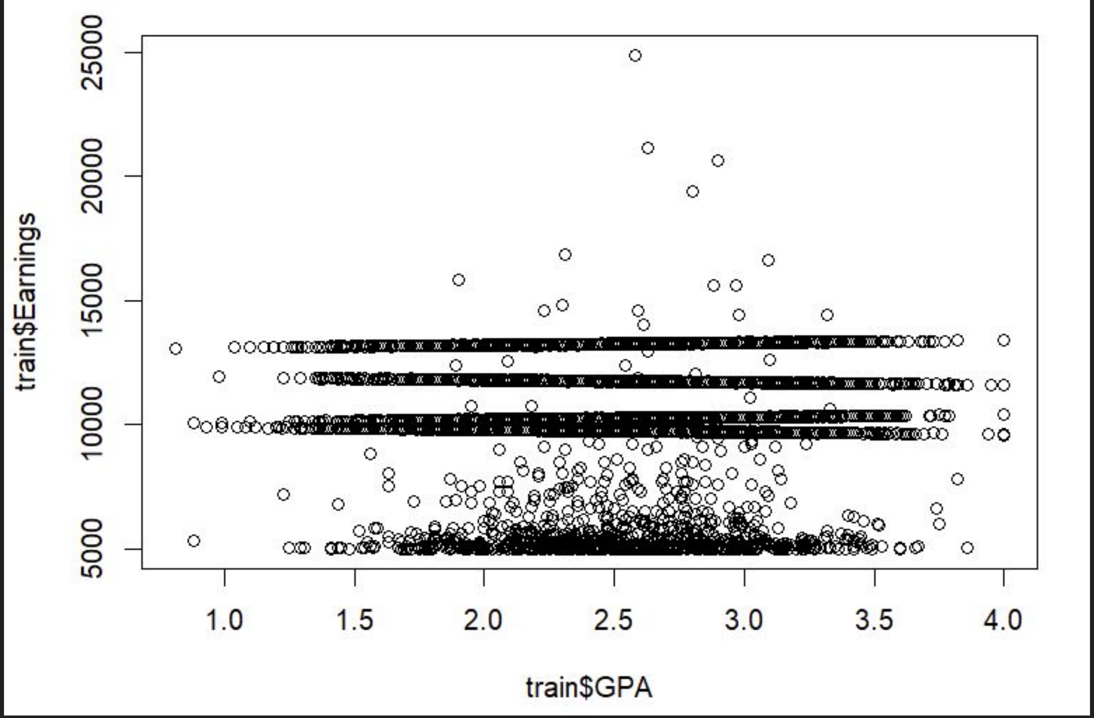
Prediction Challenge 3

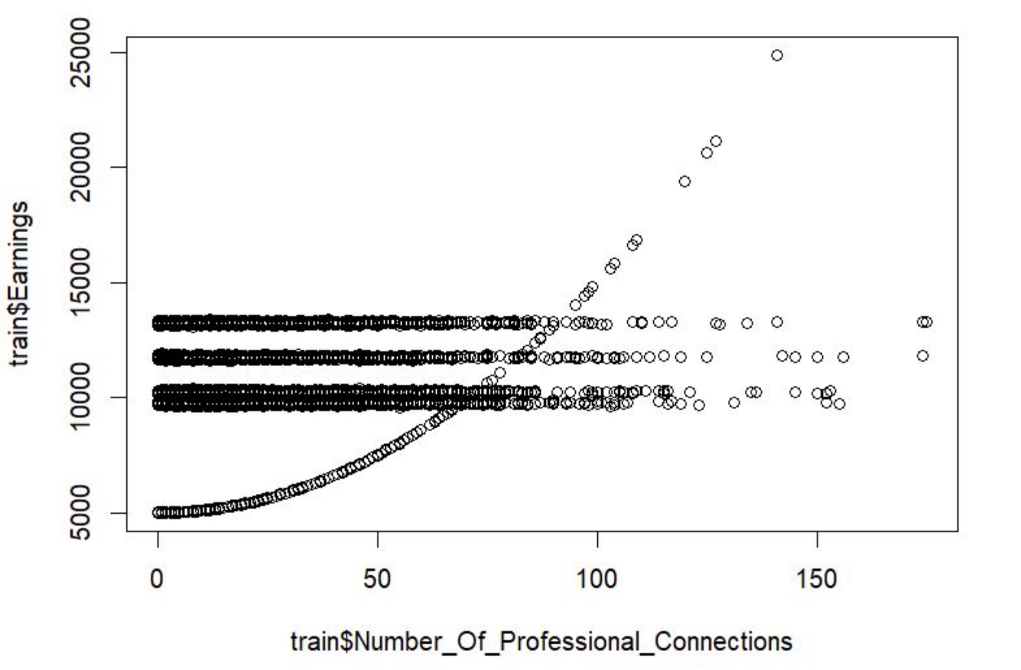
Valerie Le

First, I tried to look for a pattern using plot()

There are some relationship between **GPA and Earnings.** There are certain visible lines, but overall it is not so strong…

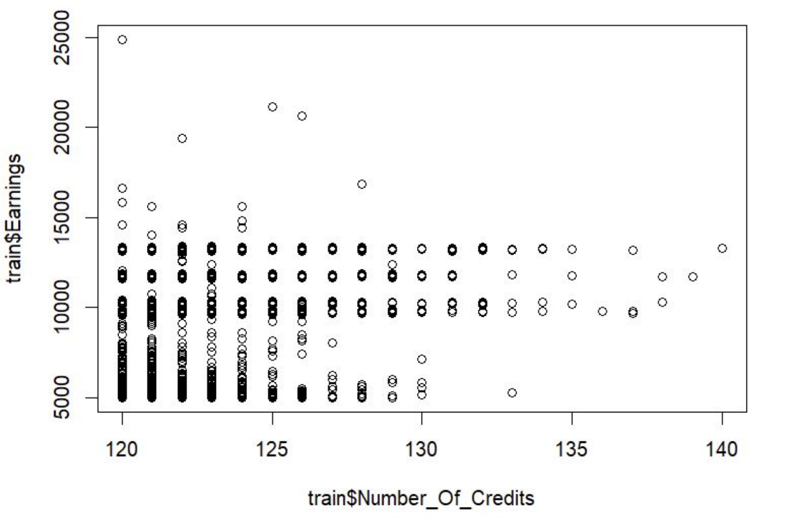


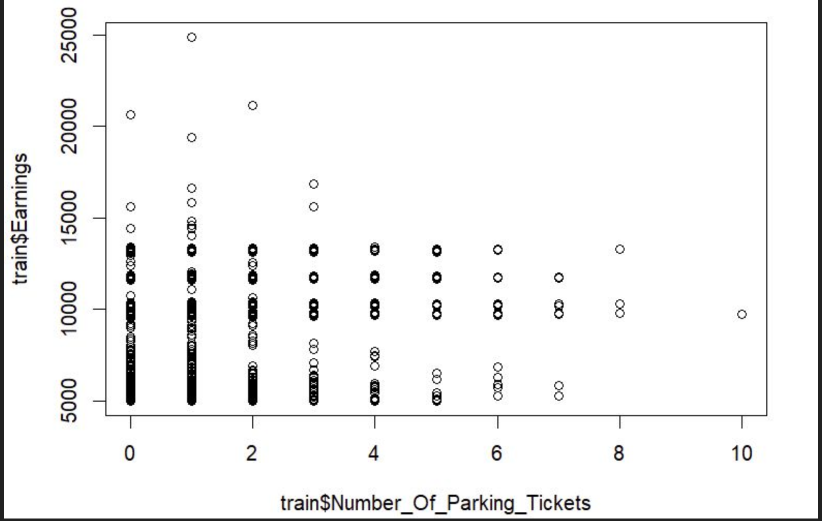
Using plot() to find relationships



There are some correlations between, but it is not strong enough. Specifically, the quadratic line (curved line) is interesting; there might be some relationship between Earnings and Professional Connections

