Project Proposal 1 - [R]EDA

## INTRODUCTION

This data is from the GitHub repository, from **Fivethirtyeight.com**. All data is from American Community Survey 2010 -2012 Public Use Microdata 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-studets.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.

**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

Just to start with our Assignment we need one basic Library, which is “tidyverse” in our case.

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()

## DATABASE INFORMATION

For this Assignment our database has different form and we have to convert it to tibble form before we use it, and for this Assignment we are going to make our Working Directory.

setwd("D:\\SJSU\_HW\\201\\Data\_sets\\college\_majors")  
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>

## INFORMATION OF THE MISSING VALUES.

Now, here I performing the checks to determine the quality of my “storms” data.

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

## Total\_Count  
## 1 173

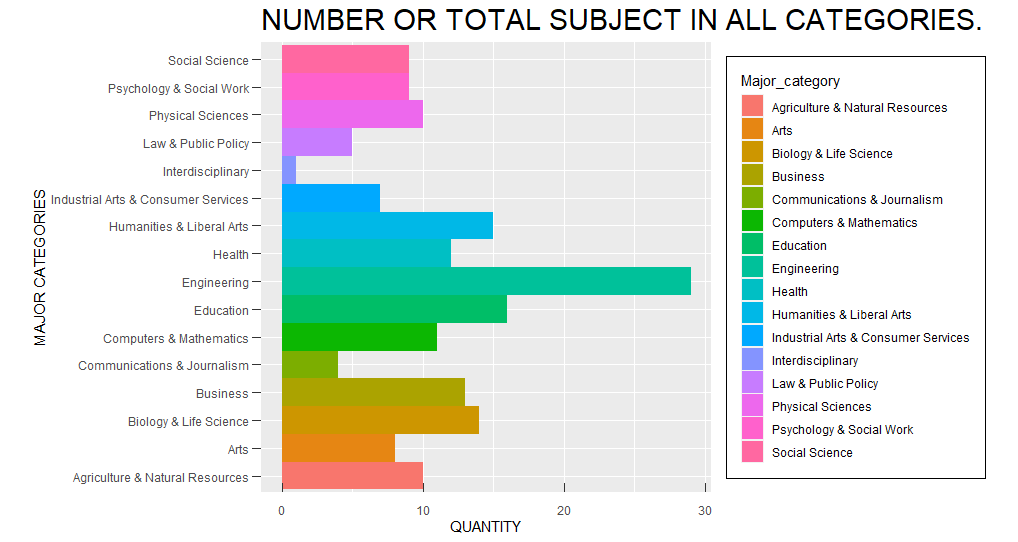
filter(recent\_grade, is.na(Total) | is.na(Men) | is.na(Women) | is.na(ShareWomen)) %>%  
 summarise(Missing\_Count = n())

## Missing\_Count  
## 1 1

As you can see, we have one missing values for these 4 columns. Although, we are not deleting this data entry because, other data of this row is very much important for the future investigation.

