Project EDA

Meghna Bajoria

2022-11-03

# Introduction

The data selected for this project consists of majors choosen by recent college graduates along with the salary that these graduates earn just after graduation. The data also consists of number of graduates that were men and women, their median salary etc.

I am keen on working with this data as I will also be a graduate soon and this data will help me know about the past trends that can be expected.

# About the data

### Data Source

The data has been taken from a github repository which is maintained by Aaron Bycoff, Jay Boice, Neil Paine, Ryan Best. Citation : A.Bycoff, J.Boice, N.Paine, R.Best (Apr 3, 2018) special-elections. link : <https://github.com/fivethirtyeight/data/blob/master/college-majors/recent-grads.csv>

### Data collection

Data was collected using Ballotpedia and American Community Survey.Ballotpedia was used to compile the list of elections between Jan. 20, 2017 and March 27, 2018. Income and education data comes from the American Community Survey’s five-year estimates for 2012–2016. Presidential results by district were collected from Daily Kos Elections (Florida results are from Matthew Isbell).

### Units of observation

### Variables

The variables present in the dataset are:  
Rank - Rank by median earnings  
Major\_code - Major code, FO1DP in ACS PUMS  
xMajor - 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 (used for earnings)  
Men - Male graduates  
Women - Female graduates  
ShareWomen - Women as share of total  
Employed - Number employed (ESR == 1 or 2)  
Full\_time - Employed 35 hours or more  
Part\_time - Employed less than 35 hours  
Full\_time\_year\_round - Employed at least 50 weeks (WKW == 1) and at least 35 hours (WKHP >= 35)  
Unemployed - Number unemployed (ESR == 3)  
Unemployment\_rate - Unemployed / (Unemployed + Employed)  
Median - Median earnings of full-time, year-round workers  
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

I will be studying multiple variables like Major, Full\_time, Part\_time, Men, Women etc.

### Data cleanup

library("readxl")  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.2 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

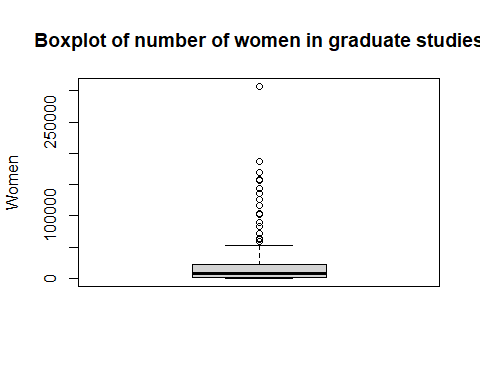
setwd("C:\\Users\\Meghna\\OneDrive\\Documents\\Fall'22\\ISO-201\\project proposal\\")  
raw\_data <- read\_excel("data1.xlsx")

sum(is.na(raw\_data))

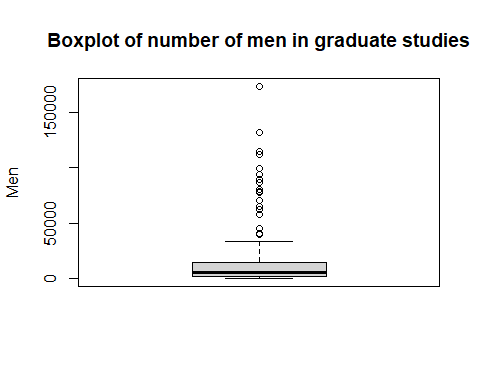
## [1] 4

Let’s identify outliers and remove them

#finding outliers in women  
women\_outliers <- boxplot(raw\_data$Women,  
 ylab = "Women",  
 main = "Boxplot of number of women in graduate studies")$out



#finding outliers in men  
men\_outliers <- boxplot(raw\_data$Men,  
 ylab = "Men",  
 main = "Boxplot of number of men in graduate studies")$out



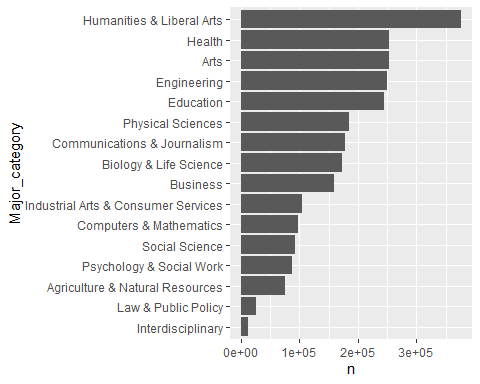
#removing women outliers  
data <- raw\_data  
data <- data[-which(data$Women %in% women\_outliers),]  
  
#removing men outliers  
data <- data[-which(data$Men %in% men\_outliers),]

