# ISE – 201 Final Project

Meghna Bajoria

2022-11-29

Topic : Study of earnings by recent graduates

## Contents

## Introduction

The dataset recent-grads.csv 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.

Since this dataset gives insights about the earnings of a graduate students, I feel current students will get usefull insights from this analysis.

## 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

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.

### Type of study

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

### Data clean up

```
getwd()

## [1] "C:/Users/Meghna/OneDrive/Documents/Fall'22/ISO-201"

#setwd("C://Users//Meghna//OneDrive//Documents//Fall'22//ISO-201//project pro
posal")

library("readxl")
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

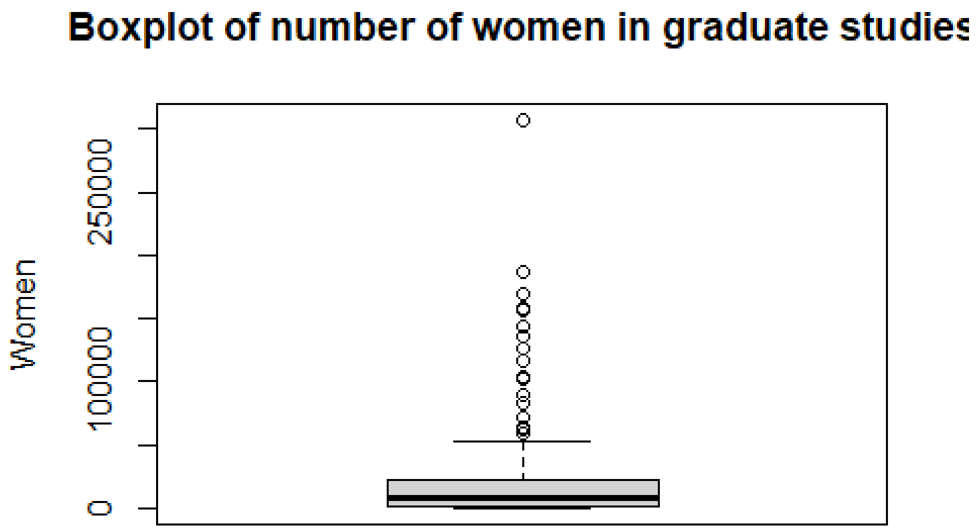
```
library(ggplot2)
library(tidyverse)

## ── Attaching packages
## ───────────────────────────────────────
## tidyverse 1.3.2 ──

## ✓ tibble  3.1.8     ✓ purrr   0.3.4
## ✓ tidyr   1.2.1     ✓ stringr 1.4.1
## ✓ readr   2.1.2     ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()    masks stats::lag()

raw_data <- read_excel("C:\\Users\\Meghna\\OneDrive\\Documents\\Fall'22\\ISO-
201\\project proposal\\data1.xlsx")

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

## [1] 4
```

There are 4 missing values in the dataset. Let’s remove it

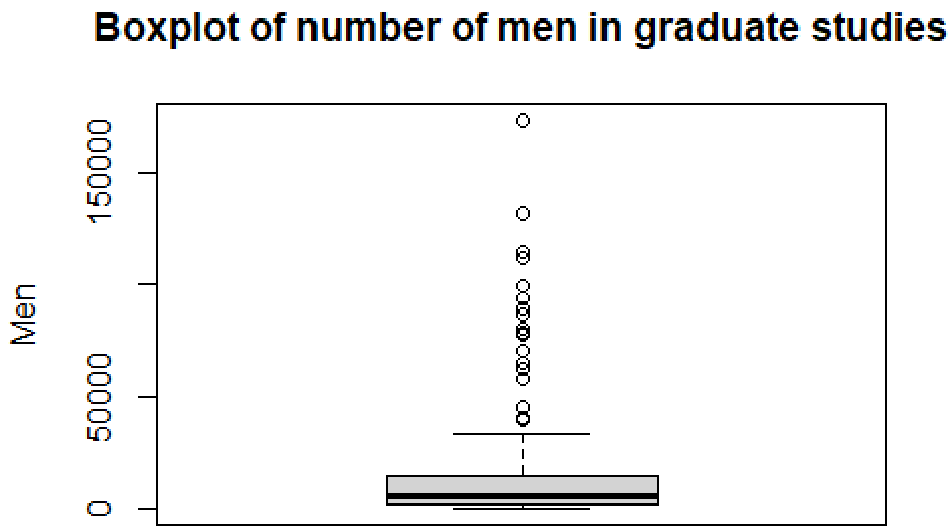
```
raw_data <- na.omit(raw_data)

sum(duplicated(raw_data))

