Data606 Project

Joe Rovalino

12/8/2019

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Part 1 - Introduction

What is your research question?

Which majors/Major categories should a young female choose to increase earning potential and face less competition as a female in the field.

Why do you care?

The project was prompted from reading the following article "The Economic Guide to Picking A College Major" Ths was interesting to me on a personal level due to the fact that my niece entered college this year and chose to go to a 2 year college to reduce the loans that she would incur. She did come to me for advice on her next steps and choosing a major. I wasn't really sure what would be the best major she should chose. This was a very interesting article that prompted my curiousity to get some data driven answers to her. I have shared the article with her and will hopefully find similiar conclusions from the article.

Why should others care?

There are numerous articles in the papers of our young population having huge debt in the pursuit of a good paying career via education. I feel that if you are going to take away 4-8 years away from your job earning years-it should have a great return on investment (ROI). As a society, having our young people in debt for something that just gives a very poor ROI is a recipe for a weak and unhapy country in the future.

Part 2 - Data

The data found on fivethirtyeight.com article: https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/ Data is found here: https://github.com/fivethirtyeight/data/tree/master/college-majors

Three main data files: - all-ages.csv - recent-grads.csv (ages <28) - grad-students.csv (ages 25+)

All contain 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-students.csv contains details on graduate school attendees. The data set most intereting to my need to get back some info to my niece is the recent-grads.csv.

For the data in datafile - recent grads under the age of 28, there are 173 majors in the data set. The response variable is the major code, major and the major category. It is qualitative. The quantitative independet variables are total, Sample_size, ShareWomen, Employed, Full-time, Part-time, Full_time_year_round, Unemployed, Unemployment_rate, Median, P25th, P75th, College, jobs, Non-college_jobs, Low_wage_job

```r
library(tidyr)
library(dplyr)
library(tidyverse)
library(ggplot2)
```

```r
# load data
majors <- read_csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/college-majors/all-ag
```

```
## Parsed with column specification:
## cols(
##   Major_code = col_double(),
##   Major = col_character(),
##   Major_category = col_character(),
##   Total = col_double(),
##   Employed = col_double(),
##   Employed_full_time_year_round = col_double(),
##   Unemployed = col_double(),
##   Unemployment_rate = col_double(),
##   Median = col_double(),
##   P25th = col_double(),
##   P75th = col_double()
## )
```

```r
head(majors)
```

```
## # A tibble: 6 x 11
##   Major_code Major Major_category  Total Employed Employed_full_t~ Unemployed
##        <dbl> <chr> <chr>           <dbl>    <dbl>            <dbl>      <dbl>
## 1       1100 GENE~ Agriculture &~ 128148    90245            74078       2423
## 2       1101 AGRI~ Agriculture &~  95326    76865            64240       2266
## 3       1102 AGRI~ Agriculture &~  33955    26321            22810        821
## 4       1103 ANIM~ Agriculture &~ 103549    81177            64937       3619
## 5       1104 FOOD~ Agriculture &~  24280    17281            12722        894
## 6       1105 PLAN~ Agriculture &~  79409    63043            51077       2070
## # ... with 4 more variables: Unemployment_rate <dbl>, Median <dbl>,
## #   P25th <dbl>, P75th <dbl>
```

```r
#women_stem <- read_csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/college-majors/w
#head(women_stem)
recentgrads <- read_csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/college-majors/r
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Major = col_character(),
##   Major_category = col_character()
## )

## See spec(...) for full column specifications.
```

```r
head(recentgrads)
```

```
## # A tibble: 6 x 21
##    Rank Major_code Major Total   Men Women Major_category ShareWomen Sample_size
##   <dbl>      <dbl> <chr> <dbl> <dbl> <dbl> <chr>               <dbl>       <dbl>
## 1     1       2419 PETR~  2339  2057   282 Engineering         0.121          36
## 2     2       2416 MINI~   756   679    77 Engineering         0.102           7
## 3     3       2415 META~   856   725   131 Engineering         0.153           3
## 4     4       2417 NAVA~  1258  1123   135 Engineering         0.107          16
## 5     5       2405 CHEM~ 32260 21239 11021 Engineering         0.342         289
## 6     6       2418 NUCL~  2573  2200   373 Engineering         0.145          17
## # ... with 12 more variables: Employed <dbl>, Full_time <dbl>, 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>
```

