College Data

Code

##This data contain earnings and information of college graduate based on their fields of study, major. In this project, I will try to do some exploratory analysis.I will try do a prediction on the salary if the time allow me

let’s start by loading this below packages and our data

Hide

```
library(tidyverse)
library(ggplot2)

college_grad <- read.csv("C:/Users/Amara Diallo/Desktop/college_data.txt")
```

##Let’s convert our column in capital letter

Hide

```
college_grad <- college_grad %>%
  set_names(names(.) %>%
              str_to_title())


glimpse(college_grad)
```

```
Rows: 173
Columns: 21
$ Rank                 [3m[38;5;246m<int>[39m[23m 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,
14, 15, 16...
$ Major_code           [3m[38;5;246m<int>[39m[23m 2419, 2416, 2415, 2417, 2405, 2418, 6202,
5001, 2414,...
$ Major                [3m[38;5;246m<fct>[39m[23m PETROLEUM ENGINEERING, MINING AND MINERAL
ENGINEERING...
$ Total                [3m[38;5;246m<int>[39m[23m 2339, 756, 856, 1258, 32260, 2573, 3777, 1
792, 91227,...
$ Men                  [3m[38;5;246m<int>[39m[23m 2057, 679, 725, 1123, 21239, 2200, 2110, 8
32, 80320, ...
$ Women                [3m[38;5;246m<int>[39m[23m 282, 77, 131, 135, 11021, 373, 1667, 960,
10907, 1601...
$ Major_category       [3m[38;5;246m<fct>[39m[23m Engineering, Engineering, Engineering, Eng
ineering, E...
$ Sharewomen           [3m[38;5;246m<dbl>[39m[23m 0.12056434, 0.10185185, 0.15303738, 0.1073
1320, 0.341...
$ Sample_size          [3m[38;5;246m<int>[39m[23m 36, 7, 3, 16, 289, 17, 51, 10, 1029, 631,
399, 147, 7...
$ Employed             [3m[38;5;246m<int>[39m[23m 1976, 640, 648, 758, 25694, 1857, 2912, 15
26, 76442, ...
$ Full_time            [3m[38;5;246m<int>[39m[23m 1849, 556, 558, 1069, 23170, 2038, 2924, 1
085, 71298,...
$ Part_time            [3m[38;5;246m<int>[39m[23m 270, 170, 133, 150, 5180, 264, 296, 553, 1
3101, 12695...
$ Full_time_year_round [3m[38;5;246m<int>[39m[23m 1207, 388, 340, 692, 16697, 1449, 2482, 82
7, 54639, 4...
$ Unemployed           [3m[38;5;246m<int>[39m[23m 37, 85, 16, 40, 1672, 400, 308, 33, 4650,
3895, 2275,...
$ Unemployment_rate    [3m[38;5;246m<dbl>[39m[23m 0.018380527, 0.117241379, 0.024096386, 0.0
50125313, 0...
$ Median               [3m[38;5;246m<int>[39m[23m 110000, 75000, 73000, 70000, 65000, 65000,
62000, 620...
$ P25th                [3m[38;5;246m<int>[39m[23m 95000, 55000, 50000, 43000, 50000, 50000,
53000, 3150...
$ P75th                [3m[38;5;246m<int>[39m[23m 125000, 90000, 105000, 80000, 75000, 10200
0, 72000, 1...
$ College_jobs         [3m[38;5;246m<int>[39m[23m 1534, 350, 456, 529, 18314, 1142, 1768, 97
2, 52844, 4...
$ Non_college_jobs     [3m[38;5;246m<int>[39m[23m 364, 257, 176, 102, 4440, 657, 314, 500, 1
6384, 10874...
$ Low_wage_jobs        [3m[38;5;246m<int>[39m[23m 193, 50, 0, 0, 972, 244, 259, 220, 3253, 3
170, 980, 3...
```