## [1] 0
```

We have successfully removed the missing values. Now, let’s identify outliers and remove them

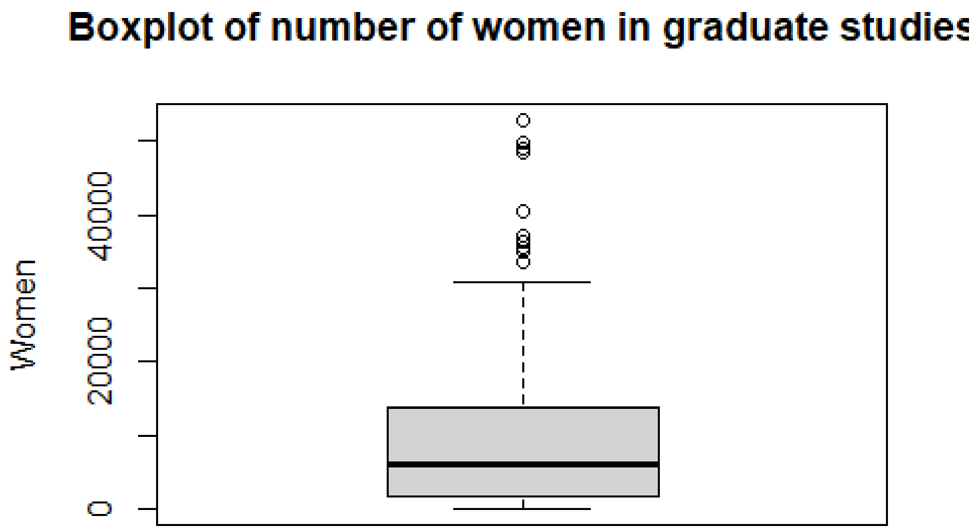
```
#finding outliers in women
women_outliers <- boxplot(raw_data$Women,
 ylab = "Women",
 main = "Boxplot of number of women in graduate studies")$out
```



```r
#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),]

boxplot(data$Women,
 ylab = "Women",
 main = "Boxplot of number of women in graduate studies")$out
```



```
##  [1] 33607 48883 35037 40300 35411 35004 49030 49654 52835 48415 37054 364
22
```

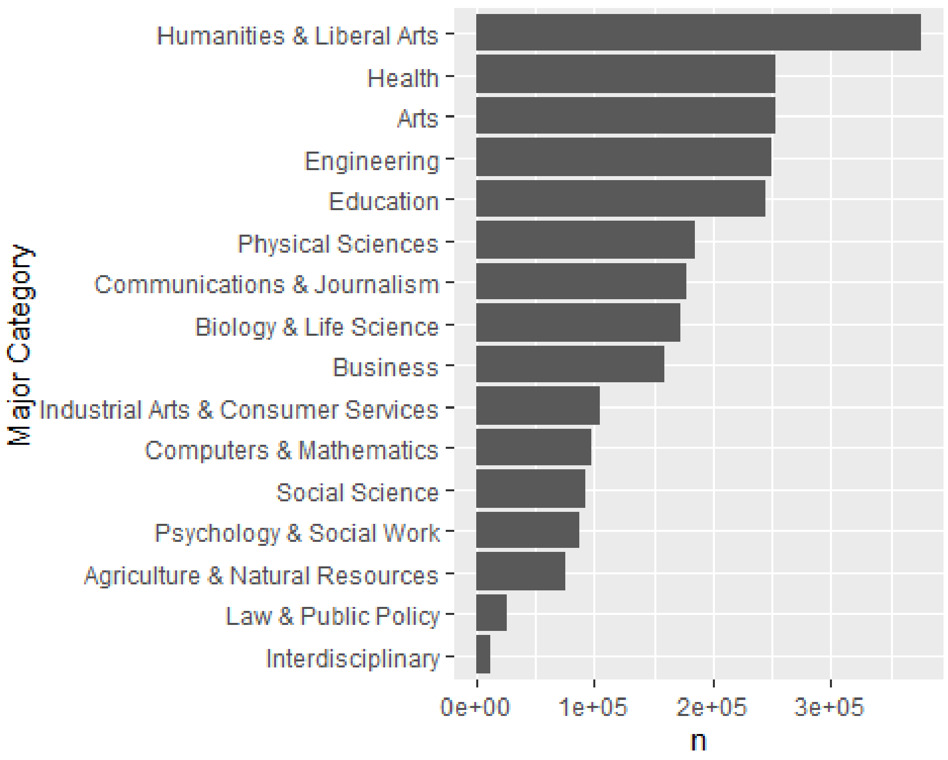
There are no duplicate rows in the dataset.

# Exploratory Data Analysis

## Data visualization

Let's find out which is the most common major

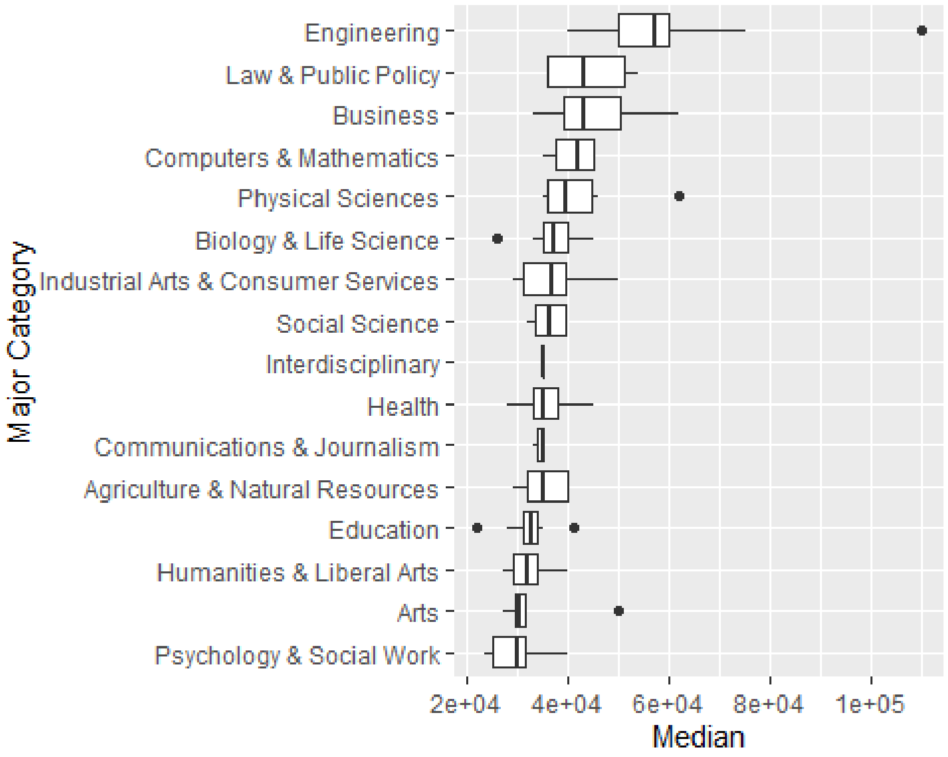
```
data %>%
  count(Major_category, wt = Total, sort = TRUE) %>%
  mutate(Major_category = fct_reorder(Major_category,n)) %>%
  ggplot(aes(Major_category,n)) +
  geom_col() +
  coord_flip() +
  xlab("Major Category")
```



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() +
  xlab("Major Category")
```