Part 3 - Exploratory data analysis

```
summary(recentgrads)
```

```
##       Rank       Major_code      Major               Total       
##  Min.   :  1   Min.   :1100   Length:173         Min.   :   124  
##  1st Qu.: 44   1st Qu.:2403   Class :character   1st Qu.:  4550  
##  Median : 87   Median :3608   Mode  :character   Median : 15104  
##  Mean   : 87   Mean   :3880                      Mean   : 39370  
##  3rd Qu.:130   3rd Qu.:5503                      3rd Qu.: 38910  
##  Max.   :173   Max.   :6403                      Max.   :393735  
##                                                  NA's   :1       
##       Men             Women        Major_category       ShareWomen    
##  Min.   :   119   Min.   :     0   Length:173         Min.   :0.0000  
##  1st Qu.:  2178   1st Qu.:  1778   Class :character   1st Qu.:0.3360  
##  Median :  5434   Median :  8386   Mode  :character   Median :0.5340  
##  Mean   : 16723   Mean   : 22647                      Mean   :0.5222  
##  3rd Qu.: 14631   3rd Qu.: 22554                      3rd Qu.:0.7033  
##  Max.   :173809   Max.   :307087                      Max.   :0.9690  
##  NA's   :1        NA's   :1                           NA's   :1       
##   Sample_size        Employed        Full_time        Part_time     
##  Min.   :   2.0   Min.   :     0   Min.   :   111   Min.   :     0  
##  1st Qu.:  39.0   1st Qu.:  3608   1st Qu.:  3154   1st Qu.:  1030  
##  Median : 130.0   Median : 11797   Median : 10048   Median :  3299  
##  Mean   : 356.1   Mean   : 31193   Mean   : 26029   Mean   :  8832  
##  3rd Qu.: 338.0   3rd Qu.: 31433   3rd Qu.: 25147   3rd Qu.:  9948  
##  Max.   :4212.0   Max.   :307933   Max.   :251540   Max.   :115172  
##                                                                     
##  Full_time_year_round   Unemployed    Unemployment_rate     Median      
##  Min.   :   111       Min.   :    0   Min.   :0.00000   Min.   : 22000  
##  1st Qu.:  2453       1st Qu.:  304   1st Qu.:0.05031   1st Qu.: 33000  
##  Median :  7413       Median :  893   Median :0.06796   Median : 36000  
##  Mean   : 19694       Mean   : 2416   Mean   :0.06819   Mean   : 40151  
##  3rd Qu.: 16891       3rd Qu.: 2393   3rd Qu.:0.08756   3rd Qu.: 45000  
##  Max.   :199897       Max.   :28169   Max.   :0.17723   Max.   :110000  
##                                                                         
##      P25th           P75th          College_jobs    Non_college_jobs
##  Min.   :18500   Min.   : 22000   Min.   :     0   Min.   :     0  
##  1st Qu.:24000   1st Qu.: 42000   1st Qu.:  1675   1st Qu.:  1591  
##  Median :27000   Median : 47000   Median :  4390   Median :  4595  
##  Mean   :29501   Mean   : 51494   Mean   : 12323   Mean   : 13284  
##  3rd Qu.:33000   3rd Qu.: 60000   3rd Qu.: 14444   3rd Qu.: 11783  
##  Max.   :95000   Max.   :125000   Max.   :151643   Max.   :148395  
##                                                                   
```

```
##  Low_wage_jobs
##  Min.   :    0
##  1st Qu.:  340
##  Median : 1231
##  Mean   : 3859
##  3rd Qu.: 3466
##  Max.   :48207
##
```

```
top5majorsbyPay<- top_n(recentgrads, 5, Median)
top5majorsbyPay
```

```
## # A tibble: 6 x 21
##    Rank Major_code Major Total   Men Women Major_category ShareWomen Sample_size
##   <dbl>      <dbl> <chr> <dbl> <dbl> <dbl> <chr>               <dbl>       <dbl>
## 1     1       2419 PETR~  2339  2057   282 Engineering         0.121          36
## 2     2       2416 MINI~   756   679    77 Engineering         0.102           7
## 3     3       2415 META~   856   725   131 Engineering         0.153           3
## 4     4       2417 NAVA~  1258  1123   135 Engineering         0.107          16
## 5     5       2405 CHEM~ 32260 21239 11021 Engineering         0.342         289
## 6     6       2418 NUCL~  2573  2200   373 Engineering         0.145          17
## # ... with 12 more variables: Employed <dbl>, Full_time <dbl>, 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>
```

```
top5majorswhighunempl <- top_n(recentgrads, 5, Unemployment_rate)
top5majorswhighunempl
```

```
## # A tibble: 5 x 21
##    Rank Major_code Major Total   Men Women Major_category ShareWomen Sample_size
##   <dbl>      <dbl> <chr> <dbl> <dbl> <dbl> <chr>               <dbl>       <dbl>
## 1     6       2418 NUCL~  2573  2200   373 Engineering         0.145          17
## 2    30       5402 PUBL~  5978  2639  3339 Law & Public ~      0.559          55
## 3    85       2107 COMP~  7613  5291  2322 Computers & M~      0.305          97
## 4    90       5401 PUBL~  5629  2947  2682 Law & Public ~      0.476          46
## 5   171       5202 CLIN~  2838   568  2270 Psychology & ~      0.800          13
## # ... with 12 more variables: Employed <dbl>, Full_time <dbl>, 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>
```

```
# using the sammary function on recent grads. Will create a data frame to only include the majors with
overmeanpayingmajors <- filter(recentgrads, Median > 40151)
#reduces the number of majors to focus on from 173 down to 56
top5wMedian <- top_n(overmeanpayingmajors, 5, Median)
top5wMedian
```

```
## # A tibble: 6 x 21
##    Rank Major_code Major Total   Men Women Major_category ShareWomen Sample_size
##   <dbl>      <dbl> <chr> <dbl> <dbl> <dbl> <chr>               <dbl>       <dbl>
## 1     1       2419 PETR~  2339  2057   282 Engineering         0.121          36
## 2     2       2416 MINI~   756   679    77 Engineering         0.102           7
## 3     3       2415 META~   856   725   131 Engineering         0.153           3
## 4     4       2417 NAVA~  1258  1123   135 Engineering         0.107          16
## 5     5       2405 CHEM~ 32260 21239 11021 Engineering         0.342         289
## 6     6       2418 NUCL~  2573  2200   373 Engineering         0.145          17
## # ... with 12 more variables: Employed <dbl>, Full_time <dbl>, 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>
```

