ASSIGNMENT 11: PREDICTION CHALLENGE 3
(NUMERICAL VARIABLE)

DATA 101

Seek Yim

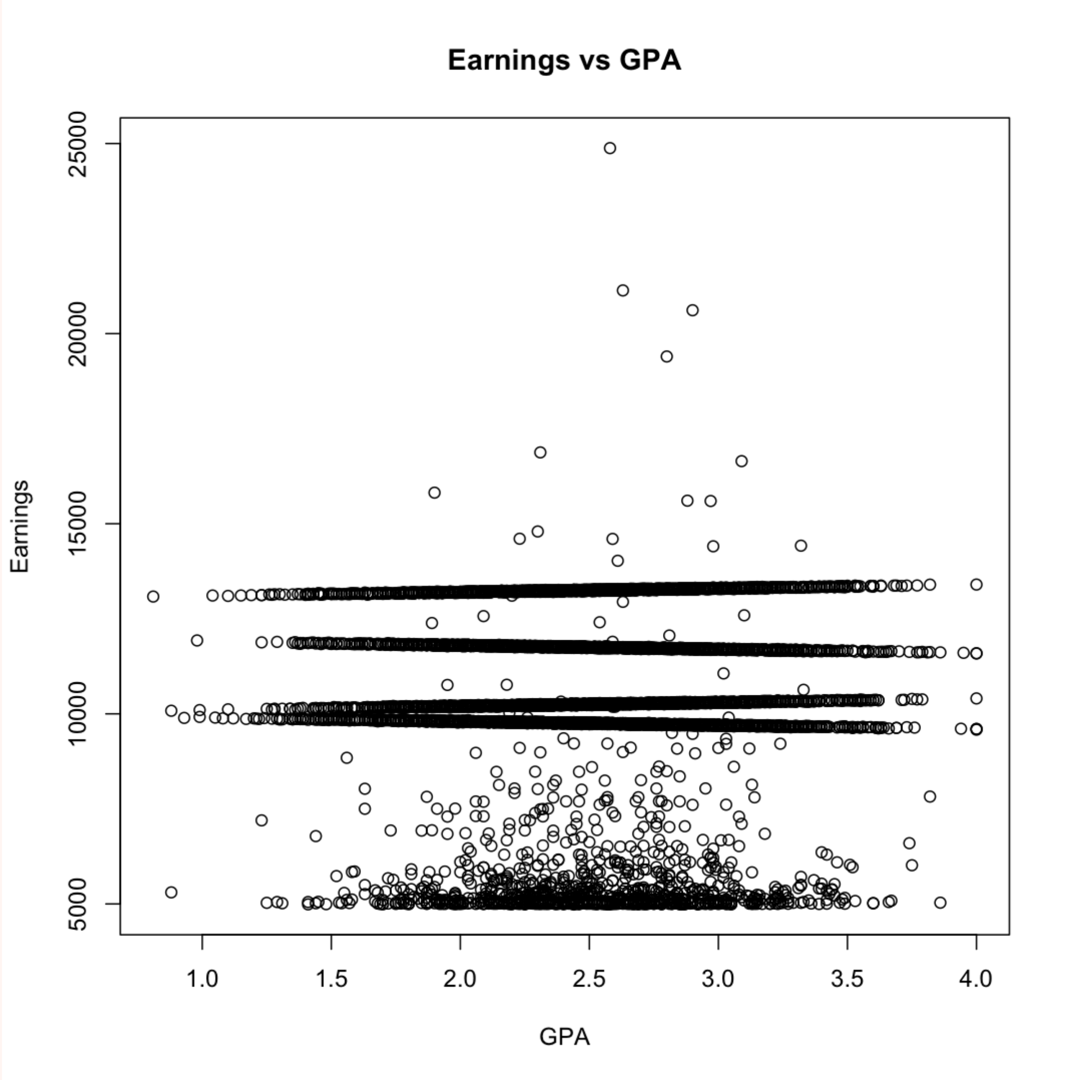
APRIL 23, 2021

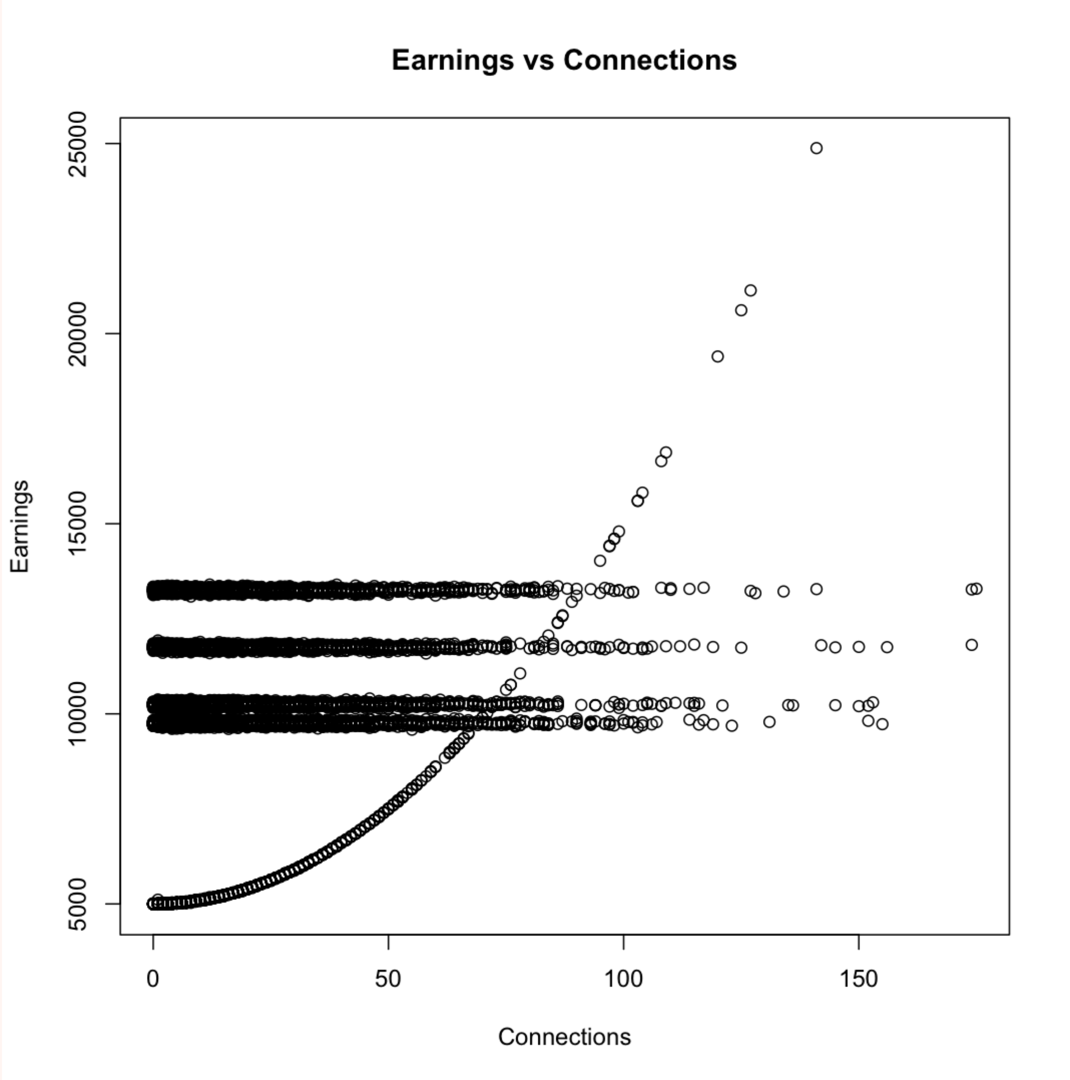
LOOKING AT THE DATA

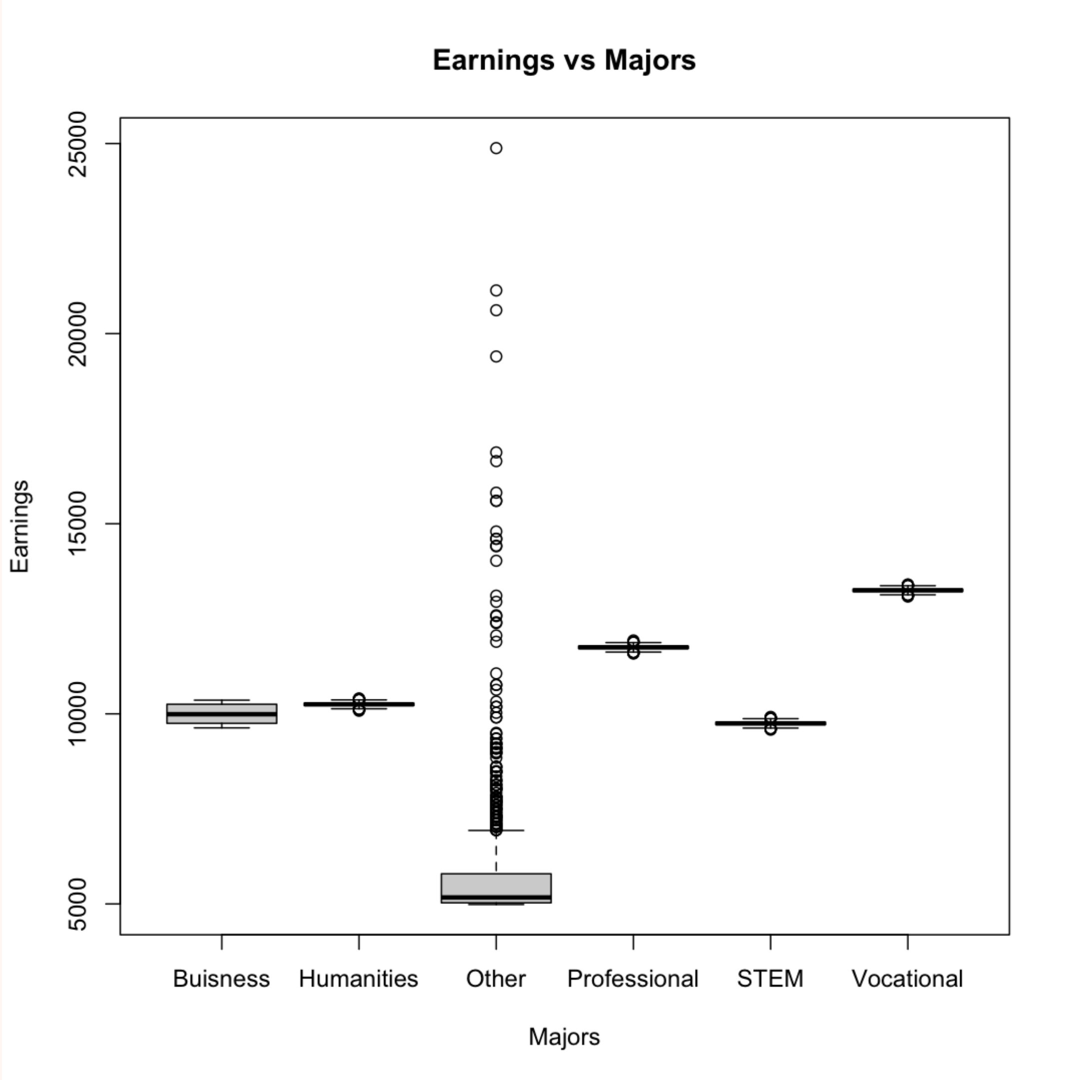
	GPA	Number_Of_Professional_Connections	Earnings	Major	Graduation_Year	Height	Number_Of_Credits	Number_Of_Parking_Tickets
1	2.50	1	9756.15	STEM	2001	64.22	124	1
2	2.98	1	9709.03	STEM	2001	69.55	120	0
3	2.98	23	9711.37	STEM	1996	68.98	120	1
4	3.35	5	9656.15	STEM	2008	69.23	124	1
5	2.47	37	9751.92	STEM	1981	70.45	123	0
6	2.75	2	9728.30	STEM	2000	65.26	121	0
7	1.66	17	9847.59	STEM	2001	65.91	121	0
8	2.59	10	9743.36	STEM	1990	66.35	123	0
9	1.89	7	9793.38	STEM	1975	70.42	121	1
10	1.89	22	9810.38	STEM	1997	65.18	122	0
11	2.80	39	9714.16	STEM	1972	71.19	120	1
12	2.29	16	9788.13	STEM	1968	68.97	120	0
13	2.39	4	9754.40	STEM	1999	70.01	122	0
14	3.61	4	9632.89	STEM	1999	67.55	125	1
15	2.68	58	9723.80	STEM	1971	66.61	127	0