Hide

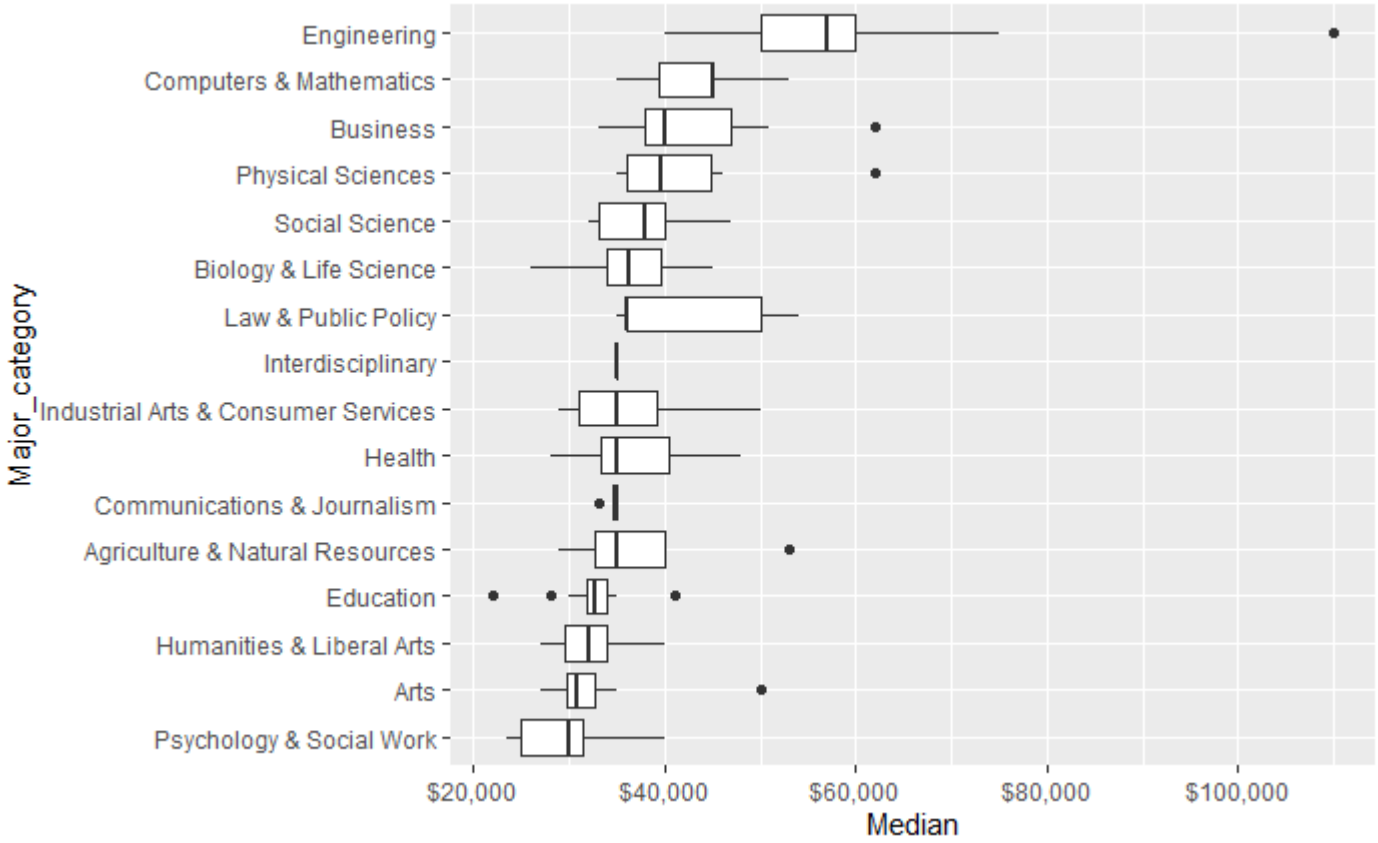
```
#view(college_grad)
```

Here we will look for the Major categories that make more money upon graduation

In this section I will do a visualizations of the data in order to find out what major category is leading in term of salary in the job market.

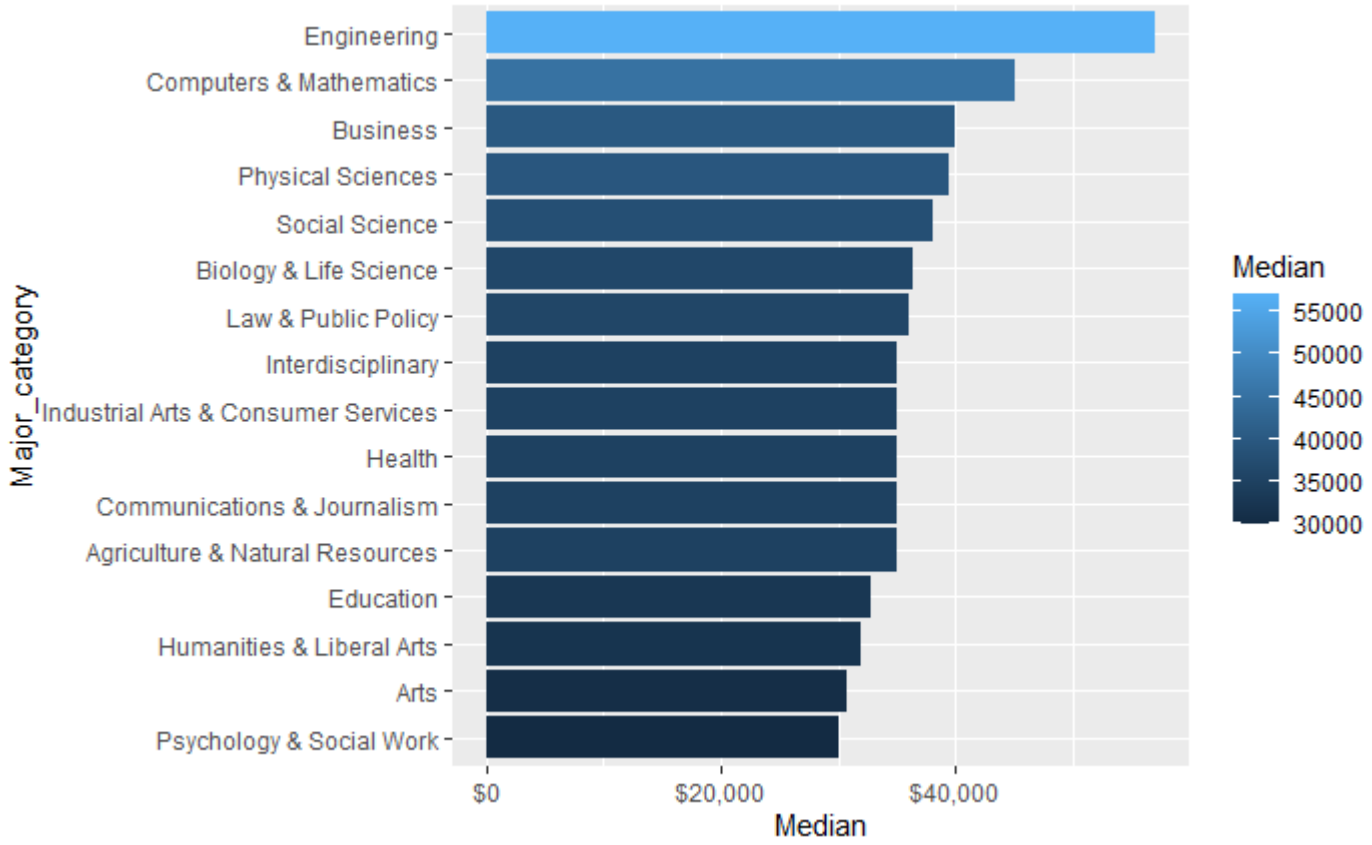
Hide

```
college_grad %>%
    mutate(Major_category = fct_reorder(Major_category, Median)) %>%
    ggplot(aes(Major_category, Median)) +
    geom_boxplot() +
    scale_y_continuous(labels = scales::dollar_format()) +
    coord_flip()
```