It looks like median unemployement is approximately 20% less for the top 5 majors compared to overall recent grads (0.05561/0.6819).

It also looks like median Share of Women is approximately 25% less in the top 5 paid majors then compared to overall recent grads (.1328/.5340)

```
summary(recentgrads$Unemployment_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.05031 0.06796 0.06819 0.08756 0.17723
```

```
summary(overmeanpayingmajors$Unemployment_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.04667 0.06151 0.06562 0.08801 0.17723
```

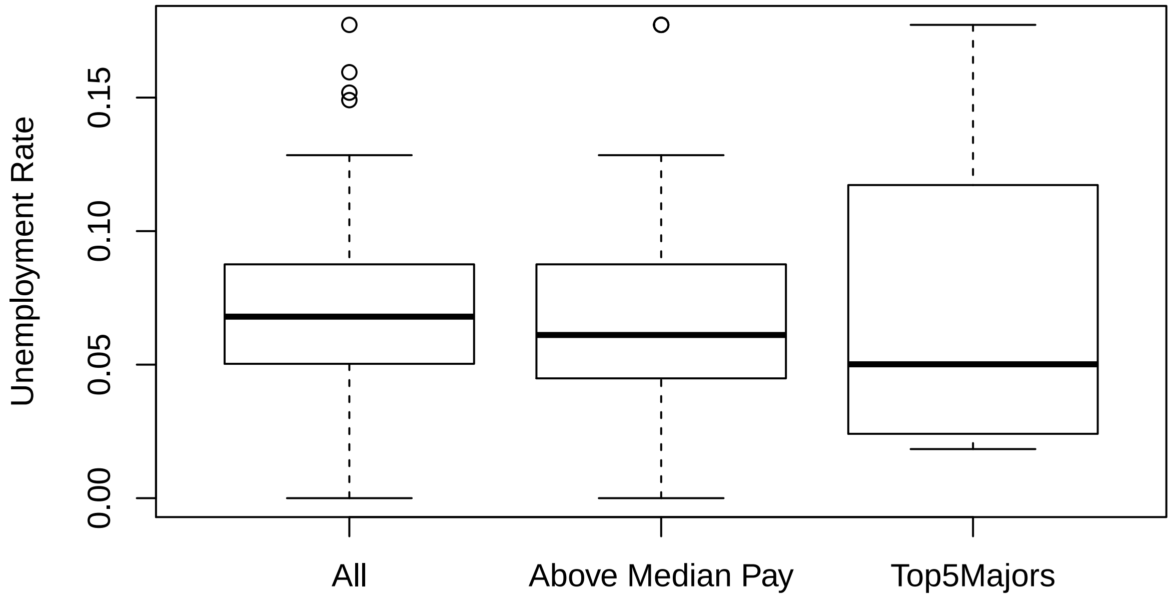
```
summary(top5wMedian$Unemployment_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01838 0.03060 0.05561 0.07469 0.10321 0.17723
```

```
ChanceofUnemployment <- cbind(recentgrads$Unemployment_rate, overmeanpayingmajors$Unemployment_rate, top
```

```
## Warning in cbind(recentgrads$Unemployment_rate,
## overmeanpayingmajors$Unemployment_rate, : number of rows of result is not a
## multiple of vector length (arg 2)
```

```
boxplot(ChanceofUnemployment,names = c("All", "Above Median Pay", "Top5Majors"), ylab = "Unemployment Ra
```



```
summary(recentgrads$ShareWomen)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##  0.0000  0.3360  0.5340  0.5222  0.7033  0.9690       1
```

```
summary(overmeanpayingmajors$ShareWomen)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
## 0.09071 0.21105 0.32078 0.35872 0.43960 0.92781       1
```