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

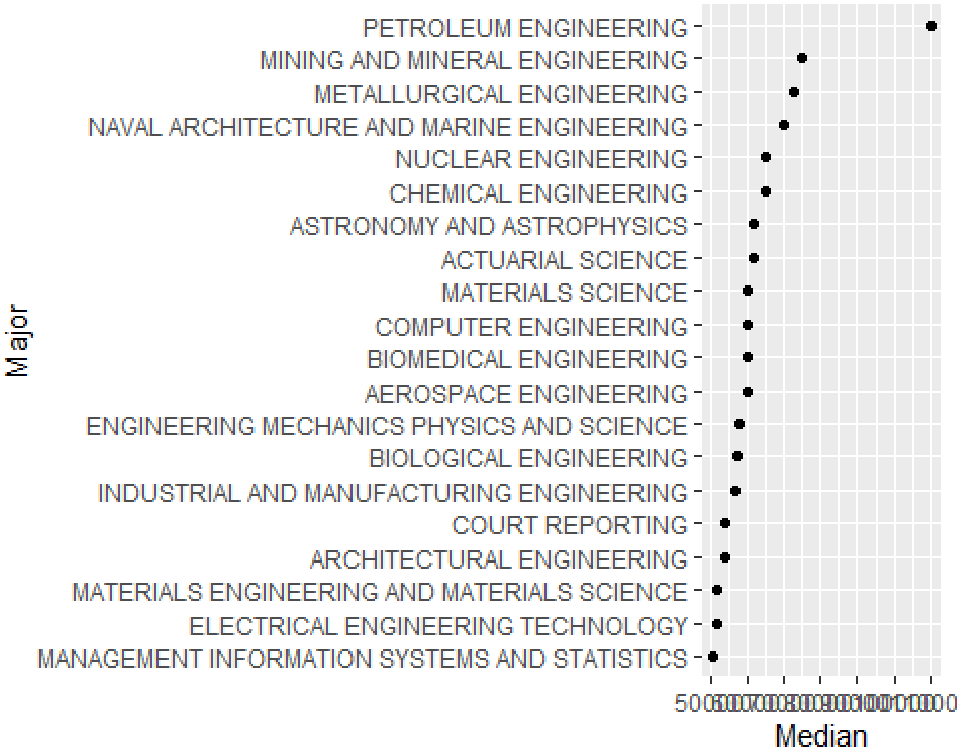
```
top_majors <- head(
  arrange(
    data,
    desc(data$Median)
  ),n=3
)
top_majors

## # A tibble: 3 × 21
##    Rank Major_…¹ Major Total   Men Women Major…² Share…³ Sampl…⁴ Emplo…⁵ F
ull_…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 ¹Major_code,
## #   ²Major_category, ³ShareWomen, ⁴Sample_size, ⁵Employed, ⁶Full_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

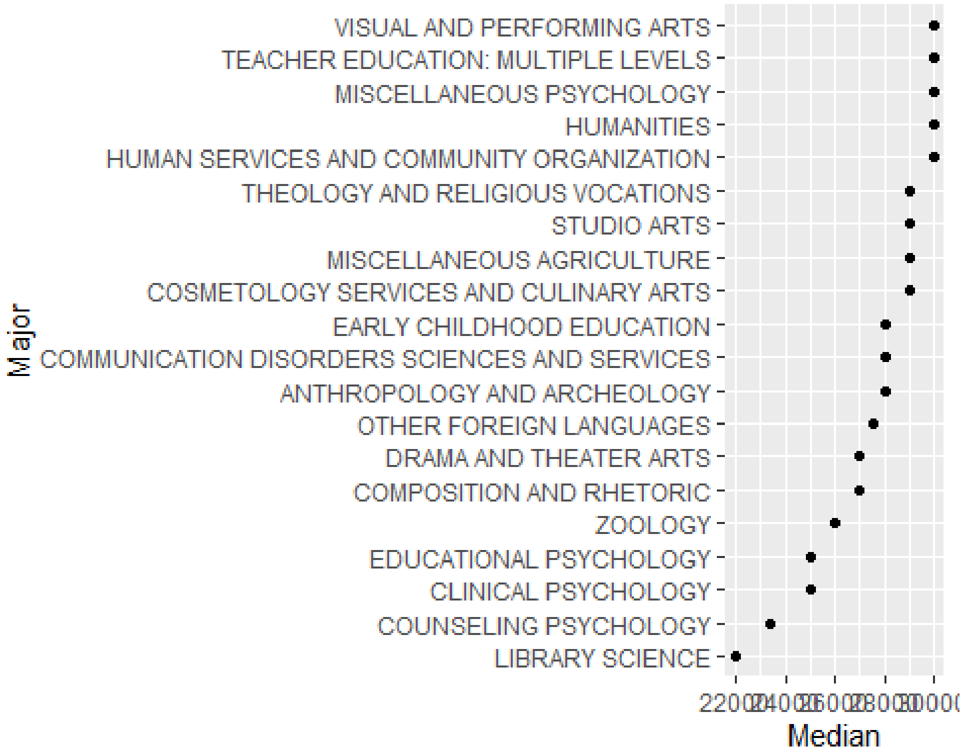
```
data %>%
   arrange(desc(data$Median)) %>%
  select(Major, Median) %>%
  head(20) %>%
  mutate(Major = fct_reorder(Major,Median)) %>%
  ggplot(aes(Major,Median)) +
  geom_point() +
