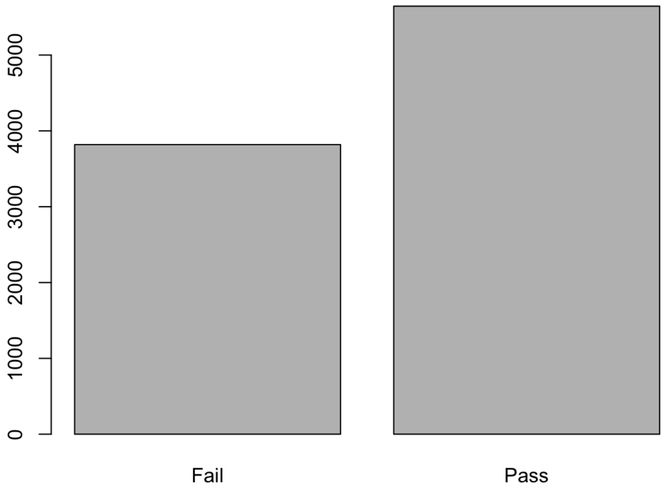
Data 101 Assignment 8 - Prediction Challenge

Seok Yim

April 2nd, 2021

Taking a Quick Look at the Data

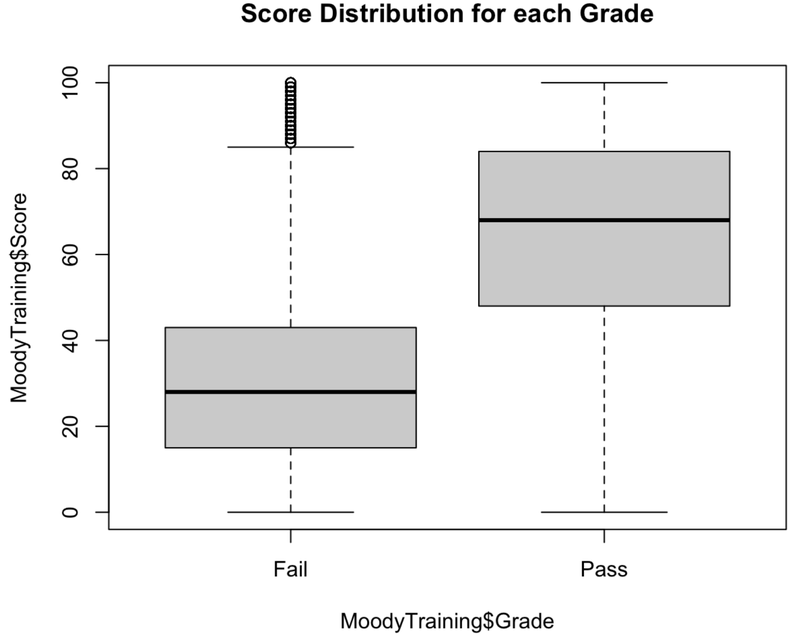
The frequency distribution of grades



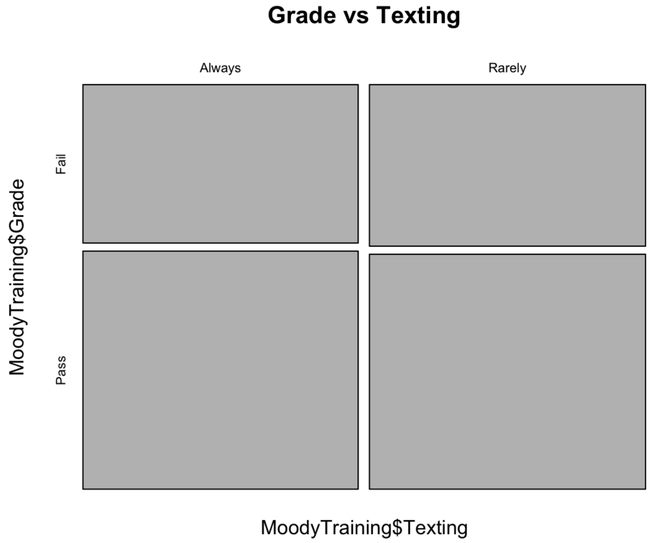
barplot(table(MoodyTraining$Grade))

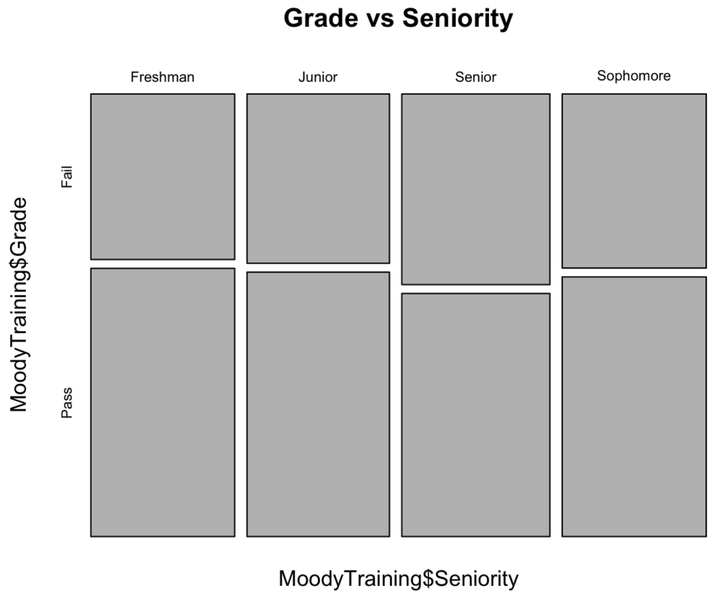
- There are more passes than there are fails

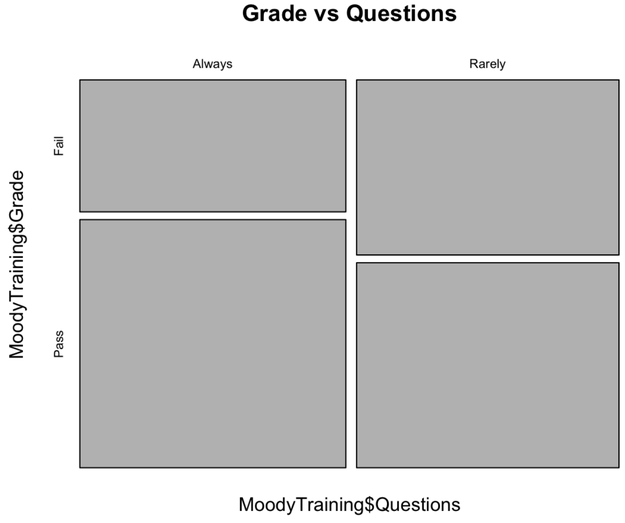
Score Distribution

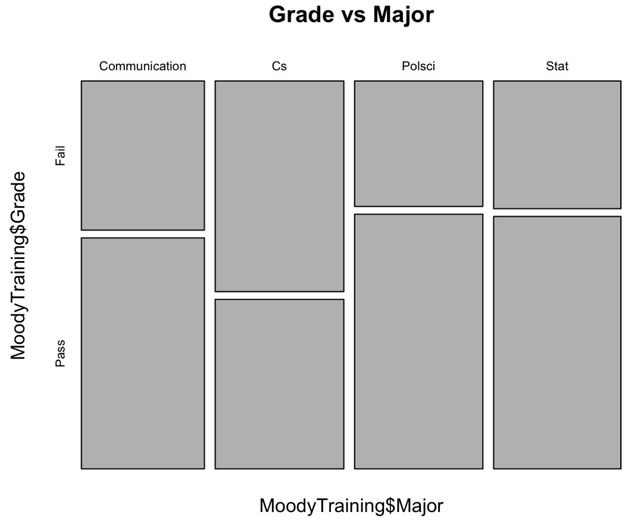


- The average scores for each grade value is different, yet there exists big overlaps between the two grades pass and fail





