[Project] Exploratory Data Analysis

Priyank Thakkar

20/10/2021

## INTRODUCTION

This data set is all about the students who graduate from the Universities of USA. I am curious to know that does your Major in your study matter or not for your economic success. What is the general trend right then? And, what are the steps we can take before selecting the Major to boost your odds for the same. And mostly what is the share of the women in the different categories, because from the observations, their ration is very negligible in various Majors.

## DATA

This data is from the GitHub repository, from **Fivethirtyeight.com**. All data is from American Community Survey **2010-2012 Public Use Micro data series**. All Three files in the repository contains basic earnings and labor force information. “**recent-grads.csv**” contains a more detailed breakdown, including by sex and by the type of job they got. “grad-students.csv” contains details on graduate school attendees. Here, we have used the data from the file of “recent-grades.csv”. And here is the bifurcation of all the Headers and their Descriptions. Our data contains **174x21 size** of entries.

Where, we can see that 174 are rows, they represent distinct 174 majors, from all Major Categories available in the data. Moreover, all 21 columns and their representation are below.

**Header Description**

Rank: Rank by median earnings

Major\_code: Major code

Major: Major description

Major\_category: Category of major from Carnevale et al

Total: Total number of people with major

Sample\_size: Sample size (unweighted) of full-time, year-round ONLY

Men: Male graduates

Women: Female graduates

ShareWomen: Women as share of total

Employed: Number employed

Full\_time: Employed 35 hours or more

Part\_time: Employed less than 35 hours

Full\_time\_year\_round: Employed at least 50 weeks and at least 35 hours

Unemployed: Number unemployed

Unemployment\_rate: Unemployed / (Unemployed + Employed)

Median: Median earnings of full-time, year-round workers

(Normalized)

P25th: 25th percentile of earnings

P75th: 75th percentile of earnings

College\_jobs: Number with job requiring a college degree

Non\_college\_jobs: Number with job not requiring a college degree

Low\_wage\_Jobs: Number in low-wage service jobs

For our **variable study**, we may need all the variables later on for more explorations. But here are few of the variables that we can focus on.

* Total, Men, Women
* ShareWomen
* Mean
* Employed
* Unemployed
* Unemployment\_rate
* College\_jobs
* Non\_college\_jobs
* Low\_wage\_jobs

EXPLORATORY DATA ANALYSIS

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.4 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

setwd("D:\\SJSU\_HW\\GitHubSJSU\\RStudio\_Learning\\Data\_set")  
recent\_grade <- read.csv("recent\_grads.csv")  
as\_tibble(recent\_grade)

## # A tibble: 173 x 21  
## Rank Major\_code Major Total Men Women Major\_category ShareWomen  
## <int> <int> <chr> <int> <int> <int> <chr> <dbl>  
## 1 1 2419 PETROLEUM ENGIN~ 2339 2057 282 Engineering 0.121  
## 2 2 2416 MINING AND MINE~ 756 679 77 Engineering 0.102  
## 3 3 2415 METALLURGICAL E~ 856 725 131 Engineering 0.153  
## 4 4 2417 NAVAL ARCHITECT~ 1258 1123 135 Engineering 0.107  
## 5 5 2405 CHEMICAL ENGINE~ 32260 21239 11021 Engineering 0.342  
## 6 6 2418 NUCLEAR ENGINEE~ 2573 2200 373 Engineering 0.145  
## 7 7 6202 ACTUARIAL SCIEN~ 3777 2110 1667 Business 0.441  
## 8 8 5001 ASTRONOMY AND A~ 1792 832 960 Physical Scie~ 0.536  
## 9 9 2414 MECHANICAL ENGI~ 91227 80320 10907 Engineering 0.120  
## 10 10 2408 ELECTRICAL ENGI~ 81527 65511 16016 Engineering 0.196  
## # ... with 163 more rows, and 13 more variables: Sample\_size <int>,  
## # Employed <int>, Full\_time <int>, Part\_time <int>,  
## # Full\_time\_year\_round <int>, Unemployed <int>, Unemployment\_rate <dbl>,  
## # Median <int>, P25th <int>, P75th <int>, College\_jobs <int>,  
## # Non\_college\_jobs <int>, Low\_wage\_jobs <int>

# NA values  
recent\_grade %>% summarise(Total\_Count = n())

## Total\_Count  
## 1 173

filter(recent\_grade, is.na(Total) | is.na(Men) | is.na(Women) | is.na(ShareWomen)) %>%  
 summarise(Missing\_Count = n())

## Missing\_Count  
## 1 1

new\_recent\_grade <- drop\_na(recent\_grade)  
  
batch <- select(new\_recent\_grade, Major:ShareWomen, Employed, Unemployed,Unemployment\_rate, Median, College\_jobs, Non\_college\_jobs, Low\_wage\_jobs)  
#head(batch)