  coord_flip()
```



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

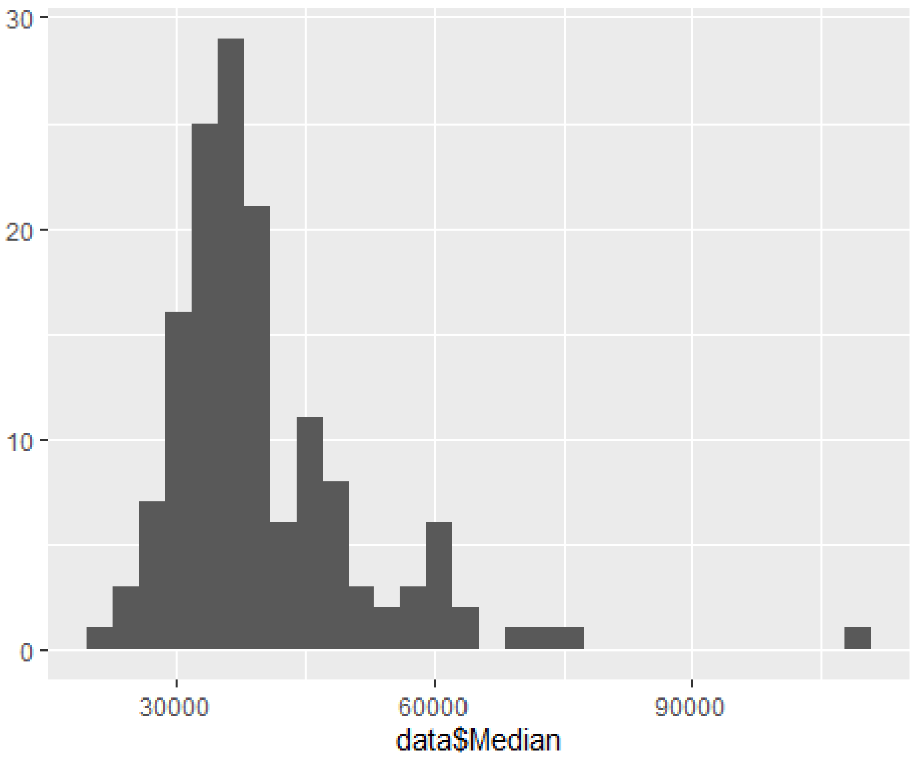
```
data %>%
   arrange(desc(data$Median)) %>%
  select(Major, Median) %>%
  tail(20) %>%
  mutate(Major = fct_reorder(Major,Median)) %>%
  ggplot(aes(Major,Median)) +
  geom_point() +
  coord_flip()
```



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

Let’s plot the median salary of all the graduate students

```
qplot(data$Median, geom = 'histogram')

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The above graph is left skewed and it appears that the mean median salary should be somewhere between 30000 to 40000. Let's try to verify it using hypothesis testing.