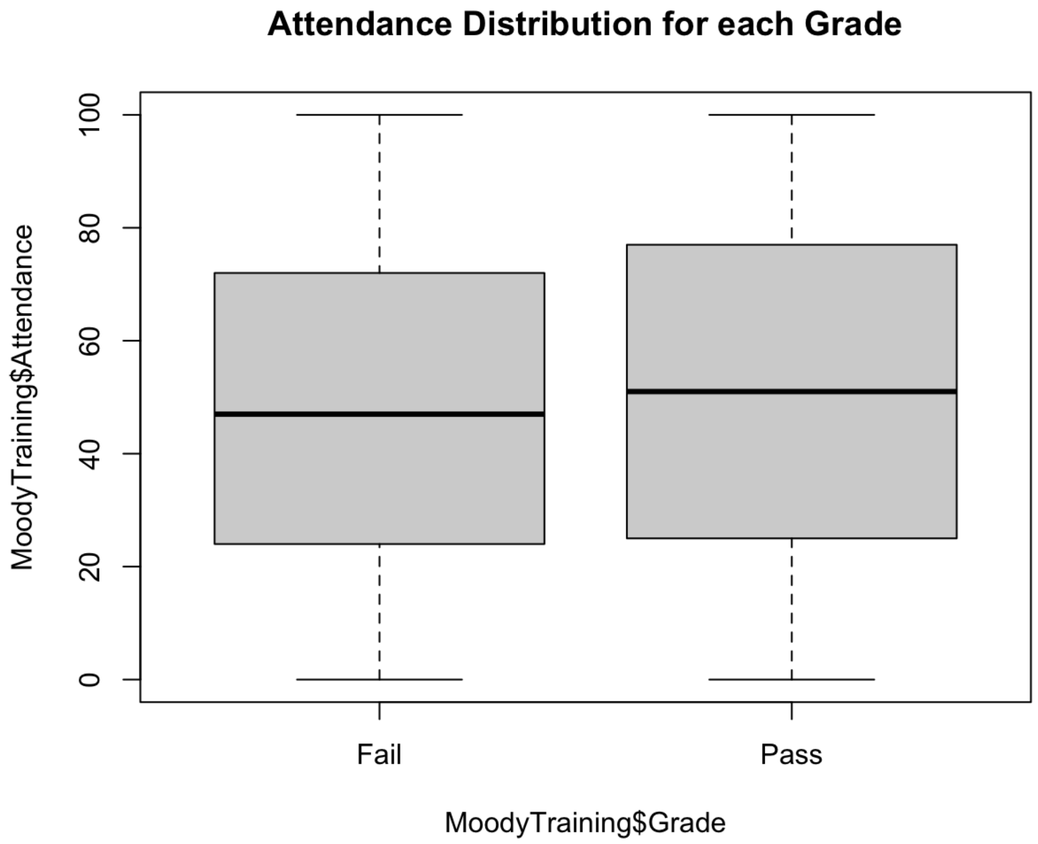




Plots for categorical variables

- These plots show the general pattern of the entire data set
- However, we need to keep in mind that different score ranges might possess different patterns
    - Subsetting needed!

Attendance



- Attendance does not seem too important here
- No big difference shown between the two values of grade
    - I personally did not use attendance as an important variable

Prediction Models

All of them were created with numerous subsetting and slicing and dicing

- Since there were too many steps involved in creating each prediction model, I will briefly showcase the codes for the early models without showing actual plots
- However, for the last (final) prediction model which I submitted to Kaggle, I will explain in detail the process of creating the model using plots and R code
    - If you are only interested in the finalized model, you can skip the short description of the early models

Forming Subsets and Prediction Models: First Model

Through Trial and Error

First Prediction Model:

```
#subsetting the data sets

Fail.belowMean <- MoodyTraining[MoodyTraining$Score < mean(MoodyTraining[MoodyTraining$Grade == “Fail",]$Score),]
Fail.aboveMean <- MoodyTraining[MoodyTraining$Score >= mean(MoodyTraining[MoodyTraining$Grade == "Fail",]$Score) & MoodyTraining$Score < mean(MoodyTraining[MoodyTraining$Grade == "Pass",]$Score),]
Pass.aboveMean <- MoodyTraining[MoodyTraining$Score >= mean(MoodyTraining[MoodyTraining$Grade == "Pass",]$Score),]

mean(MoodyTraining[MoodyTraining$Grade == "Fail",]$Score)
#value is 30.78686
mean(MoodyTraining[MoodyTraining$Grade == "Pass",]$Score)
#value is 63.31851
```

Below is the model I came up with after plotting and analyzing plots:

```
#mypredict model #1
  #permuting the order of the rows in the data set
randomSubset <- MoodyTraining[sample(1:nrow(MoodyTraining)),]
  #taking the first 500 rows from the result of the permutation as the subset to test my model on
testingSubset <- randomSubset[1:1000,]

myprediction <- testingSubset
decision <- rep("Fail",nrow(myprediction))
decision[myprediction$Score < 30.78686 & myprediction$Questions == "Always" & myprediction$Seniority == "Junior"] <- "Pass"
decision[myprediction$Score >= 30.78686 & myprediction$Score <63.31851 & myprediction$Questions == "Always" & (myprediction$Major == "Communication" | myprediction$Major == "Polsci")] <- "Pass"
decision[myprediction$Score >= 63.31851] <- "Pass"
decision[myprediction$Score >= 63.31851 & myprediction$Seniority == "Senior" & myprediction$Questions == "Rarely" & myprediction$Major != "Cs"] <- "Fail"
myprediction$Grade <- decision

error <- mean(myprediction$Grade != testingSubset$Grade)
error

#error percentage for 5 random subsets tested: 0.262, 0.246, 0.282, 0.254, 0.324
#average error: 0.2736
```

Second Prediction Model

New Approach: using 1st Qu. from “Passed” and 3rd Qu. From “Failed”

```
#new approach: subsetting into two data frames: pass and fail
Passed <- MoodyTraining[MoodyTraining$Grade == "Pass",]
Failed <- MoodyTraining[MoodyTraining$Grade == "Fail",]

summary(Passed)
summary(Failed)
```

```
> summary(Passed)
   Studentid       Attendance        Major            Questions            Score         Seniority
 Min.   :29998   Min.   :  0.00   Length:5645        Length:5645        Min.   :  0.00   Length:5645
 1st Qu.:32354   1st Qu.: 25.00   Class :character   Class :character   1st Qu.: 48.00   Class :character
 Median :34736   Median : 51.00   Mode  :character   Mode  :character   Median : 68.00   Mode  :character
 Mean   :34728   Mean   : 51.07                                        Mean   : 63.32
 3rd Qu.:37062   3rd Qu.: 77.00                                        3rd Qu.: 84.00
 Max.   :39460   Max.   :100.00                                        Max.   :100.00
   Texting             Grade
 Length:5645        Length:5645
 Class :character   Class :character
 Mode  :character   Mode  :character
```

```
> summary(Failed)
   Studentid       Attendance        Major            Questions            Score         Seniority
 Min.   :29999   Min.   :  0.00   Length:3819        Length:3819        Min.   :  0.00   Length:3819
 1st Qu.:32380   1st Qu.: 24.00   Class :character   Class :character   1st Qu.: 15.00   Class :character
 Median :34715   Median : 47.00   Mode  :character   Mode  :character   Median : 28.00   Mode  :character
 Mean   :34731   Mean   : 48.36                                        Mean   : 30.79
 3rd Qu.:37132   3rd Qu.: 72.00                                        3rd Qu.: 43.00
 Max.   :39461   Max.   :100.00                                        Max.   :100.00
   Texting             Grade
 Length:3819        Length:3819
 Class :character   Class :character
 Mode  :character   Mode  :character
```

```
barplot(table(Passed[Passed$Score < 48,]$Major))
barplot(table(Passed[Passed$Score < 48,]$Questions))
barplot(table(Passed[Passed$Score < 48,]$Seniority))
barplot(table(Passed[Passed$Score < 48,]$Texting))


barplot(table(Failed[Failed$Score >= 48,]$Major))
barplot(table(Failed[Failed$Score >= 48,]$Questions))
barplot(table(Failed[Failed$Score >= 48,]$Seniority))
barplot(table(Failed[Failed$Score >= 48,]$Texting))

barplot(table(Failed[Failed$Score >= 84,]$Major))
barplot(table(Failed[Failed$Score >= 84,]$Questions))
barplot(table(Failed[Failed$Score >= 84,]$Seniority))
barplot(table(Failed[Failed$Score >= 84,]$Texting))
```