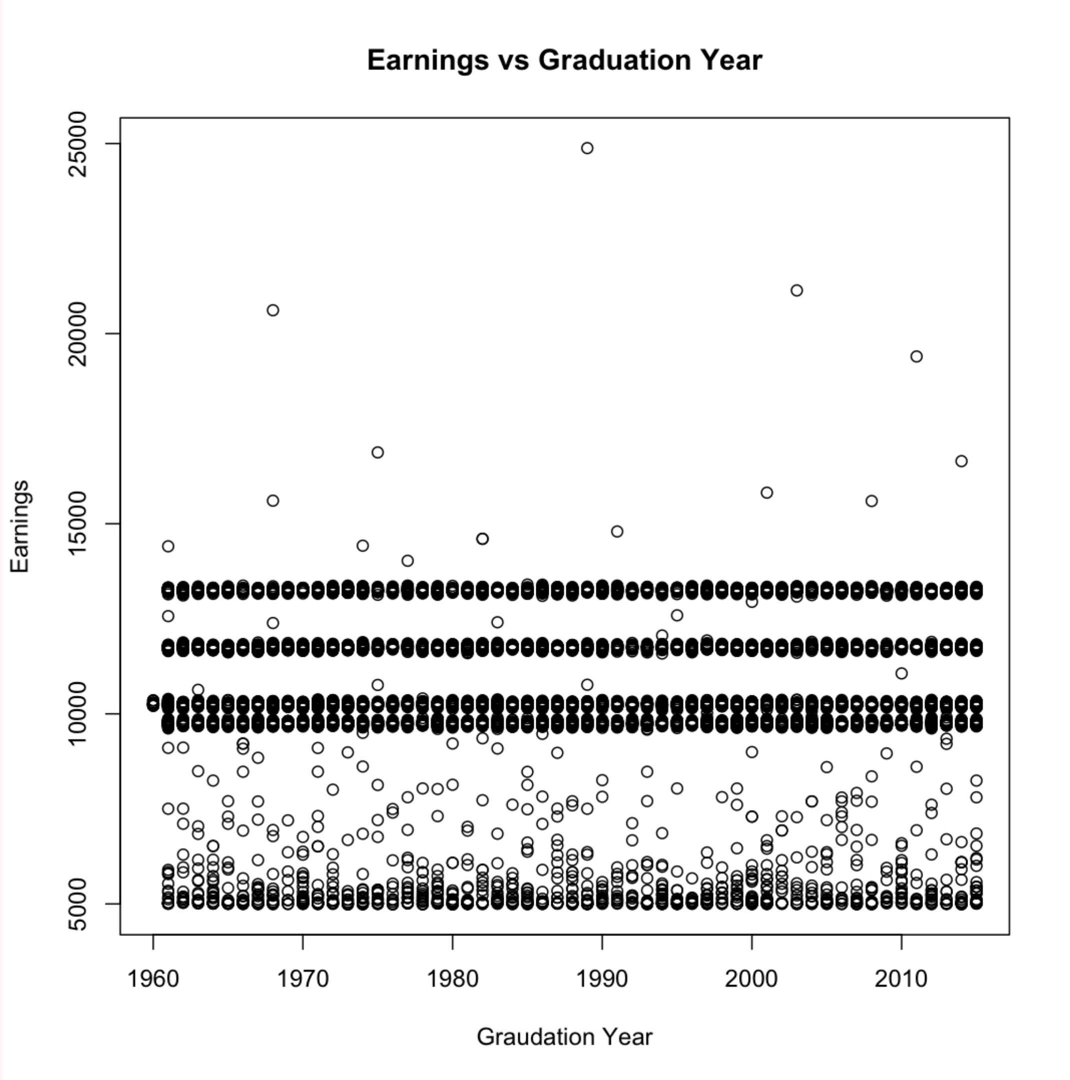
- **The attribute that we are predicting for the testing data set is Earnings**
  - **Dependent Variable: Earnings**
  - **Independent Variable(s): the ones we select out of all the other attributes**

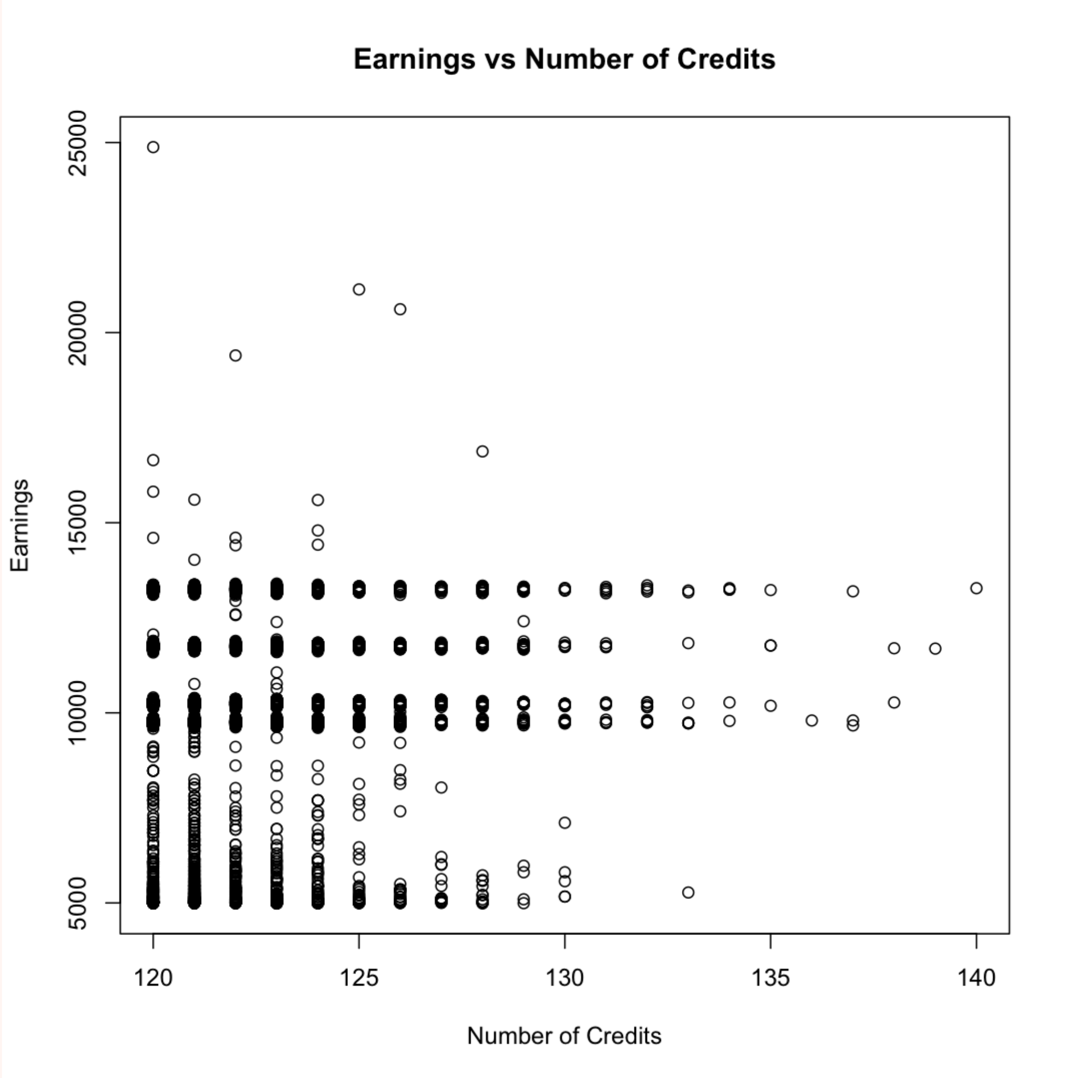
LOOKING AT THE DATA SET (ENTIRE DATA SET)