## QUESTION 1: Which Major categories has Highest/Lowest numbers of Subject selection, and what are those subjects.

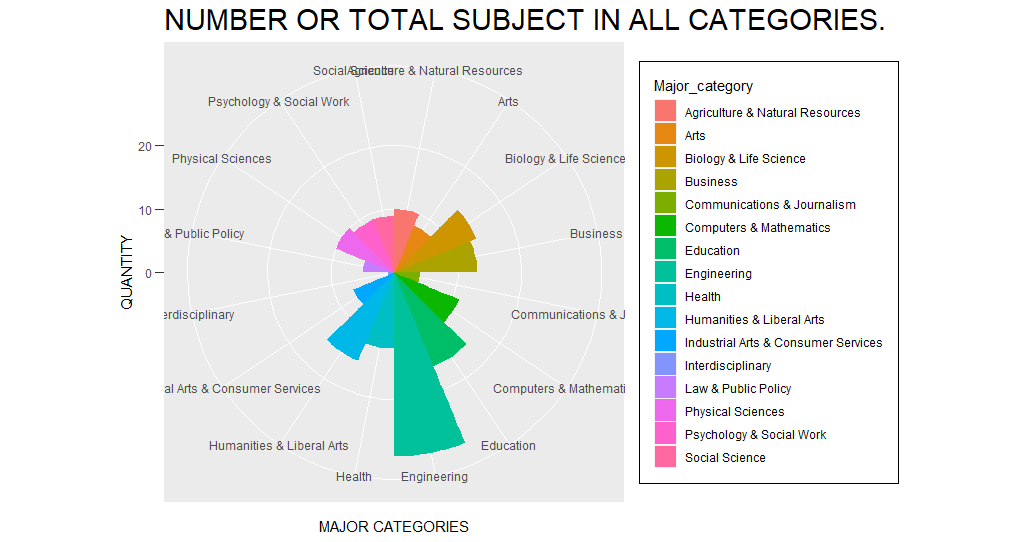
bar <- ggplot(data = recent\_grade)+  
 geom\_bar(  
 mapping = aes(x = Major\_category,   
 fill = Major\_category),  
 group = 1,  
 show.legend = TRUE, width = 1  
 ) +  
 theme(aspect.ratio = 1) +  
 labs(x=NULL,y=NULL)  
  
bar + coord\_flip() +   
 theme(  
 legend.box.background = element\_rect(),  
 legend.box.margin = margin(6, 6, 6, 6)  
 ) +   
 labs(  
 title = "NUMBER OR TOTAL SUBJECT IN ALL CATEGORIES. ",  
 x = "MAJOR CATEGORIES",  
 y = "QUANTITY"  
 ) +  
 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"))  
 )



Here, as you can see, we can clearly see from the BAR graph that Engineering department has the highest variety or category of the different subject. Just like that we can easily identify Interdisciplinary department with only having one subject in it.

We can see this graph different way to make these things clearer, and to get good idea about Major categories.

bar + coord\_polar() +  
 theme(  
 legend.box.background = element\_rect(),  
 legend.box.margin = margin(6, 6, 6, 6)  
 )+   
 labs(  
 title = "NUMBER OR TOTAL SUBJECT IN ALL CATEGORIES. ",  
 x = "MAJOR CATEGORIES",  
 y = "QUANTITY"  
 ) +  
 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"))  
 )



Even after plotting this, we can easily identify those subjects. Now, just to see what kind of subject those Highest and Lowest categories has. We, are doing filtering just to give you a good idea.

Engineering <- filter(recent\_grade, Major\_category == "Engineering")  
select(Engineering, Major:Major\_category)

## Major Total Men Women Major\_category  
## 1 PETROLEUM ENGINEERING 2339 2057 282 Engineering  
## 2 MINING AND MINERAL ENGINEERING 756 679 77 Engineering  
## 3 METALLURGICAL ENGINEERING 856 725 131 Engineering  
## 4 NAVAL ARCHITECTURE AND MARINE ENGINEERING 1258 1123 135 Engineering  
## 5 CHEMICAL ENGINEERING 32260 21239 11021 Engineering  
## 6 NUCLEAR ENGINEERING 2573 2200 373 Engineering  
## 7 MECHANICAL ENGINEERING 91227 80320 10907 Engineering  
## 8 ELECTRICAL ENGINEERING 81527 65511 16016 Engineering  
## 9 COMPUTER ENGINEERING 41542 33258 8284 Engineering  
## 10 AEROSPACE ENGINEERING 15058 12953 2105 Engineering  
## 11 BIOMEDICAL ENGINEERING 14955 8407 6548 Engineering  
## 12 MATERIALS SCIENCE 4279 2949 1330 Engineering  
## 13 ENGINEERING MECHANICS PHYSICS AND SCIENCE 4321 3526 795 Engineering  
## 14 BIOLOGICAL ENGINEERING 8925 6062 2863 Engineering  
## 15 INDUSTRIAL AND MANUFACTURING ENGINEERING 18968 12453 6515 Engineering  
## 16 GENERAL ENGINEERING 61152 45683 15469 Engineering  
## 17 ARCHITECTURAL ENGINEERING 2825 1835 990 Engineering  
## 18 ELECTRICAL ENGINEERING TECHNOLOGY 11565 8181 3384 Engineering  
## 19 MATERIALS ENGINEERING AND MATERIALS SCIENCE 2993 2020 973 Engineering  
## 20 CIVIL ENGINEERING 53153 41081 12072 Engineering  
## 21 MISCELLANEOUS ENGINEERING 9133 7398 1735 Engineering  
## 22 ENVIRONMENTAL ENGINEERING 4047 2662 1385 Engineering  
## 23 ENGINEERING TECHNOLOGIES 3600 2695 905 Engineering  
## 24 GEOLOGICAL AND GEOPHYSICAL ENGINEERING 720 488 232 Engineering  
## 25 INDUSTRIAL PRODUCTION TECHNOLOGIES 4631 3477 1154 Engineering  
## 26 ENGINEERING AND INDUSTRIAL MANAGEMENT 2906 2400 506 Engineering  
## 27 ARCHITECTURE 46420 25463 20957 Engineering  
## 28 MISCELLANEOUS ENGINEERING TECHNOLOGIES 8804 7043 1761 Engineering  
## 29 MECHANICAL ENGINEERING RELATED TECHNOLOGIES 4790 4419 371 Engineering