## Hypothesis testing

### Question 1 : An official claims that mean salary for all graduate students is $50,390 Is that possible ?

**Authoritive source :** An article by Maurie Backman in USA Today mentioned that the average starting salary for the class of 2018 is $50,390. Link : https://www.usatoday.com/story/money/careers/getting-started/2018/06/15/average-starting-salary-class-of-2018/35867859/

**Null hypothesis :** The average starting salary for class of 2018 is 50390. H0: mu = 50390

**Alternate hypothesis :** The average starting salary for class of 2018 is less than $50,390. H1: mu < 50390

**Test statistic :** Test statistic would be sample average of salary earned by graduate students irrespective of the major they belong to.

**Reference distribution :** t-distribution

```
t.test(data$Median, mu = 50390)
```

```
##
##  One Sample t-test
##
## data:  data$Median
## t = -10.785, df = 146, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 50390
## 95 percent confidence interval:
##  37945.54 41800.04
## sample estimates:
## mean of x
##  39872.79
```

**Rejection criteria :** p-value is less than 0.05 indicates rejection of null hypothesis, we will reject our null hypothesis. Also, we can observe that mean of x is equal to 39961.49 which differes from the mean in null hypothesis by 10428.51. This is a big difference and hence we will reject the null hypothesis.

## Question 2 : 1. Are there more male graduate students than women ?

**Null hypothesis :** Mean number of men graduates is greater than mean number of females graduates H0:

**Alternate hypothesis :** Mean number of men graduates is less than or equal to mean number of females graduates H1:

**Test statistic :** Test statistic would be the difference between the sample average of men and women graduates

**Reference distribution :** Welch Two Sample t-test

```
t.test(data$Men, data$Women, alternative = "less")

##
##  Welch Two Sample t-test
##
## data:  data$Men and data$Women
## t = -2.6978, df = 246.15, p-value = 0.003731
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1239.284
## sample estimates:
## mean of x mean of y
##  7155.667 10349.741
```

**Rejection criteria :** Since p-value is less than 0.05, we will reject the null hypothesis and accept the alternative hypothesis.

## PCA

Let’s find the correlation matrix. Since, the dataset has categorical values too we need to select numerical values only to find the correlation matrix.

```
num_data <- data[, sapply(data, is.numeric)]
cor <- cor(num_data)
round(cor)
```

```
##                      Rank Major_code Total Men Women ShareWomen Sample_siz
e
## Rank                    1          0     0   0     0          1
0
## Major_code              0          1     0   0     0          0
0
## Total                   0          0     1   1     1          0
1
## Men                     0          0     1   1     1          0
1
## Women                   0          0     1   1     1          1
1
## ShareWomen              1          0     0   0     1          1
0
## Sample_size             0          0     1   1     1          0
1
## Employed                0          0     1   1     1          0
1
## Full_time               0          0     1   1     1          0
1
## Part_time               0          0     1   1     1          0
1
## Full_time_year_round    0          0     1   1     1          0
1
## Unemployed              0          0     1   1     1          0
1
## Unemployment_rate       0          0     0   0     0          0
0
## Median                 -1          0     0   0     0         -1
0
## P25th                  -1          0     0   0     0          0
0
## P75th                  -1          0     0   0     0         -1
0
## College_jobs            0          0     1   1     1          0
1
## Non_college_jobs        0          0     1   1     1          0
1
## Low_wage_jobs           0          0     1   1     1          0
1
##                      Employed Full_time Part_time Full_time_year_round
## Rank                        0         0         0                    0
## Major_code                  0         0         0                    0
## Total                       1         1         1                    1
## Men                         1         1         1                    1
## Women                       1         1         1                    1
## ShareWomen                  0         0         0                    0
## Sample_size                 1         1         1                    1
## Employed                    1         1         1                    1
## Full_time                   1         1         1                    1
## Part_time                   1         1         1                    1
## Full_time_year_round        1         1         1                    1
## Unemployed                  1         1         1                    1
## Unemployment_rate           0         0         0                    0
## Median                      0         0         0                    0
## P25th                       0         0         0                    0
## P75th                       0         0         0                    0
## College_jobs                1         1         1                    1
## Non_college_jobs            1         1         1                    1
## Low_wage_jobs               1         1         1                    1
##                      Unemployed Unemployment_rate Median P25th P75th
## Rank                          0                 0     -1    -1    -1
## Major_code                    0                 0      0     0     0
## Total                         1                 0      0     0     0
## Men                           1                 0      0     0     0
## Women                         1                 0      0     0     0
## ShareWomen                    0                 0     -1     0    -1
## Sample_size                   1                 0      0     0     0
## Employed                      1                 0      0     0     0
## Full_time                     1                 0      0     0     0
## Part_time                     1                 0      0     0     0
## Full_time_year_round          1                 0      0     0     0
## Unemployed                    1                 0      0     0     0
## Unemployment_rate             0                 1      0     0     0
## Median                        0                 0      1     1     1
## P25th                         0                 0      1     1     1
## P75th                         0                 0      1     1     1
## College_jobs                  1                 0      0     0     0
## Non_college_jobs              1                 0      0     0     0
## Low_wage_jobs                 1                 0      0     0     0
##                      College_jobs Non_college_jobs Low_wage_jobs
## Rank                            0                0             0
## Major_code                      0                0             0
## Total                           1                1             1
## Men                             1                1             1
## Women                           1                1             1
## ShareWomen                      0                0             0
## Sample_size                     1                1             1
## Employed                        1                1             1
## Full_time                       1                1             1
## Part_time                       1                1             1
## Full_time_year_round              1                1              1
## Unemployed                        1                1              1
## Unemployment_rate                 0                0              0
## Median                            0                0              0
## P25th                             0                0              0
## P75th                             0                0              0
## College_jobs                      1                1              1
## Non_college_jobs                  1                1              1
## Low_wage_jobs                     1                1              1
```