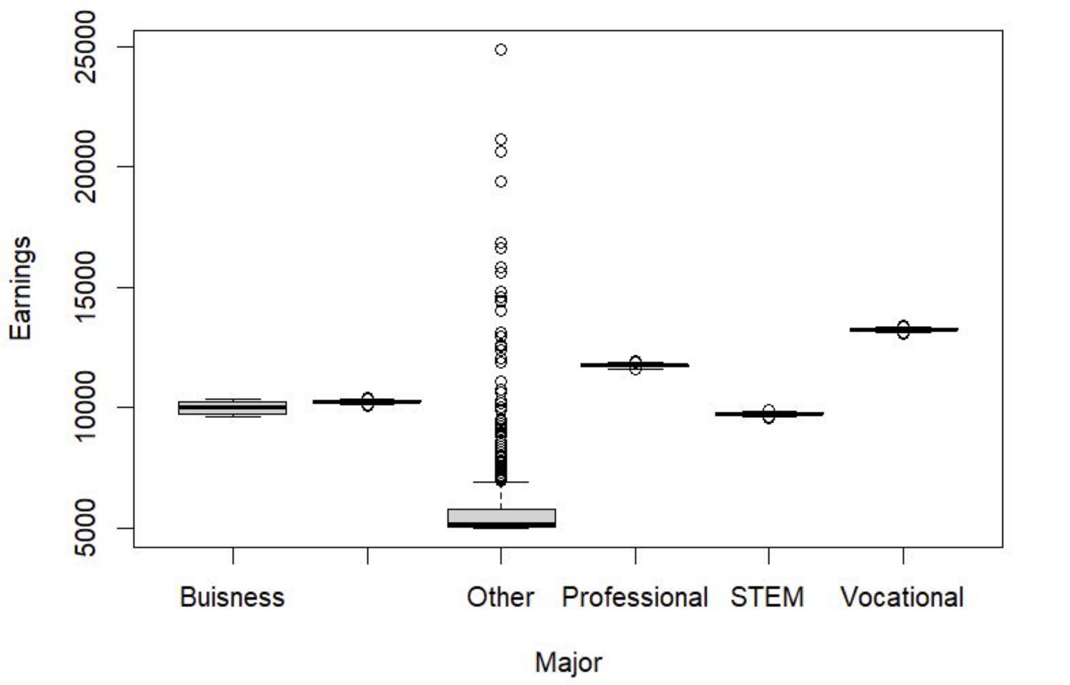
Using plot() to find relationships

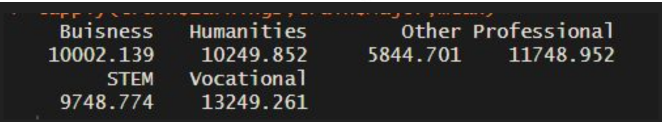




There are no identifiable relationship between the two variables

Earnings ~ Major



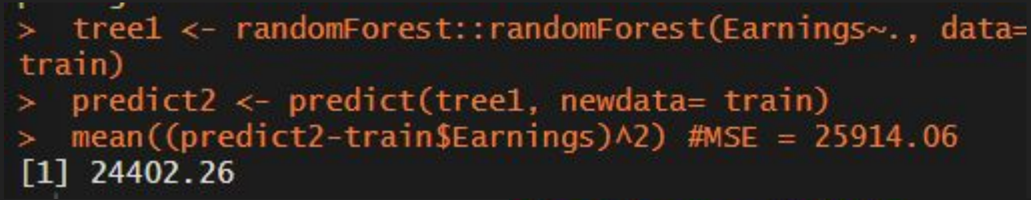


```
  Buisness  Humanities       Other Professional
 10002.139   10249.852    5844.701    11748.952
      STEM  Vocational
  9748.774   13249.261
```

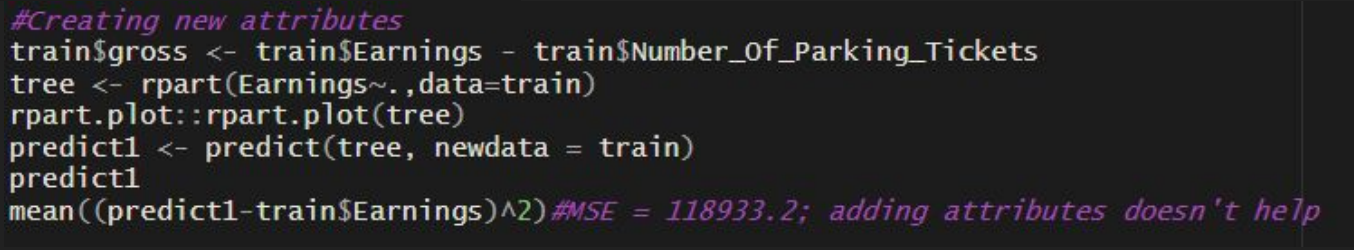
- There is clearly a separation between major and earnings, which means that there is a correlation between earnings and major since the means of income based on major is vastly different

Tried to use different type of machine learning before lm()

- I tried different ML tools that is introduced in the active textbook, such as SVM and randomForest
- So far, randomforest gives the best result. Its mse is around 24402.06, which is still fairly large compared to the ‘excellent’ benchmark (<200)
- I also tried creating new attributes for rpart, but it doesn’t really do anything either (MSE is much HIGHER than using randomForest)



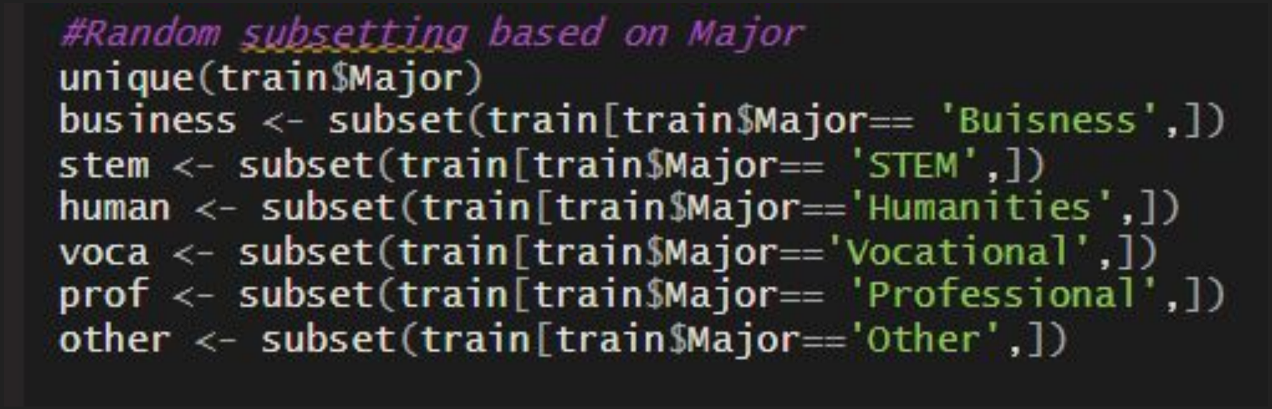
```
>  tree1 <- randomForest::randomForest(Earnings~., data=
train)
>  predict2 <- predict(tree1, newdata= train)
>  mean((predict2-train$Earnings)^2) #MSE = 25914.06
[1] 24402.26
```



```
#Creating new attributes
train$gross <- train$Earnings - train$Number_Of_Parking_Tickets
tree <- rpart(Earnings~.,data=train)
rpart.plot::rpart.plot(tree)
predict1 <- predict(tree, newdata = train)
predict1
mean((predict1-train$Earnings)^2)#MSE = 118933.2; adding attributes doesn't help
```