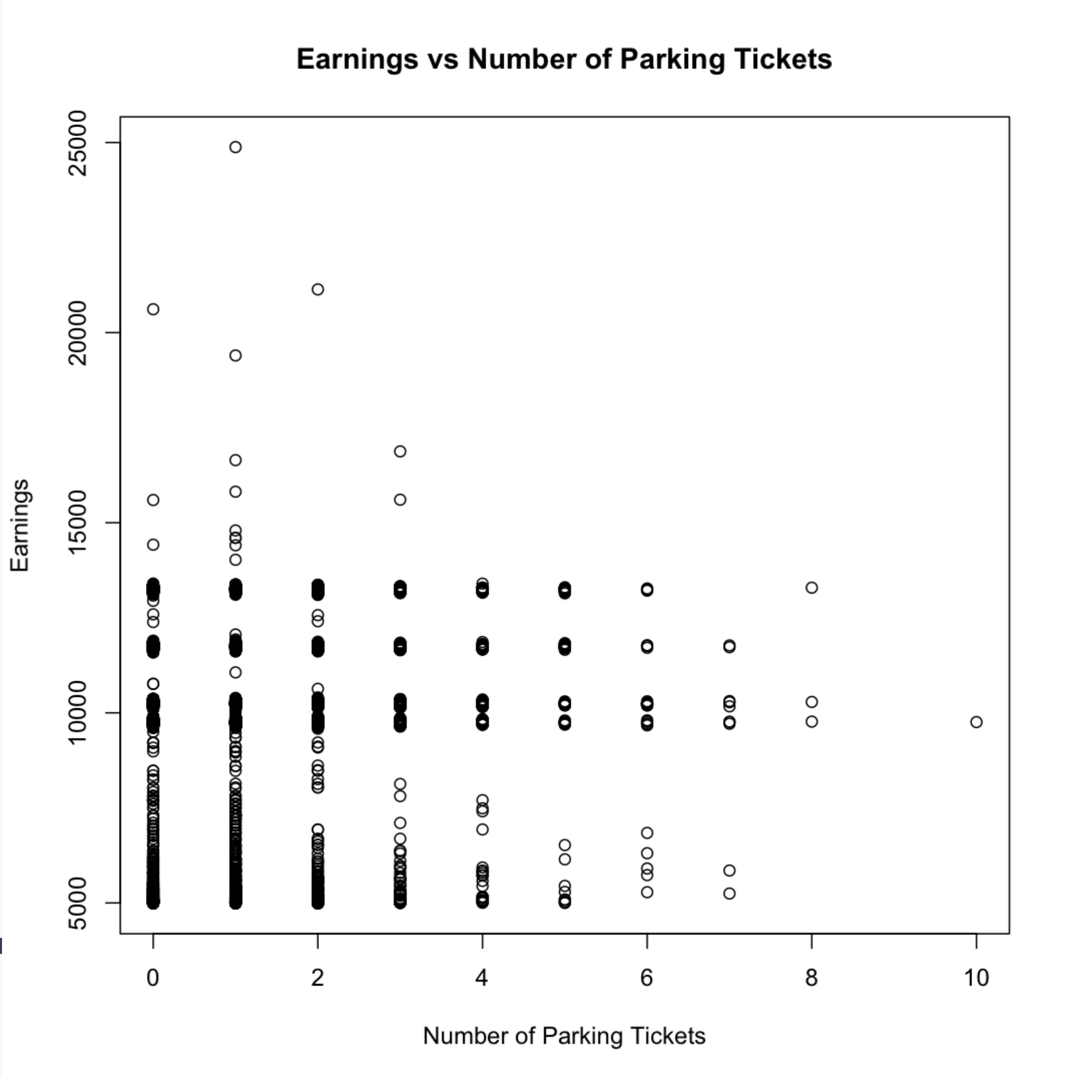


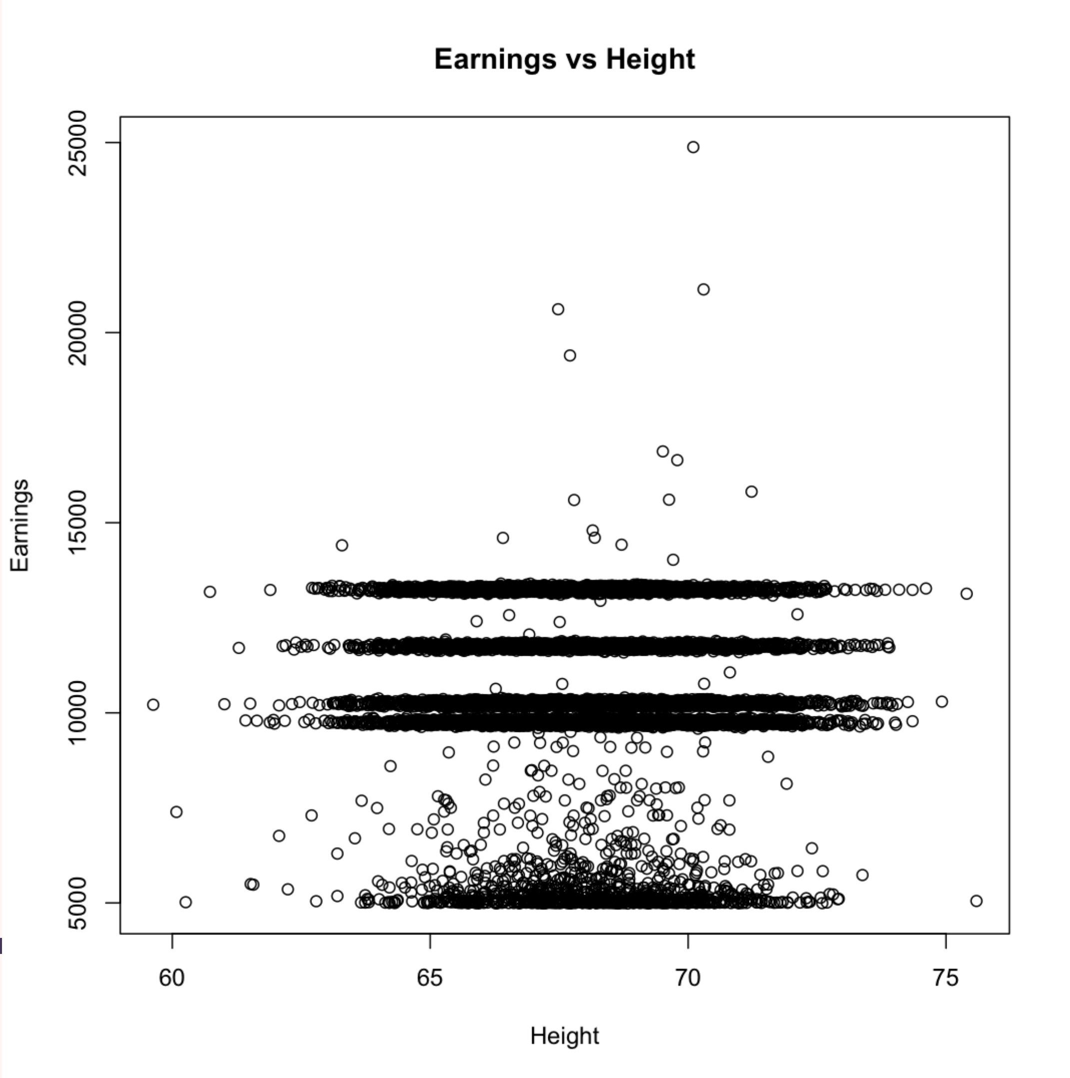












- **It seems like Major is the most significant attribute in determining the Earnings**
  - **Clear difference in Earnings between Majors**
- **GPA and Connections seem to have a pattern for specific subsets of the entire data set**
  - **Subsetting the data set might help understanding underlying correlations**
- **All the other attributes do not seem interesting**

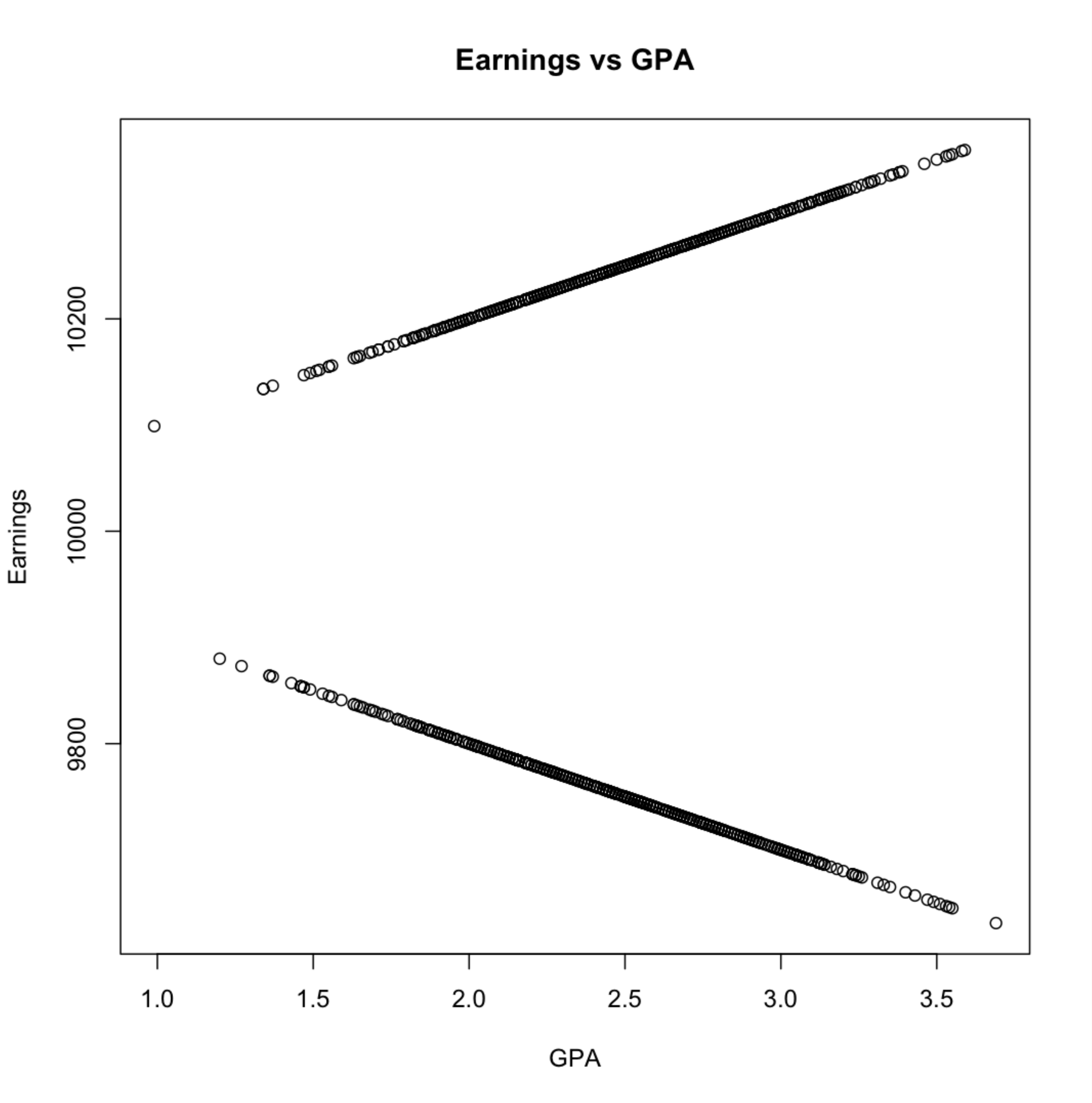
SUBSETTING BASED ON MAJORS

- I sliced the entire data set into smaller subsets based on Major
  - 6 total subsets created
  - Observed plots for each subset
    - The interesting plots will be listed in the following slides
- GENERAL APPROACH: Carrying out predictions on these subsets based on majors and later combining all the predictions

```
#subsetting the data set based on major
Earnings.Business <- Earnings[Earnings$Major == "Buisness",]
Earnings.Humanintíes <- Earnings[Earnings$Major == "Humanities",]
Earnings.Professional <- Earnings[Earnings$Major == "Professional",]
Earnings.STEM <- Earnings[Earnings$Major == "STEM",]
Earnings.Vocational <- Earnings[Earnings$Major == "Vocational",]
Earnings.Other <- Earnings[Earnings$Major == "Other",]
```

