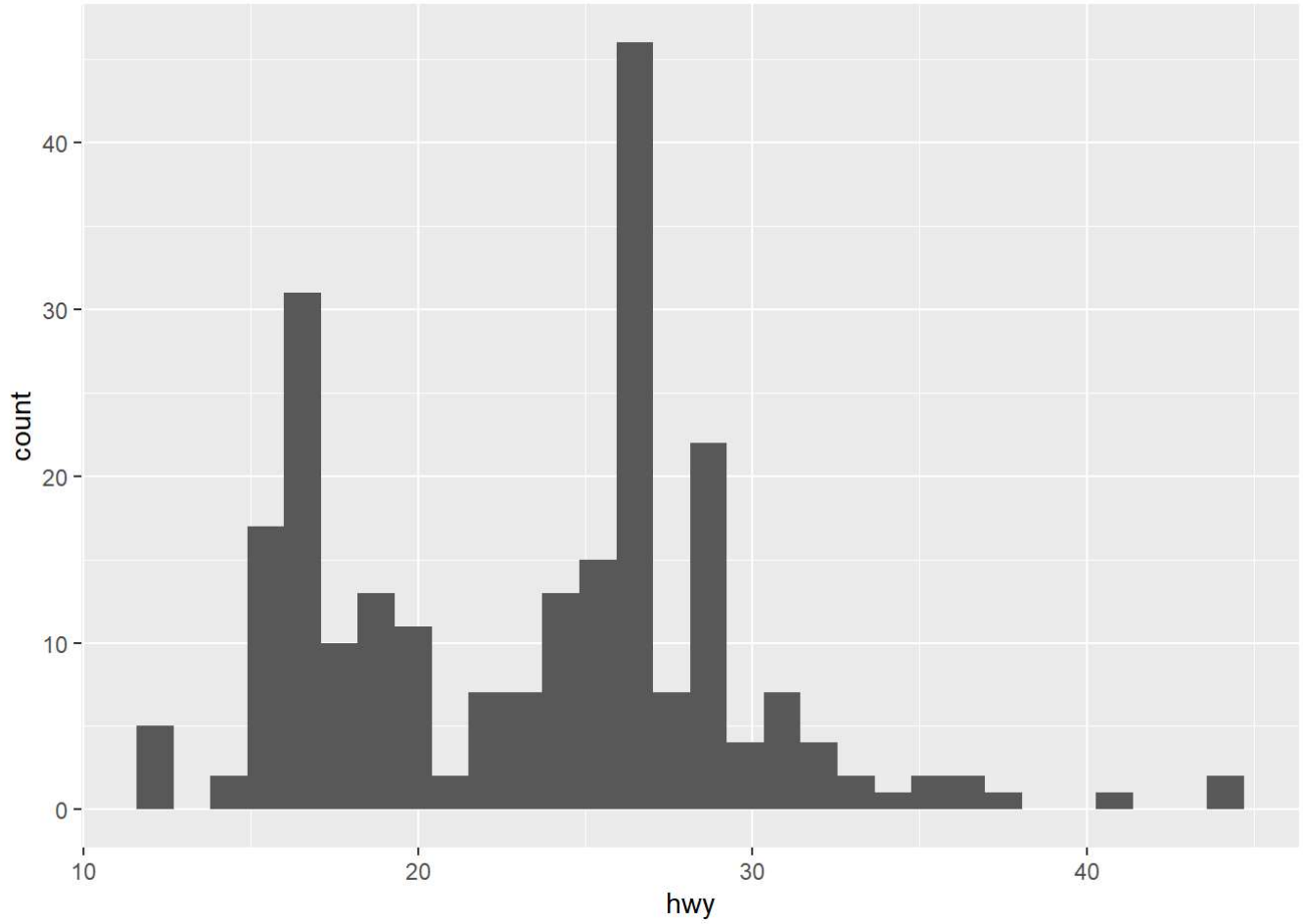
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
## v purrr   0.3.4
```

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

```
ggplot(data=mpg,mapping=aes(hwy))+geom_histogram()
```