Trying to see the correlation between majors and grade as well:

```
summary(MoodyTraining[MoodyTraining$Major == "Cs" & MoodyTraining$Grade == "Pass",])
summary(MoodyTraining[MoodyTraining$Major == "Cs" & MoodyTraining$Grade == "Fail",])

summary(MoodyTraining[MoodyTraining$Major == "Polsci" & MoodyTraining$Grade == "Pass",])
summary(MoodyTraining[MoodyTraining$Major == "Polsci" & MoodyTraining$Grade == "Fail",])

summary(MoodyTraining[MoodyTraining$Major == "Stat" & MoodyTraining$Grade == "Pass",])
summary(MoodyTraining[MoodyTraining$Major == "Stat" & MoodyTraining$Grade == "Fail",])

summary(MoodyTraining[MoodyTraining$Major == "Communication" & MoodyTraining$Grade == "Pass",])
summary(MoodyTraining[MoodyTraining$Major == "Communication" & MoodyTraining$Grade == "Fail",])
```

Second Prediction Model Results

The Error percentage

```
#mypredict model #2
  #permuting the order of the rows in the data set
randomSubset <- MoodyTraining[sample(1:nrow(MoodyTraining)),]
  #taking the first 500 rows from the result of the permutation as the subset to test my model on
testingSubset <- randomSubset[1:1000,]

myprediction <- testingSubset
decision <- rep("Fail",nrow(myprediction))
decision[myprediction$Score < 30.78686 & myprediction$Questions == "Always" & myprediction$Seniority == "Junior"] <- "Pass"
decision[myprediction$Score >= 30.78686 & myprediction$Score <63.31851 & myprediction$Questions == "Always" & (myprediction$Major == "Communication" |
myprediction$Major == "Polsci")] <- "Pass"
decision[myprediction$Score >= 63.31851] <- "Pass"
decision[myprediction$Score >= 63.31851 & myprediction$Seniority == "Senior" & myprediction$Questions == "Rarely" & myprediction$Major != "Cs"] <- "Fail"
decision[myprediction$Score >= 48 & myprediction$Major == "Cs"] <- "Fail"
decision[myprediction$Score >= 84 & myprediction$Seniority == "Junior" & myprediction$Questions == "Always"] <- "Fail"
myprediction$Grade <- decision

error <- mean(myprediction$Grade != testingSubset$Grade)
error
#error percentage for 5 randonm subsets tested: 0.364, 0.36, 0.346, 0.404, 0.37
#not so good compared to the first model
```

Third Prediction Model

New Approach: trying out more things randomly to find patterns and adding them to my first model

```
barelyPassed <- Passed[Passed$Score < 48,]
barelyFailed <- Failed[Failed$Score > 43,]

summary(barelyPassed)
summary(barelyFailed)

barplot(table(barelyFailed$Major))
barplot(table(barelyFailed$Questions))
barplot(table(barelyFailed$Seniority))
barplot(table(barelyFailed$Texting))

barplot(table(barelyPassed$Major))
barplot(table(barelyPassed$Questions))
barplot(table(barelyPassed$Seniority))
barplot(table(barelyPassed$Texting))
```

```
#mypredict model #3
  #permuting the order of the rows in the data set
randomSubset <- MoodyTraining[sample(1:nrow(MoodyTraining)),]
  #taking the first 500 rows from the result of the permutation as the subset to test my model on
testingSubset <- randomSubset[1:1000,]

myprediction <- testingSubset
decision <- rep("Fail",nrow(myprediction))
decision[myprediction$Score < 30.78686 & myprediction$Questions == "Always" & myprediction$Seniority == "Junior"] <- "Pass"
decision[myprediction$Score >= 30.78686 & myprediction$Score <63.31851 & myprediction$Questions == "Always" & (myprediction$Major ==
"Communication" | myprediction$Major == "Polsci")] <- "Pass"
decision[myprediction$Score >= 63.31851] <- "Pass"
decision[myprediction$Score >= 63.31851 & myprediction$Seniority == "Senior" & myprediction$Questions == "Rarely" & myprediction$Major != "Cs"] <-
"Fail"
decision[myprediction$Score < 43 & myprediction$Major == "Cs"] <- "Fail"
decision[myprediction$Score > 43 & myprediction$Questions == "Rarely" & myprediction$Seniority == "Senior"] <- "Fail"
decision[myprediction$Score < 48 & myprediction$Major != "Cs" & myprediction$Questions == "Always" && myprediction$Texting == "Always"] <- "Pass"

myprediction$Grade <- decision

error <- mean(myprediction$Grade != testingSubset$Grade)
error

#error percentages for 5 random subsets tested: .302, 0.258, 0.246, 0.28, 0.378
#not so stable
```

4th Prediction Model

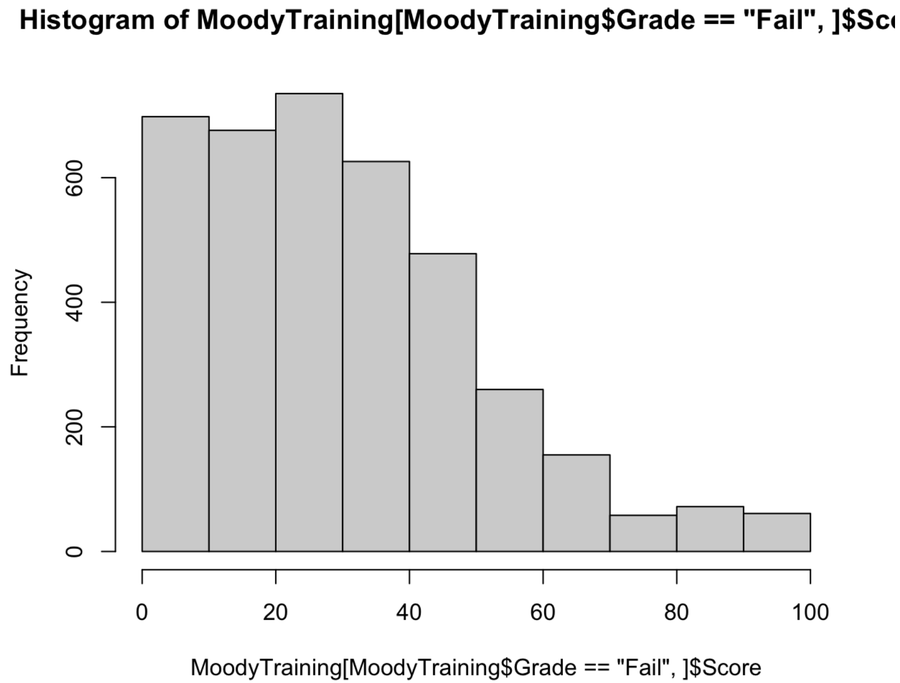
Approach: Being as desperate as I can be

```
barplot(table(MoodyTraining[MoodyTraining$Score < 40,]$Grade))
barplot(table(MoodyTraining[MoodyTraining$Score > 70,]$Grade))
barplot(table(MoodyTraining[MoodyTraining$Score >= 40 & MoodyTraining$Score < 55,]$Grade))
barplot(table(MoodyTraining[MoodyTraining$Score >= 40 & MoodyTraining$Score < 47.5,]$Grade))
barplot(table(MoodyTraining[MoodyTraining$Score >= 47.5 & MoodyTraining$Score < 55,]$Grade))
barplot(table(MoodyTraining[MoodyTraining$Score >= 55 & MoodyTraining$Score <= 70,]$Grade))
```

- The ranges of scores for each subset was as follows:
  - Score <40
  - Score >= 40 & Score <47.5
  - Score >= 47.5 & Score < 55
  - Score >= 55 & Score <=70
  - Score >70

- Basically, this time I divided the data set into 5 subsets based on the scores, since I considered the score variable to be the most impactful in the data set
  - As we saw before, the different in mean scores was quite clear between pass and fail, although the range of scores overlapped to a worrisome degree

How did I get these Numbers?




```r
hist(MoodyTraining[MoodyTraining$Grade == "Pass",]$Score)
# score < 40 -> Fail
hist(MoodyTraining[MoodyTraining$Grade == "Fail",]$Score)
# score > 70 -> Pass
```

