Homework 1

Yingshan Li (7937790)

April 03, 2022

Machine Learning Main Ideas

Question 1: In supervised learning, we have an associated response y for each observation of predictor measurements x, so the actual data of response Y can acts as the supervisor. The model is fitted to relates the response to the predictors, which aimed to accurate predicting the response for the future observations ro better understanding the relationship between the response and the predictors(textbook).

For unsupervised learning, we only know the predictors, but we don't know the observed response Y, so there is no supervisor in our analysis.

Question 2: In the context of machine learning, the response of a regression model is quantitative. On the other hand, the response of a classification model is qualitative.

Question 3: Regression: Mean squared error, expected test MSE. Classification: Error rate, Bayes error rate

Question 4: Descriptive models:Choose model to best visually emphasize a trend in data (lecture).

Inferential models: Aim is to test theories, (possibly) causal claims, state relationship between outcome & predictors (lecture).

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

Question 5: Mechanistic: We assume a parametric form for f, and we select a suitable model based on this assumption in order to estimate the set of parameters, but it won't match true unknown f. Increase parameters means more flexibility.

empirically-driven: There are no underlying assumptions about f, so require a large number of observations in order to estimate the unknown function f.

Mechanistic model requires assumption about the function f but the empirically-driven model does not require such assumption. Moreover, the empirically-driven model is generally has more flexibility than the Mechanistic model. Both types may have the problem of over fitting.

From my perspective, mechanistic model is easier to understand because we can fit the model based on our assumptions, and normally the assumptions about f are easy to understand. Moreover, due to the high flexibility of the empirically-driven model, it is very possible that it includes the fitting of random noises that we might find hard to understand.

Bias-Variance trade off:In general, more flexible statistical methods have higher variance, but also result in less bias(textbook). When we use mechanistic or empirically-driven models, we need to choose the flexibility of the model, so we have to select the one that balance between the bias and variances. For example, for a mechanistic model, if our assumption is too simple and less flexible, it may result a model with high bias and low variance, but if it is too complicated, the bias might be reduced, but the variance increases significantly.

Question 6: The first question is predictive. As stated above, the predictive model focused to predict the outcome with minimum reducible error. For this question, we want to know how likely a voter will vote a candidate given a voter's data. In other words, we try to predict how likely they will vote to a candidate based on voter's data.

The second question is inferential. An inferential model aimed to test theories and state relationship between outcome and predictors. For this question, we try to know whether or not the personal contact will affect a voter's likelihood of support for a candidate, so we want to investigate the relationship between these two things.

Exploratory Data Analysis

```
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()
```

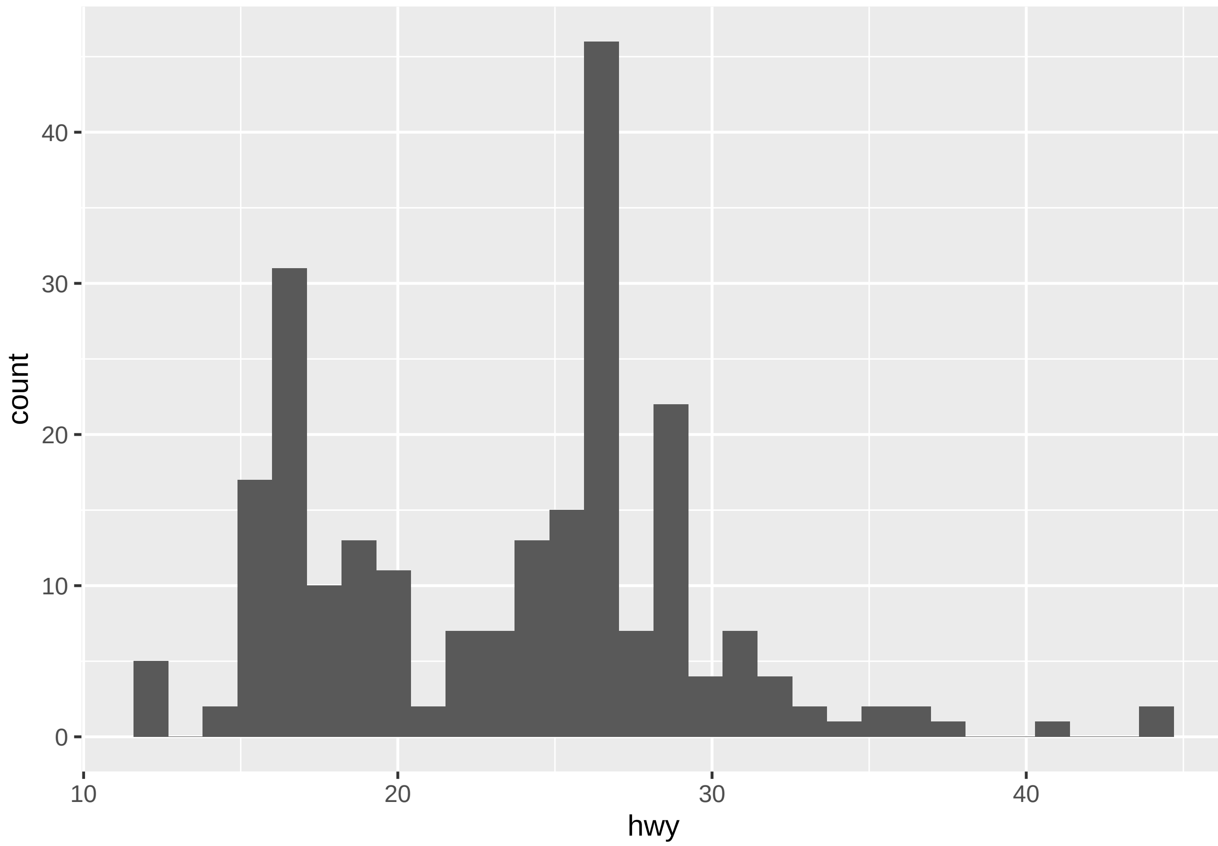
```
mpg
```

```
## # 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
```