LOOKING AT THE DATA SET(SUBSETS)

BUSINESS MAJORS



plot(Earnings.Business$Earnings~Earnings.Business$GPA, main = "Earnings vs GPA", xlab = "GPA", ylab = "Earnings")

- The pattern is much clearer than before
  - The Earnings of Business Majors seem to have a correlation with GPA
    - When Earnings > 10000, Earnings increases as GPA increases, whereas when Earnings < = 10000, Earnings decreases as GPA increases
    - The summary of the two groups did not give any interesting or groundbreaking information, so I decided to dig a big more into this major

```
> Earnings.Business.Higher <- Earnings.Business[Earnings.Business$Earnings > 10000,]
> Earnings.Business.Lower <- Earnings.Business[Earnings.Business$Earnings <= 10000,]
> summary(Earnings.Business.Higher)
      GPA        Number_Of_Professional_Connections    Earnings          Major           Graduation_Year     Height      Number_Of_Credits Number_Of_Parking_Tickets
 Min.   :0.990   Min.   :  0.00                     Min.   :10099   Length:500         Min.   :1960    Min.   :63.11   Min.   :120       Min.   :0.000            
 1st Qu.:2.210   1st Qu.:  6.00                     1st Qu.:10221   Class :character   1st Qu.:1972    1st Qu.:66.70   1st Qu.:121       1st Qu.:0.000            
 Median :2.520   Median : 14.00                     Median :10252   Mode  :character   Median :1986    Median :68.03   Median :121       Median :1.000            
 Mean   :2.508   Mean   : 21.74                     Mean   :10251                      Mean   :1987    Mean   :68.07   Mean   :122       Mean   :0.988            
 3rd Qu.:2.803   3rd Qu.: 30.00                     3rd Qu.:10280                      3rd Qu.:2000    3rd Qu.:69.46   3rd Qu.:123       3rd Qu.:1.000            
 Max.   :3.590   Max.   :150.00                     Max.   :10359                      Max.   :2014    Max.   :74.92   Max.   :138       Max.   :7.000            
   Competence   
 Min.   :    0  
 1st Qu.:  726  
 Median : 1743  
 Mean   : 2656  
 3rd Qu.: 3632  
 Max.   :18150  
> summary(Earnings.Business.Lower)
      GPA        Number_Of_Professional_Connections    Earnings        Major           Graduation_Year     Height      Number_Of_Credits Number_Of_Parking_Tickets
 Min.   :1.200   Min.   :  0.00                     Min.   :9631   Length:500         Min.   :1961    Min.   :61.42   Min.   :120       Min.   :0.000            
 1st Qu.:2.160   1st Qu.:  5.75                     1st Qu.:9724   Class :character   1st Qu.:1973    1st Qu.:66.93   1st Qu.:121       1st Qu.:0.000            
 Median :2.500   Median : 12.00                     Median :9750   Mode  :character   Median :1987    Median :68.08   Median :121       Median :1.000            
 Mean   :2.466   Mean   : 19.18                     Mean   :9753                      Mean   :1988    Mean   :68.06   Mean   :122       Mean   :0.938            
 3rd Qu.:2.760   3rd Qu.: 26.25                     3rd Qu.:9784                      3rd Qu.:2001    3rd Qu.:69.15   3rd Qu.:123       3rd Qu.:1.000            
 Max.   :3.690   Max.   :123.00                     Max.   :9880                      Max.   :2015    Max.   :74.35   Max.   :136       Max.   :6.000            
   Competence   
 Min.   :    0  
 1st Qu.:  705  
 Median : 1452  
 Mean   : 2342  
 3rd Qu.: 3267  
 Max.   :14760  
```

BUSINESS MAJORS: TWO CONFLICTING PATTERNS

```
#trying to understand the two conflicting patterns for Buisness majors
Earnings.Business$PASS <- '0'
Earnings.Business[Earnings.Business$Earnings > 10000,]$PASS <- '1'
Earnings.Business[Earnings.Business$Earnings <= 10000,]$PASS <- '0'

Earnings.Business

r_model <- rpart(PASS~GPA+Height+Number_Of_Professional_Connections+Graduation_Year+Number_Of_Parking_Tickets+Number_Of_Credits,data = Earnings.Business)
r_model

predicted <- predict(r_model,newdata = Earnings.Business,type = "class")
mean(Earnings.Business$PASS != predicted)
#PATTERN FOUND!! EVEN AND ODD GRAUDATION_YEARS ARE SIGNIFICANT!!!
#YAYAYAYAYAYAY!!! AHAHAHAHAHAHAHA FOUND IT!!!! HAHAHAHAHAJHAHAAHHAHAHA! TOLD YOU I COULD DO ITTTTT!
```

- **At first sight, the difference between the two groups (Group with Earnings > 10,000 and Group with Earnings <= 10,000) was very obscure**
  - **I tested numerous formulas, such as GPA * Number of Professional Connections or GPA * Height* Number of Parking Tickets**
    - **None of them gave fruitful insight**
- **After trying for hours, I decided to use rPart just to figure out the differences between the two groups by assigning the group with higher Earnings a value of '1' for the attribute PASS, and the other group a value of '0' (PASS is a newly created attribute)**

PATTERN BEHIND BUSINESS MAJORS: GRADUATION YEARS

```
> r_model <- rpart(PASS~GPA+Height+Number_Of_Professional_Connections+Graduation_Year+Number_Of_Parking_Tickets+Number_Of_Credits,data = Earnings.Business)
> r_model
n= 1000

node), split, n, loss, yval, (yprob)
      * denotes terminal node

        1) root 1000 500 0 (0.5000000 0.5000000)
          2) Graduation_Year>=1960.5 982 482 0 (0.5091650 0.4908350)
            4) Graduation_Year< 1961.5 18   0 0 (1.0000000 0.0000000) *
            5) Graduation_Year>=1961.5 964 482 0 (0.5000000 0.5000000)
             10) Graduation_Year>=1962.5 946 464 0 (0.5095137 0.4904863)
               20) Graduation_Year< 1963.5 18   0 0 (1.0000000 0.0000000) *
               21) Graduation_Year>=1963.5 928 464 0 (0.5000000 0.5000000)
                 42) Graduation_Year>=1964.5 910 446 0 (0.5098901 0.4901099)
                   84) Graduation_Year< 1965.5 18   0 0 (1.0000000 0.0000000) *
                   85) Graduation_Year>=1965.5 892 446 0 (0.5000000 0.5000000)
                    170) Graduation_Year>=1966.5 874 428 0 (0.5102975 0.4897025)
                      340) Graduation_Year< 1967.5 18   0 0 (1.0000000 0.0000000) *
                      341) Graduation_Year>=1967.5 856 428 0 (0.5000000 0.5000000)
                        682) Graduation_Year>=1968.5 838 410 0 (0.5107399 0.4892601)
                         1364) Graduation_Year< 1969.5 18   0 0 (1.0000000 0.0000000) *
                         1365) Graduation_Year>=1969.5 820 410 0 (0.5000000 0.5000000)
                           2730) Graduation_Year>=1970.5 802 392 0 (0.5112219 0.4887781)
                             5460) Graduation_Year< 1971.5 18   0 0 (1.0000000 0.0000000) *
                             5461) Graduation_Year>=1971.5 784 392 0 (0.5000000 0.5000000)
                              10922) Graduation_Year>=1972.5 766 374 0 (0.5117493 0.4882507)
                                21844) Graduation_Year< 1973.5 18   0 0 (1.0000000 0.0000000) *
                                21845) Graduation_Year>=1973.5 748 374 0 (0.5000000 0.5000000)
                                  43690) Graduation_Year>=1974.5 730 356 0 (0.5123288 0.4876712)
                                    87380) Graduation_Year< 1975.5 18   0 0 (1.0000000 0.0000000) *
                                    87381) Graduation_Year>=1975.5 712 356 0 (0.5000000 0.5000000)
                                     174762) Graduation_Year>=1976.5 694 338 0 (0.5129683 0.4870317)
                                       349524) Graduation_Year< 1977.5 18   0 0 (1.0000000 0.0000000) *
                                       349525) Graduation_Year>=1977.5 676 338 0 (0.5000000 0.5000000)
                                         699050) Graduation_Year>=1978.5 658 320 0 (0.5136778 0.4863222)
                                          1398100) Graduation_Year< 1979.5 18   0 0 (1.0000000 0.0000000) *
                                          1398101) Graduation_Year>=1979.5 640 320 0 (0.5000000 0.5000000)
                                            2796202) Graduation_Year>=1980.5 622 302 0 (0.5144695 0.4855305)
                                              5592404) Graduation_Year< 1981.5 18   0 0 (1.0000000 0.0000000) *
                                              5592405) Graduation_Year>=1981.5 604 302 0 (0.5000000 0.5000000)
                                               11184810) Graduation_Year>=1982.5 586 284 0 (0.5153584 0.4846416)
                                                 22369620) Graduation_Year< 1983.5 18   0 0 (1.0000000 0.0000000) *
                                                 22369621) Graduation_Year>=1983.5 568 284 0 (0.5000000 0.5000000)
                                                   44739242) Graduation_Year>=1984.5 550 266 0 (0.5163636 0.4836364)
                                                     89478484) Graduation_Year< 1985.5 18   0 0 (1.0000000 0.0000000) *
                                                     89478485) Graduation_Year>=1985.5 532 266 0 (0.5000000 0.5000000)
                                                      178956970) Graduation_Year>=1986.5 514 248 0 (0.5175097 0.4824903)
                                                        357913940) Graduation_Year< 1987.5 18   0 0 (1.0000000 0.0000000) *
                                                        357913941) Graduation_Year>=1987.5 496 248 0 (0.5000000 0.5000000)
                                                          715827882) Graduation_Year>=1988.5 478 230 0 (0.5188285 0.4811715) *
                                                          715827883) Graduation_Year< 1988.5 18   0 1 (0.0000000 1.0000000) *
                                                      178956971) Graduation_Year< 1986.5 18   0 1 (0.0000000 1.0000000) *
                                                   44739243) Graduation_Year< 1984.5 18   0 1 (0.0000000 1.0000000) *
                                               11184811) Graduation_Year< 1982.5 18   0 1 (0.0000000 1.0000000) *
                                            2796203) Graduation_Year< 1980.5 18   0 1 (0.0000000 1.0000000) *
                                         699051) Graduation_Year< 1978.5 18   0 1 (0.0000000 1.0000000) *
                                     174763) Graduation_Year< 1976.5 18   0 1 (0.0000000 1.0000000) *
                                  43691) Graduation_Year< 1974.5 18   0 1 (0.0000000 1.0000000) *
                              10923) Graduation_Year< 1972.5 18   0 1 (0.0000000 1.0000000) *
                           2731) Graduation_Year< 1970.5 18   0 1 (0.0000000 1.0000000) *
                        683) Graduation_Year< 1968.5 18   0 1 (0.0000000 1.0000000) *
                    171) Graduation_Year< 1966.5 18   0 1 (0.0000000 1.0000000) *
```

