Question 1:

In supervised learning, for each observation of the predictor measurement X, there is an associated response measurement Y (AKA supervisor). In unsupervised learning, for each observation, we observe a vector of measurements X but no associated response Y. The difference between these two is that unsupervised learning have to learn without a supervisor and it is not posible to fit certain models because there is no response variable to predict (From page #26 of book)

Question 2:

In regression model, Y is a quantitative response, and takes on numerical values like price and blood pressure while in classification model, Y is a qualitative response, and takes on categorical values like survived/died, etc.

However, the intriguing part is that Least squares linear regression used with a quantitative response, whereas logistic regression is typically used with a qualitative (two-class, or binary) response. (From page #28 of book)

Question 3: None

Question 4:

- Descriptive models: choose model to visually emphasize a trend in data (using a line on a scatterplot).
- Inferential models: aim to test theories, find the possible correlation and causality, and state relationship between outcome and predictor.
- Predictive models: aim to predict Y with minimum reducible error and not focused on hypothesis test.(from the lecture)

Question 5:

- mechanistic: mechanistic modeling assume a parametric form for f. The more parameters, the more flexibility
- empirically-driven: assume no parametric form for f. Instead, it requires a larger number of observation for the prediction. The original default is more fexible.

The similarity between these two model is that they both have the problem of overfitting. (from the lecture)

Question 6:

I reckon the first one is a inferential question because data can be used to form a implicit analysis based on the voter's past voting proclivity. This is to test a theory that how likely these voters will vote for the candidate by looking at their old data. It is something of evaluating the relationship between the predictor (past data) and response variables (how likely now).

The second one is predictive. Since there is no hypothetical assumption embedded in this one, I think this question is trying to predict something that is yet to happen, which is voter's likelihood of support for the candidate whether or not is going to change.

```
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
```