Hide

```
college_grad %>%
    group_by(Major_category) %>%
    summarize(Median = median(Median)) %>%
    mutate(Major_category = fct_reorder(Major_category, Median)) %>%
    ggplot(aes(Major_category, Median, fill =Median )) +
    geom_bar(stat="identity") +
    scale_y_continuous(labels = scales::dollar_format()) +
    coord_flip()
```



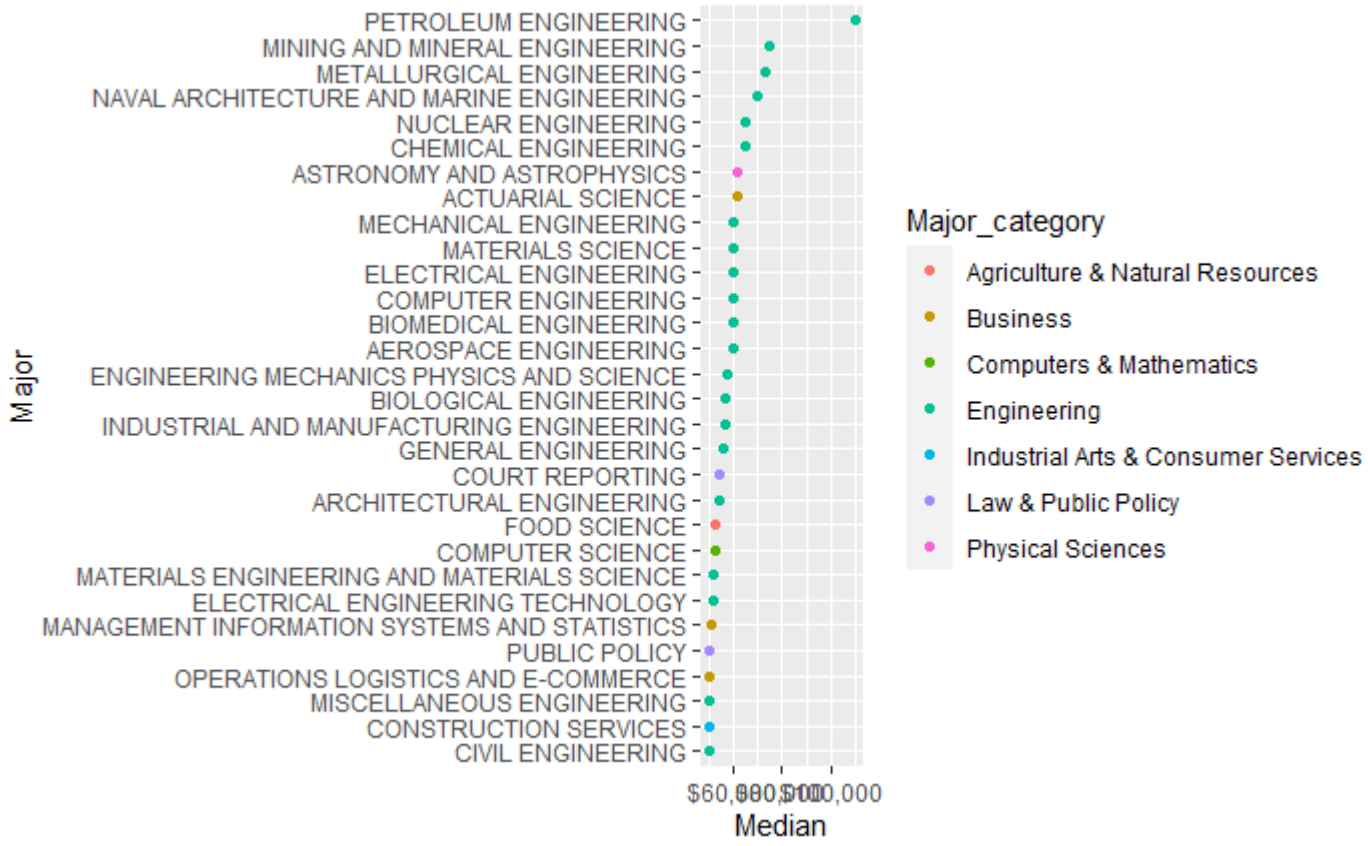
As we can see above, Engineering student are the one who get more money after graduation and follow by computer&Math students. With Arts & Journalism as the lowest paying job. This finding is based on the current data we have, I am also assuming that this is probably for junior position. But, on the first graph we could point out an outlier, which mean that there is a field in the **Engineering Major** that makes a lot of money than the other field in the ***Engineering***