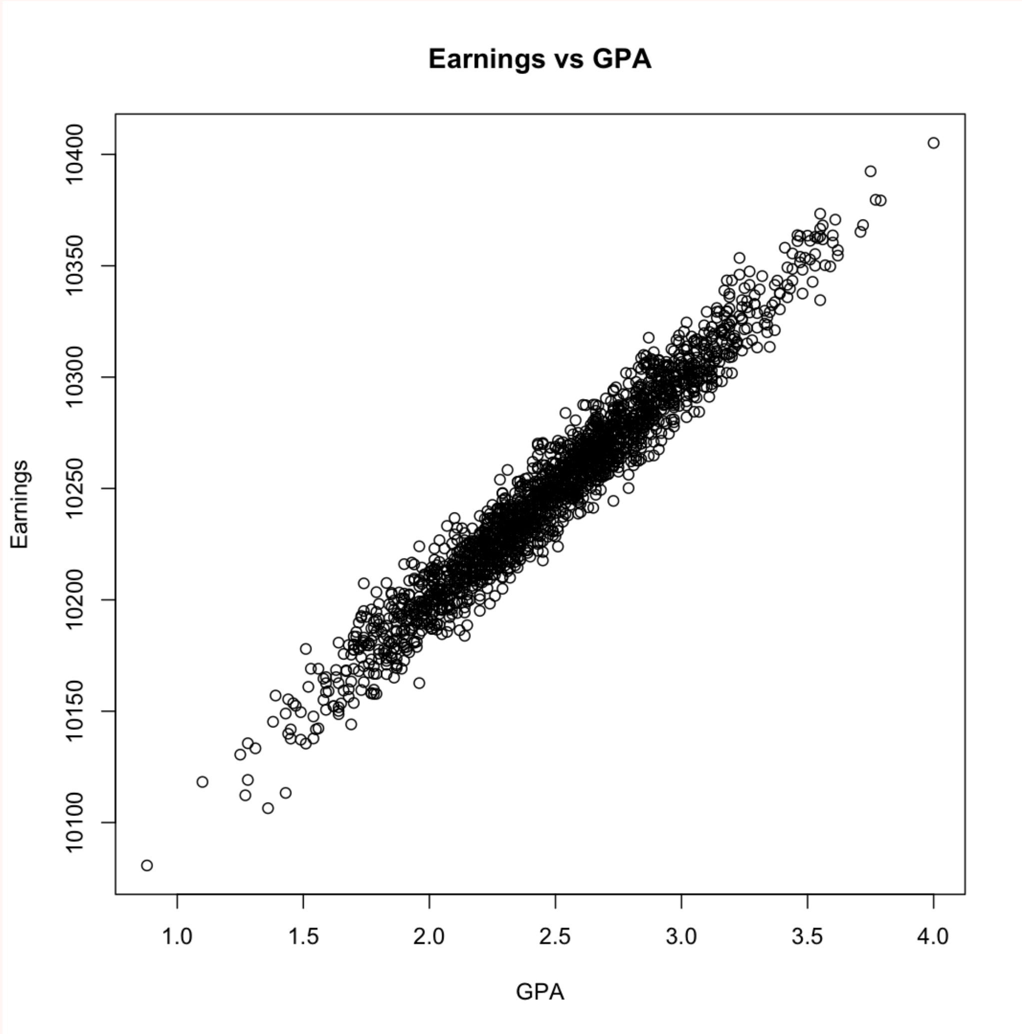
- rPart ended up giving me this **ENORMOUS TREE** that showed that basically even graduation year for business majors meant Earnings > 10,000, whereas odd meant Earnings <= 10,000
- This explained the two conflicting patterns in the Earnings VS GPA plot for Business Majors
  - In the final prediction model, I multiplied -1 to the GPA of rows with graduation year %% 2 != 0 (the odd GPAs)

For a better picture of the method, this is how I did it:

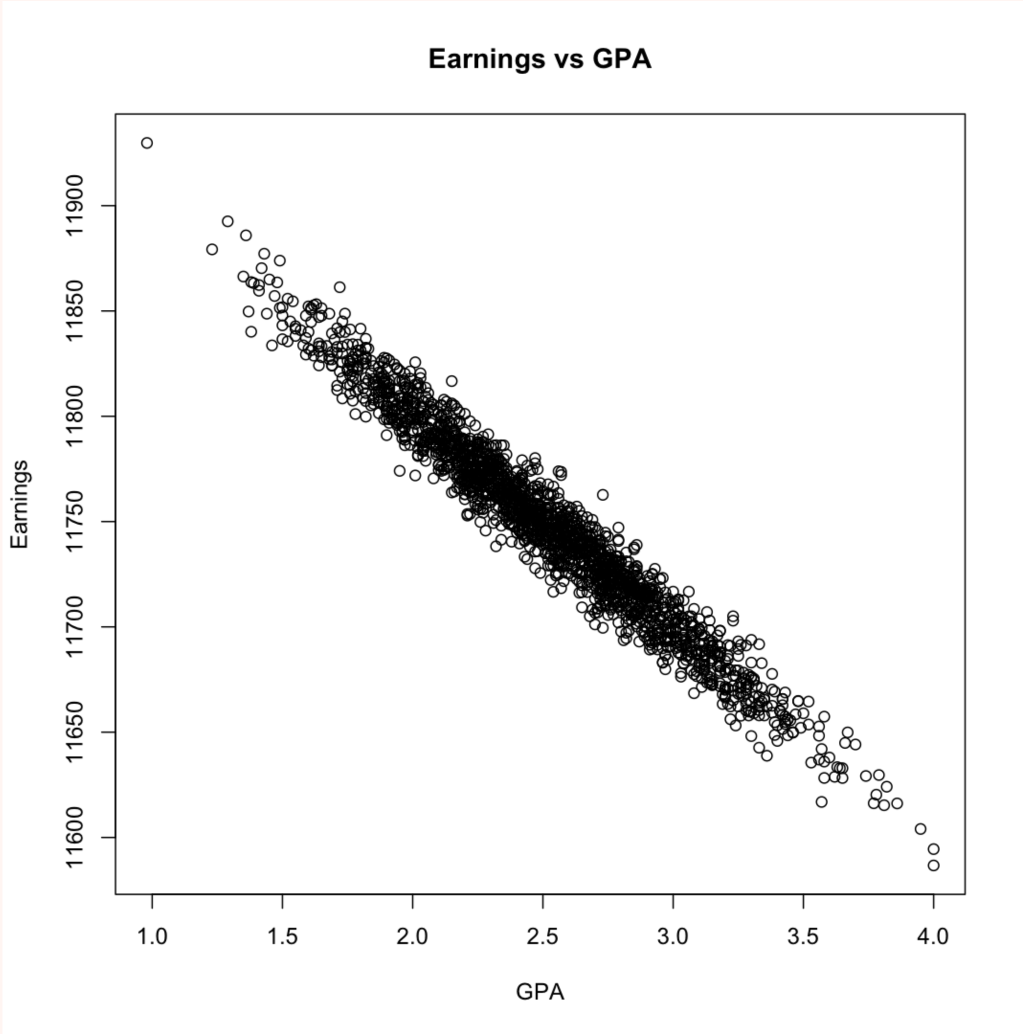
```
Earnings[Earnings$Major == "Buisness" & Earnings$Graduation_Year %% 2 != 0,]$GPA <- Earnings[Earnings$Major == "Buisness" & Earnings$Graduation_Year %% 2 != 0,]$GPA * (-1)
Earnings[Earnings$Major != "Buisness",]$GPA <- 0
```

LOOKING AT THE DATA SET(SUBSETS)

