Nuran Ghoneim - Final Project

2022-12-09

# How Salary Affects Students’ College Major Decisions

I’ll be comparing two data sets to see if students enroll in certain majors because of the salary associated with the degree. The first data set called ‘studentsEnrolledByMajor’, and it shows the majors grouped by type and the estimate number of students enrolled along with the margin of error. The second data set is called ‘salaryDegree’ which it shows each major and it’s starting salary, mid-career salary and percentiles.

## Cleaning the Data

First I’m going to read in my first data set from my desktop. As you can see there’s some cleaning I need to do. I need to get rid of rows that only present as sub titles, unnecessary characters, and convert the second column to integers.

studentsEnrolled = read.csv("C:/Users/nuran/Desktop/STA 9750/Final Project/studentsEnrolledByMajor.csv")  
  
head(studentsEnrolled)

## ï..Degree Estimate MarginOfError  
## 1 Science and Engineering   
## 2 Computers, Math and Stats 4,276,538 Â±36,911  
## 3 Environmental Sciences 5,520,366 Â±41,974  
## 4 Physical Sciences 2,660,755 Â±24,800  
## 5 Psychology 4,456,492 Â±35,208  
## 6 Social Sciences 6,965,234 Â±49,423

I dropped the sub heading rows as well as the third column since I won’t be using it. Then I got rid of the commas that appear in column 2 and converted it to integer. Finally, I renamed the first column to get rid of the characters before “degree”. Once I did that, this data set was cleaned and ready for use.

studentsEnrolled <- studentsEnrolled[-c(1, 12), ]  
studentsEnrolled = select(studentsEnrolled, -3)

studentsEnrolled$Estimate<-gsub(",","",as.character(studentsEnrolled$Estimate))  
  
studentsEnrolled$Estimate = strtoi(studentsEnrolled$Estimate)

names(studentsEnrolled)[names(studentsEnrolled) == 'ï..Degree'] <- 'Degrees'

My second data set was originally called “salaryByDegree” but the majors were individually listed out while the other file is listed by the group types of majors. So in order to make them the same since that’s the column I’ll be merging on, I grouped the majors into their respective categories and averaged their salaries for each group. So below is the second data set I’ll be using called “salaryDegree”, which doesn’t need much cleaning as you can see below.

salaryDegree = read.csv("C:/Users/nuran/Desktop/STA 9750/Final Project/salaryDegree.csv")  
head(salaryDegree)

## ï..Undergraduate.Major Starting.Median.Salary Mid.Career.Median.Salary  
## 1 Computers, Math and Stats 49900.00 86250.00  
## 2 Environmental Sciences 40425.00 66200.00  
## 3 Physical Sciences 45466.67 85566.67  
## 4 Psychology 36200.00 59300.00  
## 5 Social Sciences 39166.67 65000.00  
## 6 Engineering 58957.14 99257.14  
## Percent.change.from.Starting.to.Mid.Career.Salary  
## 1 73.48  
## 2 63.73  
## 3 87.93  
## 4 63.85  
## 5 65.80  
## 6 68.30  
## Mid.Career.10th.Percentile.Salary Mid.Career.25th.Percentile.Salary  
## 1 47750.00 64075.00  
## 2 38550.00 49700.00  
## 3 48766.67 64833.33  
## 4 31150.00 41250.00  
## 5 36300.00 48433.33  
## 6 65114.29 80128.57  
## Mid.Career.75th.Percentile.Salary Mid.Career.90th.Percentile.Salary  
## 1 113675.00 153000.0  
## 2 89950.00 132000.0  
## 3 113666.67 160666.7  
## 4 84350.00 122500.0  
## 5 91266.67 135066.7  
## 6 128857.14 167000.0

I renamed the first column to Degrees, to match the first data set, so I wouldn’t have to specify which columns I want to merge on. Then I proceeded to merge the two data sets.

names(salaryDegree)[names(salaryDegree) == 'ï..Undergraduate.Major'] <- 'Degrees'

combinedDf = merge(salaryDegree, studentsEnrolled)

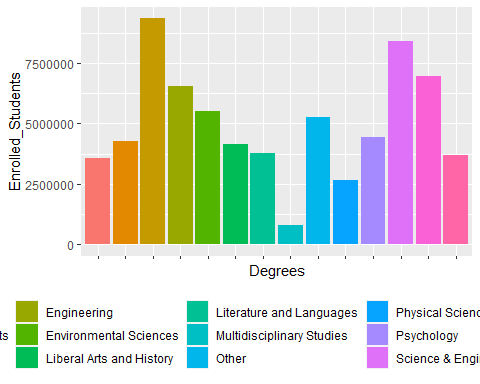
I also renamed the last column to a more useful and helpful title.

names(combinedDf)[names(combinedDf) == 'Estimate'] <- 'Enrolled\_Students'

## Taking a look at the data

First I wanted to see what majors most students are enrolled in so I graphed it. As you can see below, most students are enrolled in Education followed by Science and Engineering Related fields with Social sciences and Engineering coming in third and fourth place. All other majors after engineering are lower in numbers, making those the top 4 degrees students are interested in.