- I got it from these histograms!
  - Most students that passed have a score > =40
  - Most students that failed have a score <= 70
    - So first give people below the score of 40 a fail and the people above a 70 a pass!
  - After coming up with 40 and 70, I further divided up the range between 40 and 70 into smaller intervals to see if there were distinct patterns in the distribution of grades

```r
hist(middle[middle$Grade == "Fail",]$Score)


barplot(table(MoodyTraining[MoodyTraining$Score >= 40 & MoodyTraining$Score < 55 & MoodyTraining$Grade == "Pass",]$Major))
barplot(table(MoodyTraining[MoodyTraining$Score >= 40 & MoodyTraining$Score < 55 & MoodyTraining$Grade == "Fail",]$Major))
# communication -> pass, cs -> fail

barplot(table(MoodyTraining[MoodyTraining$Score >= 40 & MoodyTraining$Score < 55 & MoodyTraining$Grade == "Pass",]$Questions))
barplot(table(MoodyTraining[MoodyTraining$Score >= 40 & MoodyTraining$Score < 55 & MoodyTraining$Grade == "Fail",]$Questions))
# always -> pass, rarely -> fail
# these aren't that big of a deal though


barplot(table(MoodyTraining[MoodyTraining$Score >= 40 & MoodyTraining$Score < 55 & MoodyTraining$Grade == "Pass",]$Texting))
barplot(table(MoodyTraining[MoodyTraining$Score >= 40 & MoodyTraining$Score < 55 & MoodyTraining$Grade == "Fail",]$Texting))
#not interesting


barplot(table(MoodyTraining[MoodyTraining$Score >= 40 & MoodyTraining$Score < 55 & MoodyTraining$Grade == "Pass",]$Seniority))
barplot(table(MoodyTraining[MoodyTraining$Score >= 40 & MoodyTraining$Score < 55 & MoodyTraining$Grade == "Fail",]$Seniority))
#senior more likely to fail, freshman more likely to pass
#not too big of a deal though

#try always questions & freshman -> pass, rarely question & senior -> fail
barplot(table(MoodyTraining[MoodyTraining$Score >= 55 & MoodyTraining$Score <= 70 & MoodyTraining$Grade == "Pass",]$Major))
barplot(table(MoodyTraining[MoodyTraining$Score >= 55 & MoodyTraining$Score <= 70 & MoodyTraining$Grade == "Fail",]$Major))
#not cs -> more likely to pass

barplot(table(MoodyTraining[MoodyTraining$Score >= 55 & MoodyTraining$Score <= 70 & MoodyTraining$Grade == "Pass",]$Questions))
barplot(table(MoodyTraining[MoodyTraining$Score >= 55 & MoodyTraining$Score <= 70 & MoodyTraining$Grade == "Fail",]$Questions))
#not interesting

barplot(table(MoodyTraining[MoodyTraining$Score >= 55 & MoodyTraining$Score <= 70 & MoodyTraining$Grade == "Pass",]$Texting))
barplot(table(MoodyTraining[MoodyTraining$Score >= 55 & MoodyTraining$Score <= 70 & MoodyTraining$Grade == "Fail",]$Texting))
#not interesting

barplot(table(MoodyTraining[MoodyTraining$Score >= 55 & MoodyTraining$Score <= 70 & MoodyTraining$Grade == "Pass",]$Seniority))
barplot(table(MoodyTraining[MoodyTraining$Score >= 55 & MoodyTraining$Score <= 70 & MoodyTraining$Grade == "Fail",]$Seniority))
#freshman or sophomore -> pass, else -> fail
```

55:1 (Top Level) R Script

- And then what was left was a whole bunch of plots to make to find patterns...

4th Prediction Model Result

Better Error percentages than before!

```
#desperate model
  #permuting the order of the rows in the data set
randomSubset <- MoodyTraining[sample(1:nrow(MoodyTraining)),]
  #taking the first 500 rows from the result of the permutation as the subset to test my model on
testingSubset <- randomSubset[1:1000,]

myprediction <- testingSubset
decision <- rep("Fail",nrow(myprediction))
decision[myprediction$Score < 40] <- "Fail"
decision[myprediction$Score > 70] <- "Pass"
decision[myprediction$Score >= 40 & myprediction$Score <47.5] <- "Fail"
decision[myprediction$Score >=47.5 & myprediction$Score <55] <- "Pass"
decision[myprediction$Score < 40 & (myprediction$Major == "Cs" | myprediction$Major == "Communication")] <- "Fail"
decision[myprediction$Score < 40 & myprediction$Major != "Cs" & myprediction$Major != "Communication"] <- "Pass"
decision[myprediction$Score < 40 & myprediction$Seniority == "Junior" & myprediction$Questions == "Always"] <- "Pass"
decision[myprediction$Score >=40 & myprediction$Score <47.5 & myprediction$Major == "Communication"] <- "Pass"
decision[myprediction$Score >=40 & myprediction$Score <47.5 & myprediction$Major != "Cs" & myprediction$Questions == "Always" & myprediction$Seniority !=
"Senior"] <- "Pass"
decision[myprediction$Score >= 55 & myprediction$Score <= 70] <- "Pass"
decision[myprediction$Score >= 55 & myprediction$Score <= 70 & myprediction$Major == "Cs" & (myprediction$Seniority == "Junior" | myprediction$Seniority ==
"Senior")] <- "Fail"
decision[myprediction$Score >=47.5 & myprediction$Score <55 & (myprediction$Major == "Cs" |myprediction$Major == "Stat") & myprediction$Seniority ==
"Senior"] <- "Fail"

myprediction$Grade <- decision


error <- mean(myprediction$Grade != testingSubset$Grade)
error
#error percentage for the random subsets tested:0.23, 0.201, 0.192, 0.213, 0.222
Average error percentage: 0.2116
#MMM!!! Better error percentages now!! BUT DO I STOP HERE? NO!
```

Result for my 4th prediction model

Improvements in my Prediction Model

- Compared to the average error of 0.2736 that I got from my first model, the final prediction model gave 0.2116, which is quite smaller!
    - The error values seemed to be stable as well
    - However, this is only the result that I got from testing using randomized subsets from the training dataset
    - Also, I want to **lower** my error percentage by finding more patterns!
        - I want to rank high in the leaderboard for the prediction challenge!
            - Which means… I'll be putting in more work!

FINAL PREDICTION MODEL

Revising and Adding and Subtracting from my 4th Prediction Model

- I took only the last prediction model design from the original document and then included it in my newly created R code in order to focus on it more
    - Created more plots and analyzed them
    - There are a crazy number of lines, so I apologize if it is hard to read every line.