Correlation indicates both the strength and direction of the linear relationship between two variables. Numbers closer to 1 indicate high correlation. For example - women, College_jobs, Non_college_jobs are highly correlated.

```
cov <- cov(num_data)
round(cov)
```

The covariance matrix indicates the direction of linear relationship between variables. From the above result, we can say that the variables which change positively with each other are

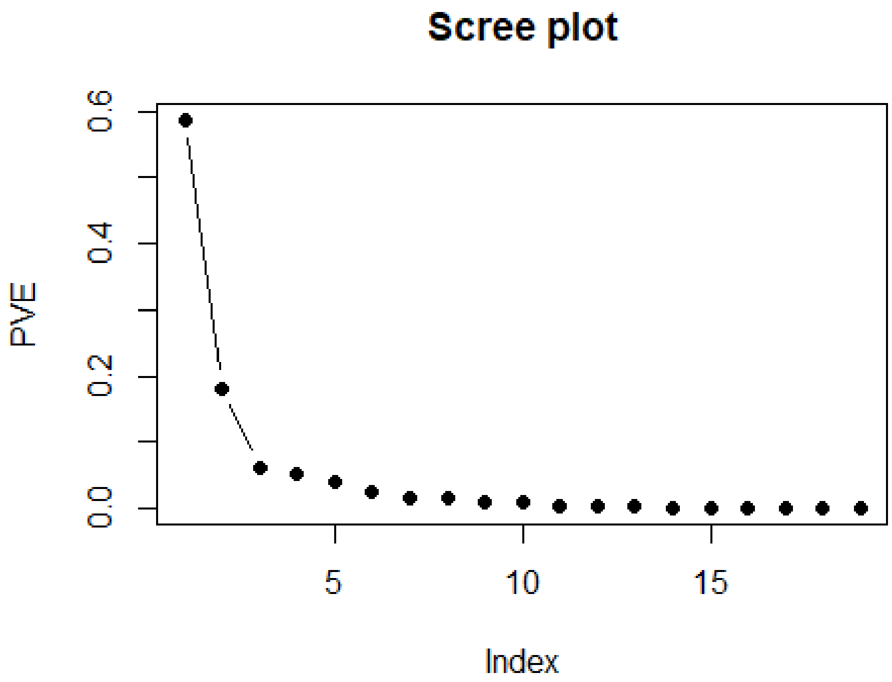
Eigen vector and values for covariance matrix

```
eigen_covariance = eigen(cov)
eigen_covariance$values

##  [1] 1.001890e+09 3.718001e+08 4.014190e+07 3.615536e+07 1.812836e+07
##  [6] 7.326324e+06 6.469552e+06 2.444670e+06 1.245000e+06 2.978059e+05
## [11] 2.447787e+05 1.483707e+05 1.071242e+05 4.125039e+04 4.878230e+02
## [16] 4.337420e+02 1.611260e-02 5.763235e-04 4.204398e-08

eigen_covariance$vectors

##                [,1]          [,2]          [,3]          [,4]          [,5
]
##  [1,] -7.614574e-04  1.847046e-03 -8.313346e-05  1.182942e-03  5.979931e-0
4
##  [2,] -6.272185e-03  7.304818e-03 -3.821667e-02  3.323412e-02  1.015240e-0
1
##  [3,] -5.395117e-01 -1.775542e-01 -4.451004e-02 -6.247434e-02  2.403964e-0
2
##  [4,] -1.779905e-01 -1.369126e-01 -2.581162e-01 -6.828157e-01 -2.374602e-0
2
##  [5,] -3.615212e-01 -4.064160e-02  2.136061e-01  6.203414e-01  4.778566e-0
2
##  [6,] -2.981153e-06  5.737103e-06  6.568175e-06  1.788422e-05 -2.304641e-

…………….

#for correlation matrix
eigen_correlation <- eigen(cor)
```