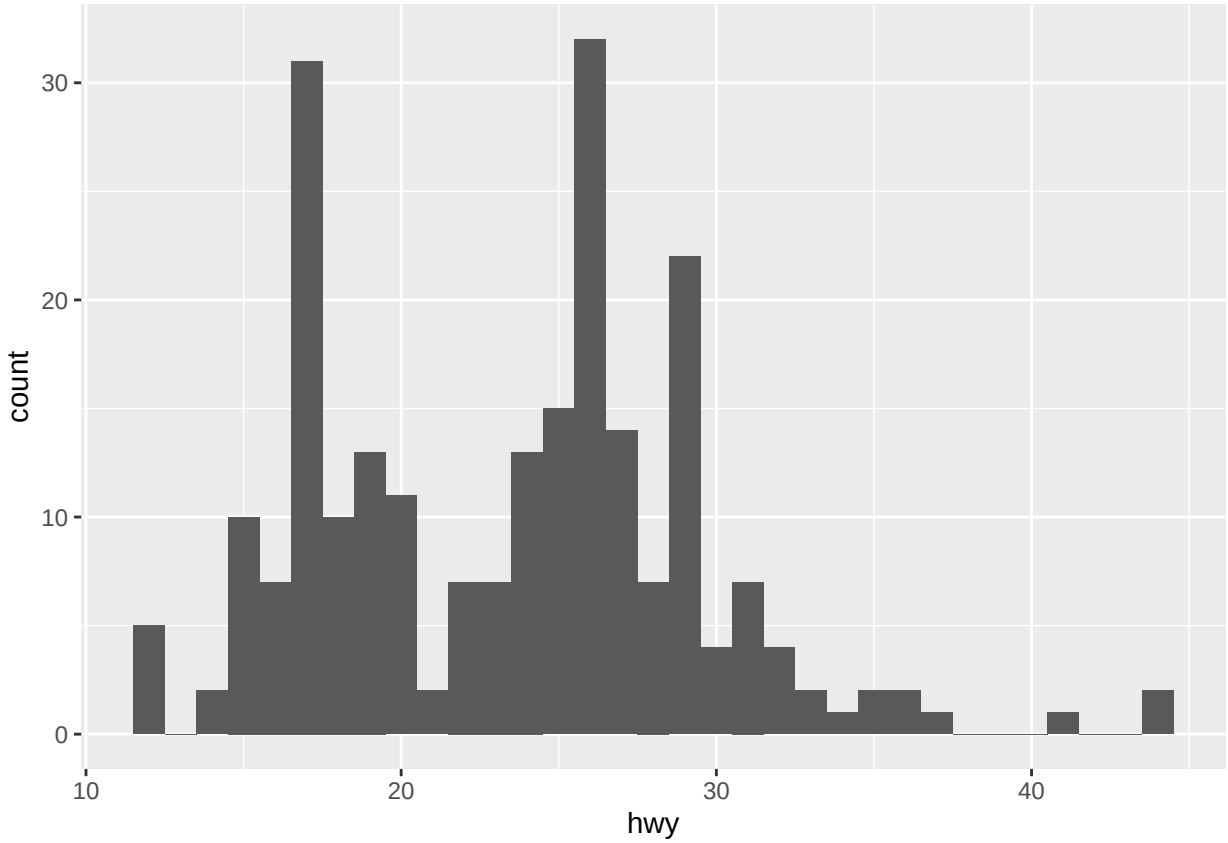
```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
mpg
```

```
## # A tibble: 234 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~ f        18    29 p     comp~
##  2 audi         a4           1.8  1999     4 manu~ f        21    29 p     comp~
##  3 audi         a4           2    2008     4 manu~ f        20    31 p     comp~
##  4 audi         a4           2    2008     4 auto~ f        21    30 p     comp~
##  5 audi         a4           2.8  1999     6 auto~ f        16    26 p     comp~
##  6 audi         a4           2.8  1999     6 manu~ f        18    26 p     comp~
##  7 audi         a4           3.1  2008     6 auto~ f        18    27 p     comp~
##  8 audi         a4 quattro   1.8  1999     4 manu~ 4        18    26 p     comp~
##  9 audi         a4 quattro   1.8  1999     4 auto~ 4        16    25 p     comp~
## 10 audi         a4 quattro   2    2008     4 manu~ 4        20    28 p     comp~
## # ... with 224 more rows
```

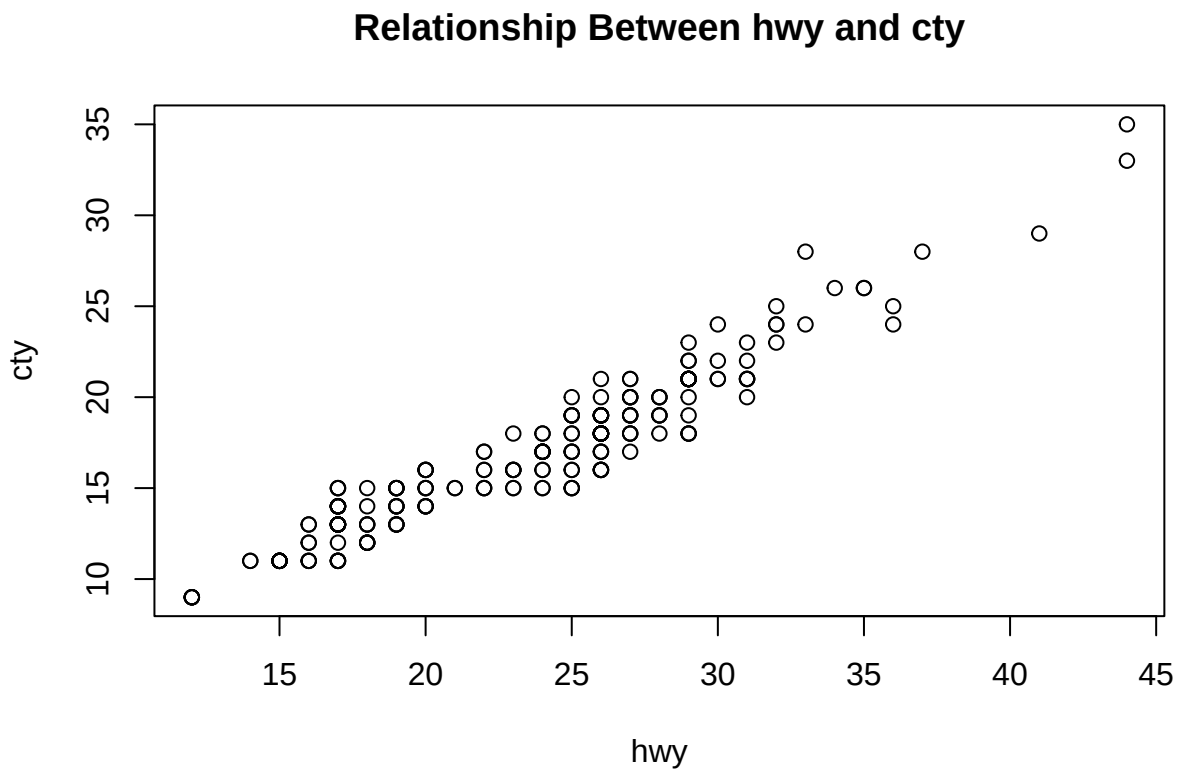
```
?mpg
```

```
# EX1
# create a histogram
qplot(mpg$hwy, geom="histogram",bins=40,binwidth=1, xlab = 'hwy', ylab = 'count')
```