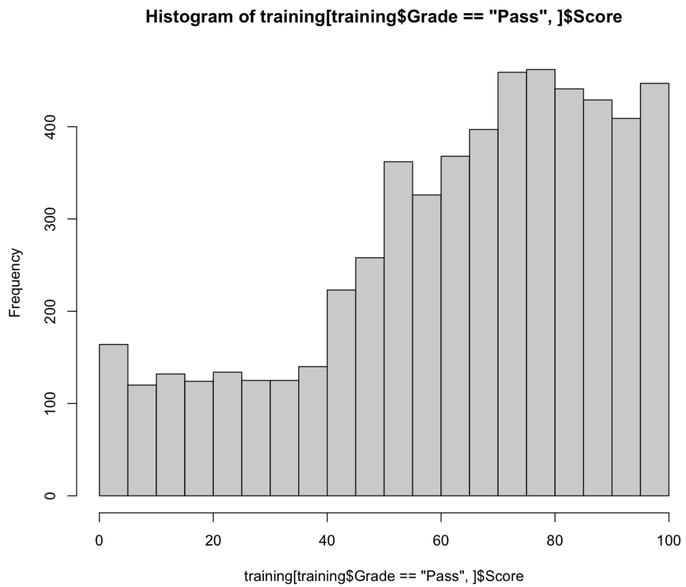
R Code of the finalized model

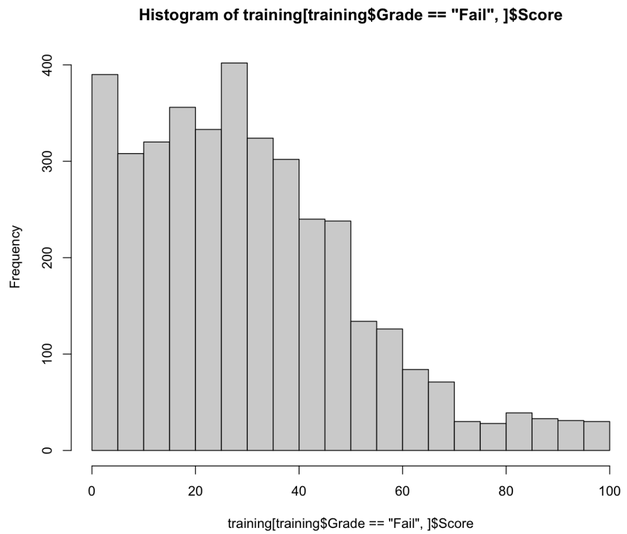
```
#FINAL MODEL
#permuting the order of the rows in the data set
randomSubset <- training[sample(1:nrow(training)),]
#taking the first 500 rows from the result of the permutation as the subset to test my model on
testingSubset <- randomSubset[1:1000,]

myprediction <- testingSubset
decision <- rep("Fail",nrow(myprediction))
decision[myprediction$Score < 40] <- "Fail"
decision[myprediction$Score > 70] <- "Pass"
decision[myprediction$Score > 60 & myprediction$Major == "Stat"] <- "Pass"
decision[myprediction$Score > 50 & myprediction$Major == "Polsci"] <- "Pass"
decision[myprediction$Score >=47.5 & myprediction$Score <55] <- "Pass"
decision[myprediction$Score < 40 & myprediction$Major != "Cs" & myprediction$Major != "Communication" & myprediction$Seniority != "Senior" & myprediction$Seniority != "Freshman" & myprediction$Questions == "Always"] <- "Pass"
decision[myprediction$Score >=40 & myprediction$Score <47.5 & myprediction$Major == "Communication" & myprediction$Texting == "Always"] <- "Pass"
decision[myprediction$Score >=40 & myprediction$Score <47.5 & myprediction$Major == "Communication" & myprediction$Questions== "Always"] <- "Pass"
decision[myprediction$Score >=40 & myprediction$Score <47.5 & myprediction$Major != "Cs" & myprediction$Questions == "Always" & myprediction$Seniority != "Senior"] <- "Pass"
decision[myprediction$Score >= 55 & myprediction$Score <= 70] <- "Pass"
decision[myprediction$Score >= 55 & myprediction$Score <= 70 & myprediction$Major == "Cs" & (myprediction$Seniority == "Junior" | myprediction$Seniority == "Senior")] <- "Fail"
decision[myprediction$Score >=47.5 & myprediction$Score <55 & (myprediction$Major == "Cs" |myprediction$Major == "Stat") & myprediction$Seniority == "Senior"] <- "Fail"
decision[myprediction$Score <50 & myprediction$Score >= 0 & myprediction$Major == "Cs" & myprediction$Seniority != "Junior"  & myprediction$Questions == "Rarely"] <- "Fail"
decision[myprediction$Score <40 & myprediction$Major == "Polsci" & myprediction$Questions == "Always" & myprediction$Seniority != "Senior" & myprediction$Seniority != "Junior"] <- "Pass"
decision[myprediction$Score <40 & myprediction$Score >= 0 & myprediction$Major != "Stat" & myprediction$Major != "Polsci" & myprediction$Major != "Cs" & (myprediction$Seniority == "Senior"|myprediction$Seniority=="Freshman")&
myprediction$Texting=="Rarely"] <- "Fail"
decision[myprediction$Score <45 & myprediction$Score >= 40 & myprediction$Major != "Stat" & myprediction$Major != "Polsci" & myprediction$Major != "Cs" & (myprediction$Seniority == "Senior"|myprediction$Seniority=="Freshman")&
myprediction$Texting=="Rarely"] <- "Pass"
decision[myprediction$Score >= 47.5 & myprediction$Score <55 & myprediction$Seniority=="Senior" & (myprediction$Major == "Cs" | myprediction$Major == "Stat")& myprediction$Questions == "Always"] <- "Fail"
#ALL FAILS
decision[myprediction$Score < 70 & myprediction$Major == "Cs" & myprediction$Seniority == "Senior"] <- "Fail"
decision[myprediction$Score <50 & myprediction$Score >= 0 & myprediction$Major == "Polsci" & myprediction$Seniority != "Senior"  & myprediction$Questions == "Rarely"] <- "Fail"
decision[myprediction$Score > 0 & myprediction$Score < 50 & myprediction$Major == "Polsci" & myprediction$Questions == "Rarely" ] <- "Fail"
#ALL PASSES
decision[myprediction$Score > 70 & myprediction$Major == "Cs"] <- "Pass"
decision[myprediction$Score <70 & myprediction$Score >= 50 & myprediction$Major == "Polsci" & myprediction$Seniority == "Freshman"  & myprediction$Questions == "Rarely" & myprediction$Texting == "Always"] <- "Pass"
decision[myprediction$Score > 40 & myprediction$Score < 50 & myprediction$Major == "Communication" & myprediction$Seniority == "Sophomore"] <- "Pass"


myprediction$Grade <- decision


error <- mean(myprediction$Grade != testingSubset$Grade)
error
#error percentage for the random subsets tested: 0.163, 0.158,  0.145, 0.135, 0.154
#WOW!! MUCH BETTER RESULTS COMPARED TO THE PREVIOUS MODELS!! Average error: 0.151
```

What was the process?

How I got these lines for my model

```
hist(training[training$Grade == "Pass",]$Score, breaks = 19)
hist(training[training$Grade == "Fail",]$Score, breaks = 19)
#score > 70 -> pass
```





- Basic Steps:
  - Giving “Fail” to every row first
  - Afterwards, I attempted to understand to which score intervals of the entire data set I should give a “Pass”
    - It seemed that giving a pass to everyone above a score of 70 would be a wise choice
      - This is because of how low the population was for score> 70 for those who failed
      - Means that most of the people with score > 70 will get the grade that they correspond to

What are these numbers that I got?

40, 47.5, 55, 70?

```
summary(training[training$Grade == "Pass",])
summary(training[training$Grade == "Fail",])
summary(training[training$Score >= 40 & training$Score < 70,])

summary(training[training$Score >= 40 & training$Score < 55,])
#almost half if pass
summary(training[training$Score >= 55 & training$Score < 70,])
#around 75% is pass


barplot(table(training[training$Score>= 40 & training$Score < 55,]$Major))
barplot(table(training[training$Score>= 40 & training$Score < 55,]$Questions))
barplot(table(training[training$Score>= 40 & training$Score < 55,]$Seniority))
barplot(table(training[training$Score>= 40 & training$Score < 55,]$Texting))

summary(training[training$Score >= 40 & training$Score < 47.5,])
summary(training[training$Score >= 47.5 & training$Score < 51,])
summary(training[training$Score >=  51& training$Score < 55,])
summary(training[training$Score >=  51& training$Score < 55 & training$Major != "Cs",])
```

- Used the summary() function multiple times on many different subsets
  - Groups that I tried out:
    - Pass and fail, score between 40 and 47.5, stat and non stat majors, attendance > score, score > 70 for Fail, etc...
    - Had to try completely random subsets, since I did not know which variables were more important(weighted more) than others!
    - After spending hours, I came up with the following score intervals:
      - Score <40, score >= 40 & score <47.5, score >= 47.5 & score <55, score >=55 & score<=70, score > 70
      - How I got 40: to the right of 40 in the histogram for those who passed was where the majority was located, and to the left was the minority
      - 47.5, and 55 are just numbers I got by trying to subset the interval between 40 and 70 into smaller parts
      - Assigned pass to scores >= 47.5

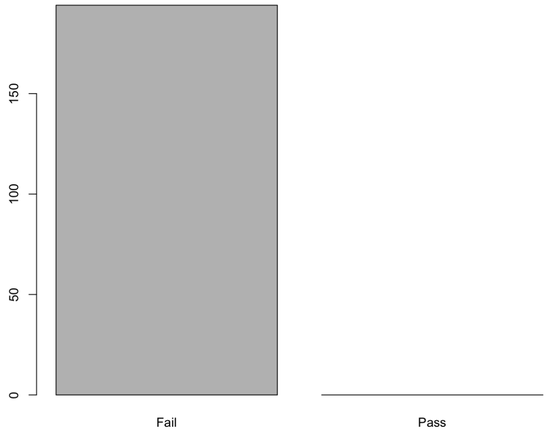
Rcode Explained

- When assigning the grades for the distinct subsets, I made sure that each subset (for instance, score <40) received the grade which was the most populated within that particular subset
    - This way, every subset has a low error percentage for their grades
    - Afterwards, I made big or small exceptions to the fundamental rules (the grade assignment to each subset at the beginning) based on my findings