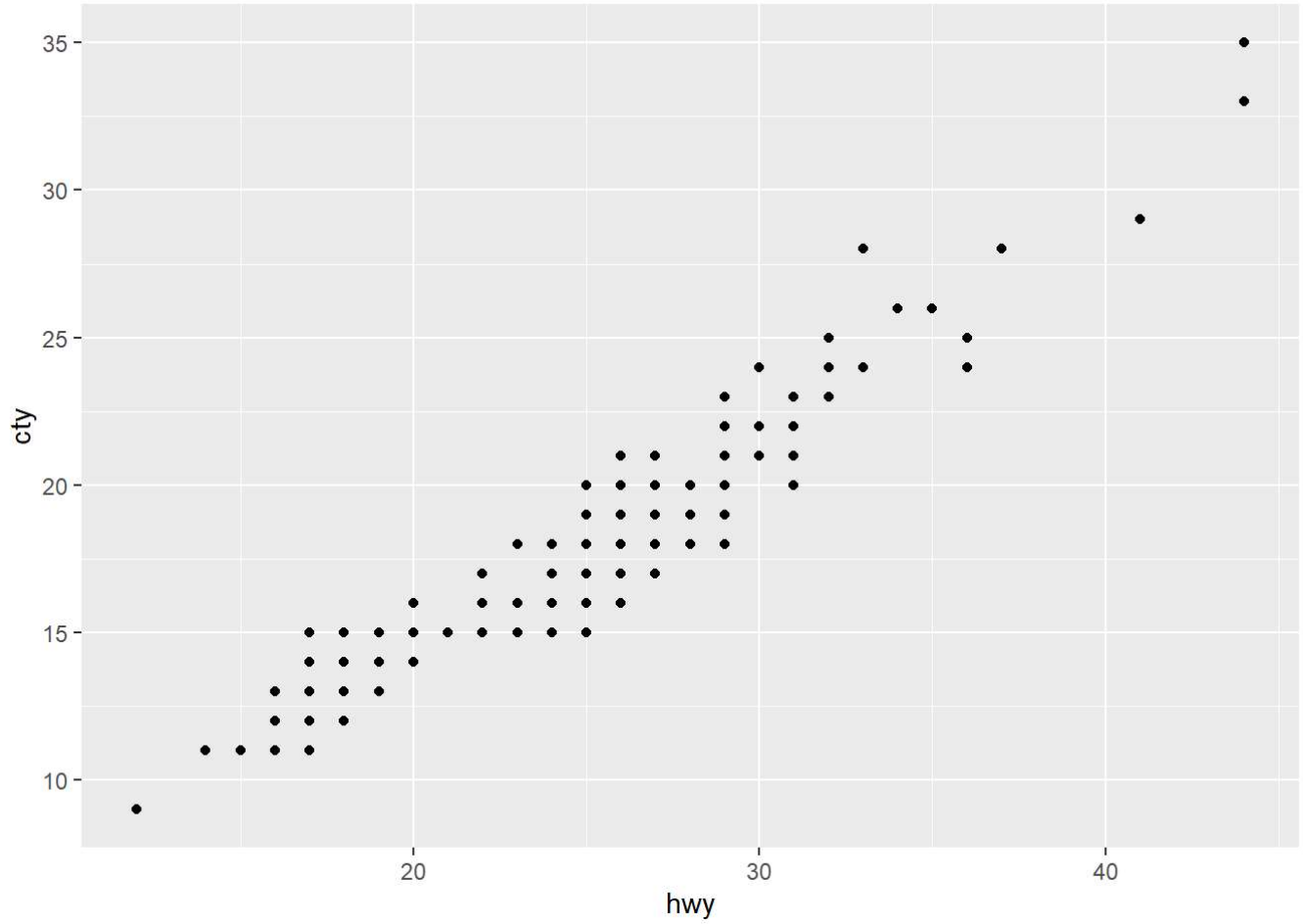
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The histogram has two peaks: one centered at hwy=17 and another at hwy=26. Highway miles per gallon has the greatest counts when it equals to 26.

2.