In this section, we will find the highest top earning majors in all major category.

This will help us extrapolate what is the highest major amont all major category, but it will also allow us to find out what the *OUTLIER IN ENGINEERING* we found in our early graph: The field that makes more money than all other **ENGINEERING** fields.

Hide

```
Majors <- college_grad %>%
    arrange(desc(Median)) %>%
    select(Major, Major_category, Median, P25th, P75th, Sample_size)  %>%
    mutate(Major = fct_reorder(Major, Median))

Majors %>% head(30) %>%
    ggplot(aes(Major, Median, color = Major_category)) +
    geom_point() +
    scale_y_continuous(labels = scales::dollar_format()) +
    coord_flip()
```



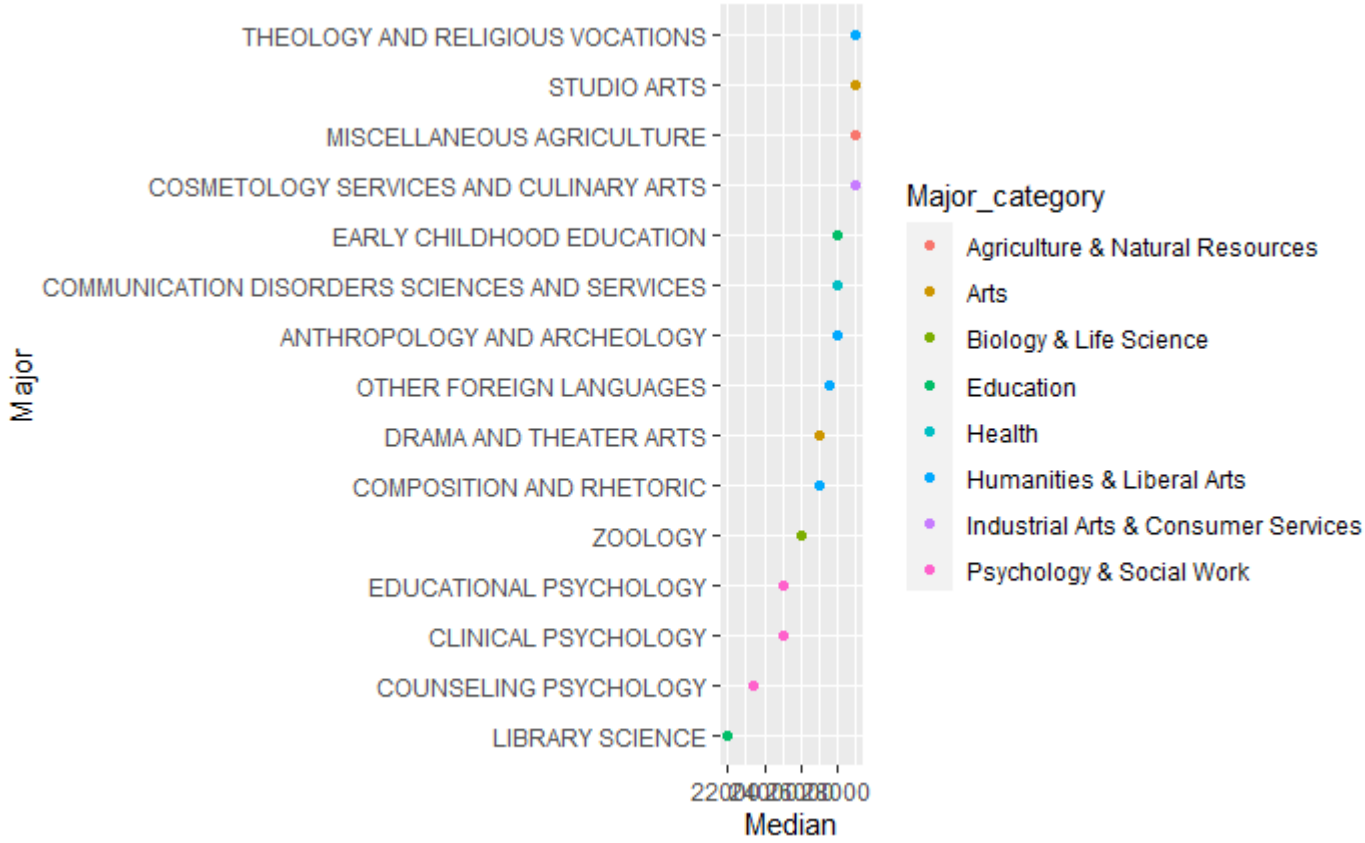
Most the highest earning majors are from **ENGINEERING** field. We also realized that *PETROLEUM ENGINEERING* is not only the highest paying position in **ENGINEERING** field, but it is also the highest paying job in all major category…according to this dataset.