Creating new subsets

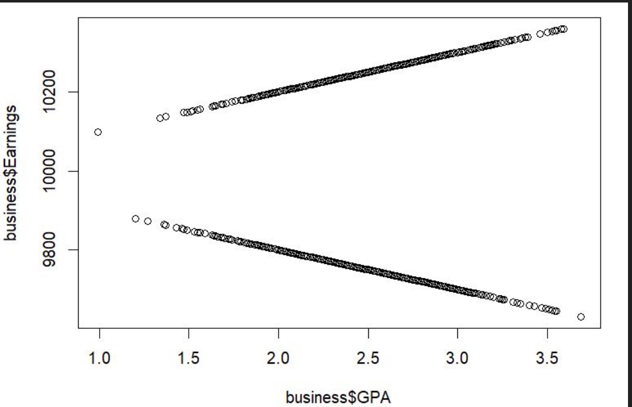
- After trying different machine learning techniques, it is clear that lm() is the only thing that might lower the MSE
- First, I will start dividing data to smaller subsets

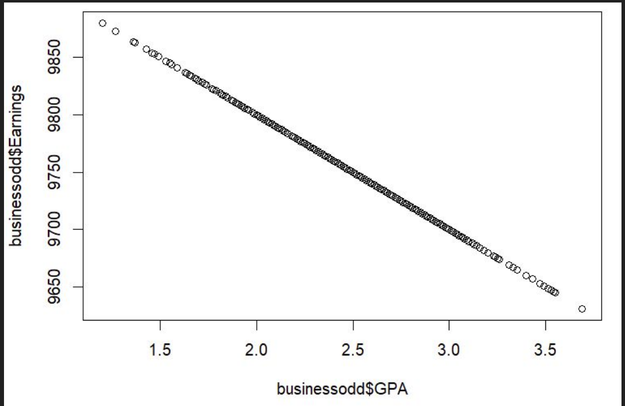


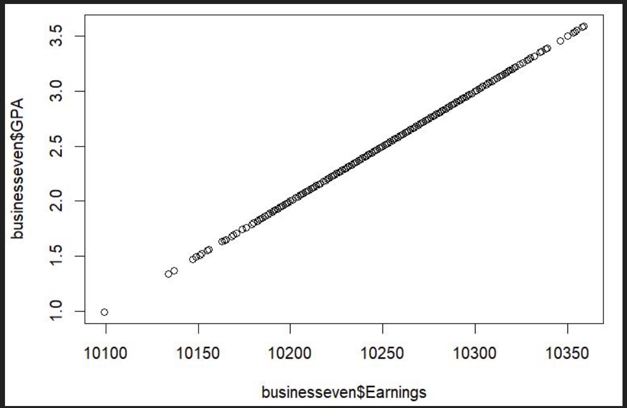
```
#Random subsetting based on Major
unique(train$Major)
business <- subset(train[train$Major== 'Buisness',])
stem <- subset(train[train$Major== 'STEM',])
human <- subset(train[train$Major=='Humanities',])
voca <- subset(train[train$Major=='Vocational',])
prof <- subset(train[train$Major== 'Professional',])
other <- subset(train[train$Major=='Other',])
```

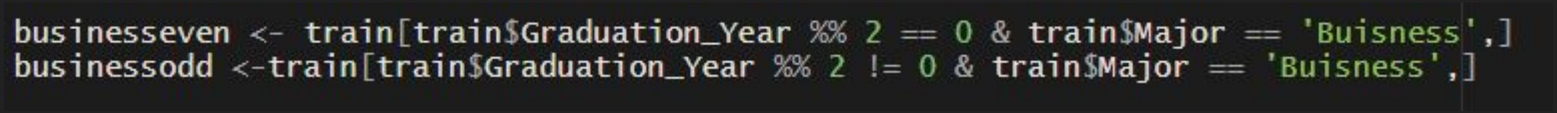
Regression for *Business* major

- Interestingly, this is there are two distinct strong relationships
- After using rpart to analyze this, it seems like this is due to the year of graduating (evens and odds)
- After subsetting the data again and plot, we have two distinct strong correlations





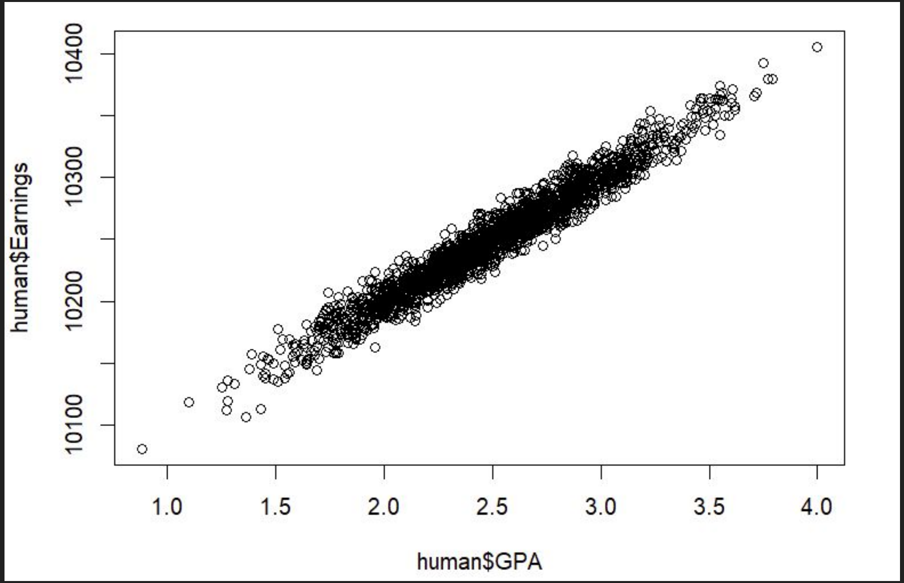


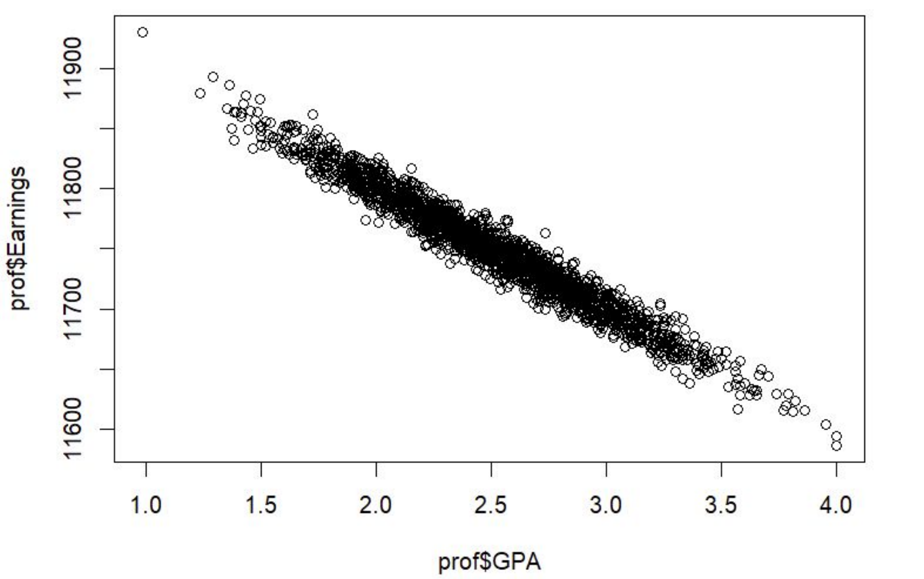


```
businesseven <- train[train$Graduation_Year %% 2 == 0 & train$Major == 'Buisness',]
businessodd <-train[train$Graduation_Year %% 2 != 0 & train$Major == 'Buisness',]
```