# 

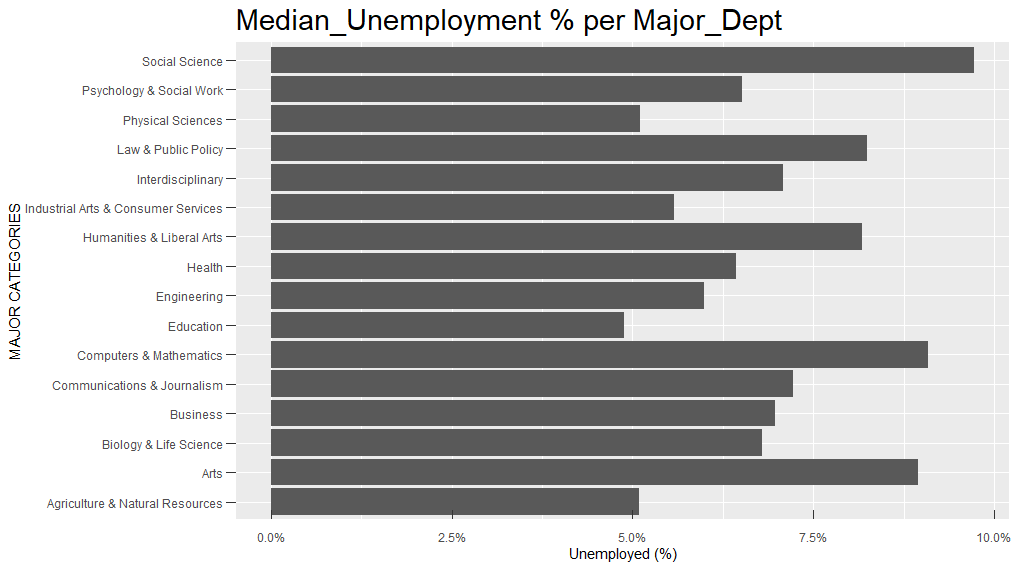
# Part 1: Unemployment Rate

From this area of the data, we will get the information about the Median Unemployment rate for department major, from all the different department, regardless of their majors. Which will give us some of the insights to predict about the Department Major itself, that which one is better over another in the view for Employment.

(Unemp <- group\_by(batch, Major\_category) %>%  
 summarise(Avg\_Unemployed\_Rate = median(Unemployment\_rate)))

## # A tibble: 16 x 2  
## Major\_category Avg\_Unemployed\_Rate  
## <chr> <dbl>  
## 1 Agriculture & Natural Resources 0.0509  
## 2 Arts 0.0895  
## 3 Biology & Life Science 0.0680  
## 4 Business 0.0697  
## 5 Communications & Journalism 0.0722  
## 6 Computers & Mathematics 0.0908  
## 7 Education 0.0488  
## 8 Engineering 0.0598  
## 9 Health 0.0643  
## 10 Humanities & Liberal Arts 0.0817  
## 11 Industrial Arts & Consumer Services 0.0557  
## 12 Interdisciplinary 0.0709  
## 13 Law & Public Policy 0.0825  
## 14 Physical Sciences 0.0511  
## 15 Psychology & Social Work 0.0651  
## 16 Social Science 0.0972

bar <- ggplot(data = Unemp)+  
 geom\_col(mapping = aes(x= Major\_category, y = Avg\_Unemployed\_Rate))  
  
bar + coord\_flip() +   
 theme(  
 legend.box.background = element\_rect(),  
 legend.box.margin = margin(6, 6, 6, 6)  
 ) +   
 labs(  
 title = "Median\_Unemployment % per Major\_Dept",  
 x = "MAJOR CATEGORIES",  
 y = "Unemployed (%)"  
 ) +  
 theme(plot.title = element\_text(size = rel(2))) +  
 theme(  
 axis.ticks.length.y = unit(.25, "cm"),  
 axis.ticks.length.x = unit(-.25, "cm"),  
 axis.text.x = element\_text(margin = margin(t = .3, unit = "cm"))  
 ) + scale\_y\_continuous(labels = scales::percent)



Part 2: Best Wages for Department Major

Here, we will explore exactly opposite that, which department has the highest paying scale in general, regardless of their major. Which ever department has the highest median income, we will draft the same for working women in that field. Just to give the support the to these upcoming data calculations.