ASTRONOMY & ASTROPHYSICS is the second highest paying job, coming from the *COMPUTER & MATH* Major category; **ACTUARIAL SCIENCE** is the third and the first in the Business department;

The lowest earning Majors

Hide

```
college_grad %>%
    select(Major, Major_category, Median, P25th, P75th)  %>%
    tail(15) %>%

    mutate(Major = fct_reorder(Major, Median)) %>%
    ggplot(aes(Major, Median,color = Major_category)) +
    geom_point() +
    coord_flip()
```



The above graph shows the top 15 lowest paying job ** according to this dataset. ***Library Science*** **is the lowest paying job. As we notice, no field in the** ENGINEERING** is present in this list.

Hide

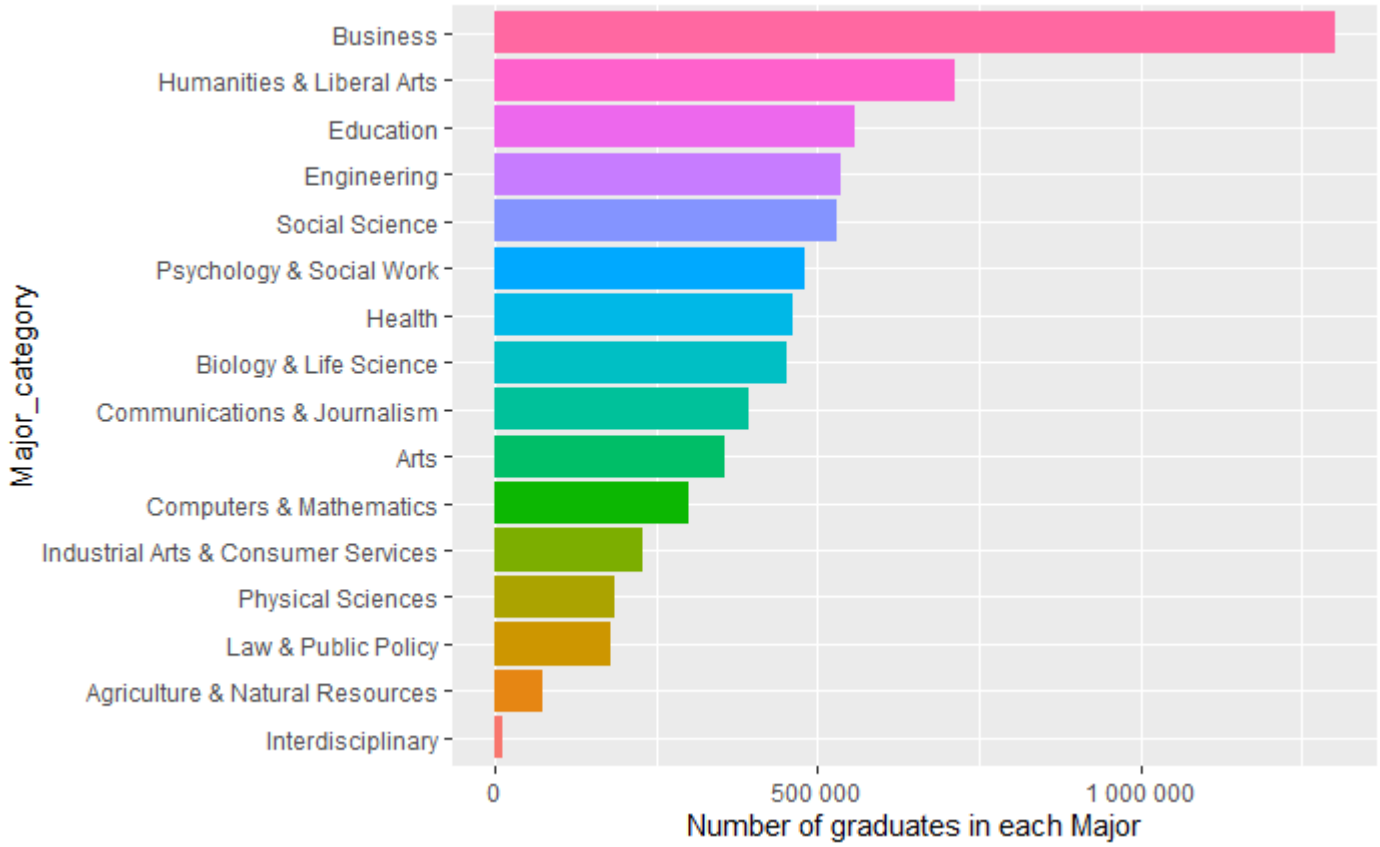
```
# Majors  %>%
#     ggplot(aes(Sample_size, Median)) +
#     geom_point() +
#     scale_x_log10()
# install.packages("tm")           # for text mining
# install.packages("SnowballC")    # for text stemming
# install.packages("wordcloud")    # word-cloud generator
# install.packages("RColorBrewer") # color palettes

# Load the packages
# library("tm")
# library("SnowballC")
# library("wordcloud")
# library("RColorBrewer")
# wordcloud(words = Majors$Major_category,
#           freq = Majors$Median,
#           min.freq = 1,
#           max.words =200,
#           random.order = TRUE,
#           rot.per = 0.35,
#           colors = brewer.pal(8, "Dark2"))
```