Regression for *Humanities* and *Professional* major

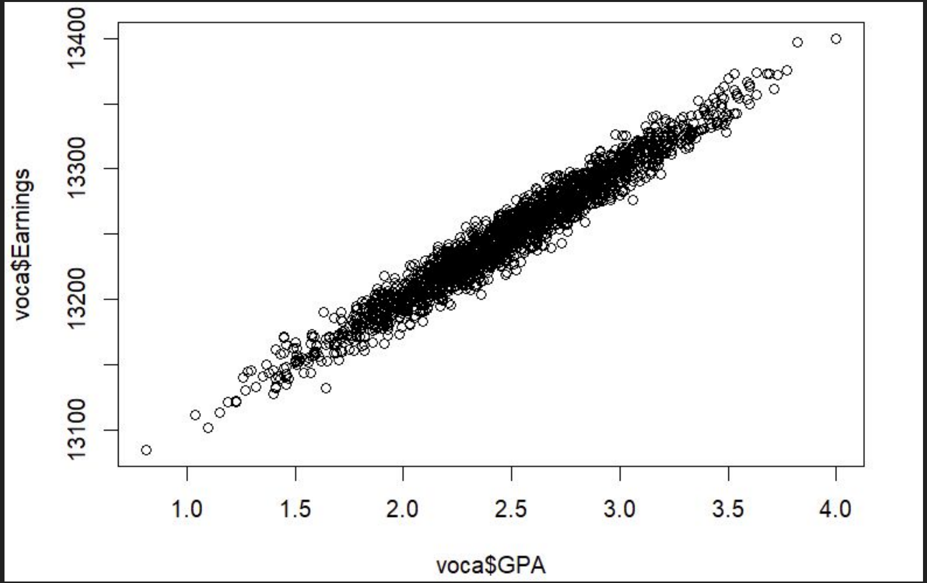
- Strong, positive relationship between GPA and Earnings with Humanities major
- Strong negative relationship with Professional major

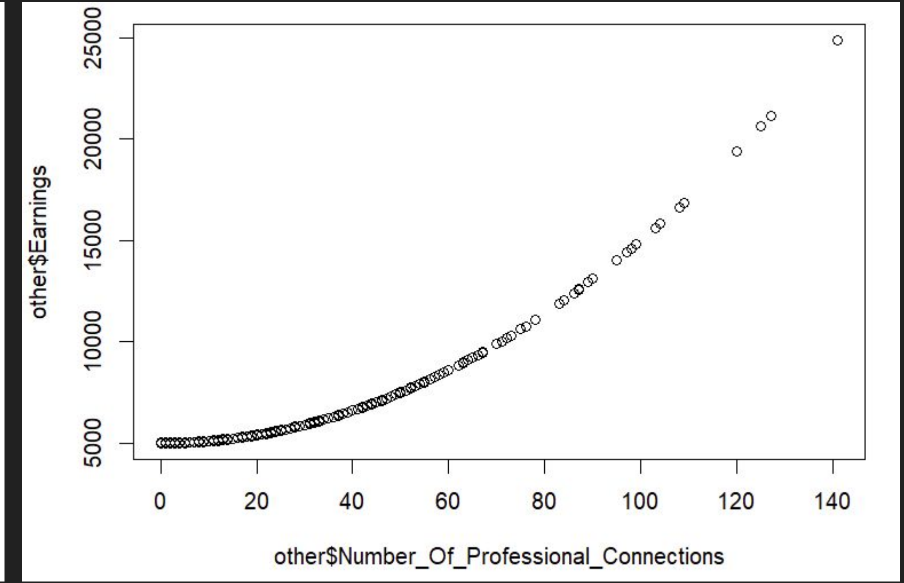




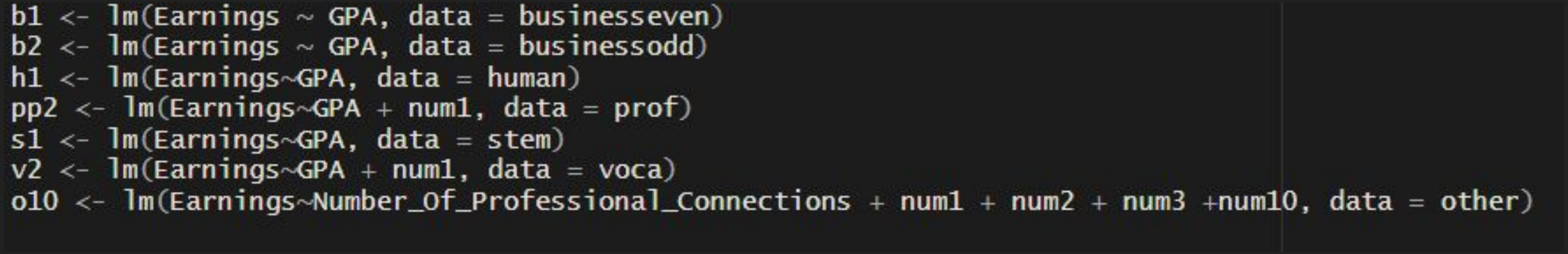
Regression for *Vocational* and *Other* major

- Strong, positive relationship between GPA and Earnings with Humanities major
- Strong positive **quadratic** relationship with Professional connections and earnings