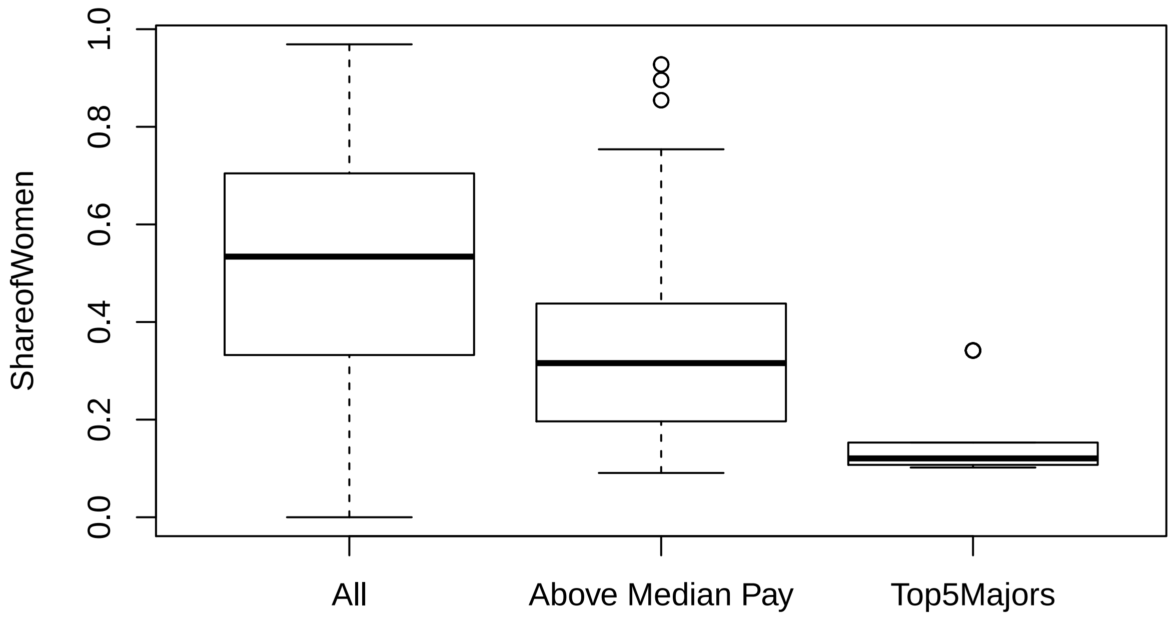
```
summary(top5wMedian$ShareWomen)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1019  0.1106  0.1328  0.1616  0.1510  0.3416
```

```
ShareofWomen <- cbind(recentgrads$ShareWomen, overmeanpayingmajors$ShareWomen, top5wMedian$ShareWomen)
```

```
## Warning in cbind(recentgrads$ShareWomen, overmeanpayingmajors$ShareWomen, :
## number of rows of result is not a multiple of vector length (arg 2)
```

```
boxplot(ShareofWomen,names = c("All", "Above Median Pay", "Top5Majors"), ylab = "ShareofWomen")
```



It looks like median unemployement is approximately 20% less for the top 5 majors compared to overall recent grads (0.05561/0.6819).