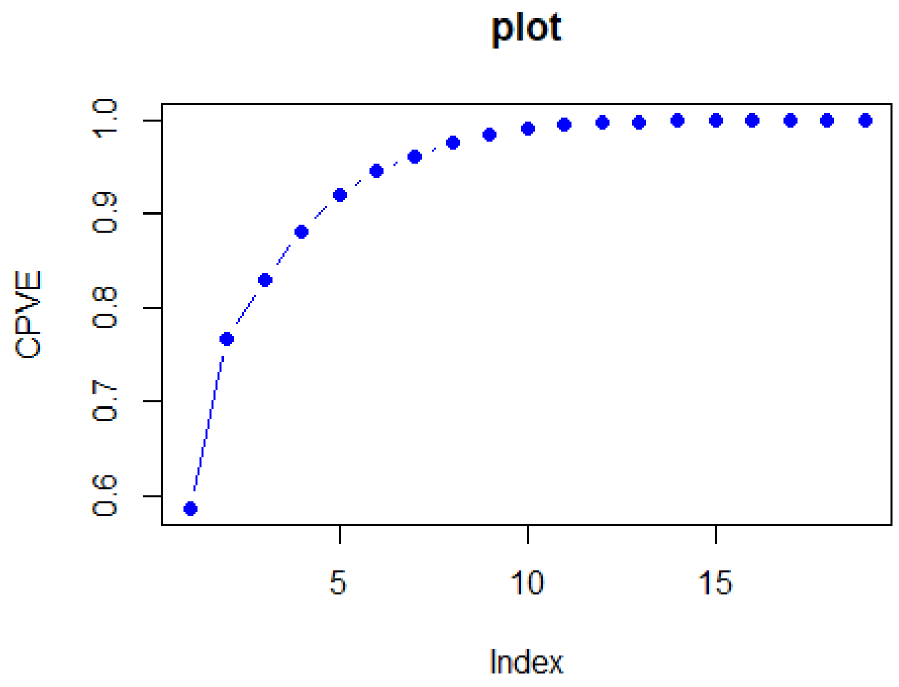
```
#Eigen values for correlation matrix are
eigen_correlation$values

##  [1] 1.115844e+01 3.442027e+00 1.181950e+00 9.836921e-01 7.282724e-01
##  [6] 4.781036e-01 2.973385e-01 2.650278e-01 1.611916e-01 1.525433e-01
## [11] 4.894005e-02 3.881609e-02 2.805157e-02 2.229135e-02 8.666067e-03
## [16] 2.660307e-03 1.618528e-03 3.660584e-04 1.100845e-15

#Eigen vectors for correlation matrix are
eigen_correlation$vectors

##              [,1]         [,2]        [,3]        [,4]         [,5]
[,6]
##  [1,] -0.14218286  0.419512204 -0.07847690 -0.07087645 -0.025134409 -0.166
95555
##  [2,] -0.04259721  0.105018688  0.61715154  0.60266129  0.369022345  0.226
52659
##  [3,] -0.29352595 -0.091884125 -0.02514566  0.03541161 -0.015452556  0.072
93408
##  [4,] -0.22435787 -0.225551344  0.04103812 -0.23190869  0.349723275  0.310
76308
##  [5,] -0.28026091  0.010139572 -0.06197982  0.19699950 -0.242567753 -0.091
05479
##  [6,] -0.11835388  0.346401782  0.04278095  0.37159556 -0.551963095  0.019
94807
##  [7,] -0.27267895 -0.118544616 -0.06412007 -0.07858987  0.043973810 -0.114
19488
##  [8,] -0.29413546 -0.089305800 -0.04835756  0.03981562 -0.018902560

........................
```

Let’s see how many principal components you would select to reduce feature dimensions yet capture atleast 85% of the variability in the data.

```
PVE = eigen_correlation$values/sum(eigen_correlation$values)
PVE

##  [1] 5.872865e-01 1.811593e-01 6.220790e-02 5.177327e-02 3.833012e-02
##  [6] 2.516335e-02 1.564939e-02 1.394883e-02 8.483770e-03 8.028593e-03
## [11] 2.575792e-03 2.042952e-03 1.476398e-03 1.173229e-03 4.561088e-04
## [16] 1.400162e-04 8.518566e-05 1.926623e-05 5.793919e-17

plot(PVE,type="b", col="black",pch=16,main = "Scree plot")
```



```
CPVE <- cumsum(PVE)
CPVE

##  [1] 0.5872865 0.7684458 0.8306537 0.8824270 0.9207571 0.9459205 0.9615699
##  [8] 0.9755187 0.9840025 0.9920311 0.9946068 0.9966498 0.9981262 0.9992994
## [15] 0.9997555 0.9998955 0.9999807 1.0000000 1.0000000

plot(CPVE, type = "b", col="blue", pch=16, main = "plot")
```



We need to find the cumulative sum to find the number of pricnipal components needed to accumulate the sum to 85%. Hence, we need 4 principal components to capture at least 85% of the variability in the data.