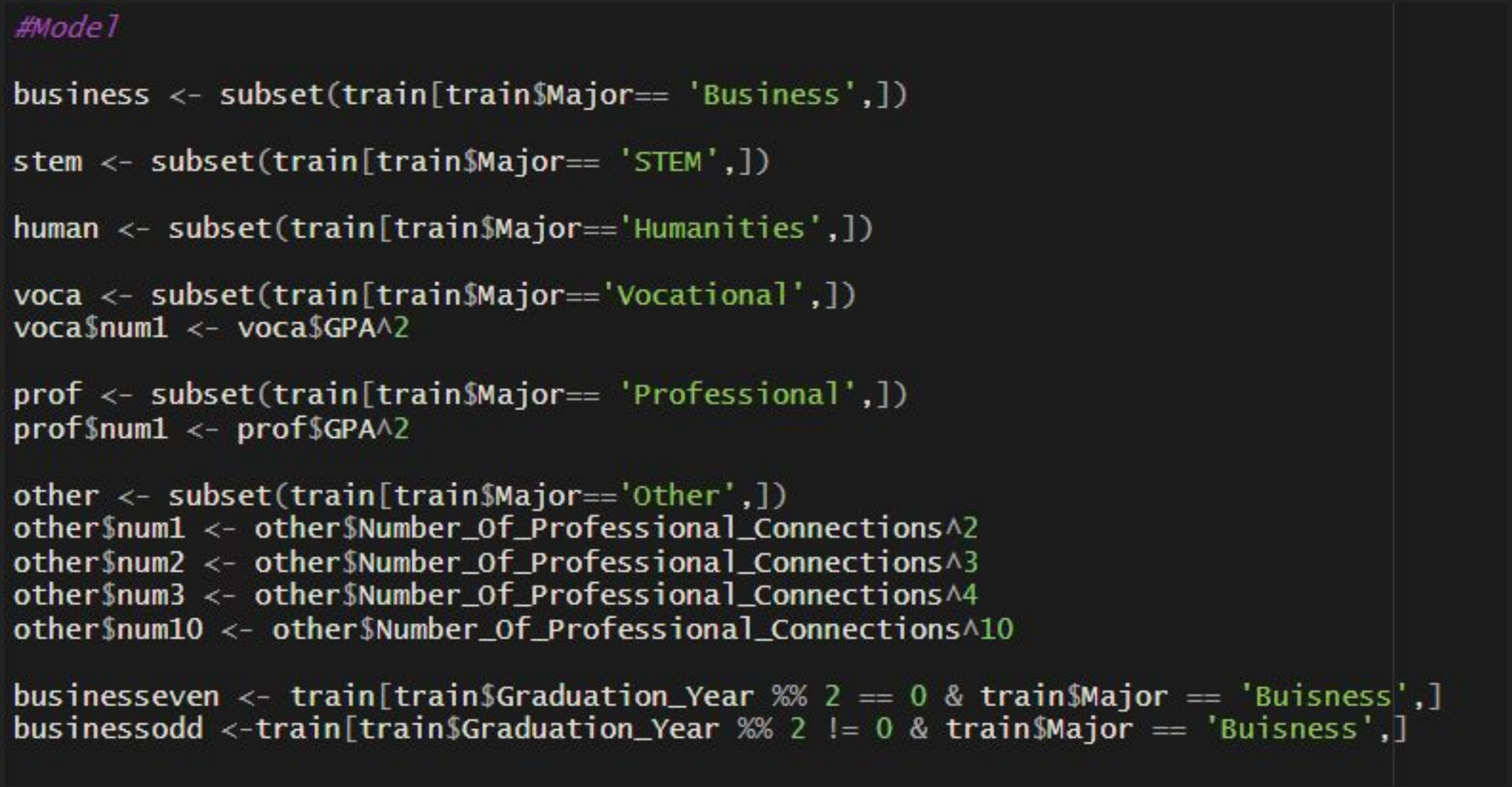


Creating prediction functions using lm() and previous findings



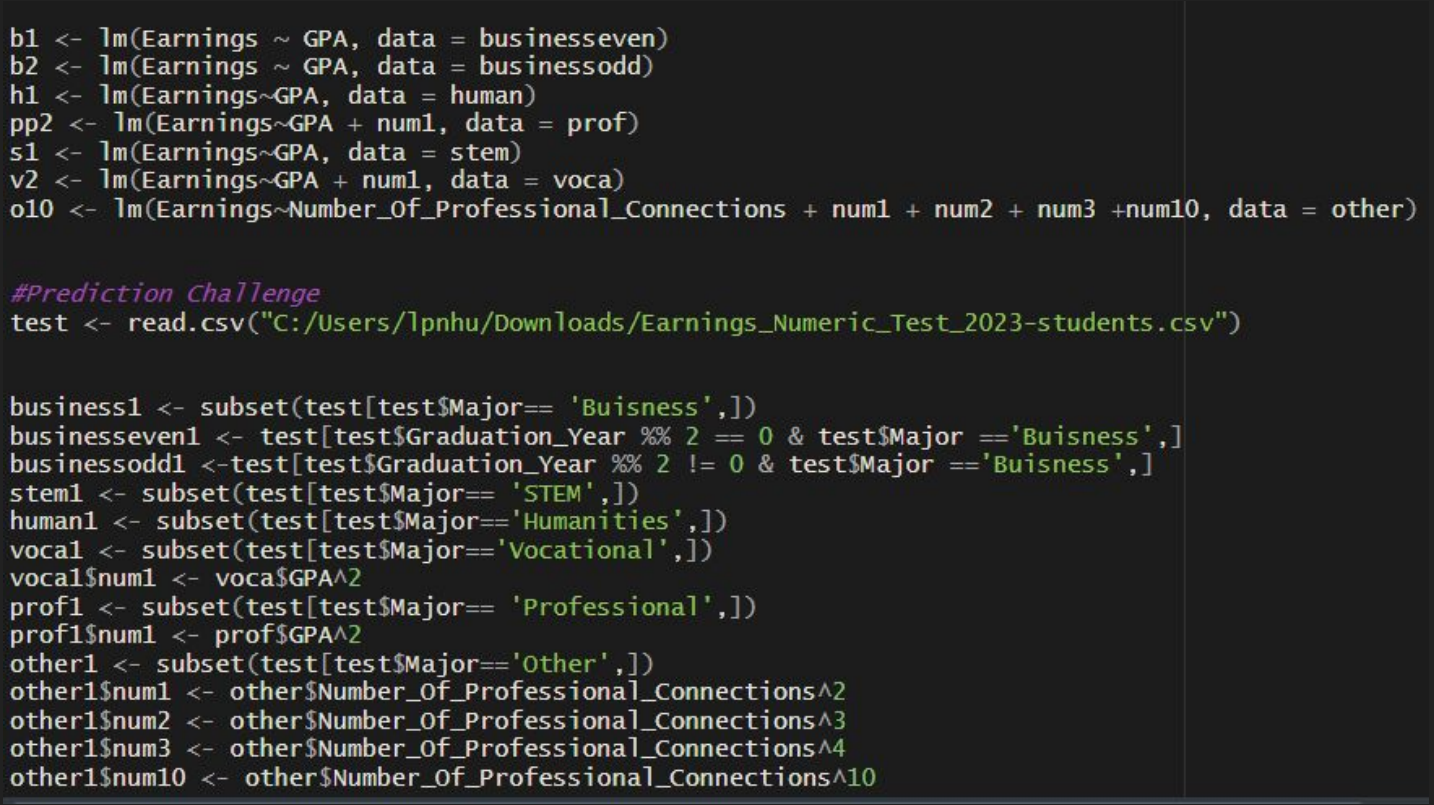
```
b1 <- lm(Earnings ~ GPA, data = businesseven)
b2 <- lm(Earnings ~ GPA, data = businessodd)
h1 <- lm(Earnings~GPA, data = human)
pp2 <- lm(Earnings~GPA + num1, data = prof)
s1 <- lm(Earnings~GPA, data = stem)
v2 <- lm(Earnings~GPA + num1, data = voca)
o10 <- lm(Earnings~Number_Of_Professional_Connections + num1 + num2 + num3 +num10, data = other)
```

Final Model



```
#Model

business <- subset(train[train$Major== 'Business',])

stem <- subset(train[train$Major== 'STEM',])

human <- subset(train[train$Major=='Humanities',])

voca <- subset(train[train$Major=='Vocational',])
voca$num1 <- voca$GPA^2

prof <- subset(train[train$Major== 'Professional',])
prof$num1 <- prof$GPA^2

other <- subset(train[train$Major=='Other',])
other$num1 <- other$Number_Of_Professional_Connections^2
other$num2 <- other$Number_Of_Professional_Connections^3
other$num3 <- other$Number_Of_Professional_Connections^4
other$num10 <- other$Number_Of_Professional_Connections^10

businesseven <- train[train$Graduation_Year %% 2 == 0 & train$Major == 'Buisness',]
businessodd <-train[train$Graduation_Year %% 2 != 0 & train$Major == 'Buisness',]
```