Exercise 1:

```
library(ggplot2)
ggplot(mpg, aes(x=hwy)) + geom_histogram()
```

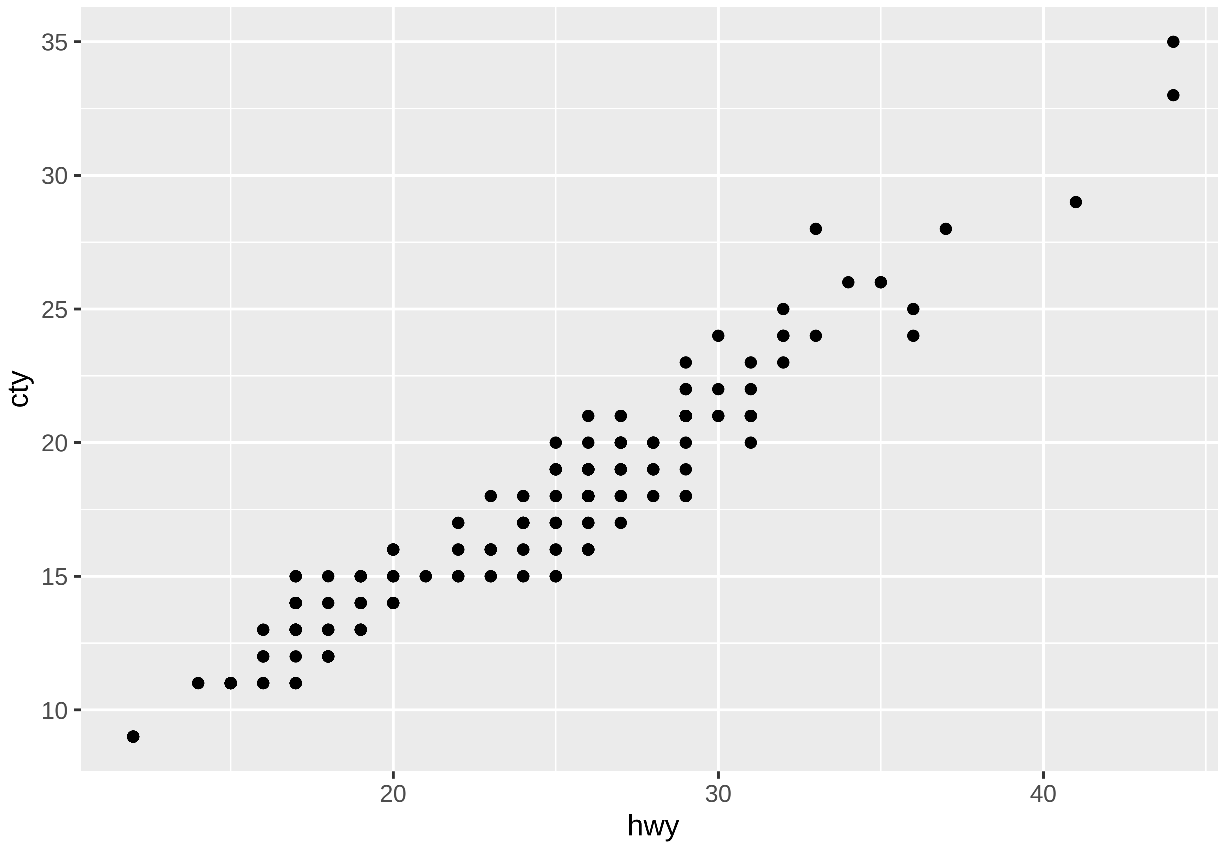
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can see from the histogram that the majority of highway miles per gallon is between 15 and 30, and the most frequent range for hwy variable is 25-30. Based on this histogram, this data set follows a non-symmetric bimodal distribution.