# Exploratory Data Analysis

### Data visualization

Let’s find out which is the most common major

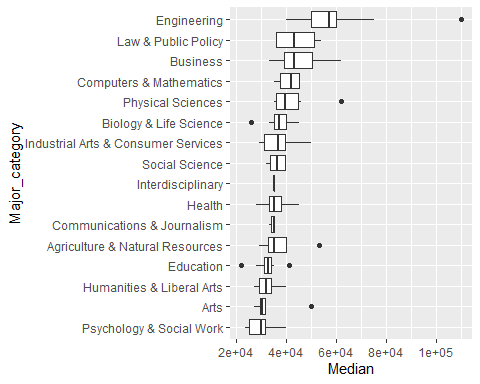
data %>%  
 count(Major\_category, wt = Total, sort = TRUE) %>%  
 mutate(Major\_category = fct\_reorder(Major\_category,n)) %>%  
 ggplot(aes(Major\_category,n)) +  
 geom\_col() +  
 coord\_flip()



We can see that Humanities & Liberal Arts is the most common major and Interdisciplinary is the least common major.

Now, let’s see which major category has the highest salary

data %>%  
 mutate(Major\_category = fct\_reorder(Major\_category, Median)) %>%  
 ggplot(aes(Major\_category,Median)) +   
 geom\_boxplot() +  
 coord\_flip()



From the above plot we can understand that  
1. Engineering students get the highest salary with median salary being around $58,000.  
2. Law & Public Policy students get second highest salary after Engineering students with median salary being $42,000.  
3. Psychology & Social work students get the lowest salary with median salary being around $30,000.

Now, let’s see which major has the highest salary

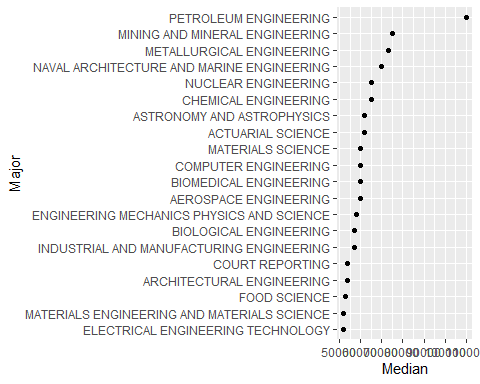
top\_majors <- head(  
 arrange(  
 data,  
 desc(data$Median)  
 ),n=3  
)  
top\_majors

## # A tibble: 3 × 21  
## Rank Major\_…¹ Major Total Men Women Major…² Share…³ Sampl…⁴ Emplo…⁵ Full\_…⁶  
## <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 1 2419 PETR… 2339 2057 282 Engine… 0.121 36 1976 1849  
## 2 2 2416 MINI… 756 679 77 Engine… 0.102 7 640 556  
## 3 3 2415 META… 856 725 131 Engine… 0.153 3 648 558  
## # … with 10 more variables: Part\_time <dbl>, Full\_time\_year\_round <dbl>,  
## # Unemployed <dbl>, Unemployment\_rate <dbl>, Median <dbl>, P25th <dbl>,  
## # P75th <dbl>, College\_jobs <dbl>, Non\_college\_jobs <dbl>,  
## # Low\_wage\_jobs <dbl>, and abbreviated variable names ¹​Major\_code,  
## # ²​Major\_category, ³​ShareWomen, ⁴​Sample\_size, ⁵​Employed, ⁶​Full\_time

We can see that Petroleum engineering has the highest median salary followed by mining and mineral engineering and then metallurgical engineering.

We can also plot this

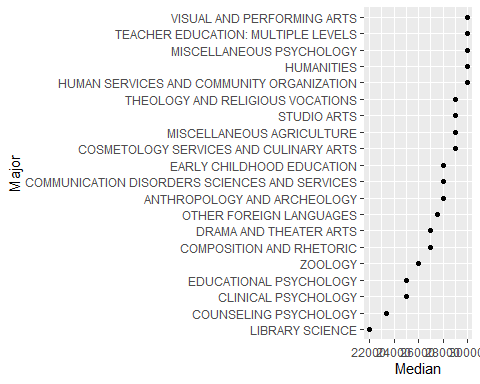
data %>%   
 arrange(desc(data$Median)) %>%  
 select(Major, Median) %>%  
 head(20) %>%  
 mutate(Major = fct\_reorder(Major,Median)) %>%  
 ggplot(aes(Major,Median)) +   
 geom\_point() +  
 coord\_flip()



From the above graph too we can see that Petroleum engineering has the highest median salary.

Now, let’s see some of the lowest earning majors

data %>%   
 arrange(desc(data$Median)) %>%  
 select(Major, Median) %>%  
 tail(20) %>%  
 mutate(Major = fct\_reorder(Major,Median)) %>%  
 ggplot(aes(Major,Median)) +   
 geom\_point() +  
 coord\_flip()



From the above graph we can conclude that library science has the lowest earning.

# Questions for next stage

1. Does engineering has more graduated men than women?
2. Does engineering jobs require most number of college degrees?