```
ggplot(data=mpg,mapping = aes(hwy,cty))+geom_point()
```



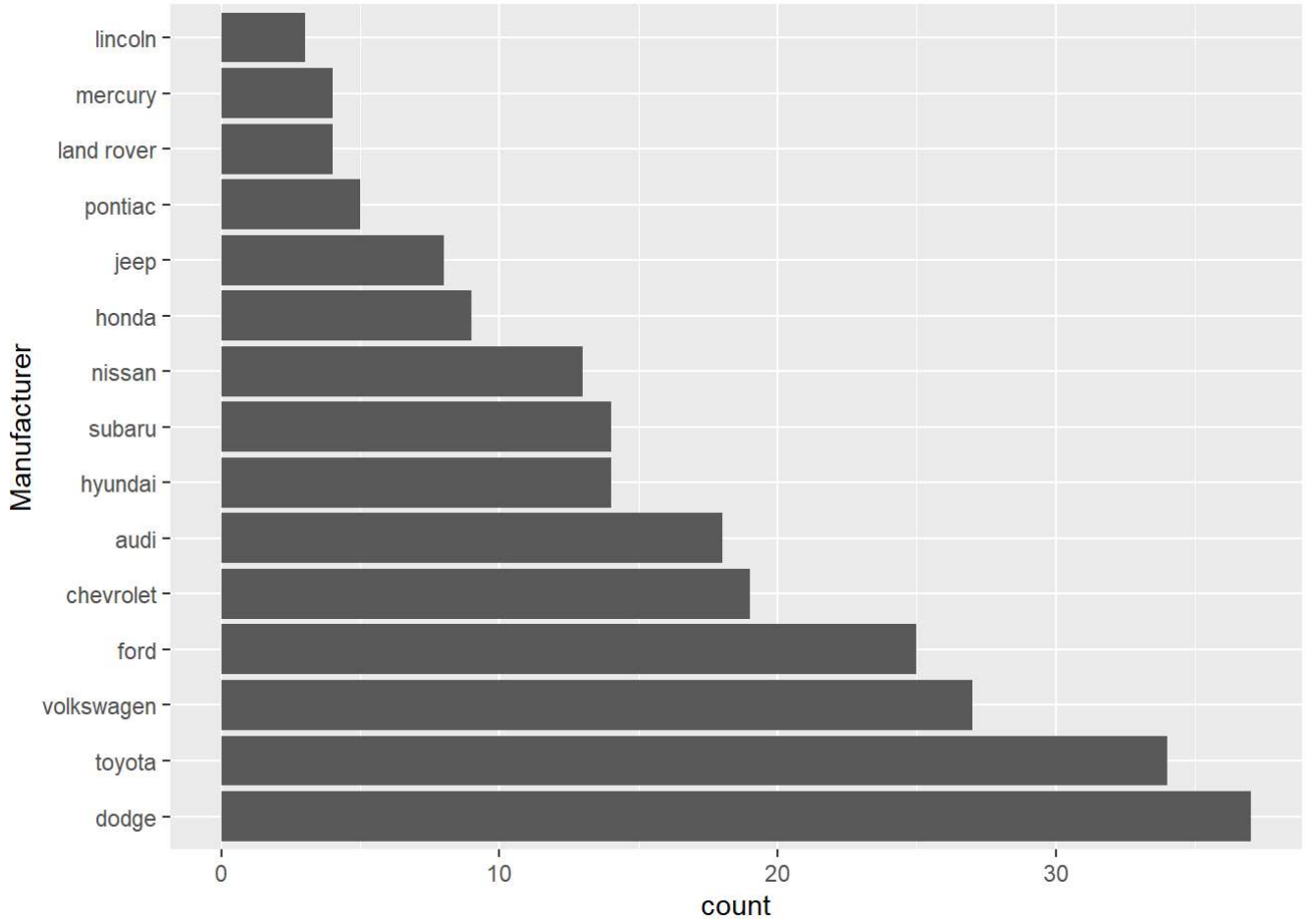
We could notice a positive trend in the scatter plot, which indicates a positive linear relationship between hwy and cty. This means that if highway miles per gallon increase, cty will also increase.