Exercise 2:

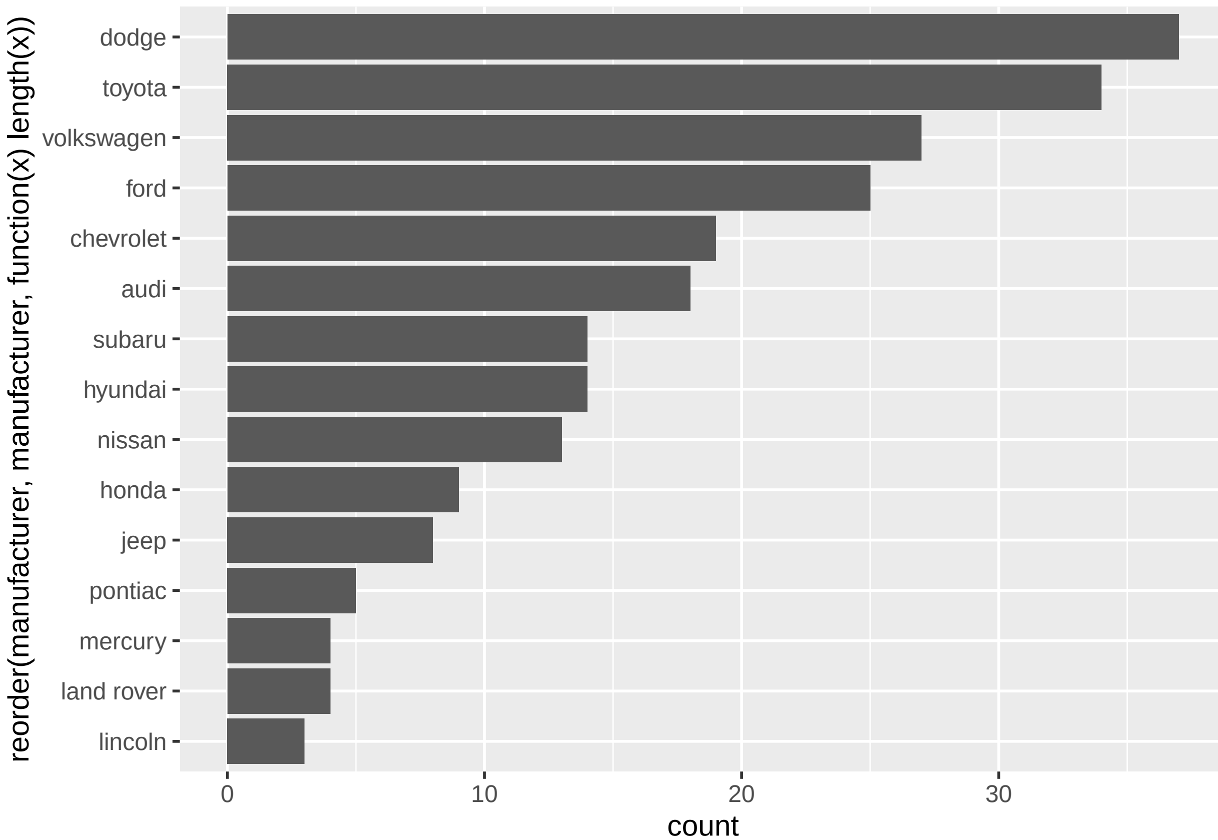
```
ggplot(mpg, aes(x=hwy, y=cty)) + geom_point()
```



We can observe a general uphill pattern in this plot, as the value of hwy increases, the value of the cty increases, so there is a positive relationship between hwy and cty.