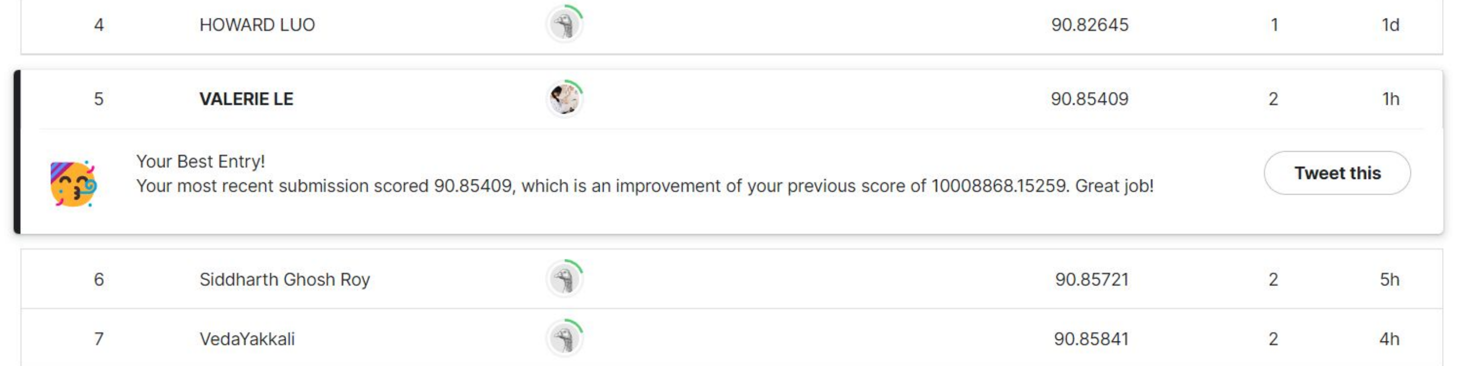
(cont.)



```
b1 <- lm(Earnings ~ GPA, data = businesseven)
b2 <- lm(Earnings ~ GPA, data = businessodd)
h1 <- lm(Earnings~GPA, data = human)
pp2 <- lm(Earnings~GPA + num1, data = prof)
s1 <- lm(Earnings~GPA, data = stem)
v2 <- lm(Earnings~GPA + num1, data = voca)
o10 <- lm(Earnings~Number_Of_Professional_Connections + num1 + num2 + num3 +num10, data = other)


#Prediction Challenge
test <- read.csv("C:/Users/lpnhu/Downloads/Earnings_Numeric_Test_2023-students.csv")


business1 <- subset(test[test$Major== 'Buisness',])
businesseven1 <- test[test$Graduation_Year %% 2 == 0 & test$Major =='Buisness',]
businessodd1 <-test[test$Graduation_Year %% 2 != 0 & test$Major =='Buisness',]
stem1 <- subset(test[test$Major== 'STEM',])
human1 <- subset(test[test$Major=='Humanities',])
voca1 <- subset(test[test$Major=='Vocational',])
voca1$num1 <- voca$GPA^2
prof1 <- subset(test[test$Major== 'Professional',])
prof1$num1 <- prof$GPA^2
other1 <- subset(test[test$Major=='Other',])
other1$num1 <- other$Number_Of_Professional_Connections^2
other1$num2 <- other$Number_Of_Professional_Connections^3
other1$num3 <- other$Number_Of_Professional_Connections^4
other1$num10 <- other$Number_Of_Professional_Connections^10
```

Final thoughts

- The final product has a pretty low MSE level
- I wasted one of my attempts because I forgot to change the df to test_df, but once I fix it, it gives me a decent result!

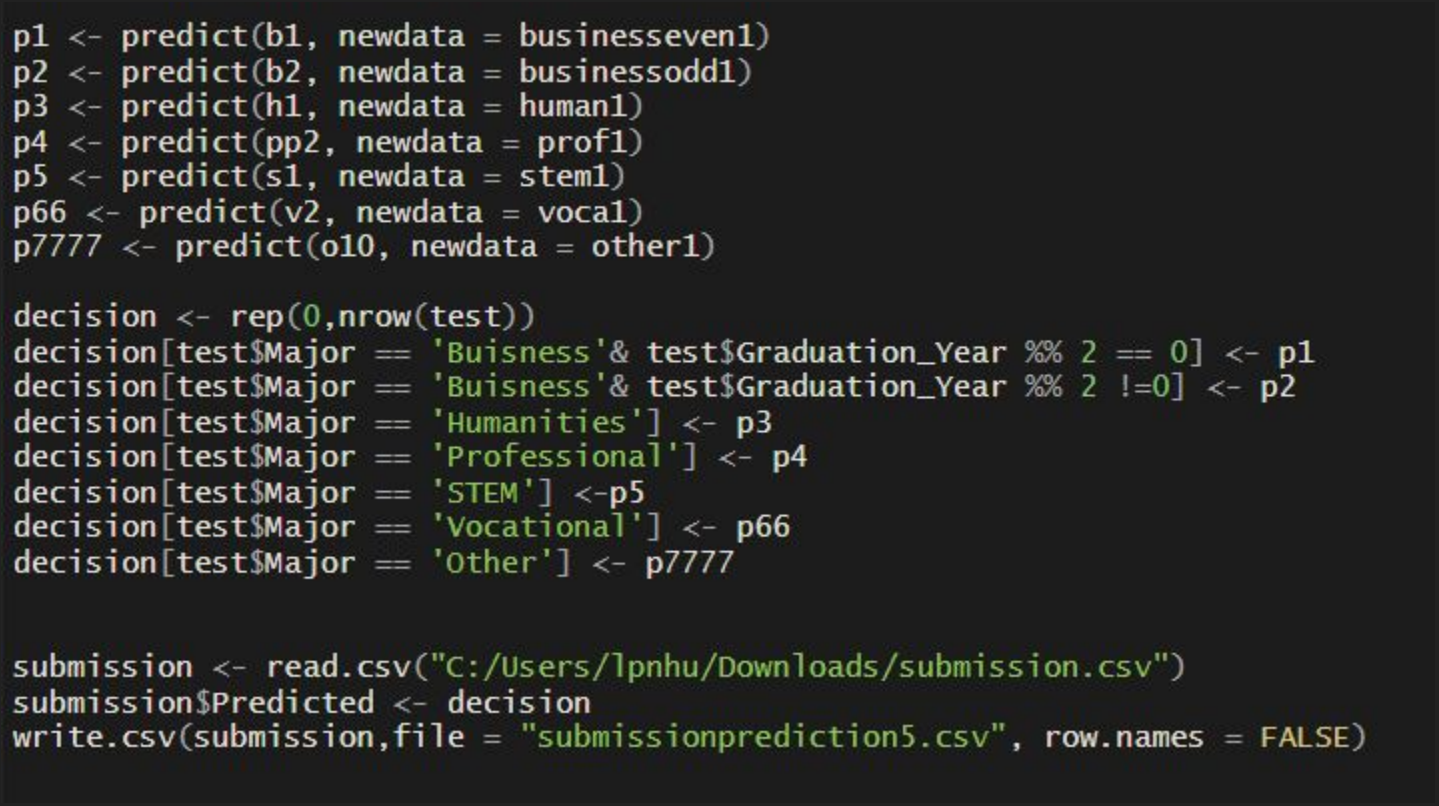


4	HOWARD LUO	90.82645	1	1d
5	**VALERIE LE**	90.85409	2	1h

Your Best Entry!
Your most recent submission scored 90.85409, which is an improvement of your previous score of 10008868.15259. Great job!

