HW 1

1. Supervised learning means that the responses and predictors are both known. In supervised learning, Y (responses) are observed data, and our model should be able to predict the response given the predictors. Unsupervised learning means that the responses are not known. In unsupervised learning, Y (responses) are not known, and we are trying to find patterns in the data.(usually by clustering)

2. In a regression model, the data are quantitative (continuous and numeric values). In a classification model, the data are qualitative (categorical values).

3. Regression: Means Squared Error(MSE), Mean Absolute Error(MAE) Classification: Accuracy, F1 Score

4. Descriptive models: used to emphasize a trend in the data. Inferential models: used to test theories and explain the relation between predictors and responses. Predictive models: used to predict the responses with minimum reducible error.

5. Mechanistic means the model has a parametric form. Empirically-driven means the model does not have a parametric form. Mechanistic models cannot perfectly match the true f, while empirically-driven models require a larger number of observations. They both have the risk of over-fitting. Empirically-driven models are generally more flexible while mechanistic models can increase flexibility by adding parameters.

   I think in general, mechanistic models are easier to understand because the parameters in the model are easy to interpret.

   The bias-variance trade off is a property of ML models which says that the variance of the parameters can be reduced by increasing the bias of the parameters. For mechanistic models, one can reduce flexibility by reducing the amount of parameters in order to increase the bias and decrease variance. For empirically-driven models, a large number of observations will result in high flexibility which leads to low bias and high variance.

6. (1)Predictive, because we are trying to predict how likely a voter will vote for the candidate. (2)Inferential, because we are trying to figure out the relation between a voter's likelihood of supporting the candidate and having a personal contact with the candidate.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     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
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

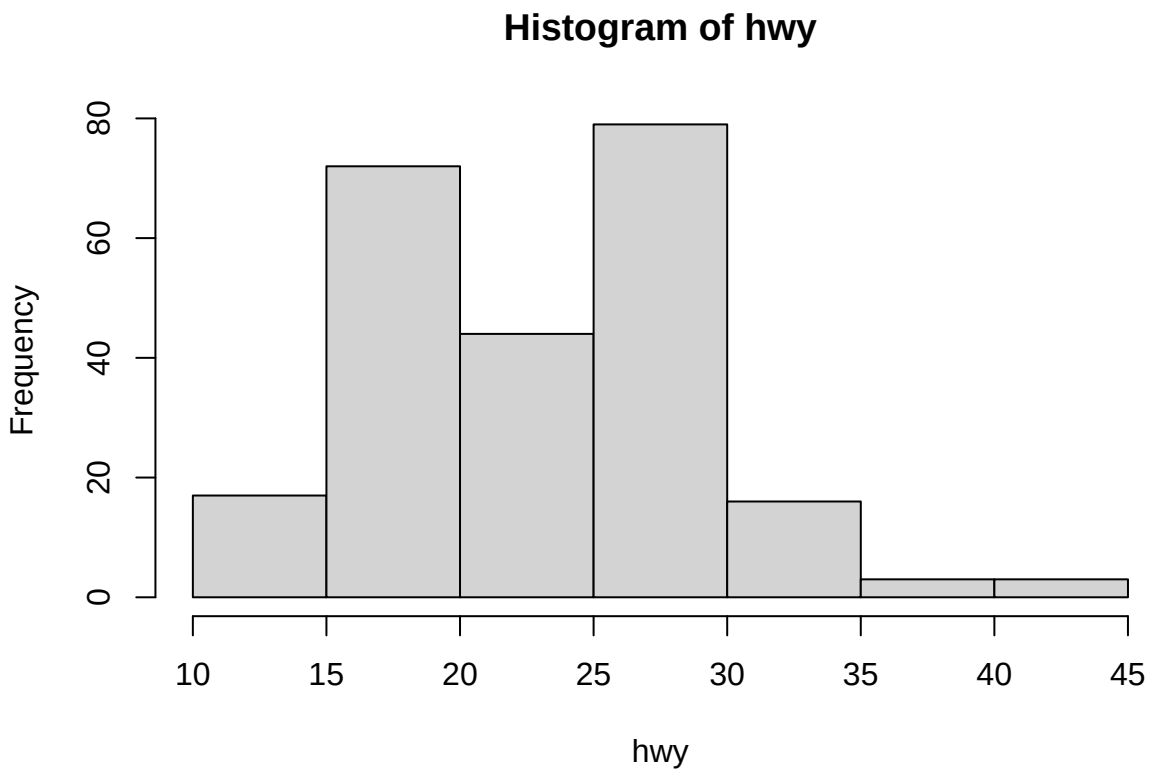
```
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
data(mpg)
```