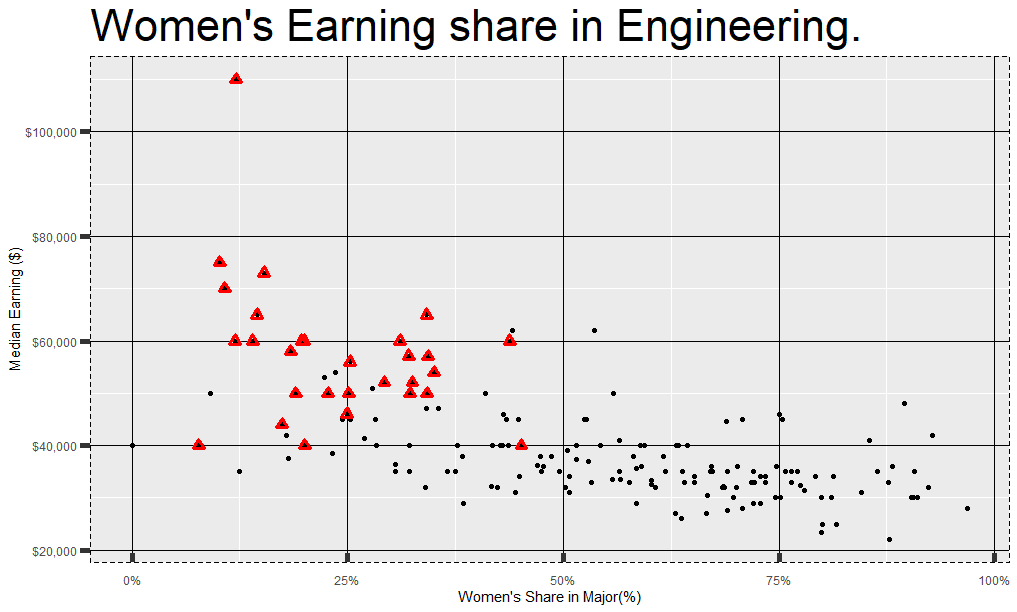
(Avg\_income <- group\_by(batch, Major\_category) %>%   
 summarise(median\_income = median(Median)))

## # A tibble: 16 x 2  
## Major\_category median\_income  
## <chr> <dbl>  
## 1 Agriculture & Natural Resources 35000  
## 2 Arts 30750  
## 3 Biology & Life Science 36300  
## 4 Business 40000  
## 5 Communications & Journalism 35000  
## 6 Computers & Mathematics 45000  
## 7 Education 32750  
## 8 Engineering 57000  
## 9 Health 35000  
## 10 Humanities & Liberal Arts 32000  
## 11 Industrial Arts & Consumer Services 35000  
## 12 Interdisciplinary 35000  
## 13 Law & Public Policy 36000  
## 14 Physical Sciences 39500  
## 15 Psychology & Social Work 30000  
## 16 Social Science 38000

This trend shows that in the 2010-2012 years, the economy is dominated by the Engineering department, with having highest amount of Median Earning. Now, we will see, the portion of the women governs in that domain with their income over all categories.

Eng <- filter(batch, Major\_category == "Engineering")  
  
w\_income <- ggplot(data = recent\_grade, mapping = aes(x = ShareWomen, y = Median)) +  
 geom\_point(shape = 20, fill = NA , size = 2, stroke = 1 ) +  
 geom\_point(data = Eng, mapping = aes(x = ShareWomen, y = Median), color = 'red', shape = 24, stroke = 2) +  
 labs(title = "Women's Earning share in Engineering.",  
 x = "Women's Share in Major(%)",  
 y = "Median Earning ($)"  
 )  
  
w\_income +   
 theme(plot.title = element\_text(size = rel(3))) +  
 theme(panel.grid.major = element\_line(colour = "black")) +  
 theme(panel.border = element\_rect(linetype = "dashed", fill = NA)) +  
 theme(axis.ticks = element\_line(size = 2)) +  
   
 theme(  
 axis.ticks.length.y = unit(.25, "cm"),  
 axis.ticks.length.x = unit(-.25, "cm"),  
 axis.text.x = element\_text(margin = margin(t = .3, unit = "cm"))  
 ) +  
   
 scale\_x\_continuous(labels = scales::percent) +  
 scale\_y\_continuous(labels = scales::dollar)

## Warning: Removed 1 rows containing missing values (geom\_point).



From the graph, it is very obvious to see that, women have peak pay/income if they are from Engineering department. Which also satisfy the general trend of domination of Engineering Department over all the departments. As, it looks more convincing when we match our previous Assignment Project proposal 1 data representation to this additional EDA, that Engineering department has the most numbers of varies majors. Moreover, they are on 2nd position with mean population of all department.

## 

## HYPOTHESIS

Mean Earning for categories from the observation, which has Median Unemployment rate grater than 8.5% are least earning of all the categories.

1. Does women earn the least in Social Science, Computer & mathematics or Art? considering it’s normally distributed between gender.
2. Does top 3 categories with highest mean earning has more College\_jobs then the rest?
3. Does non\_college jobs, and low\_wage\_jobs have direct relationship with the unemployment rate?