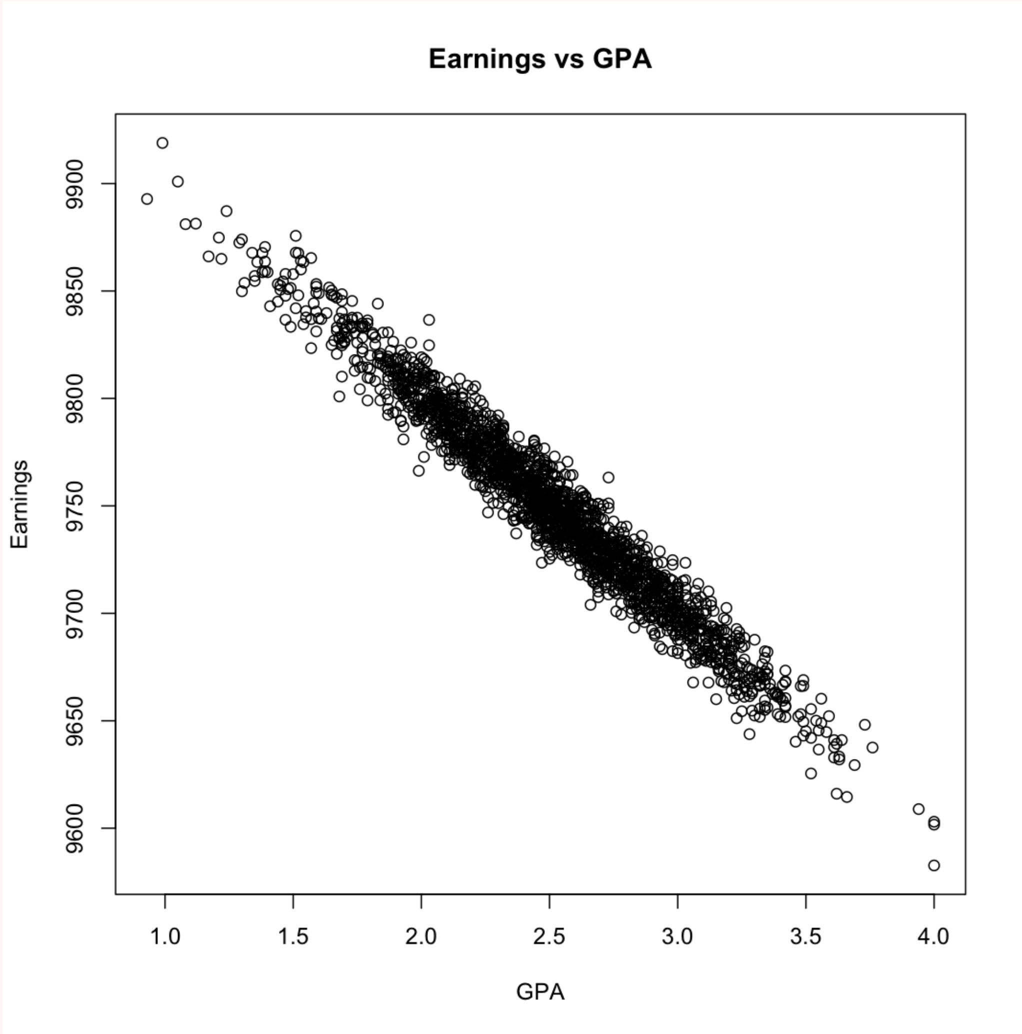
HUMANITIES, PROFESSIONAL, STEM, AND VOCATIONAL MAJORS



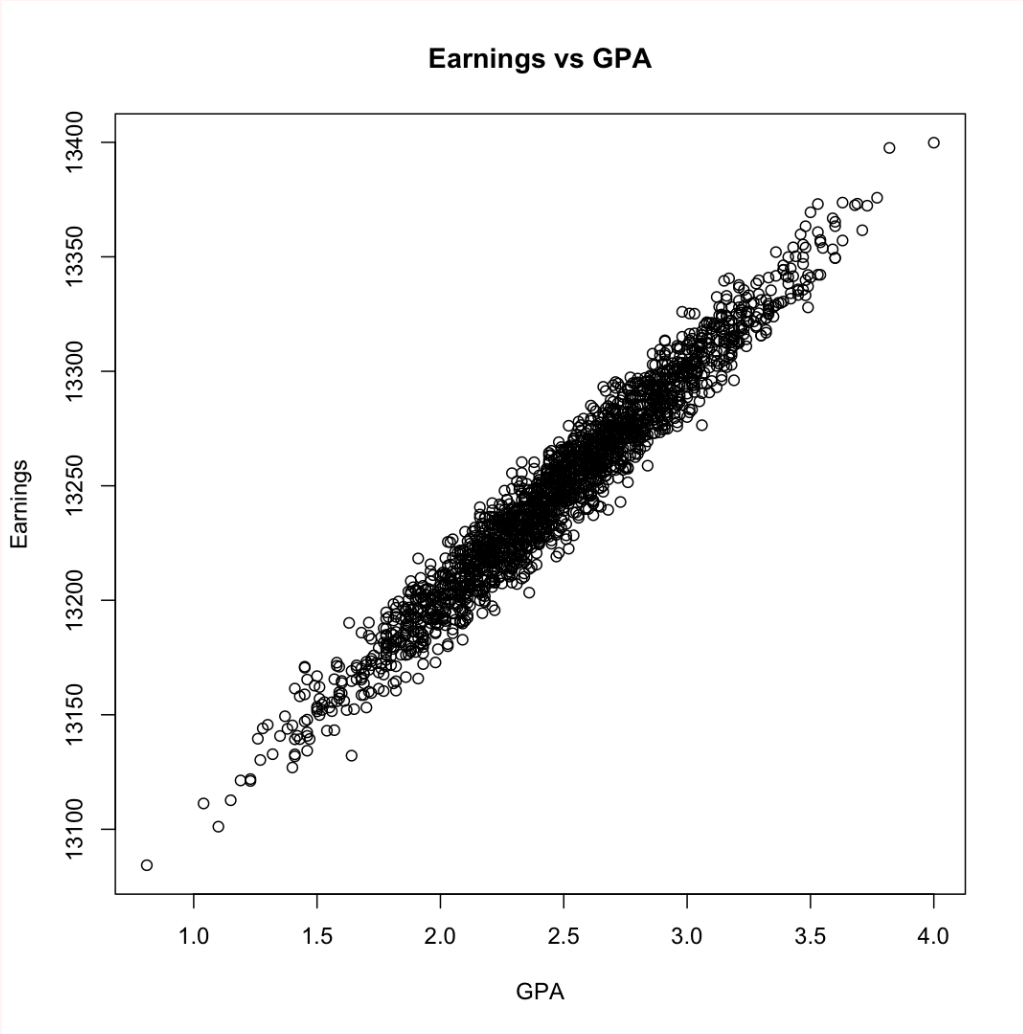
plot(Earnings.Humanintes$Earnings~Earnings.Humanintes$GPA, main = "Earnings vs GPA", xlab = "GPA", ylab = "Earnings")



plot(Earnings.Professional$Earnings~Earnings.Professional$GPA, main = "Earnings vs GPA", xlab = "GPA", ylab = "Earnings")



plot(Earnings.STEM$Earnings~Earnings.STEM$GPA, main = "Earnings vs GPA", xlab = "GPA", ylab = "Earnings")



plot(Earnings.Vocational$Earnings~Earnings.Vocational$GPA, main = "Earnings vs GPA", xlab = "GPA", ylab = "Earnings")

- **GPA seems to matter for the following Majors: Humanities, Professionals, STEM, and Vocational**
  - **GPA definitely affects Earnings for these four majors**
    - **Ex) Higher GPA for Humanities Majors means higher Earnings, Higher GPA for STEM Majors means lower Earnings**
  - **Other attributes did not create interesting plots for any of these majors**

FOR THE MAJORS OTHER THAN BUSINESS OR OTHER

CREATING NEW ATTRIBUTES

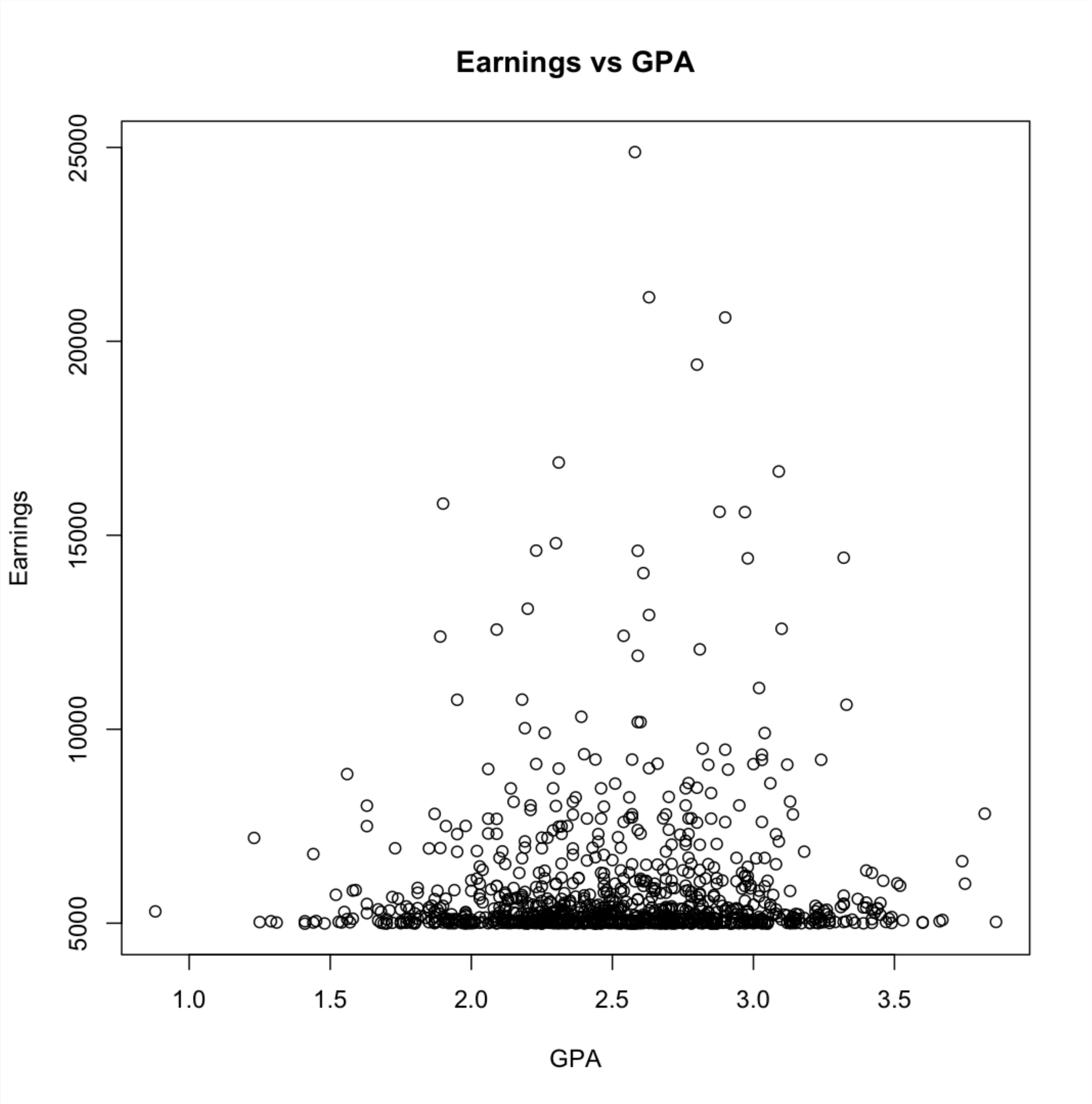
- **To elaborate on why I need new attributes, it is because each of the four majors has a distinct pattern (the slope of the general pattern in the plot Earning VS GPA)**
  - **To accommodate for their differences, I created the following new attributes:**
    - **GOH (GPA Of Humanities)**
    - **GOV (GPA Of Vocational)**
    - **GOP (GPA Of Professional)**
    - **GOS (GPA Of STEM)**

```r
Earnings$GOH <- 0
Earnings[Earnings$Major == "Humanities",]$GOH <- Earnings[Earnings$Major == "Humanities",]$GPA

Earnings$GOV <- 0
Earnings[Earnings$Major == "Vocational",]$GOV <- Earnings[Earnings$Major == "Vocational",]$GPA

Earnings$GOP <- 0
Earnings[Earnings$Major == "Professional" ,]$GOP <- Earnings[Earnings$Major == "Professional" ,]$GPA

Earnings$GOS <- 0
Earnings[Earnings$Major == "STEM",]$GOS <- Earnings[Earnings$Major == "STEM",]$GPA
```