Q1

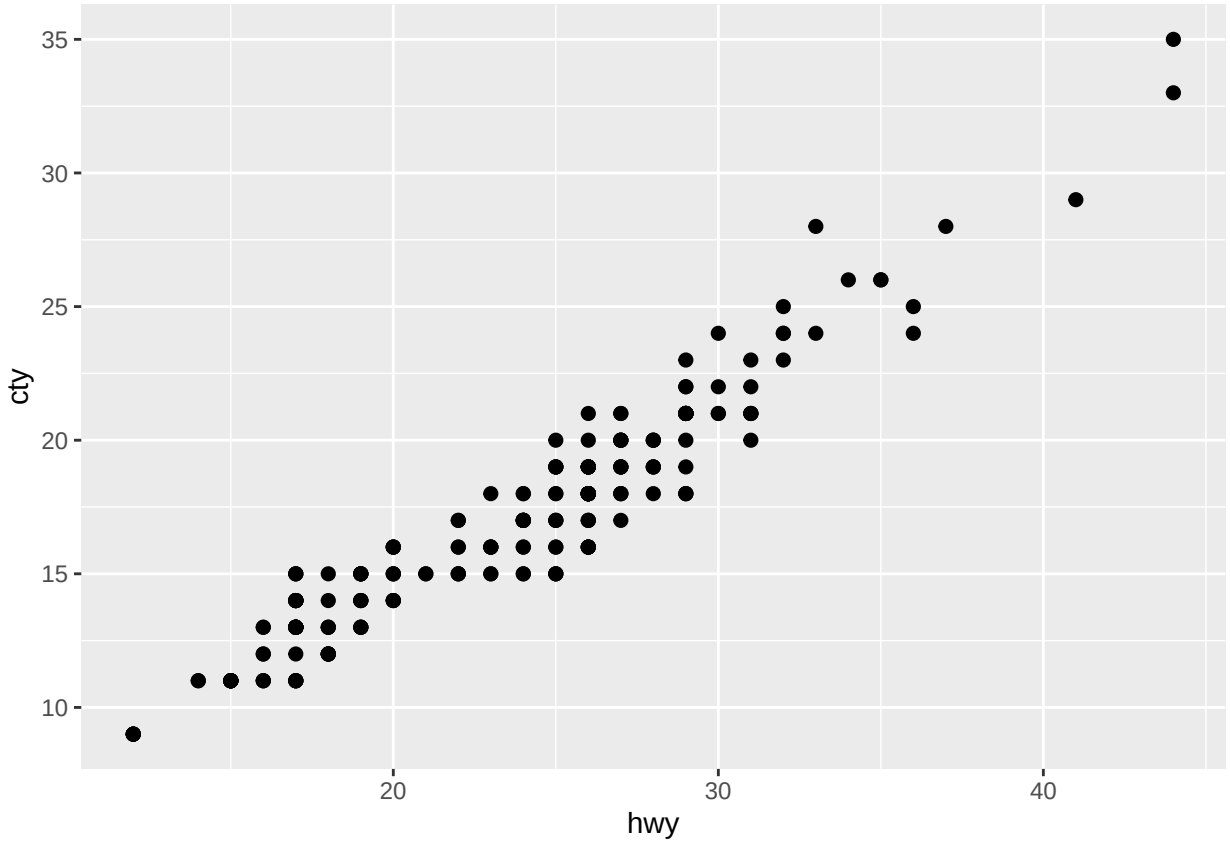
```
hwy <- mpg$hwy
hist(hwy)
```



The histogram is right-skewed, the median is between 20-25 mpg.