Interdisciplinary <- filter(recent\_grade, Major\_category =="Interdisciplinary")  
select(Interdisciplinary, Major:Major\_category)

## Major Total Men Women Major\_category  
## 1 MULTI/INTERDISCIPLINARY STUDIES 12296 2817 9479 Interdisciplinary

**Conclusion**: We, have 16 different department as per the graph and we can do a good amount of Statistical plotting. To do that we have to have idea about all the subjects, for which this graph helps us.

## QUESTION 2: Which Major Categories has Highest no of Students.

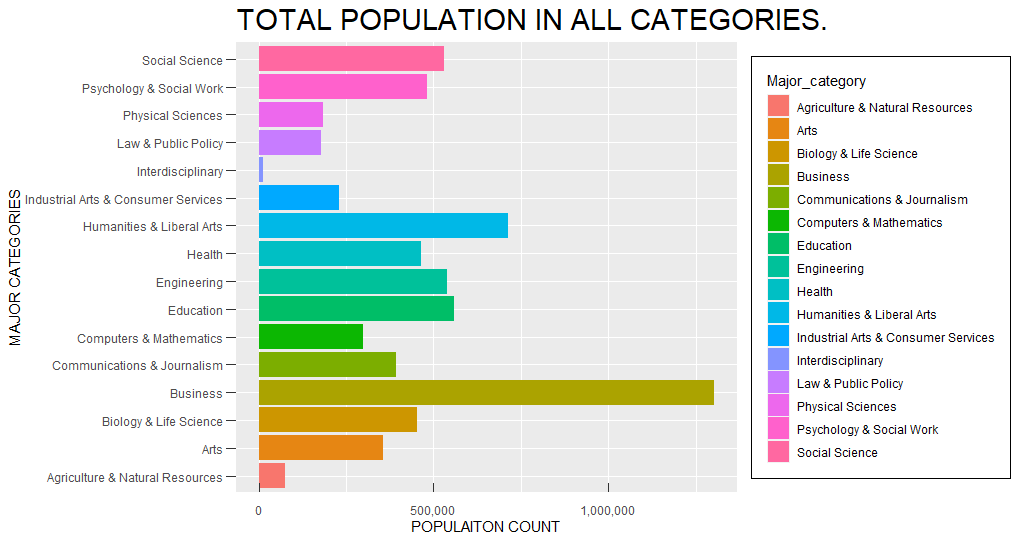
(Head\_count <-   
 group\_by(recent\_grade,  
 Major\_category  
 ) %>%   
 summarise(Sum = sum((Total),   
 na.rm =TRUE),   
 count = n()  
 ) %>%  
 arrange(Sum, order(Major\_category))  
)