**Now let's try to computing principal component vectors based on the above selection.**

```
data_new <- subset(data, select=-c(Major, Major_category))
evecs = eigen_correlation$vectors[,1:4]
colnames(evecs) = c("e1", "e2", "e3", "e4")
row.names(evecs) = colnames(data_new)
evecs
```

```
##                             e1           e2          e3          e4
## Rank                -0.14218286  0.419512204 -0.07847690 -0.07087645
## Major_code          -0.04259721  0.105018688  0.61715154  0.60266129
## Total               -0.29352595 -0.091884125 -0.02514566  0.03541161
## Men                 -0.22435787 -0.225551344  0.04103812 -0.23190869
## Women               -0.28026091  0.010139572 -0.06197982  0.19699950
## ShareWomen          -0.11835388  0.346401782  0.04278095  0.37159556
## Sample_size         -0.27267895 -0.118544616 -0.06412007 -0.07858987
## Employed            -0.29413546 -0.089305800 -0.04835756  0.03981562
## Full_time           -0.28840458 -0.112574163 -0.07755282  0.01659251
## Part_time           -0.28370831 -0.023789738  0.06216783  0.06471433
## Full_time_year_round -0.28519334 -0.118303322 -0.08277668  0.02239769
## Unemployed          -0.28219895 -0.087934624  0.14734603 -0.12399312
## Unemployment_rate   -0.05074673  0.009225210  0.67849125 -0.53011680
## Median                0.13430260 -0.458260040  0.04266182  0.15718073
## P25th                 0.12083253 -0.409001406  0.01317639  0.17344626
## P75th                 0.11798801 -0.432069542  0.13357669  0.14992450
## College_jobs         -0.24221332 -0.105852410 -0.23789929  0.11226348
## Non_college_jobs     -0.28345204 -0.036192053  0.10246124 -0.01742699
## Low_wage_jobs        -0.27434783 -0.001963236  0.11315534  0.02345216
```

```
PC1 <- as.matrix(data_new) %*% evecs[,1]
PC2 <- as.matrix(data_new) %*% evecs[,2]
PC3 <- as.matrix(data_new) %*% evecs[,3]
PC4 <- as.matrix(data_new) %*% evecs[,4]
PC <- data.frame(PC1, PC2, PC3, PC4)
head(PC)
```

```
##         PC1        PC2       PC3      PC4
## 1  37587.34 -144411.23 23517.858 53943.83
## 2  26133.64  -95942.52 17306.471 36240.94
## 3  27028.61  -99476.75 19106.232 37329.01
## 4  22329.57  -84681.37 15458.468 31846.39
## 5 -20788.91  -99314.30  6647.701 33858.59
## 6  23079.41  -95535.71 18080.764 35521.36
```

- e1 values seems to represent the relation between the Median earnings of full-time year-round workers, 25th percentile of earnings and 75th percentile of earnings.

- e2 values seems to represent the relation between women in various majors to the unemployment rate.

- e3 has a mix of positive and negative values which seems to show contrast between Men and women earnings.

- e4 values seems to represent relation between men and unemployment rate among non college jobs.

# Conclusion

## Brief summary

This dataset contains information about the earnings of students from different majors after they graduate. The dataset has around 172 observations and 21 variables. During data analysis, I came across several interesting findings like petroleum engineering has the highest median salary which makes sense as the use of fuel and the need of specialized skills in the domain is increasing day by day. It was suprising to see that Computer engineering does not have the highest salary as it is usually considered as the highest paying major. We also got to know that library science has the lowest earning among all the graduate courses.

### My learnings

Finally, I also learned alot about my research question. I got to know about the most popular and highest earning major. I also learned that it is not necessary to have more men than women in graduate studies.

### Limitations

It is important to understand the difference in earning by various majors. However, there are several other factors too which impact the earnings like location of the job, the economic condition of the country, the type of job the students took after graduation and if the job that they took is related to their major or not. Unless we have more extensive data which is capable of providing more useful insights, it is difficult to come to any conclusion.

### Future scope

The future scope of this study includes gathering more data which can affect the earnings made by students and then perform an even exhaustive study.

## References

1. https://rpubs.com/
2. https://statisticsbyjim.com/regression/identifying-important-independent-variables/
3. https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/
4. https://www.graduatetutor.com/statistics-tutor/principal-component-analysis-pca-tutoring/
5. https://rpubs.com/Xns140/Grads