Q2

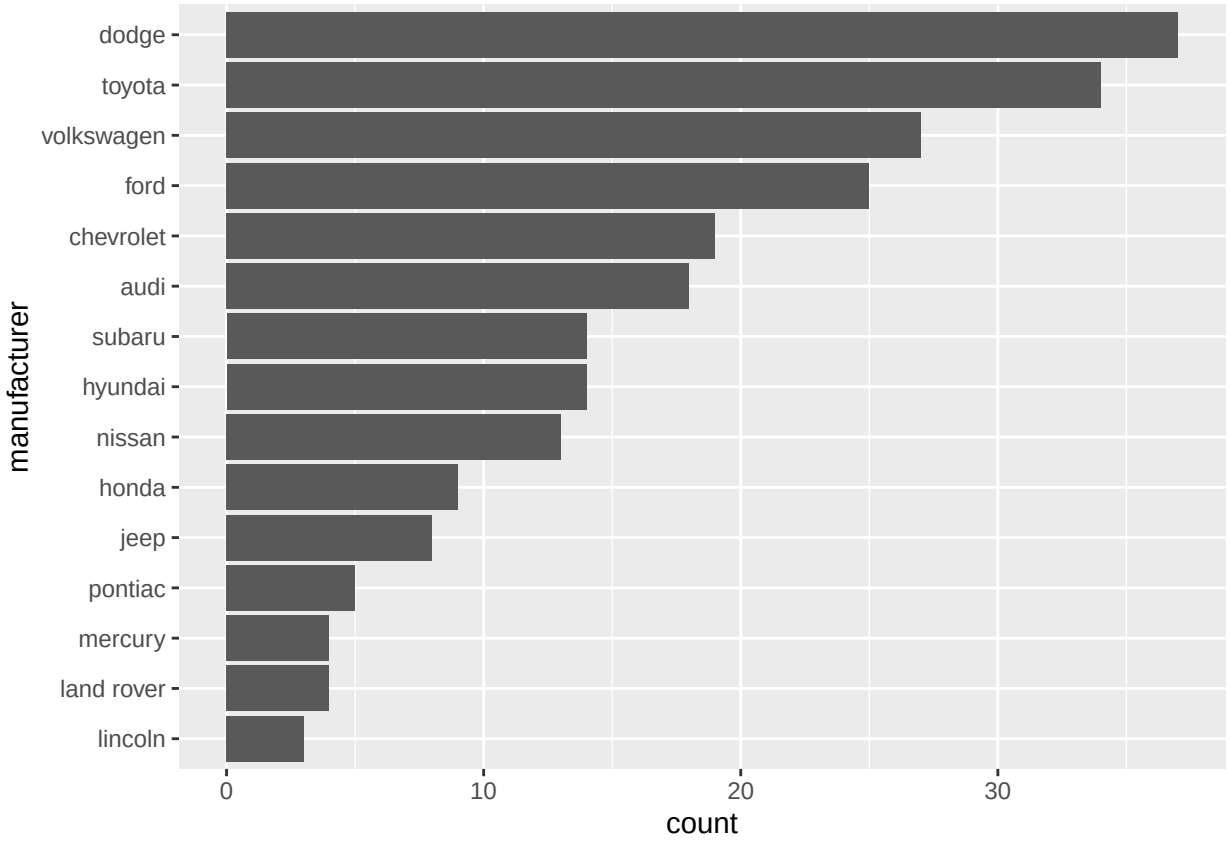
```
cty <- mpg$cty
ggplot(mpg, aes(x = hwy, y = cty)) + geom_point(size = 2)
```



The scatter plot shows a relatively strong positive linear relationship between hwy and cty which means that when hwy (x) increases, cty (y) will also increase by some proportion.

Q3

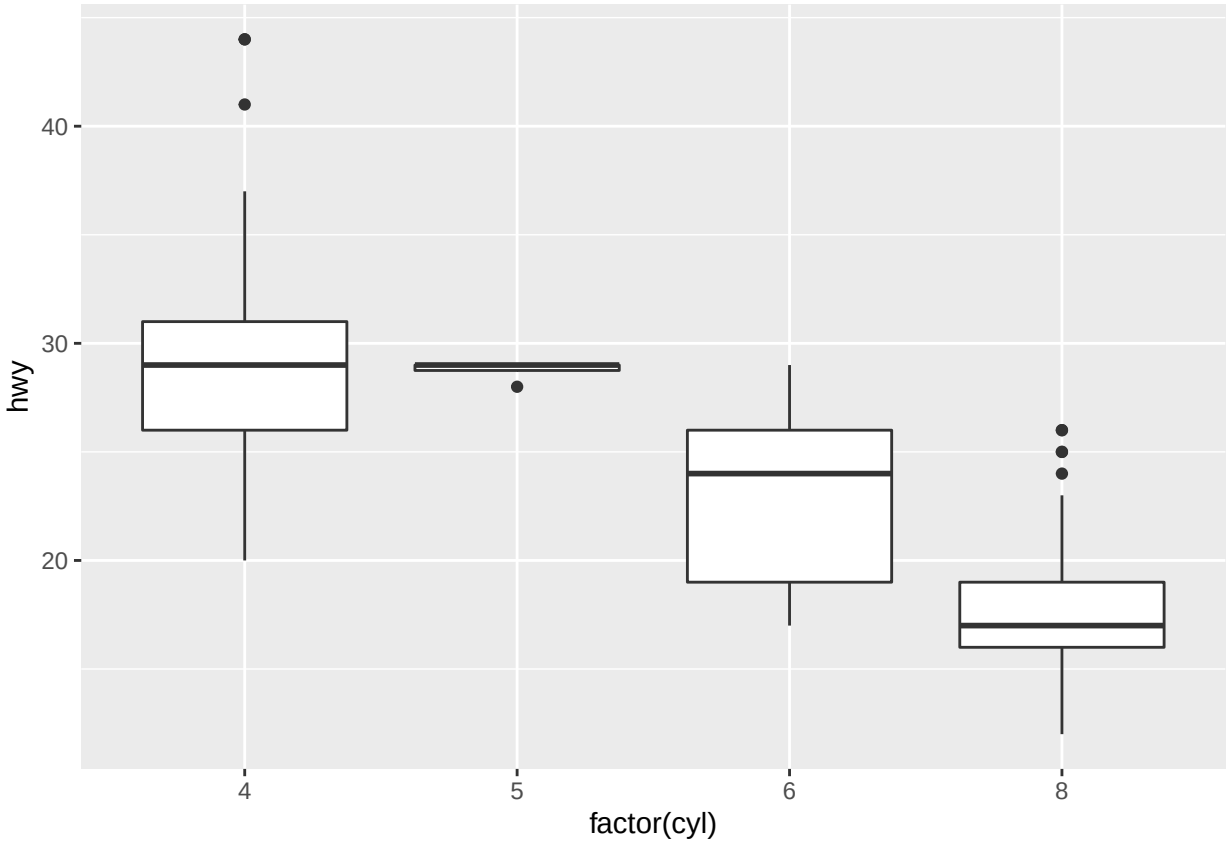
```
manufacturer <- mpg$manufacturer

mpg1 <- within(mpg, manufacturer <- factor(
manufacturer, levels = names(sort(table(manufacturer), decreasing = F))))

ggplot(mpg1, aes(x = manufacturer)) + geom_bar(stat = 'count') + coord_flip()
```