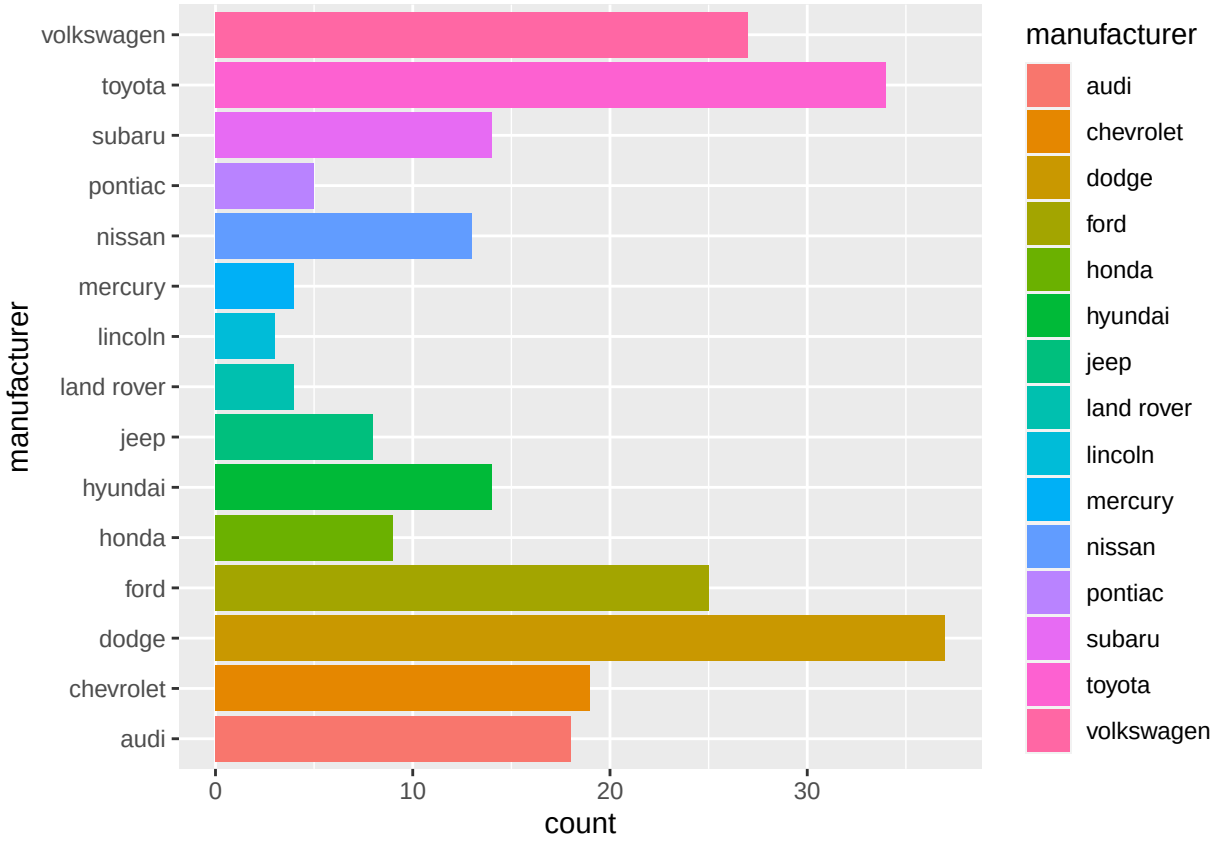
If we group by hwy, I can see that the car with 26 miles per gallon in highways has the most car, which is over 30. Next, the second most cars is the cars with 18 miles per gallon in highways. There are few outliners such as 44 hwy and 12 hwy but the number is small. Most of cars are distributed in 14 hwy to 36 hwy.

```
# EX2:
# Create a scatterplot
plot(mpg$hwy, mpg$cty, main="Relationship Between hwy and cty", xlab = 'hwy', ylab = 'cty')
```