3.

```
library(forecast)
```

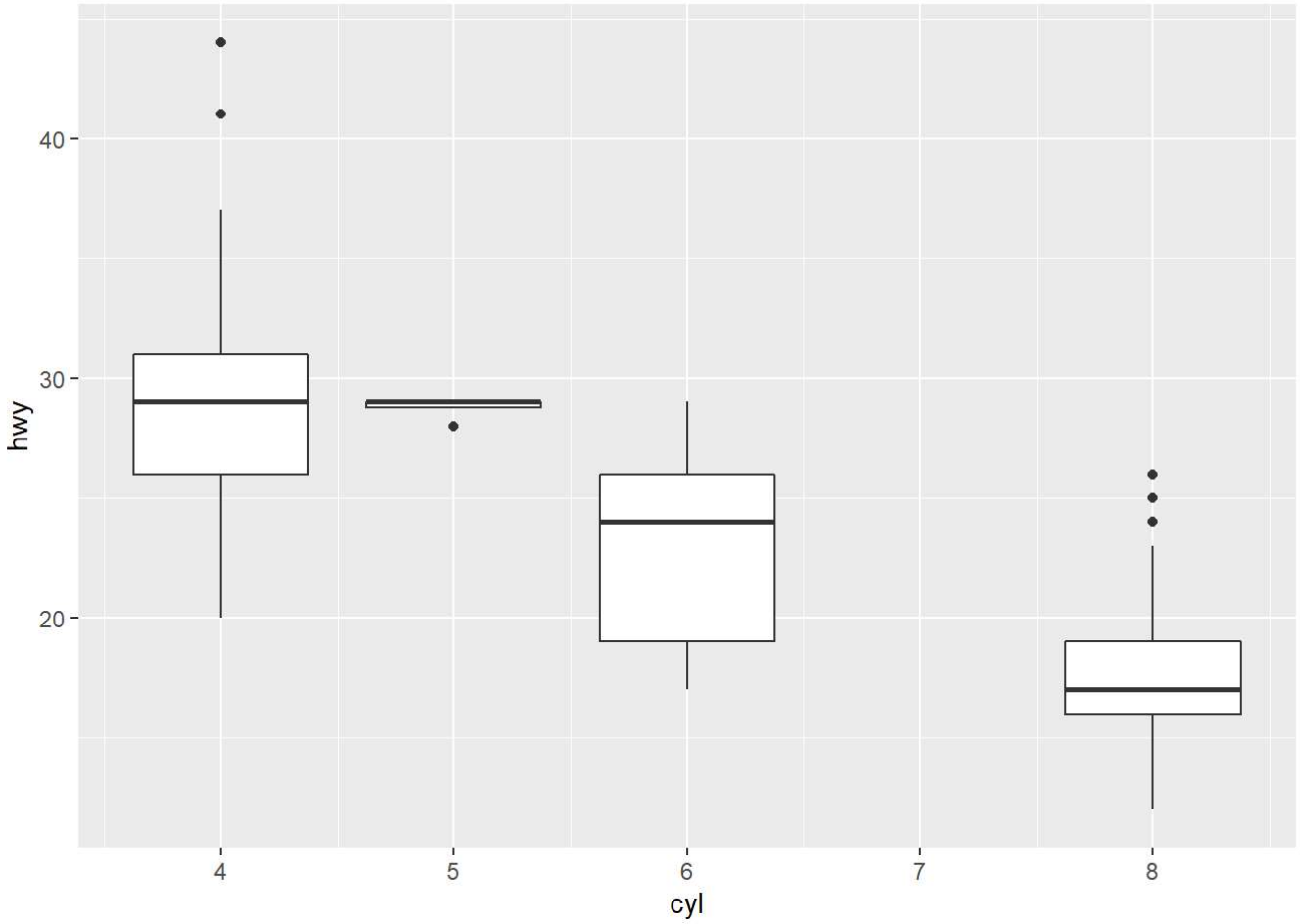
```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
ggplot(mpg,mapping=aes(fct_infreq(manufacturer)))+geom_bar()+coord_flip()+ xlab("Manufacturer")
```



Lincoln produce the least. Dodge produce the most.