Dodge produced the most cars while Lincoln produced the least.

Q4

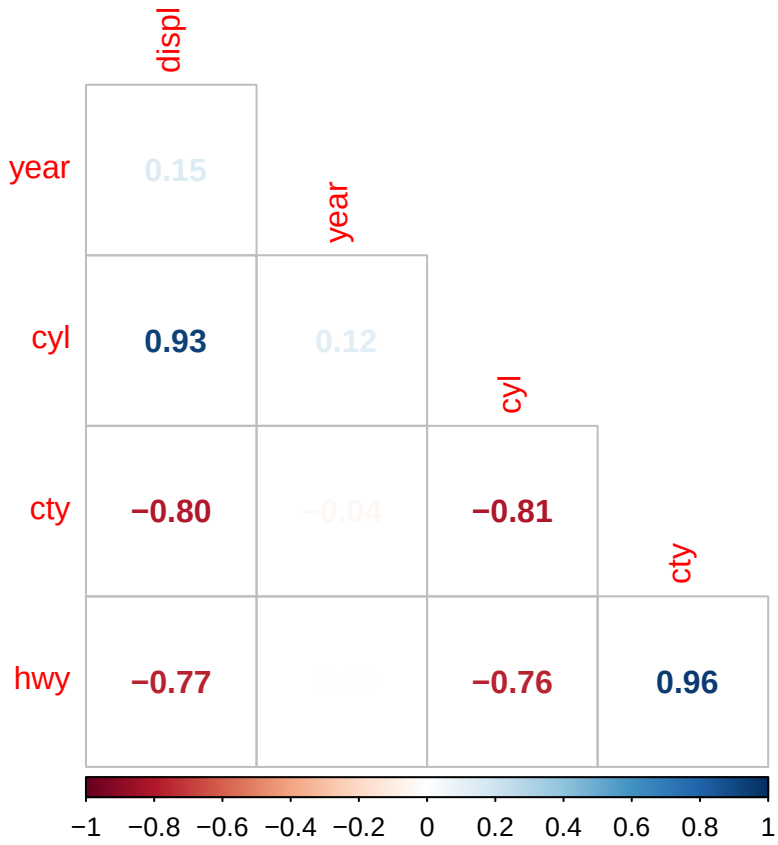
```
cyl <- mpg$cyl
ggplot(mpg, aes(x = factor(cyl), y = hwy)) + geom_boxplot()
```



The box plot demonstrated a general pattern that the more cylinder a car has, the less hwy mpg.

Q5

```
mpg2 <- mpg %>% select(displ, year, cyl, cty, hwy)
M <- cor(mpg2)
corrplot(M, method = 'number', type = 'lower', diag = F)
```



cyl and displ have strong positive correlation. cty and displ have strong negative correlation. hwy and displ have strong negative correlation. cty and cyl have strong negative correlation. hwy and cyl have strong negative correlation. hwy and cty have strong positive correlation. These relationships all make sense to me, it is reasonable that a car has 8 cylinder would have large engine displacement and less city and highway mpg compare to a 4 cylinder car.