Most common majors :

This part will tell us what is the major that attact most of students. We are not surprise to see that **BUSINESS** is by far the common major for college students. It is twice attractive than the rest of the major… Specially Engineering.

Hide

```
college_grad %>%
    count(Major_category, wt = Total, sort = TRUE) %>%
    mutate(Major_category = fct_reorder(Major_category,n)) %>%
    ggplot(aes(Major_category, n, fill = Major_category)) +
    theme(legend.position = "none") +
    geom_col() +
    coord_flip() +
    scale_y_continuous(labels = scales::number_format()) +
    labs(y = "Number of graduates in each Major")
```



Let's see if we can find the number of graduate in each major category, but based on **gender**.

Hide

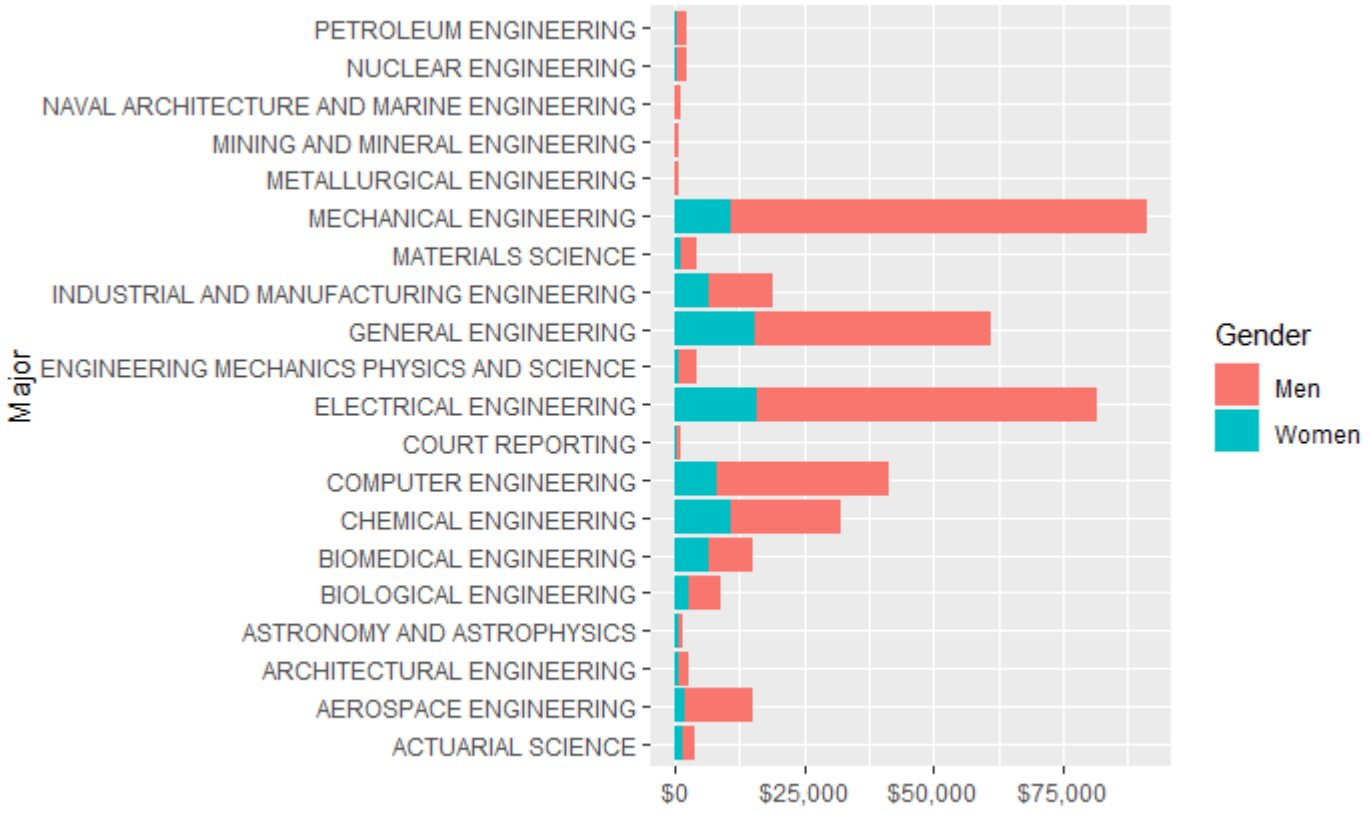
```
college_grad %>%
    mutate(Major_category = fct_reorder(Major_category, Total)) %>%
    gather(Gender,Total, Men, Women) %>%
    group_by(Major_category, Gender) %>%
    #summarize(Median = median(Median)) %>%

    ggplot(aes(Major_category,Total, fill = Gender)) +
    geom_col() +
    facet_grid(~Gender) +
    coord_flip() +
    scale_y_continuous(labels = scales::number_format()) +
    labs(y = "Total  number of graduates by gender")
```