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



Higher number of cylinders tends to has a lower value of highway miles per gallon.

5.

```
library(corrplot)
```

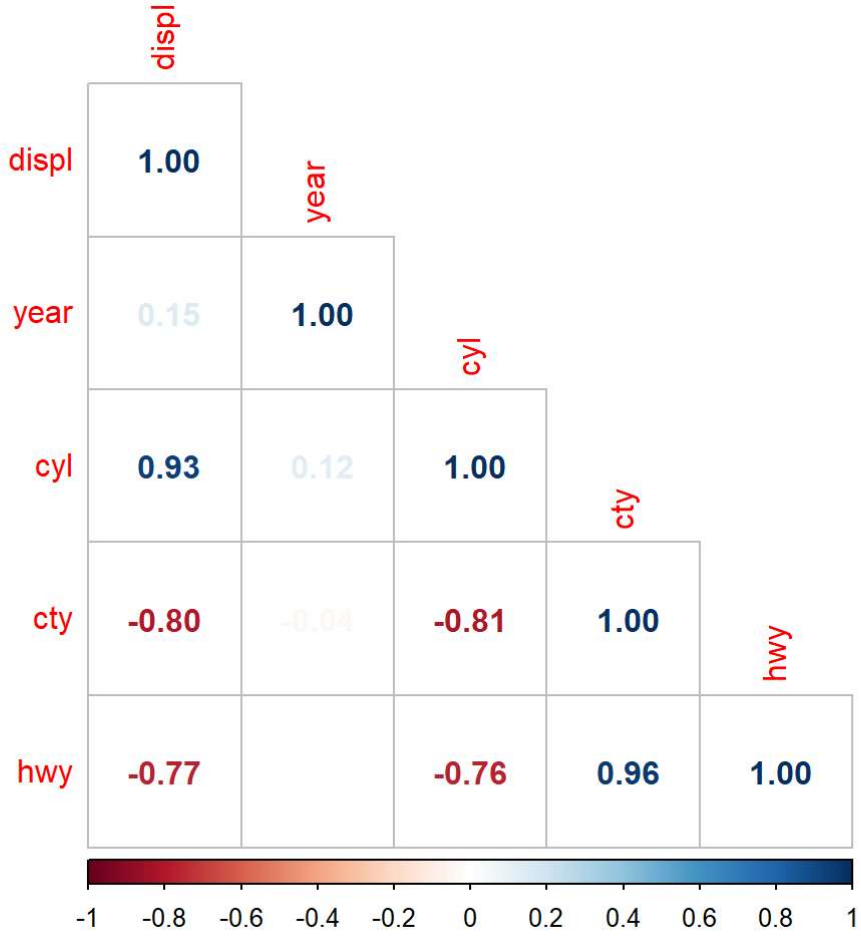
```
## Warning: 程辑包'corrplot'是用R版本4.1.3 来建造的
```

```
## corrplot 0.92 loaded
```

```
mpg2 = select_if(mpg,is.numeric)

M = cor(mpg2)

corrplot(M,method = 'number',type='lower')
```



Number of cylinders has a relatively great positive correlation with the engine displacement. Since engine displacement is determined by calculating the engine cylinder bore area multiplied by the stroke of the crankshaft, and then multiplied by the number of cylinders(from wikipedia). It makes sense that they have positive strong correlation.

cty is positively correlated with hwy, which makes sense since a car could run well in both city and highway or in neither.

cyl is negatively correlated with both cty and hwy, which makes sense. The less the number of cylinders, the car would be more cost-effective.