Explanation Continued

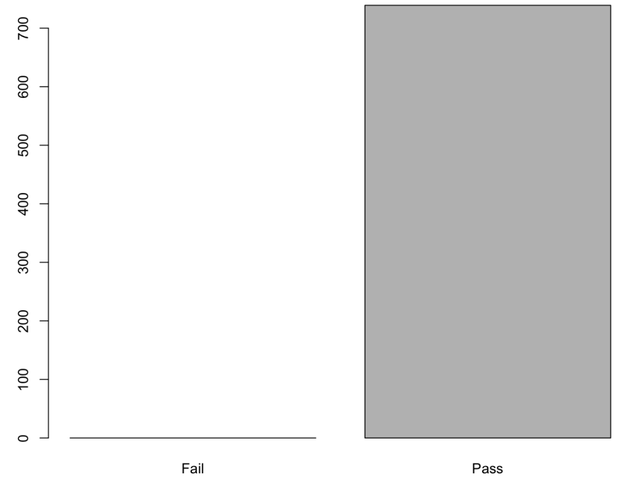
- Examples of big exceptions:
  - **1) EVERYONE** with score <70 who are Computer Science majors and are seniors and rarely text FAIL THE CLASS
  - **2) EVERYONE** with score>70 who are Cs majors PASS THE CLASS
  - **3) EVERYONE** with score <50 who are Political Science majors who aren't seniors who rarely ask questions FAIL THE CLASS

1)



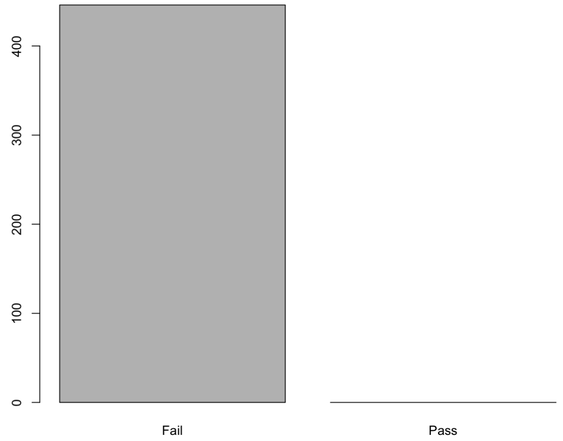
barplot(table(training[training$Score < 70 & training$Major == "Cs" & training$Seniority == "Senior" & training$Texting == "Rarely",]$Grade))

2)



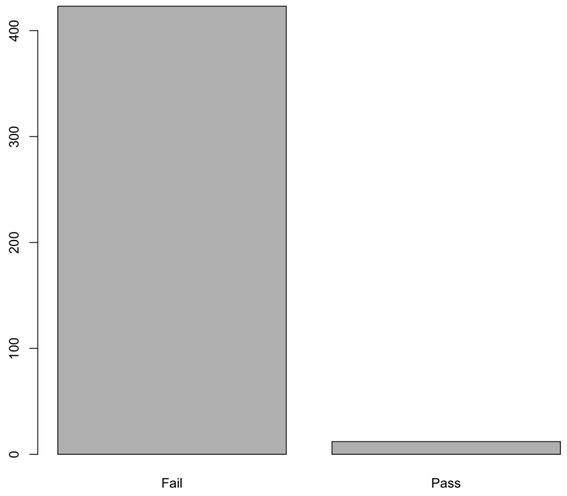
barplot(table(training[training$Score > 70 & training$Major == "Cs" ,]$Grade))

3)



barplot(table(training[training$Score <50 & training$Score >= 0 & training$Major == "Polsci" & training$Seniority != "Senior" & training$Questions == "Rarely",]$Grade))

```
barplot(table(training[training$Score > 70 & training$Major == "Cs" & training$Seniority == "Senior"  & training$Texting == "Rarely",]$Grade))
#ALL PASSES
#?????? WHAT?? WAHT ARE THESE PATTERNS???
#this one is less important, as I let everyone above 70 pass


barplot(table(training[training$Score > 70 & training$Major == "Cs" & training$Seniority == "Senior"  & training$Questions == "Rarely",]$Grade))
barplot(table(training[training$Score <70 & training$Score >= 50 & training$Major == "Polsci" & training$Seniority == "Freshman"  & training$Questions == "Rarely" & training$Texting =


#addition2
barplot(table(training[training$Score <70 & training$Score >= 50 & training$Major == "Polsci" & training$Seniority == "Freshman"  & training$Questions == "Rarely" & training$Texting =
#All Pass

#addition3
barplot(table(training[training$Score <50 & training$Score >= 0 & training$Major == "Polsci" & training$Seniority != "Senior"  & training$Questions == "Rarely",]$Grade))
#ALL FAIL


#addition4
barplot(table(training[training$Score <50 & training$Score >= 0 & training$Major == "Cs" & training$Seniority != "Junior"  & training$Questions == "Rarely",]$Grade))
#Mostly fail


barplot(table(training[training$Score <50 & training$Score >= 0 & training$Major != "Cs" & training$Seniority == "Junior"  & training$Questions == "Always" & training$Texting=="Always
#not interesting

#addition5
barplot(table(training[training$Score <40 & training$Score >= 0 & training$Major != "Stat" & training$Major != "Polsci" & training$Major != "Cs" & (training$Seniority == "Senior"|trai
#mostly fail!!

#addition6
barplot(table(training[training$Score <45 & training$Score >= 40 & training$Major != "Stat" & training$Major != "Polsci" & training$Major != "Cs" & (training$Seniority == "Senior"|tra
#mostly pass!

barplot(table(training[training$Score <50 & training$Score >= 45 & training$Major != "Stat" & training$Major != "Polsci" & training$Major != "Cs" & (training$Seniority == "Senior"|tra
```



barplot(table(training[training$Score <50 & training$Score >= 0 & training$Major == "Cs" & training$Seniority != "Junior" & training$Questions == "Rarely",]$Grade))

Explanation Continued

- I kept on adding new rules and big exceptions
  - Small exceptions were things where only most of the people either failed or passed the class, not the entirety
  - Ex) addition 4: students with score <50 who are Cs majors who aren't juniors who rarely text mostly receive a failing grade

Other Small Efforts

```r
#observing how the ratio of passes and fails changes while  the score range changes
barplot(table(training[training$Score <= 50 & training$Score >40,]$Grade))
#ratio is peculiar, almost the same amount of passes and fails
barplot(table(training[training$Score < 40 & training$Score >30,]$Grade))

barplot(table(training[training$Score < 30 & training$Score >0,]$Grade))

barplot(table(training[training$Score <= 49 & training$Score > 47.5,]$Grade))

barplot(table(training[training$Score < 50 & training$Score >45,]$Grade))


barplot(table(training[training$Score < 55 & training$Score >50,]$Grade))
```

```r
#comparison of passed groups and failed groups based on score range
summary(training[training$Score < 55 & training$Score >= 47.5 & training$Grade == "Pass",])
summary(training[training$Score < 55 & training$Score >= 47.5 & training$Grade == "Fail",])

summary(training[training$Score < 47.5 & training$Score >= 40 & training$Grade == "Pass",])
summary(training[training$Score < 47.5 & training$Score >= 40 & training$Grade == "Fail",])


summary(training[training$Score <40 &  training$Grade == "Pass",])
summary(training[training$Score > 70 &  training$Grade == "Fail",])
```

- Tried to check random intervals to see how the distribution of grades change
  - Thought it might be useful in assigning the grades to specific score intervals before I do anything
  - Didn't get to use it at the end, though
- Checked and compared the summaries of opposing groups as well
  - It gave me a good picture of what differentiates the two groups in a specific score interval

Attendance: is it important?