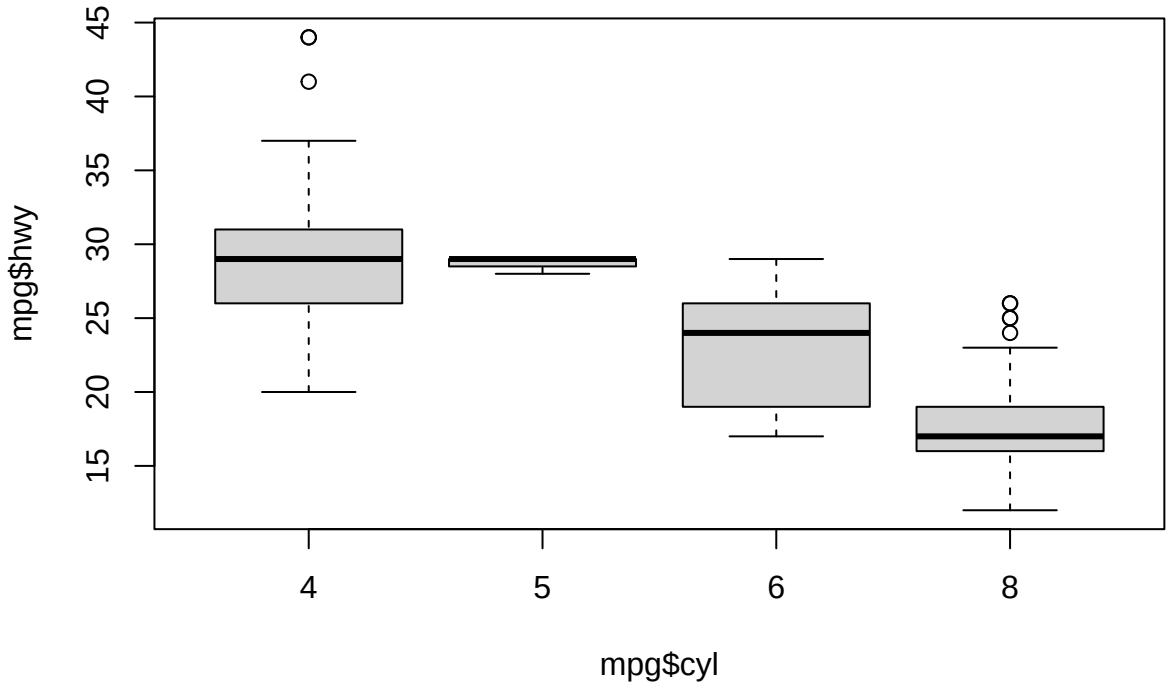
I can see that the hwy and cty have a positive relationship. As the hwy increases, tge cty also goes up.

```
# EX3:
data<-mpg
data %>% ggplot(aes(x=manufacturer))+geom_bar(aes(fill=manufacturer))+coord_flip()
```