## # A tibble: 16 x 3  
## Major\_category Sum count  
## <chr> <int> <int>  
## 1 Interdisciplinary 12296 1  
## 2 Agriculture & Natural Resources 75620 10  
## 3 Law & Public Policy 179107 5  
## 4 Physical Sciences 185479 10  
## 5 Industrial Arts & Consumer Services 229792 7  
## 6 Computers & Mathematics 299008 11  
## 7 Arts 357130 8  
## 8 Communications & Journalism 392601 4  
## 9 Biology & Life Science 453862 14  
## 10 Health 463230 12  
## 11 Psychology & Social Work 481007 9  
## 12 Social Science 529966 9  
## 13 Engineering 537583 29  
## 14 Education 559129 16  
## 15 Humanities & Liberal Arts 713468 15  
## 16 Business 1302376 13

Here, we made one other column for different object named “Head\_count”. As we can see in this data that all different 16 categories have its own total number of populations. I arranged them in Ascending Order so that we can easily identify each of its numbers.

Now, here we did plot the population.

ggplot(data = Head\_count) +  
 geom\_col(mapping=aes(x = Major\_category, y = Sum, fill = Major\_category)) +  
 coord\_flip() +  
 theme(  
 legend.box.background = element\_rect(),  
 legend.box.margin = margin(6, 6, 6, 6)  
 ) +   
 labs(  
 title = "TOTAL POPULATION IN ALL CATEGORIES. ",  
 x = "MAJOR CATEGORIES",  
 y = "POPULAITON COUNT"  
 ) +  
 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::comma)

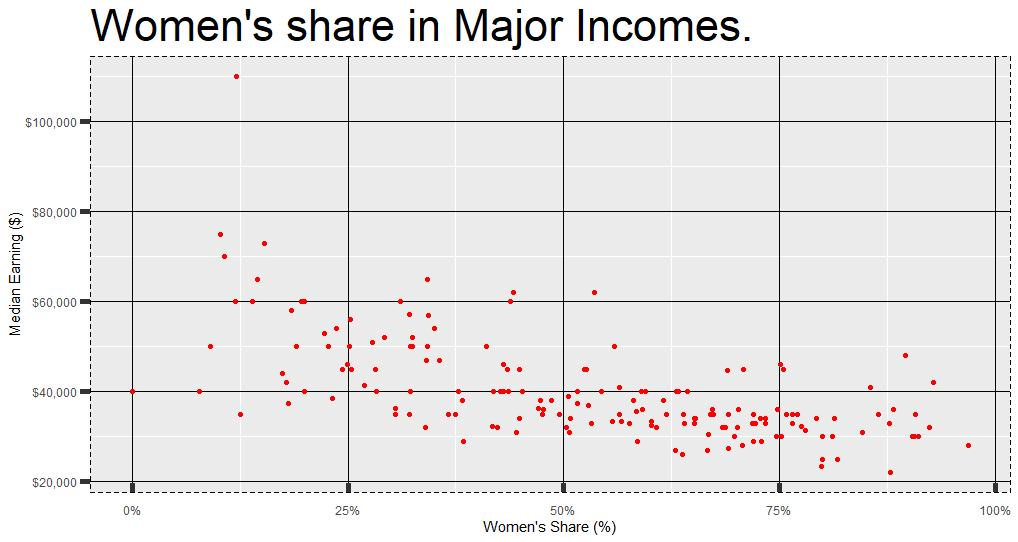


Conclusion: From both above questions, we can easily figure out that population does not always increase as the subject increase, population behaves randomly. Because, we have most subjects in Engineering, but we have highest population in Business schools.

## QUESTION 3: Find out the Woman’s Share in the Major Income sources in all the Majors.

w\_income <- ggplot(data = recent\_grade, mapping = aes(x = ShareWomen, y = Median)) +  
 geom\_point(shape = 20,colour = "red", fill = NA , size = 2, stroke = 1 ) +  
 labs(title = "Women's share in Major Incomes.",  
 x = "Women's Share (%)",  
 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 above scatter plot, we can easily give many questions about Women’s Share in the Major Income Earnings. Just from looking at the graph we can say that between 50%-75% of the women population have a huge density around earning 40,000$ per year. And only one woman among whole database earns money more than 100K Dollars.

**Conclusion**: This scatterplot gives us so much information that we can calculate this same information about Men’s earning and can compare the ration or percentage of the total income. We, can also can identify that which field has better influence on women and men, and what are the fields in which women are usually good at, we can make those kinds of predictions.