Very hard to tell

```
summary(training[training$Grade == "Fail",])
#mean attendance is higher than mean score
summary(training[training$Grade == "Pass",])
#mean attendance is lower than mean score
```

```
> summary(training[training$Grade == "Fail",])
   Studentid       Attendance              Major       Questions       Score          Seniority     Texting      Grade
 Min.   :29999   Min.   :  0.00   Communication: 901   Always:1654   Min.   :  0.00   Freshman : 898   Always:1888   Fail:3819
 1st Qu.:32380   1st Qu.: 24.00   Cs           :1330   Rarely:2165   1st Qu.: 15.00   Junior   : 910   Rarely:1931   Pass:   0
 Median :34715   Median : 47.00   Polsci       : 790                 Median : 28.00   Senior   :1065
 Mean   :34731   Mean   : 48.36   Stat         : 798                 Mean   : 30.79   Sophomore: 946
 3rd Qu.:37132   3rd Qu.: 72.00                                      3rd Qu.: 43.00
 Max.   :39461   Max.   :100.00                                      Max.   :100.00
> #mean attendance is higher than mean score
> summary(training[training$Grade == "Pass",])
   Studentid       Attendance              Major       Questions       Score          Seniority     Texting      Grade
 Min.   :29998   Min.   :  0.00   Communication:1395   Always:3110   Min.   :  0.00   Freshman :1455   Always:2837   Fail:   0
 1st Qu.:32354   1st Qu.: 25.00   Cs           :1070   Rarely:2535   1st Qu.: 48.00   Junior   :1421   Rarely:2808   Pass:5645
 Median :34736   Median : 51.00   Polsci       :1603                 Median : 68.00   Senior   :1358
 Mean   :34728   Mean   : 51.07   Stat         :1577                 Mean   : 63.32   Sophomore:1411
 3rd Qu.:37062   3rd Qu.: 77.00                                      3rd Qu.: 84.00
 Max.   :39460   Max.   :100.00                                      Max.   :100.00
>
```

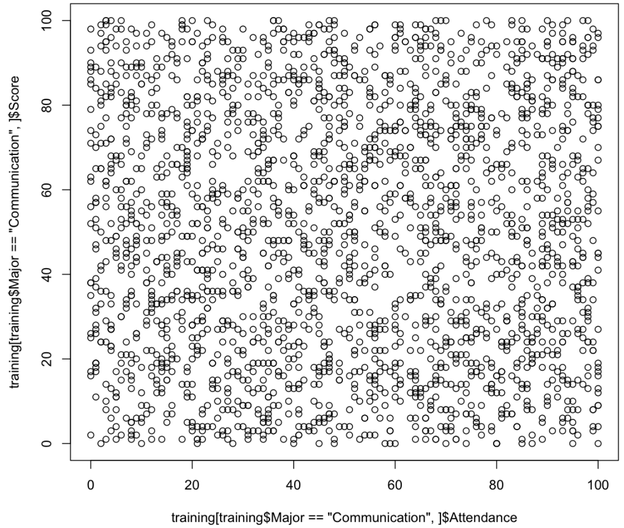
- Observed that there indeed existed some differences in terms of the attendance between the fail and pass groups
- However, the differences were very insignificant, at least to my eyes
  - Speculated that there might exist some formula that uses the attendance to make it contribute to the final grade of students, but I gave up on trying to figure it out
    - The attendance information on most subsets seemed to have no impact on the grade column of the data set

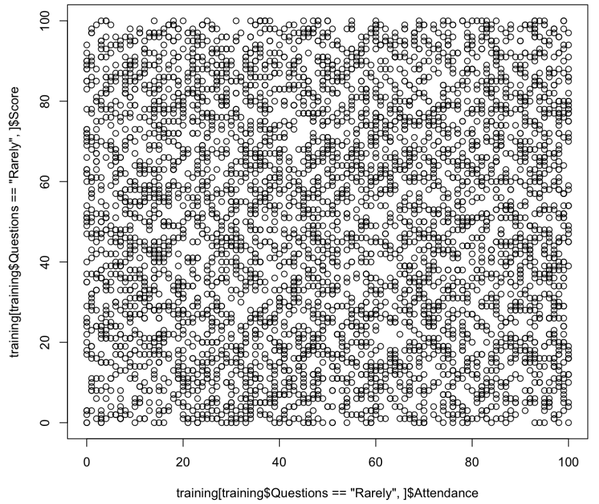
Attendance Continued

Kept on getting useless plots! So I gave up on this attribute

```
plot(training[training$Seniority == "Freshman",]$Score ~ training[training$Seniority == "Freshman",]$Attendance)
plot(training[training$Seniority == "Junior",]$Score ~ training[training$Seniority == "Junior",]$Attendance)
plot(training[training$Seniority == "Senior",]$Score ~ training[training$Seniority == "Senior",]$Attendance)
plot(training[training$Seniority == "Sophomore",]$Score ~ training[training$Seniority == "Sophomore",]$Attendance)

plot(training[training$Major == "Cs",]$Score ~ training[training$Major == "Cs",]$Attendance)
plot(training[training$Major == "Stat",]$Score ~ training[training$Major == "Stat",]$Attendance)
plot(training[training$Major == "Polsci",]$Score ~ training[training$Major == "Polsci",]$Attendance)
plot(training[training$Major == "Communication",]$Score ~ training[training$Major == "Communication",]$Attendance)

plot(training[training$Questions == "Always",]$Score ~ training[training$Questions == "Always",]$Attendance)
plot(training[training$Questions == "Rarely",]$Score ~ training[training$Questions == "Rarely",]$Attendance)

plot(training[training$Texting== "Always",]$Score ~ training[training$Texting == "Always",]$Attendance)
plot(training[training$Texting == "Always",]$Score ~ training[training$Texting == "Always",]$Attendance)
```

>plot(training[training$Major == "Communication",]$Score ~ training[training$Major == "Communication",]$Attendance)





plot(training[training$Questions == "Rarely",]$Score ~ training[training$Questions == "Rarely",]$Attendance)

More summaries, histograms, and subsetting...

Until I get something that looks useful!

```
summary(training[training$Score > 70 & training$Grade == "Fail" & training$Major == "Stat",])
summary(training[training$Score > 70 & training$Grade == "Pass" & training$Major == "Stat",])

summary(training[training$Score > 70 & training$Grade == "Pass" & training$Major == "Communication",])
summary(training[training$Score > 70 & training$Grade == "Fail" & training$Major == "Communication",])


summary(training[training$Score > 70 & training$Grade == "Pass",])
summary(training[training$Score > 70 & training$Grade == "Fail",])


summary(training[training$Major == "Stat" & training$Grade == "Pass",])
summary(training[training$Major == "Cs" & training$Grade == "Pass",])
summary(training[training$Major == "Communication" & training$Grade == "Pass",])
summary(training[training$Major == "Polsci" & training$Grade == "Pass",])

hist(training[training$Major == "Cs" & training$Grade == "Pass",]$Score)
hist(training[training$Major == "Cs" & training$Grade == "Fail",]$Score)

hist(training[training$Major == "Stat" & training$Grade == "Pass",]$Score)
hist(training[training$Major == "Stat" & training$Grade == "Fail",]$Score)

hist(training[training$Major == "Polsci" & training$Grade == "Pass",]$Score)
hist(training[training$Major == "Polsci" & training$Grade == "Fail",]$Score)

hist(training[training$Major == "Communication" & training$Grade == "Pass",]$Score)
hist(training[training$Major == "Communication" & training$Grade == "Fail",]$Score)
```

```
#
summary(training[training$Score > 80 & training$Score <=90 & training$Grade == "Pass",])
summary(training[training$Score > 80 & training$Score <=90 & training$Grade == "Fail",])

barplot(table(training[training$Score > 80 & training$Score <=90 & training$Seniority != "Junior" & training$Questions == "Rarely",]$Grade))

mosaicplot(training[training$Score > 80 & training$Score <=90 ,]$Questions~training[training$Score > 80 & training$Score <=90,]$Grade)
mosaicplot(training[training$Score > 80 & training$Score <=90 ,]$Texting~training[training$Score > 80 & training$Score <=90,]$Grade)
mosaicplot(training[training$Score > 80 & training$Score <=90 ,]$Major~training[training$Score > 80 & training$Score <=90,]$Grade)
mosaicplot(training[training$Score > 80 & training$Score <=90 ,]$Seniority~training[training$Score > 80 & training$Score <=90,]$Grade)

mean(training$Score)
```

Attempted to get some information for the higher score intervals, but no specific combination gave a large proportion of fails

Score ranges used: score > 70, score >80 & score <= 90

Subsetting for the Lower Score Intervals

Score ranges used: below 40, or below 47.5, etc.

```
summary(training[training$Score <40 & training$Grade == "Pass",])
summary(training[training$Score <40 & training$Grade == "Fail",])

summary(training[training$Score >=70 & training$Grade == "Pass",])
#Always questioning helps in the socre range of < 40
#not being Cs or communication helps

summary(training[training$Score < 40,])
```

```
barplot(table(training[training$Score <40 &training$Score >= 20 & training$Seniority == "Sophomore"&training$Questions == "Always" & training$Major != "Cs" & training$Major != "Commur
#maybe interesting?
```

```
summary(training[training$Score <40,])
summary(training[training$Score <40 & training$Major == "Polsci" & training$Questions == "Always",])
barplot(table(training[training$Score <40 & training$Major == "Polsci" & training$Questions == "Always",]$Grade))
#interesting
```

```
#addition 10
barplot(table(training[training$Score <40 & training$Major == "Polsci" & training$Questions == "Always" & training$Seniority != "Senior" & training$Seniority != "Junior",]$Grade))
#mostly passing!!
```

```
summary(training[training$Major == "Communication",])
summary(training[training$Major == "Stat",])
summary(training[training$Major == "Polsci",])
```

- Added small exceptions(rules) based on these findings
  - Having a significantly high proportion of passes in the score range where most people fail, having half and half where most people fail, etc
  - Tried subsetting purely based on the majors as well, but did not end up with any interesting results

Almost Near the END!