By inspection, the Dodge manufacturer produced the most cars and Lincoln produced the least.

```
# EX4:
boxplot(mpg$hwy~ mpg$cyl)
```



It seems that as the cylinder increases, the boxplot of hwy moves downwards.

```
# EX5:
library(corrplot)
```

```
## corrplot 0.92 loaded
```

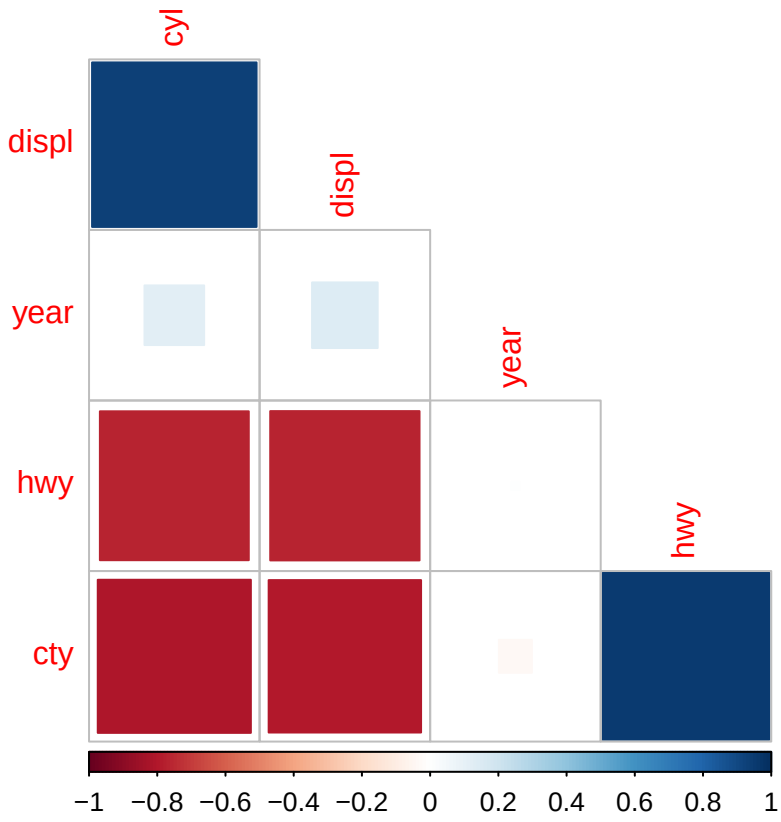
```
?mpg
```

```
library(tidyverse)

#Because cor(mpg) x has to be numerical so I removed the categorical variables myself
data<- mpg %>% select(is.numeric)
```

```
## Warning: Predicate functions must be wrapped in `where()`.
##
##   # Bad
##   data %>% select(is.numeric)
##
##   # Good
##   data %>% select(where(is.numeric))
##
## i Please update your code.
## This message is displayed once per session.
```

```
m=cor(data)
corrplot(m, method = 'square', order ='FPC',type = 'lower', diag = FALSE)
```



Blue represents positively correlated and red represents negatively correlated. I use "+" to say positive relationship, "-" to say negative relationship.

// hwy cty + (stands for variables hwy and cty have positive relationship) // year hwy - (stands for variables year and hwy have negative relationship) // displ cty - // displ hwy - // displ year + // cyl cty - // cyl hwy - // cyl year + // cyl displ+

I think the relationship between hwy and cty match with my common sense, but I am surprised that in fact the more cylinders, the less hwy and the more engine displacement, the less hwy. These two puzzled me for a while.