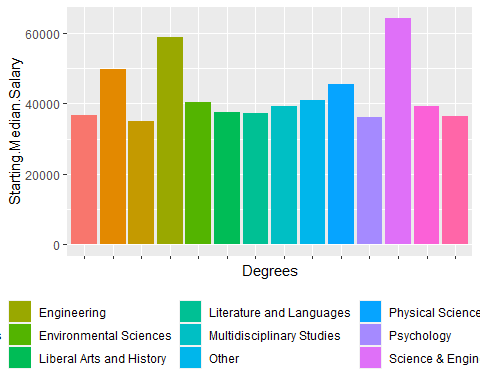
barChart<-ggplot(data=combinedDf, aes(x=Degrees, y=Enrolled\_Students, fill=Degrees)) +  
 geom\_bar(stat="identity")  
barChart+theme(axis.text.x = element\_blank(), legend.position="bottom")



Next I wanted to see which degrees earn you the most, so following the same steps as above, I developed this graph below. As you can see, the top earning field is Science and Engineering Related fields with Engineering coming in second and Computers, Math & Stats coming in third.

Now this is very interesting because while Education was the most popular degree, it’s not in the top 3 regarding starting salary. Science and Engineering Related fields being the top earner does make sense and it does fall in line with the previous results. Lastly, Computers, Math and Stats is the third top earner. Despite it not being a popular subject as previously seen, it could make sense that the starting salary is pretty high, to account for the lack of people in this field. Regardless, these two graphs do seem to demonstrate that despite certain degrees being top earners, not all may be sought after by students.

barChart<-ggplot(data=combinedDf, aes(x=Degrees, y=Starting.Median.Salary, fill=Degrees)) +  
 geom\_bar(stat="identity")  
barChart+theme(axis.text.x = element\_blank(), legend.position="bottom")



Now that I’ve gotten an idea of what the data looks like, I decided to run a multi-linear regression to see how these variables correlate with each other. As you can see, none of the p-values appear to be statistically significant, the r-squared value, it’s 0.49, meaning it’s not a strong linear relationship and the low t-values suggest they’re not highly significant.

multi\_fit = lm(combinedDf$Enrolled\_Students ~ Starting.Median.Salary + Mid.Career.Median.Salary + Mid.Career.10th.Percentile.Salary + Mid.Career.25th.Percentile.Salary + Mid.Career.75th.Percentile.Salary + Mid.Career.90th.Percentile.Salary, data = combinedDf)  
summary(multi\_fit)

##   
## Call:  
## lm(formula = combinedDf$Enrolled\_Students ~ Starting.Median.Salary +   
## Mid.Career.Median.Salary + Mid.Career.10th.Percentile.Salary +   
## Mid.Career.25th.Percentile.Salary + Mid.Career.75th.Percentile.Salary +   
## Mid.Career.90th.Percentile.Salary, data = combinedDf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4007693 -614897 23967 868406 2615267   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.000e+07 1.215e+07 0.823 0.438  
## Starting.Median.Salary 1.516e+02 4.069e+02 0.373 0.721  
## Mid.Career.Median.Salary -7.330e+02 7.604e+02 -0.964 0.367  
## Mid.Career.10th.Percentile.Salary -1.632e+02 7.074e+02 -0.231 0.824  
## Mid.Career.25th.Percentile.Salary 6.587e+02 1.050e+03 0.627 0.550  
## Mid.Career.75th.Percentile.Salary 2.205e+02 3.328e+02 0.663 0.529  
## Mid.Career.90th.Percentile.Salary -6.667e+01 1.604e+02 -0.416 0.690  
##   
## Residual standard error: 2202000 on 7 degrees of freedom  
## Multiple R-squared: 0.4957, Adjusted R-squared: 0.06347   
## F-statistic: 1.147 on 6 and 7 DF, p-value: 0.4253

I created indexes to match the Degrees so I could run a linear regression on the degrees themselves.

degreeIndex = c(1,2,3,4,5,6,7,8,9,10,11,12,13, 14)  
combinedDf$degreeIndex = degreeIndex

As you can see below, there is no linear relationship, as indicated by the correlation which is -0.02, between the degrees and how many students enrolled in said degrees as predicted by the previous regression. You can also see that the data doesn’t fit the model at all, which is also indicated by the r-squared value which is 0.0005. So I fit the model against the residuals and you can see there’s a lot of dispersion between the data points and is definitely not linear, more of a cosine graph.