INTRODUCING the CRAZIEST COMBINATIONS FOUND!

- Below is the list of the weirdest/craziest combinations of attributes that were very impactful towards the final grade of students:

- #**ALL FAILS**
- decision[myprediction$Score < 70 & myprediction$Major == "Cs" & myprediction$Seniority == "Senior"] <- "Fail"
- decision[myprediction$Score <50 & myprediction$Score >= 0 & myprediction$Major == "Polsci" & myprediction$Seniority != "Senior" & myprediction$Questions == "Rarely"] <- "Fail"
- decision[myprediction$Score > 0 & myprediction$Score < 50 & myprediction$Major == "Polsci" & myprediction$Questions == "Rarely"] <- "Fail"
- #**ALL PASSES**
- decision[myprediction$Score > 70 & myprediction$Major == "Cs"] <- "Pass"
- decision[myprediction$Score <70 & myprediction$Score >= 50 & myprediction$Major == "Polsci" & myprediction$Seniority == "Freshman" & myprediction$Questions == "Rarely" & myprediction$Texting == "Always"] <- "Pass"
- decision[myprediction$Score > 40 & myprediction$Score < 50 & myprediction$Major == "Communication" & myprediction$Seniority == "Sophomore"] <- "Pass"

- The above combinations either results in all students failing or all of them passing!

Efforts that did not end up well

Digging into the Statistics major

- Since I got weird combinations with the majors Cs, Polsci, and Communication, I thought statistics would also have some interesting correlation with other variables
- However, no matter what I tried, I couldn't find anything that was abnormal or crazy
- Definitely possible that I simply missed the juicy parts and only encountered the boring combinations for Statistics major

```
barplot(table(training[training$Score < 47.5 & training$Score >= 0 & training$Major == "Stat" ,]$Grade))

summary(training[ training$Score >60 & training$Major == "Stat" & training$Grade == "Pass",])
summary(training[training$Score >60 &training$Major == "Stat" & training$Grade == "Fail",])

barplot(table(training[training$Score > 55 & training$Score <70 & training$Major == "Stat" & training$Texting == "Always",]$Grade))


summary(training[training$Score > 55 & training$Score <70 & training$Major == "Stat" & training$Grade == "Pass",])
summary(training[training$Score > 55 & training$Score <70 & training$Major == "Stat" & training$Grade == "Fail",])

#very hard to find any correlation between major Stat and other attributes!!
```

Going back to: R Code of the finalized model

```
#FINAL MODEL
#permuting the order of the rows in the data set
randomSubset <- training[sample(1:nrow(training)),]
#taking the first 500 rows from the result of the permutation as the subset to test my model on
testingSubset <- randomSubset[1:1000,]

myprediction <- testingSubset
decision <- rep("Fail",nrow(myprediction))
decision[myprediction$Score < 40] <- "Fail"
decision[myprediction$Score > 70] <- "Pass"
decision[myprediction$Score > 60 & myprediction$Major == "Stat"] <- "Pass"
decision[myprediction$Score > 50 & myprediction$Major == "Polsci"] <- "Pass"
decision[myprediction$Score >=47.5 & myprediction$Score <55] <- "Pass"
decision[myprediction$Score < 40 & myprediction$Major != "Cs" & myprediction$Major != "Communication" & myprediction$Seniority != "Senior" & myprediction$Seniority != "Freshman" & myprediction$Questions == "Always"] <- "Pass"
decision[myprediction$Score >=40 & myprediction$Score <47.5 & myprediction$Major == "Communication" & myprediction$Texting == "Always"] <- "Pass"
decision[myprediction$Score >=40 & myprediction$Score <47.5 & myprediction$Major == "Communication" & myprediction$Questions== "Always"] <- "Pass"
decision[myprediction$Score >=40 & myprediction$Score <47.5 & myprediction$Major != "Cs" & myprediction$Questions == "Always" & myprediction$Seniority != "Senior"] <- "Pass"
decision[myprediction$Score >= 55 & myprediction$Score <= 70] <- "Pass"
decision[myprediction$Score >= 55 & myprediction$Score <= 70 & myprediction$Major == "Cs" & (myprediction$Seniority == "Junior" | myprediction$Seniority == "Senior")] <- "Fail"
decision[myprediction$Score >=47.5 & myprediction$Score <55 & (myprediction$Major == "Cs" |myprediction$Major == "Stat") & myprediction$Seniority == "Senior"] <- "Fail"
decision[myprediction$Score <50 & myprediction$Score >= 0 & myprediction$Major == "Cs" & myprediction$Seniority != "Junior"  & myprediction$Questions == "Rarely"] <- "Fail"
decision[myprediction$Score <40 & myprediction$Major == "Polsci" & myprediction$Questions == "Always" & myprediction$Seniority != "Senior" & myprediction$Seniority != "Junior"] <- "Pass"
decision[myprediction$Score <40 & myprediction$Score >= 0 & myprediction$Major != "Stat" & myprediction$Major != "Polsci" & myprediction$Major != "Cs" & (myprediction$Seniority == "Senior"|myprediction$Seniority=="Freshman")&
myprediction$Texting=="Rarely"] <- "Fail"
decision[myprediction$Score <45 & myprediction$Score >= 40 & myprediction$Major != "Stat" & myprediction$Major != "Polsci" & myprediction$Major != "Cs" & (myprediction$Seniority == "Senior"|myprediction$Seniority=="Freshman")&
myprediction$Texting=="Rarely"] <- "Pass"
decision[myprediction$Score >= 47.5 & myprediction$Score <55 & myprediction$Seniority=="Senior" & (myprediction$Major == "Cs" | myprediction$Major == "Stat")& myprediction$Questions == "Always"] <- "Fail"
#ALL FAILS
decision[myprediction$Score < 70 & myprediction$Major == "Cs" & myprediction$Seniority == "Senior"] <- "Fail"
decision[myprediction$Score <50 & myprediction$Score >= 0 & myprediction$Major == "Polsci" & myprediction$Seniority != "Senior"  & myprediction$Questions == "Rarely"] <- "Fail"
decision[myprediction$Score > 0 & myprediction$Score < 50 & myprediction$Major == "Polsci" & myprediction$Questions == "Rarely" ] <- "Fail"
#ALL PASSES
decision[myprediction$Score > 70 & myprediction$Major == "Cs"] <- "Pass"
decision[myprediction$Score <70 & myprediction$Score >= 50 & myprediction$Major == "Polsci" & myprediction$Seniority == "Freshman"  & myprediction$Questions == "Rarely" & myprediction$Texting == "Always"] <- "Pass"
decision[myprediction$Score > 40 & myprediction$Score < 50 & myprediction$Major == "Communication" & myprediction$Seniority == "Sophomore"] <- "Pass"


myprediction$Grade <- decision


error <- mean(myprediction$Grade != testingSubset$Grade)
error
#error percentage for the random subsets tested: 0.163, 0.158,  0.145, 0.135, 0.154
```

#WOW!! MUCH BETTER RESULTS COMPARED TO THE PREVIOUS MODELS!! Average error: 0.151 ! Very stable as well

AT THIS POINT, I decided to use this prediction model as my final one and used it to predict the grades for the prediction challenge.

Conclusion

The results from KAGGLE

2	Seok Yim	0.84769	1	~10s

- Fortunately, the error percentage was around 15.231% for the actual testing data!
- This isn't so bad, but what I will get for the entire data set is yet unknown
- However, I'm still happy with this result
    - It seems that my hard work(probably more than 12+ hours) paid off :)

Thank you!
(And sorry if the R code looks messy)