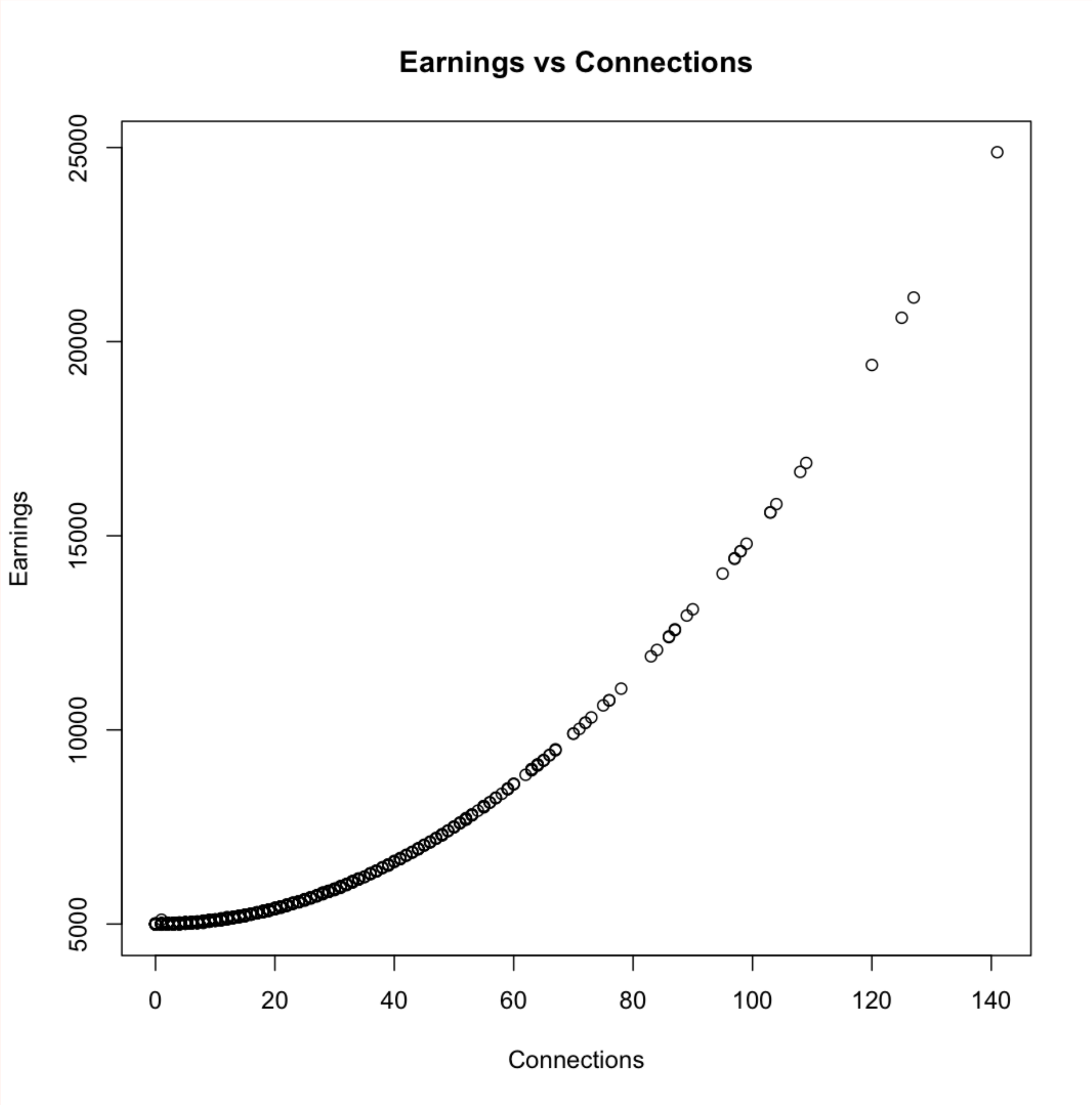
- **Doing so would let me have in the final equation (for the linear regression model) variables each of which only affects a single major**

LOOKING AT THE DATA SET(SUBSETS)

OTHER MAJORS



plot(Earnings.Other$Earnings~Earnings.Other$GPA, main = "Earnings vs GPA", xlab = "GPA", ylab = "Earnings")

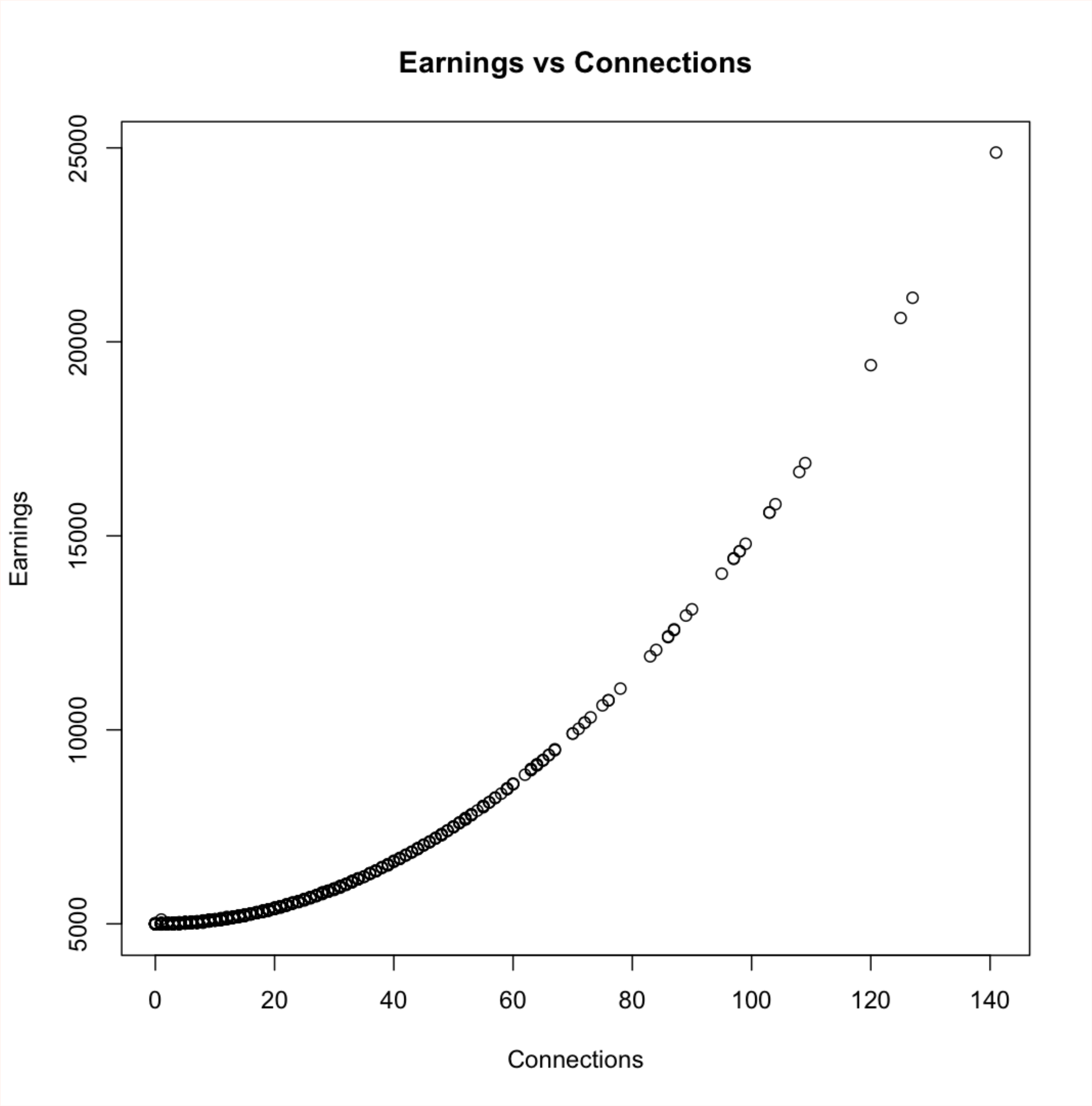


plot(Earnings.Other$Earnings~Earnings.Other$Number_Of_Professional_Connections, main = "Earnings vs Connections", xlab = "Connections", ylab = "Earnings")

- **Major Other did not have any pattern with GPA, but rather had a pattern for the number of Professional Connections**
  - **More Connections roughly seemed to mean higher Earnings**
  - **Plots using other attributes were not intriguing**

OTHER MAJORS: POLYNOMIAL RELATIONSHIP !