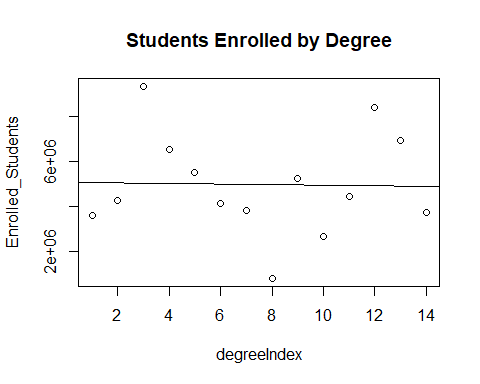
cor(combinedDf$degreeIndex, combinedDf$Enrolled\_Students)

## [1] -0.02340986

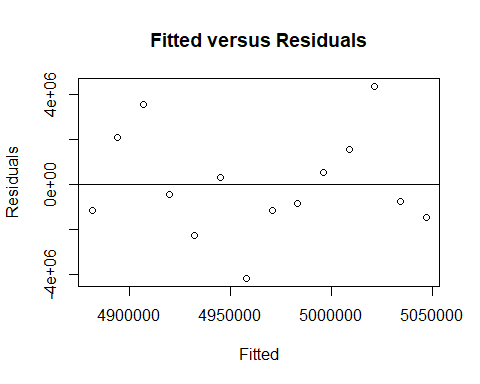
combinedDf\_fit = lm(Enrolled\_Students ~ degreeIndex, data = combinedDf)  
summary(combinedDf\_fit)

##   
## Call:  
## lm(formula = Enrolled\_Students ~ degreeIndex, data = combinedDf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4175733 -1173429 -610419 1296335 4339358   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5059702 1336624 3.785 0.0026 \*\*  
## degreeIndex -12734 156979 -0.081 0.9367   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2368000 on 12 degrees of freedom  
## Multiple R-squared: 0.000548, Adjusted R-squared: -0.08274   
## F-statistic: 0.00658 on 1 and 12 DF, p-value: 0.9367

plot(Enrolled\_Students ~ degreeIndex,  
data = combinedDf,  
main = "Students Enrolled by Degree")  
abline(combinedDf\_fit)



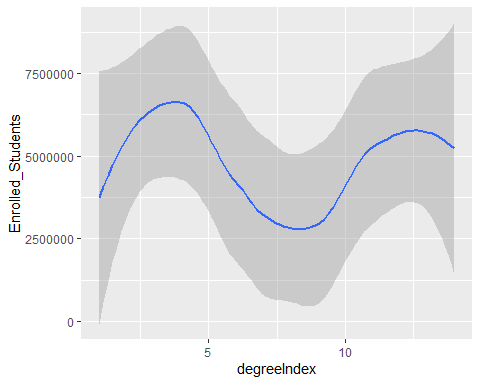
plot(fitted(combinedDf\_fit), resid(combinedDf\_fit),  
xlab = "Fitted", ylab = "Residuals", main = "Fitted versus Residuals")  
abline(h = 0)



I plotted this same variables below just to get a better representation of how the data is dispersed. As you can see, the degrees are numbered in the same order of the majors listed previously. It rises and dips depending on the degrees, and while some do make sense, some rises and dips occur around majors you wouldn’t expect, since certain majors do earn you a higher starting salary.

ggplot(combinedDf) +  
geom\_smooth(aes(x = degreeIndex, y = Enrolled\_Students))

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



## Final Thoughts

Based on what I’ve seen, I believe that occupational salary does not dictate what students major in. The multi-linear regression revealed there’s no correlation between salary and enrolled students by degrees. The bar graphs I plotted also showed that while some majors may earn you a higher salary, not all of them are fields that students seek out. Finally, the last the linear regression I ran along with the plot I displayed only further justified that there’s no linear correlation. Incoming students do not base their majors or field of study on how much they’ll make once they graduate. It could be one factor that they take into consideration but it may not necessarily be the deciding factor.