Exercise 3:

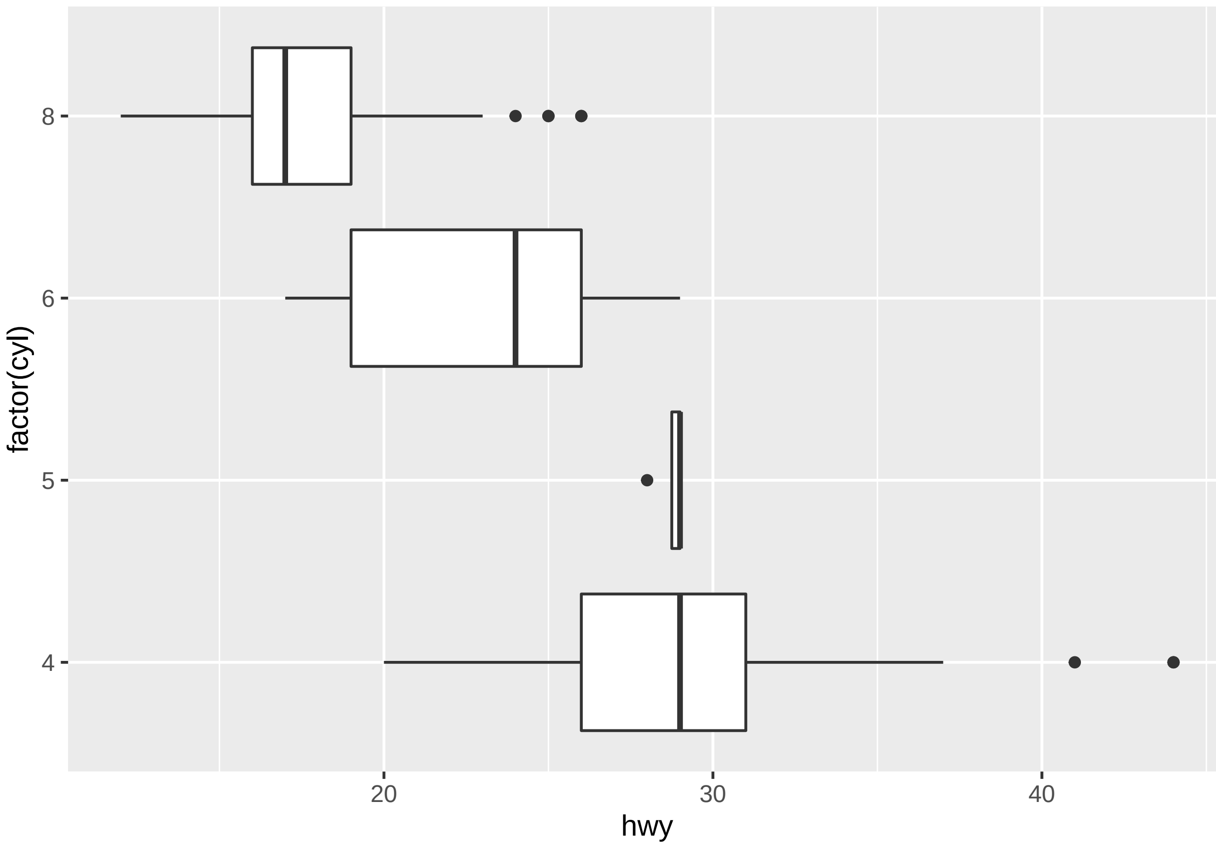
```
ggplot(mpg, aes(x=reorder(manufacturer, manufacturer, function(x) length(x)))) + geom_bar(stat="count")
```



Manufacturer dodge produced the most cars, and lincoln produced least cars.

Exercise 4:

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



We can observe that the smaller the cyl correspond with the higher value of the hwy.