- **At first, I tried using SVM or a purely linear model of Linear Regression to predict the values of Earnings for the subset with major “Other”**
  - **This didn’t work well, and after time I realized the plot was hinting at a parabola, not a straight line**
    - **This meant that I needed to be using a polynomial equation for my linear regression model**
      - **Ex) y = x^2 +2 instead of y = x +2**
    - **Implementing a term with degree = 2 gave a much better average MSE for the subset Other (group with only Other Majors)**
    - **The MSE is shown in the next slide…**




```r
Others <- Earnings[Earnings$Major == "Other",]
```

```r
#Other
Others_lr_model <- lm(Earnings~poly(Number_Of_Professional_Connections,2,raw = TRUE),data = Others, x= TRUE, y = TRUE)
Others_lr_model
cv.lm(Others_lr_model,m =3)
```

```
> Others_lr_model <- lm(Earnings~poly(Number_Of_Professional_Connections,2,raw = TRUE),data = Others, x= TRUE, y = TRUE)
> Others_lr_model

Call:
lm(formula = Earnings ~ poly(Number_Of_Professional_Connections,
    2, raw = TRUE), data = Others, x = TRUE, y = TRUE)

Coefficients:
                                            (Intercept)  poly(Number_Of_Professional_Connections, 2, raw = TRUE)1
                                              5.000e+03                                                -9.213e-03
poly(Number_Of_Professional_Connections, 2, raw = TRUE)2
                                              1.000e+00

> cv.lm(Others_lr_model,m =3)
Mean absolute error        :  7.922931
Sample standard deviation  :  0.7228003

Mean squared error         :  109.2974
Sample standard deviation  :  43.88113

Root mean squared error    :  10.31153
Sample standard deviation  :  1.816523
```

- **As you can see, the MSE is only around 109.2974**
- **This means… the pattern for Major Other is finally found!**

CHOOSING THE ATTRIBUTES FOR MY PREDICTION MODEL

- After observing all the previous plots, the following points were shown:
  - Each Major has a distinct pattern for Earnings
    - Ex) Earnings increases as GPA increases for Humanities majors
  - Business Majors (Although in the data set it is shown as "Business" , not "Business", a typo) have their Earnings being affected by their GPAs
    - The Graduation Year being Even or Odd influences how GPA affects Earnings (even means positive influence, odd means negative influence)
  - Number of Professional Connections is the only factor that affects the Earnings for "Other" Majors
    - A parabolic graph was plotted
  - Thus, I will build my prediction model around the above findings!!
    - Prediction Model to be used: LINEAR REGRESSION (ALSO USING A POLYNOMIAL FEATURE)

PREDICTION MODEL

```r
#BUILDING MY MODEL AFTER FINDING THE PATTERNS!!
Earnings <- read.csv("Earnings_Train2021.csv")

#GPA OF Humanity Majors
Earnings$GOH <- 0
Earnings[Earnings$Major == "Humanities",]$GOH <- Earnings[Earnings$Major == "Humanities",]$GPA

#GPA OF Vocational Majors
Earnings$GOV <- 0
Earnings[Earnings$Major == "Vocational",]$GOV <- Earnings[Earnings$Major == "Vocational",]$GPA

#GPA OF Professional Majors
Earnings$GOP <- 0
Earnings[Earnings$Major == "Professional" ,]$GOP <- Earnings[Earnings$Major == "Professional" ,]$GPA

#GPA OF STEM Majors
Earnings$GOS <- 0
Earnings[Earnings$Major == "STEM",]$GOS <- Earnings[Earnings$Major == "STEM",]$GPA

#GPA OF Other Majors; GPA DOES NOT MATTER
Earnings[Earnings$Major != "Other",]$Number_Of_Professional_Connections <- 0


#GPA of Buisness majors; even graudation year means increasing with gpa, odd graduation year means decreasing with gpa!!
Earnings[Earnings$Major == "Buisness" & Earnings$Graduation_Year %% 2 != 0,]$GPA <- Earnings[Earnings$Major == "Buisness" & Earnings$Graduation_Year %% 2 != 0,]$GPA * (-1)
Earnings[Earnings$Major != "Buisness",]$GPA <- 0


#Type of Model: Linear Regression Model, including a polynomial characteristic
Earnings_lr_model <- lm(Earnings~GPA+GOH+GOV+GOP+GOS+Major+poly(Number_Of_Professional_Connections,2,raw=TRUE),data = Earnings, x= TRUE, y = TRUE)
Earnings_lr_model
#Cross Validation
cv.lm(Earnings_lr_model, m= 5)
#MSE (Mean squared error) of 91.22346 !!! Sample standard deviation: 5.222781


#LETS GOOOOOOOOOOOOOOOO!! AHAHAHAHAHAHAHHAHAHAHAHA! MUCH MUCH MUCH BETTER THAN RANDOM FOREST OR SVM!!!
```

- **To the left is my final prediction model**
- **MSE Value for 5 Cross Validations: 91.22346**
- **Sample Standard Deviation: 5.222781**
- **A very satisfactory result, in my opinion**

DETAILS OF THE RESULTING MODEL

```
> Earnings_lr_model <- lm(Earnings~GPA+GOH+GOV+GOP+GOS+Major+poly(Number_Of_Professional_Connections,2,raw=TRUE),data = Earnings, x= TRUE, y = TRUE)
> Earnings_lr_model

Call:
lm(formula = Earnings ~ GPA + GOH + GOV + GOP + GOS + Major + 
    poly(Number_Of_Professional_Connections, 2, raw = TRUE), 
    data = Earnings, x = TRUE, y = TRUE)

Coefficients:
                                            (Intercept)                                                      GPA                                                      GOH
                                              1.000e+04                                                1.000e+02                                                1.005e+02
                                                    GOV                                                      GOP                                                      GOS
                                              9.920e+01                                               -1.010e+02                                               -1.003e+02
                                        MajorHumanities                                               MajorOther                                        MajorProfessional
                                             -1.255e+00                                               -5.000e+03                                                2.002e+03
                                              MajorSTEM                                          MajorVocational  poly(Number_Of_Professional_Connections, 2, raw = TRUE)1
                                              1.097e+00                                                3.002e+03                                               -9.213e-03
poly(Number_Of_Professional_Connections, 2, raw = TRUE)2
                                              1.000e+00
```

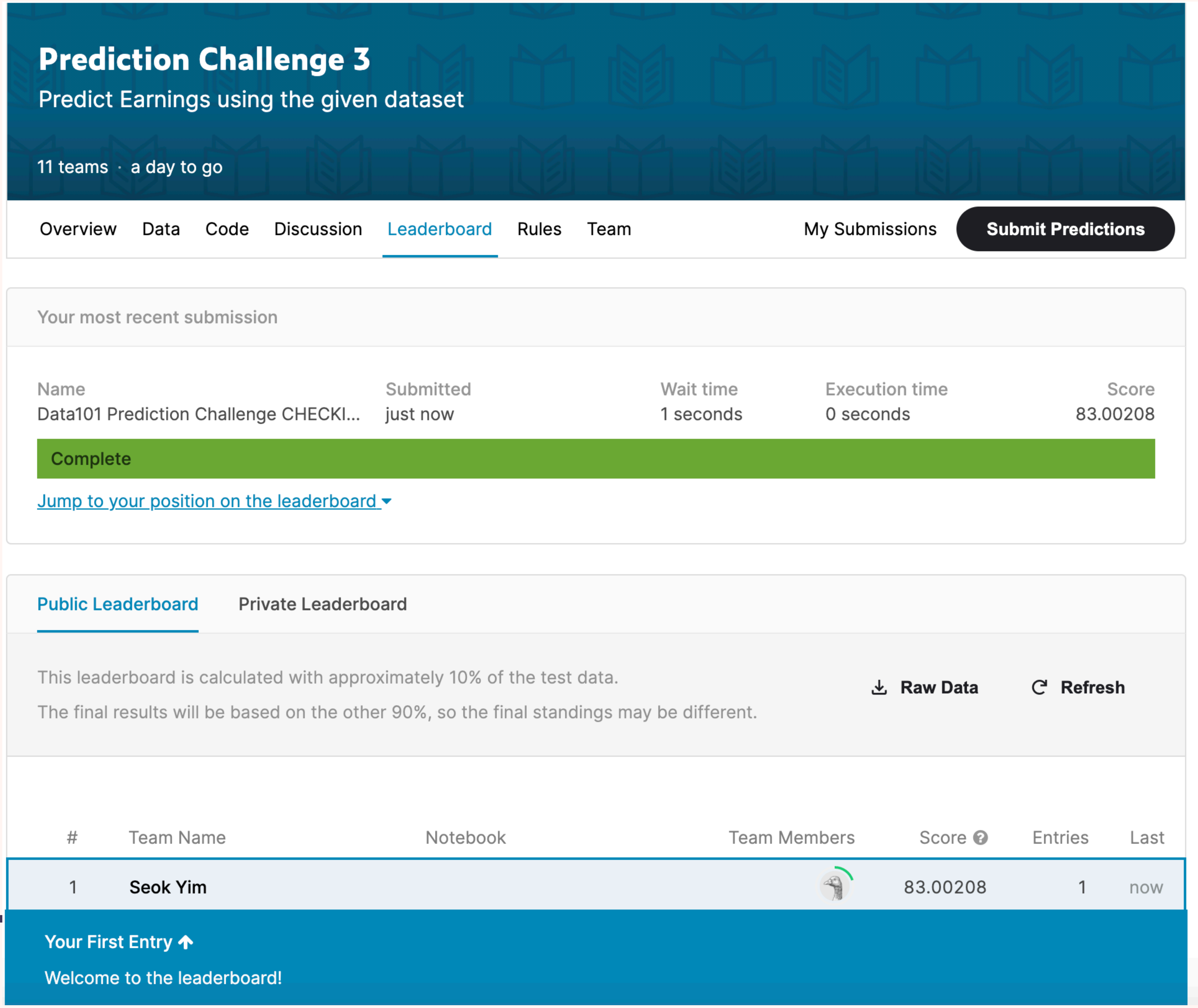
```
> #Cross Validation
> cv.lm(Earnings_lr_model, m= 5)
Mean absolute error        :  7.20674
Sample standard deviation  :  0.1427645

Mean squared error         :  91.15104
Sample standard deviation  :  5.462345

Root mean squared error    :  9.543609
Sample standard deviation  :  0.2800086
```

- **The coefficients for my equation are listed above**
- **The cross validation result is to the left**

FUN PATTERN THAT I FOUND

- It turned out that for the earnings of Business Majors, the equation was the following:
- When Graduation Year %% 2 == 0:
  - Earnings = 10000 + GPA * 100
- When Graduation Year %% 2 != 0:
  - Earnings = 10000 - GPA * 100
    - I didn’t really use this equation in my model, but it was interesting to see such a pattern

KAGGLE SUBMISSION



- Yay!

THANK YOU!