Project proposal

Meghna Bajoria

2022-10-05

# Introduction

The dataset recent-grads.xlsx contains basic earnings and labor force information of college students who are less than 28 years on the basis of sex and the type of job they are associated with. The data also contains the major these students were associated with in college.  
I really wanted to work with this data so that I can analyse and answer some questions like:  
1. What major is preffered by women and men ?  
2. Which major has the highest income ?  
3. Does any major has different salary for men and women ? This will help me analyze if there are any gender pay gap.  
4. Which major has the highest number of part time worker and full time worker ? This may help us in giving an insight about which major keeps a person most occupied.

# 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).

### Cases

The data is present as numbers, percentage and range. Each row contains basic earnings and labour force information for each type of major. It is represented more clearly by dividing the data on the basis of gender and the type of job.

### Variables

I will be studying multiple variables like:  
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

### Type of study

It is an observational study as the data has been collected without affecting the people associated with this data.

# Data quality

library("readxl")  
getwd()

## [1] "C:/Users/Meghna/OneDrive/Desktop/project proposal"

setwd("C:\\Users\\Meghna\\OneDrive\\Desktop\\project proposal")

data <- read\_excel("recent-grads.xlsx")  
data

## # A tibble: 173 × 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  
## 4 4 2417 NAVA… 1258 1123 135 Engine… 0.107 16 758 1069  
## 5 5 2405 CHEM… 32260 21239 11021 Engine… 0.342 289 25694 23170  
## 6 6 2418 NUCL… 2573 2200 373 Engine… 0.145 17 1857 2038  
## 7 7 6202 ACTU… 3777 2110 1667 Busine… 0.441 51 2912 2924  
## 8 8 5001 ASTR… 1792 832 960 Physic… 0.536 10 1526 1085  
## 9 9 2414 MECH… 91227 80320 10907 Engine… 0.120 1029 76442 71298  
## 10 10 2408 ELEC… 81527 65511 16016 Engine… 0.196 631 61928 55450  
## # … with 163 more rows, 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

dim(data)

## [1] 173 21

The data has 21 columns and 173 observations

missing <- sum(is.na(data))  
missing

## [1] 4

There are 4 missing values in the dataset and all these missing values are present in row 22 of the dataset. The column names for these missing values are Total, Men, Women and ShareWomen.

meanMen<-mean(data$Men,na.rm = TRUE)  
meanWomen<-mean(data$Women,na.rm = TRUE)  
meanTotal<-mean(data$Total,na.rm = TRUE)  
meanShareWomen<-mean(data$ShareWomen,na.rm = TRUE)  
  
data[is.na(data$Men),"Men"]<-meanMen  
data[is.na(data$Women),"Women"]<-meanWomen  
data[is.na(data$Total),"Total"]<-meanTotal  
data[is.na(data$ShareWomen),"ShareWomen"]<-meanShareWomen  
  
sum(is.na(data$Men))

## [1] 0

sum(is.na(data$Women))

## [1] 0

sum(is.na(data$Total))

## [1] 0

sum(is.na(data$ShareWomen))

## [1] 0

dim(data)

## [1] 173 21

As we can see after imputing, the data is free from any missing values and the dimension is same.

duplicated(data)

## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [169] FALSE FALSE FALSE FALSE FALSE

There are no duplicate rows in the dataset.

# References

1. In paper titled “Analysis of the Status Quo of Employment and Entrepreneurship of College Students in the New Era Based on Big Data Analysis and Discussion”, this authors first analyzes the current situation of college students’ employment and entrepreneurship in the new era based on big data analysis, and then conducts an in-depth discussion on the development path of college students in the new era of employment and entrepreneurship. With the continuous development of China’s social economy, the social demand for talent is increasing but due to the fierce competition between the markets, it also brings many problems for many college students’ employment and entrepreneurship.
2. In article “To Work or Not to Work: Student Employment, Resiliency, and Institutional Engagement of Low-Income, First-Generation College Students”, the authors do an exploratory study examines the difference between two college persistence factors—resiliency and institutional engagement—for lowincome, working, first-generation college students.
3. The article “Job Satisfaction of College Graduates with Learning Disabilities” analyses the pay range and promotion opportunities for graduates with learning disabilities. All participants graduated from a competitive midwestern university from 1987 to 1994 and represented advantaged groups when compared to both LD and non-LD populations. Data analysis indicated that the graduates with LD perceived themselves as receiving significantly less pay and promotion opportunities, and reported less total job satisfaction, than graduates without LD. However, no significant salary differences between the groups were found.
4. The study titled “The Influence of College Students’ Core Self-evaluation on Job Search Outcomes: Chain Mediating Effect of Career Exploration and Career Adaptability” was conducted to analyze the impact of core self-evaluation on job search outcomes through a chain mediation model and to discuss the role of career exploration and career adaptability in this relationship. The results indicated that core self-evaluation positively impacted job search outcomes. In addition, career exploration and career adaptability moderated the relationship between core self-evaluation and job search outcomes, respectively.
5. In article “A Comparative Analysis of Job Motivation and Career Preference of Asian Undergraduate Students” examines how various job motivators and perception toward public service affect university students’ tendencies to choose public sector jobs in a comparative context.Job security and salary are commonly important motivators for students who prefer either public or private sector jobs. Finally, the divergent characteristics of students’ career goals serve to emphasize the importance of comparative studies in identifying context-specific and context-general factors that motivate students toward public service careers.