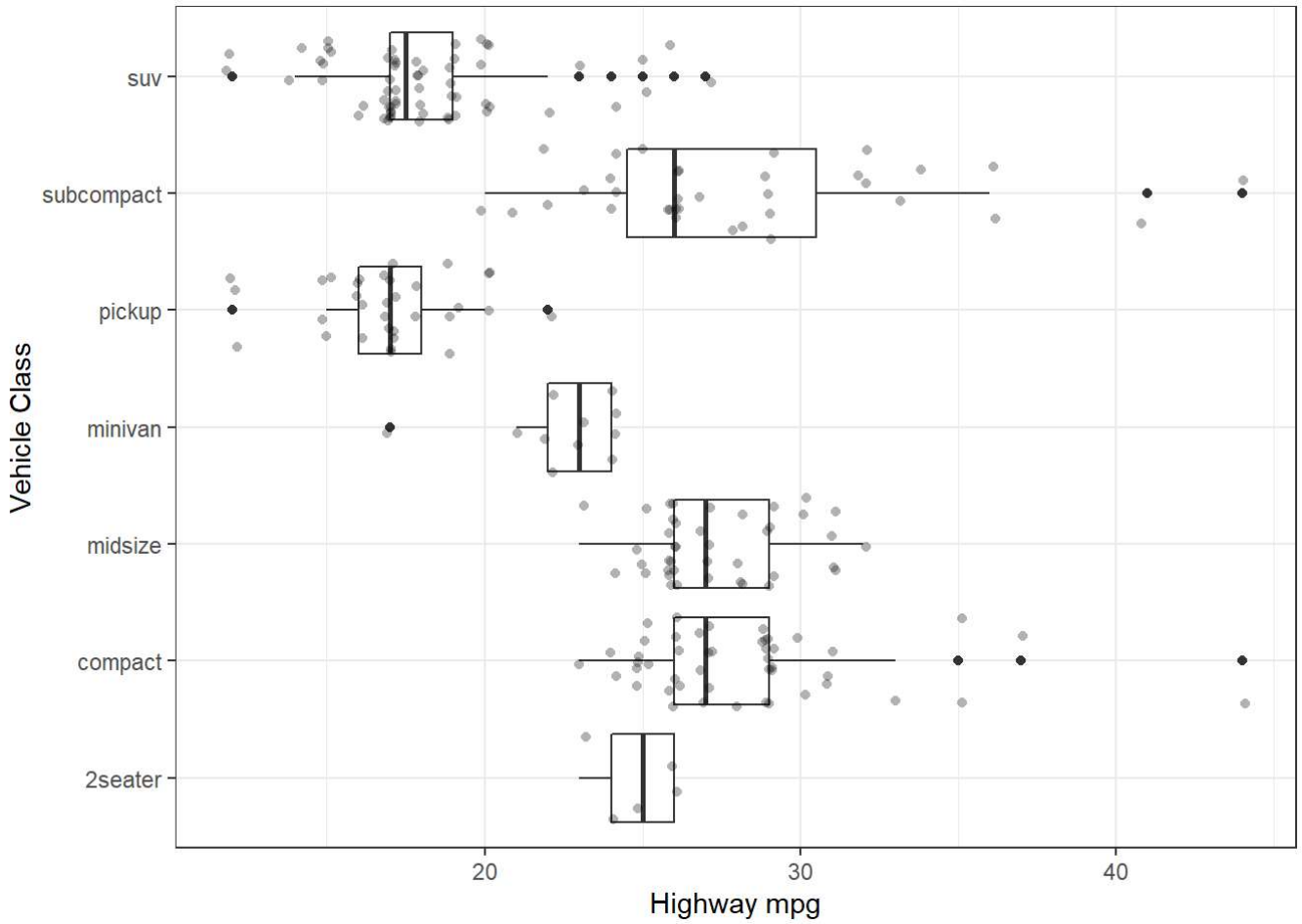
There is a negative correlation between engine displacement and highway miles per gallon, and between engine displacement and city miles per gallon. This makes sense that a greater engine displacement means that the car needs more energy, which is inefficient.

It is surprising that the correlation between year of manufacture with other variables are not significant.