It looks like *MEN* are more STERM oriented than *WOMEN*. As we can, most of the graduate students in Engineering, Computer & MATH are *MEN*. However, it is undeniable that women are ahead in health and social science related majors. Business Major is dominated by both gender.

which gender earn more money based on top 15 Major

Hide

```
college_grad %>%
mutate(Major_category = fct_reorder(Major_category, Median)) %>%
  top_n(20, Median) %>%

gather(Gender,Median, Men, Women) %>%
group_by(Major_category, Gender) %>%

  ggplot(aes(Major, Median, fill = Gender)) +
geom_col() +
scale_y_continuous(labels = scales::dollar_format()) +
labs(y = " ") +

coord_flip()
```

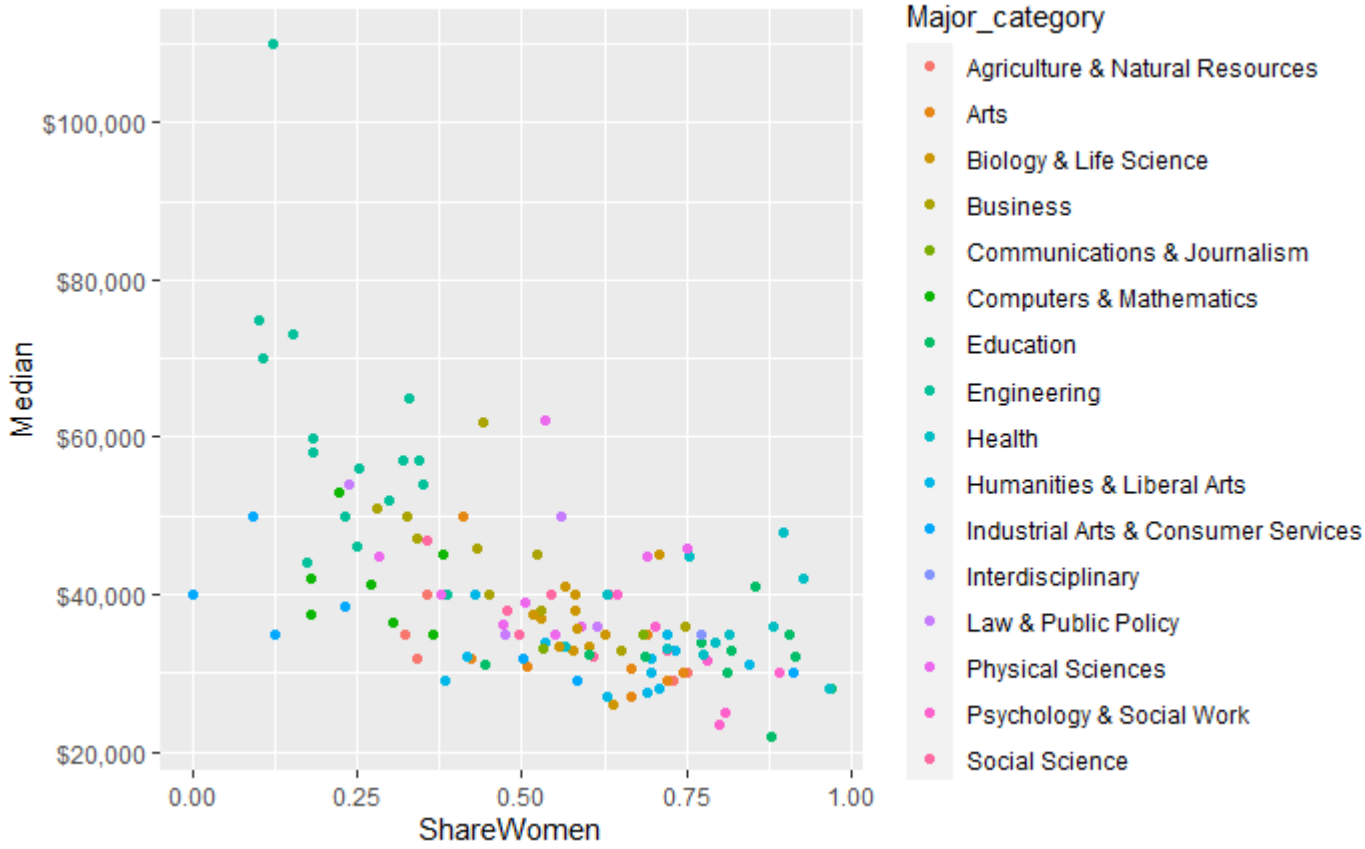


This above graph tells us that Men are pay more than women even though they are in the same major. this is just a dataset about recent graduate, so we can not totally rely on it. In order to really know the ins and out of this above question, we will need more dataset in order to get a better insight.