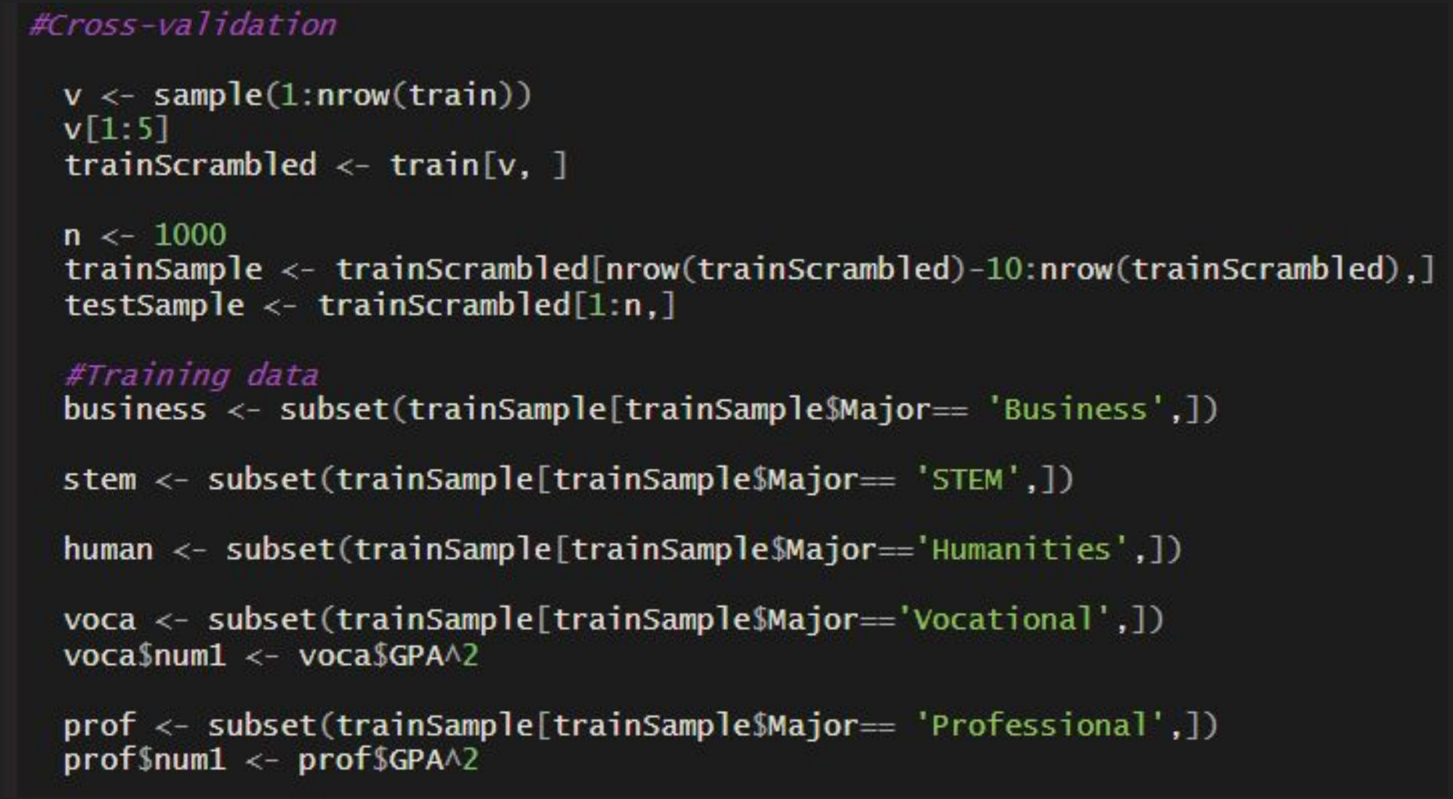
6	Siddharth Ghosh Roy	90.85721	2	5h
7	VedaYakkali	90.85841	2	4h

(cont.)



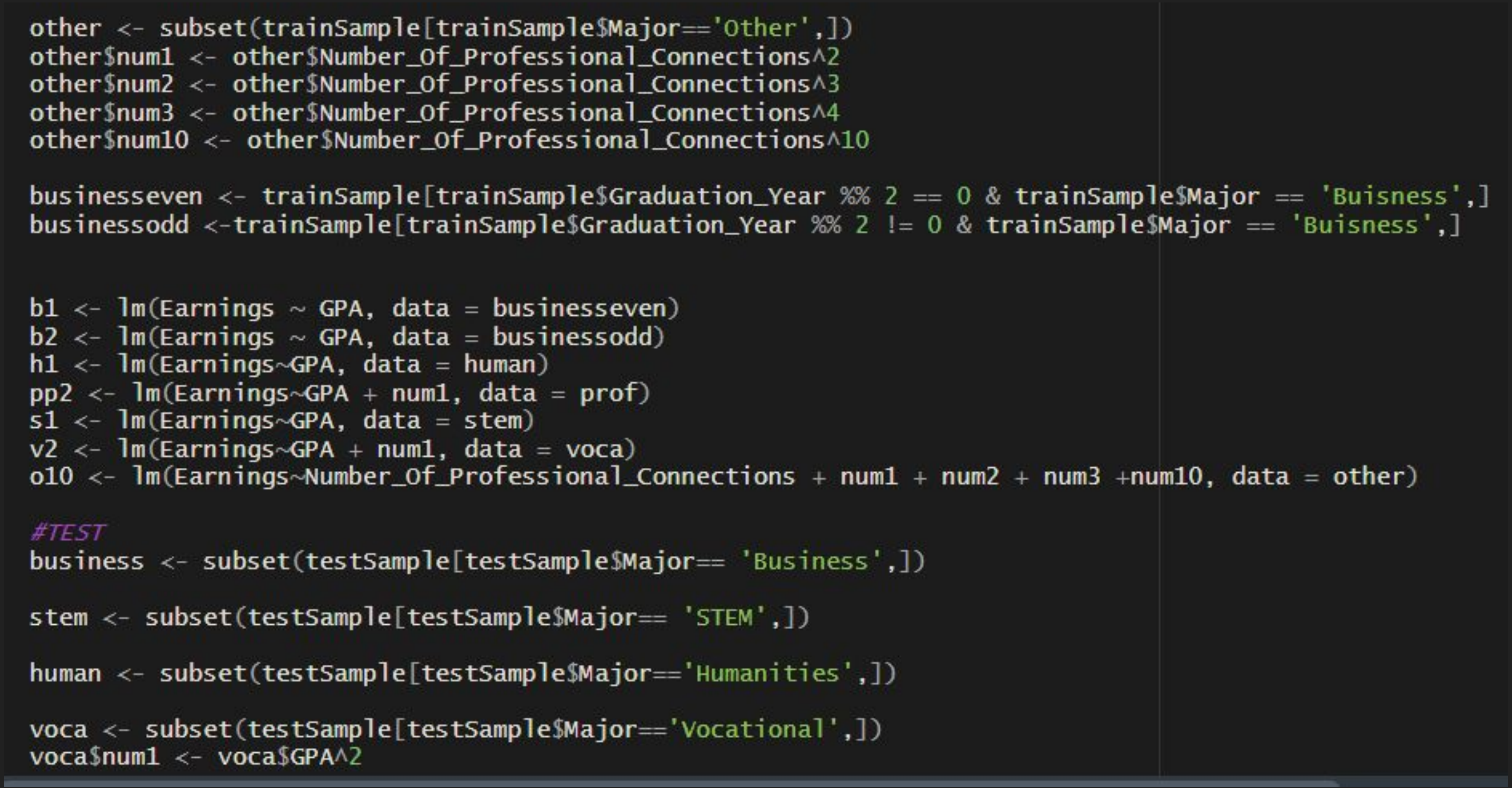
```
p1 <- predict(b1, newdata = businesseven1)
p2 <- predict(b2, newdata = businessodd1)
p3 <- predict(h1, newdata = human1)
p4 <- predict(pp2, newdata = prof1)
p5 <- predict(s1, newdata = stem1)
p66 <- predict(v2, newdata = vocal)
p7777 <- predict(o10, newdata = other1)

decision <- rep(0,nrow(test))
decision[test$Major == 'Buisness'& test$Graduation_Year %% 2 == 0] <- p1
decision[test$Major == 'Buisness'& test$Graduation_Year %% 2 !=0] <- p2
decision[test$Major == 'Humanities'] <- p3
decision[test$Major == 'Professional'] <- p4
decision[test$Major == 'STEM'] <-p5
decision[test$Major == 'Vocational'] <- p66
decision[test$Major == 'Other'] <- p7777


submission <- read.csv("C:/Users/lpnhu/Downloads/submission.csv")
submission$Predicted <- decision
write.csv(submission,file = "submissionprediction5.csv", row.names = FALSE)
```