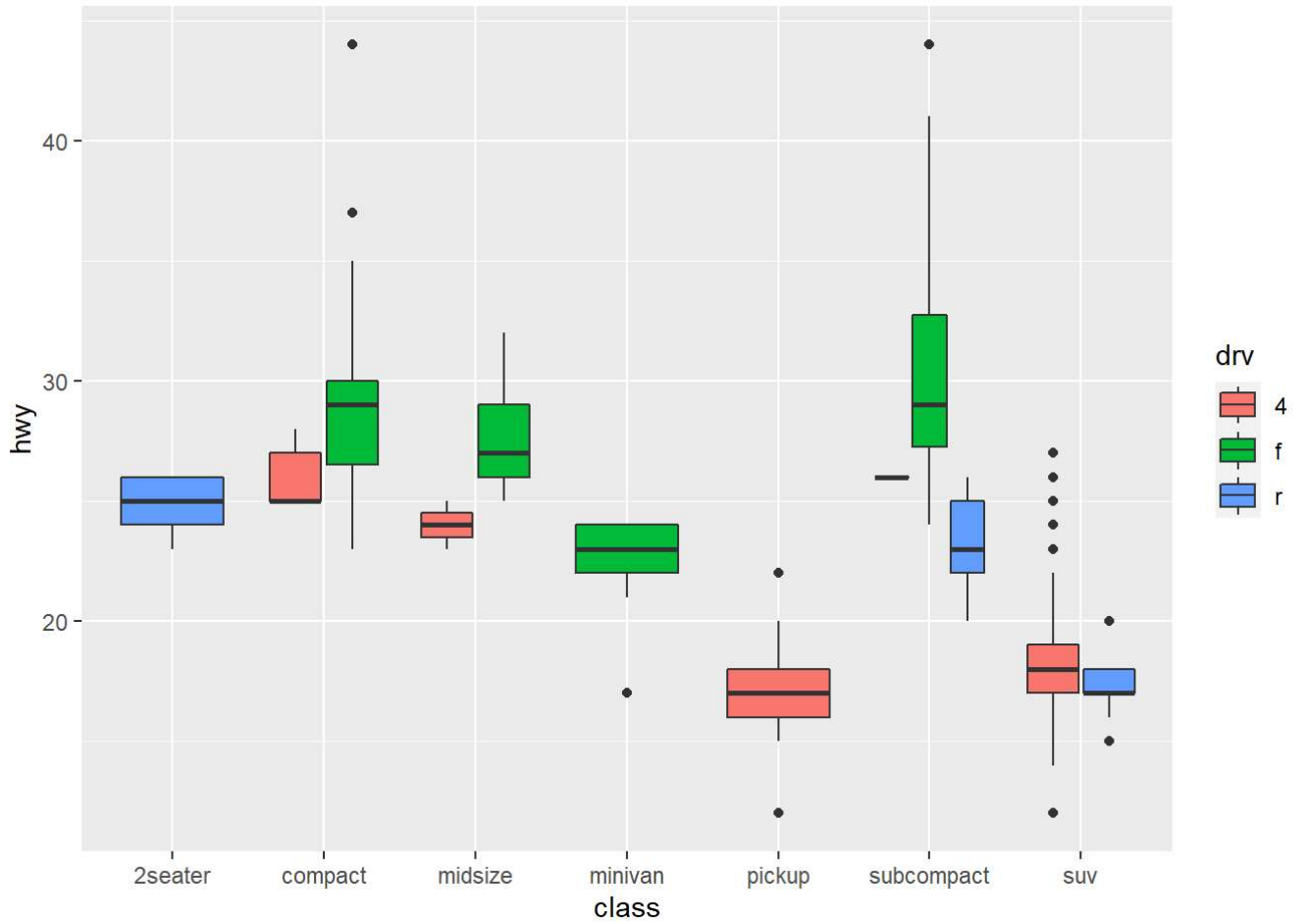
```
library(ggthemes)
```

```
## Warning: 程辑包'ggthemes'是用R版本4.1.3 来建造的
```

```
ggplot(mpg, aes(x=hwy, y=class)) + geom_boxplot() +
  geom_jitter(shape=16, position=position_jitter(0.2),alpha=0.3) + ylab("Vehicle Class") + xlab
('Highway mpg')+ theme_bw()
```




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



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

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