Share of Women in each Major

Hide

```
college_grad %>%
    group_by(Major_category, Median) %>%
    summarize_at(vars(Total, Men, Women), sum, na.rm=TRUE) %>%
    mutate(ShareWomen = Women/Total) %>%
    arrange(desc(ShareWomen)) %>%

    ggplot(aes(ShareWomen, Median, color = Major_category)) +
    geom_jitter() +
    scale_y_continuous(labels = scales::dollar_format())
```



The previous plot shows that less .25% of women make more than 50k,that is due to the fact that most of those women that are in the .25% are in the high paying major like STEM. More than .75% are below 45K, that can be explain in their major choise…like **SOCIAL SCIENCE, PSYCHOLOGY, EDUCATION, JOURNALISM ect**

Total Major in each major's category

Hide

```
college_grad %>%
    select(Major, Major_category, Total, Sharewomen, Sample_size, Median) %>%
    add_count(Major_category) %>%
    filter(n >= 10) %>%
    count(Major_category) %>%
    arrange(desc(n))
```

Major_category <fctr>	**n** <int>
Engineering	29
Education	16
Humanities & Liberal Arts	15
Biology & Life Science	14

Major_category <fctr>	n <int>
Business	13
Health	12
Computers & Mathematics	11
Agriculture & Natural Resources	10
Physical Sciences	10

9 rows

Let’s see how much money Women get in health care profession:

Hide

```
college_grad %>% filter(Major_category == 'Health' & Sharewomen >0.7) %>%
  select(Major, Median)
```

Major <fctr>	Median <int>
NURSING	48000
MEDICAL TECHNOLOGIES TECHNICIANS	45000
MEDICAL ASSISTING SERVICES	42000
MISCELLANEOUS HEALTH MEDICAL PROFESSIONS	36000
NUTRITION SCIENCES	35000
HEALTH AND MEDICAL ADMINISTRATIVE SERVICES	35000
COMMUNITY AND PUBLIC HEALTH	34000
TREATMENT THERAPY PROFESSIONS	33000
GENERAL MEDICAL AND HEALTH SERVICES	32400
COMMUNICATION DISORDERS SCIENCES AND SERVICES	28000

1-10 of 10 rows

Hide

```
#View(fresh_grad_health)
```

Majors with the lowest unemployment rate that might be interesting for students.

Hide

```
college_grad %>%
  select(Major, Unemployment_rate) %>%
  filter(Unemployment_rate<0.02)
```

Major <fctr>	Unemployment_rate <dbl>
PETROLEUM ENGINEERING	0.018380527
ENGINEERING MECHANICS PHYSICS AND SCIENCE	0.006334343
COURT REPORTING	0.011689692
MATHEMATICS AND COMPUTER SCIENCE	0.000000000
GENERAL AGRICULTURE	0.019642463
MILITARY TECHNOLOGIES	0.000000000
BOTANY	0.000000000
SOIL SCIENCE	0.000000000
MATHEMATICS TEACHER EDUCATION	0.016202835
EDUCATIONAL ADMINISTRATION AND SUPERVISION	0.000000000

1-10 of 10 rows

Hide

```
#View(fresh_grads_science)
```

Recent graduates with median salary above 40,000 USD where Women are represented by More than 50 percent.

Hide

```
college_grad %>% filter(Median >=40000 & Sharewomen >.5) %>%
  select(Major, Median)
```

Major <fctr>	Median <int>
ASTRONOMY AND ASTROPHYSICS	62000
PUBLIC POLICY	50000
NURSING	48000
NUCLEAR, INDUSTRIAL RADIOLOGY, AND BIOLOGICAL TECHNOLOGIES	46000
ACCOUNTING	45000
MEDICAL TECHNOLOGIES TECHNICIANS	45000

Major <fctr>	Median <int>
STATISTICS AND DECISION SCIENCE	45000
PHARMACOLOGY	45000
OCEANOGRAPHY	44700
MEDICAL ASSISTING SERVICES	42000

1-10 of 19 rows Previous **1** 2 Next

Hide

NA

CONCLUSION :

The above analysis hels us to understand how our major’s choice can have an impact on our financial status. this analysis also touch upon the choice of majors based on gender. For instance, we saw that *Men* are more attracted to STEM option, while *WOMEN* dominate health science profession.

Predict the median of ShareWomen