Exercise 5:

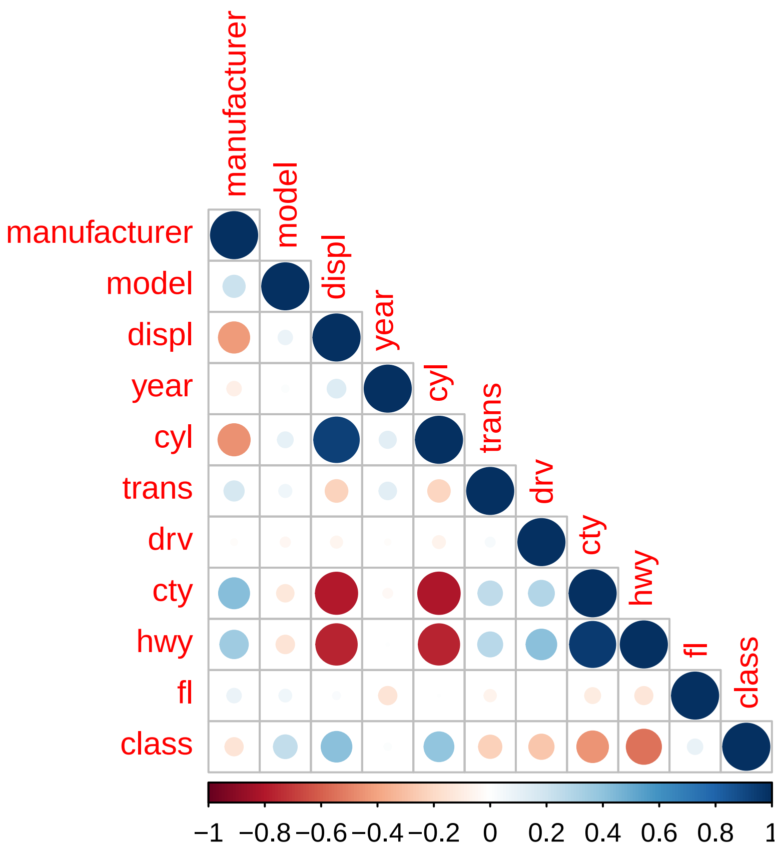
```
summary(mpg)
```

```
##  manufacturer          model               displ            year     
##  Length:234         Length:234         Min.   :1.600   Min.   :1999  
##  Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999  
##  Mode  :character   Mode  :character   Median :3.300   Median :2004  
##                                        Mean   :3.472   Mean   :2004  
##                                        3rd Qu.:4.600   3rd Qu.:2008  
##                                        Max.   :7.000   Max.   :2008  
##       cyl           trans               drv                 cty       
##  Min.   :4.000   Length:234         Length:234         Min.   : 9.00  
##  1st Qu.:4.000   Class :character   Class :character   1st Qu.:14.00  
##  Median :6.000   Mode  :character   Mode  :character   Median :17.00  
##  Mean   :5.889                                         Mean   :16.86  
##  3rd Qu.:8.000                                         3rd Qu.:19.00  
##  Max.   :8.000                                         Max.   :35.00  
##       hwy             fl               class          
##  Min.   :12.00   Length:234         Length:234        
##  1st Qu.:18.00   Class :character   Class :character  
##  Median :24.00   Mode  :character   Mode  :character  
##  Mean   :23.44                                        
##  3rd Qu.:27.00                                        
##  Max.   :44.00                                        
```

```
manufacturer<- factor(mpg$manufacturer)
model <- factor(mpg$model)
displ <- mpg$displ
year <- mpg$year
cyl <- mpg$cyl
trans <- factor(mpg$trans)
drv <- factor(mpg$drv)
cty <- mpg$cty
hwy <- mpg$hwy
fl <- factor(mpg$fl)
class <- factor(mpg$class)
D2 <- cbind(manufacturer, model, displ, year, cyl, trans, drv, cty, hwy, fl, class)
M <- cor(D2)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(M, type = "lower")
```



From this graph, we can see how variables are correlated with each other. (positive correlations are displayed in blue and negative correlations in red color) For example, cyl is positively correlated with displ, and cty is negatively correlated with cyl. There are some relationship seems make sense to me. We can see from the graph that hwy is negatively correlated with cyl, which correspond to the pattern I have discovered in the exercise 4 where smaller cyl has higher hwy. The positively correlation between hwy and cty also correspond to the relationship discovered in exercise 2.