Cross-validation before turning in the product



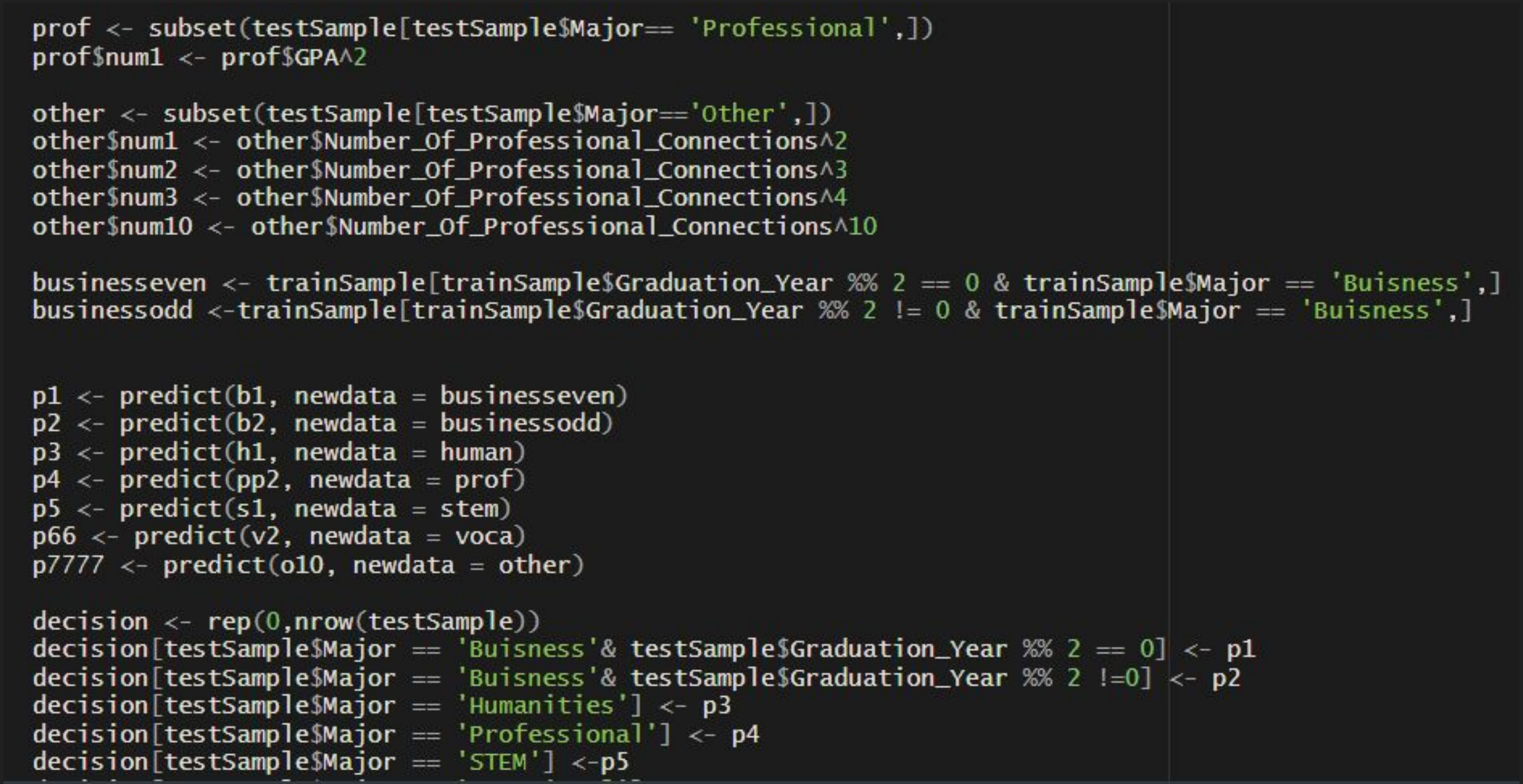
```
#Cross-validation

  v <- sample(1:nrow(train))
  v[1:5]
  trainScrambled <- train[v, ]

  n <- 1000
  trainSample <- trainScrambled[nrow(trainScrambled)-10:nrow(trainScrambled),]
  testSample <- trainScrambled[1:n,]

  #Training data
  business <- subset(trainSample[trainSample$Major== 'Business',])

  stem <- subset(trainSample[trainSample$Major== 'STEM',])

  human <- subset(trainSample[trainSample$Major=='Humanities',])

  voca <- subset(trainSample[trainSample$Major=='Vocational',])
  voca$num1 <- voca$GPA^2

  prof <- subset(trainSample[trainSample$Major== 'Professional',])
  prof$num1 <- prof$GPA^2
```

(cont.)

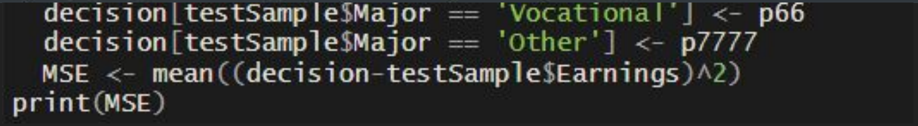


```
other <- subset(trainSample[trainSample$Major=='Other',])
other$num1 <- other$Number_Of_Professional_Connections^2
other$num2 <- other$Number_Of_Professional_Connections^3
other$num3 <- other$Number_Of_Professional_Connections^4
other$num10 <- other$Number_Of_Professional_Connections^10

businesseven <- trainSample[trainSample$Graduation_Year %% 2 == 0 & trainSample$Major == 'Buisness',]
businessodd <-trainSample[trainSample$Graduation_Year %% 2 != 0 & trainSample$Major == 'Buisness',]


b1 <- lm(Earnings ~ GPA, data = businesseven)
b2 <- lm(Earnings ~ GPA, data = businessodd)
h1 <- lm(Earnings~GPA, data = human)
pp2 <- lm(Earnings~GPA + num1, data = prof)
s1 <- lm(Earnings~GPA, data = stem)
v2 <- lm(Earnings~GPA + num1, data = voca)
o10 <- lm(Earnings~Number_Of_Professional_Connections + num1 + num2 + num3 +num10, data = other)

#TEST
business <- subset(testSample[testSample$Major== 'Business',])

stem <- subset(testSample[testSample$Major== 'STEM',])

human <- subset(testSample[testSample$Major=='Humanities',])

voca <- subset(testSample[testSample$Major=='Vocational',])
voca$num1 <- voca$GPA^2
```

(cont.)



```
prof <- subset(testSample[testSample$Major== 'Professional',])
prof$num1 <- prof$GPA^2

other <- subset(testSample[testSample$Major=='Other',])
other$num1 <- other$Number_Of_Professional_Connections^2
other$num2 <- other$Number_Of_Professional_Connections^3
other$num3 <- other$Number_Of_Professional_Connections^4
other$num10 <- other$Number_Of_Professional_Connections^10

businesseven <- trainSample[trainSample$Graduation_Year %% 2 == 0 & trainSample$Major == 'Buisness',]
businessodd <-trainSample[trainSample$Graduation_Year %% 2 != 0 & trainSample$Major == 'Buisness',]


p1 <- predict(b1, newdata = businesseven)
p2 <- predict(b2, newdata = businessodd)
p3 <- predict(h1, newdata = human)
p4 <- predict(pp2, newdata = prof)
p5 <- predict(s1, newdata = stem)
p66 <- predict(v2, newdata = voca)
p7777 <- predict(o10, newdata = other)

decision <- rep(0,nrow(testSample))
decision[testSample$Major == 'Buisness'& testSample$Graduation_Year %% 2 == 0] <- p1
decision[testSample$Major == 'Buisness'& testSample$Graduation_Year %% 2 !=0] <- p2
decision[testSample$Major == 'Humanities'] <- p3
decision[testSample$Major == 'Professional'] <- p4
decision[testSample$Major == 'STEM'] <-p5
```



```
  decision[testSample$Major == 'Vocational'] <- p66
  decision[testSample$Major == 'Other'] <- p7777
  MSE <- mean((decision-testSample$Earnings)^2)
print(MSE)
```