It also looks like median Share of Women is approximately 25% less in the top 5 paid majors then compared to overall recent grads (.1328/.5340)

```
options(scipen = 999)

gg <- ggplot(overmeanpayingmajors, aes(x=ShareWomen, y=Median)) +
  geom_point(aes(col=ShareWomen, size=Median)) +
  geom_smooth(method="loess", se=F)
  labs(subtitle="SHare of Women by Median pay",
       y="Median Pay",
       x="Share of Women",
       title="Scatterplot",
       caption = "Source: DF overmeanpayingmajors")
```

```
## $y
## [1] "Median Pay"
##
## $x
## [1] "Share of Women"
##
## $title
## [1] "Scatterplot"
##
## $subtitle
## [1] "SHare of Women by Median pay"
```

```
##
## $caption
## [1] "Source: DF overmeanpayingmajors"
##
## attr(,"class")
## [1] "labels"
```

```
require(scales)
```

```
## Loading required package: scales
```

```
##
## Attaching package: 'scales'
```

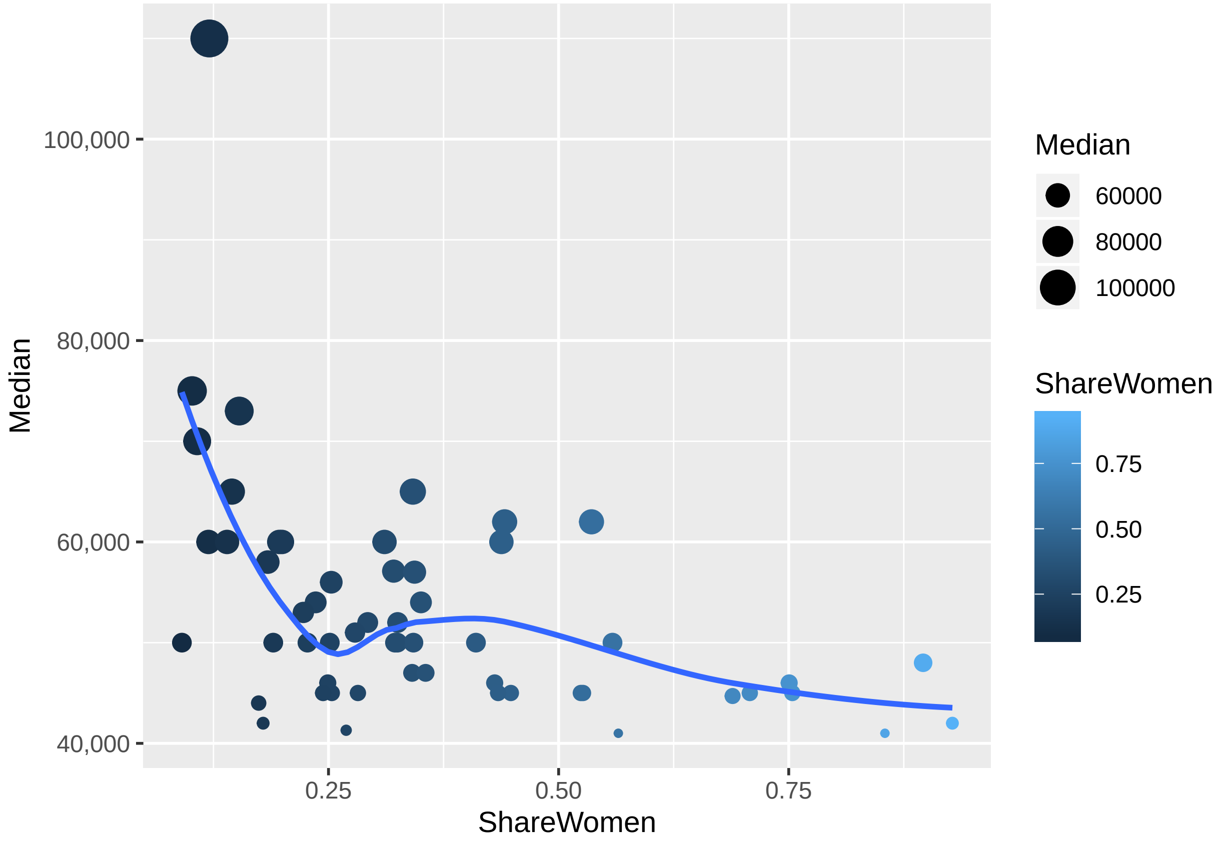
```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
gg + scale_y_continuous(labels = comma)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

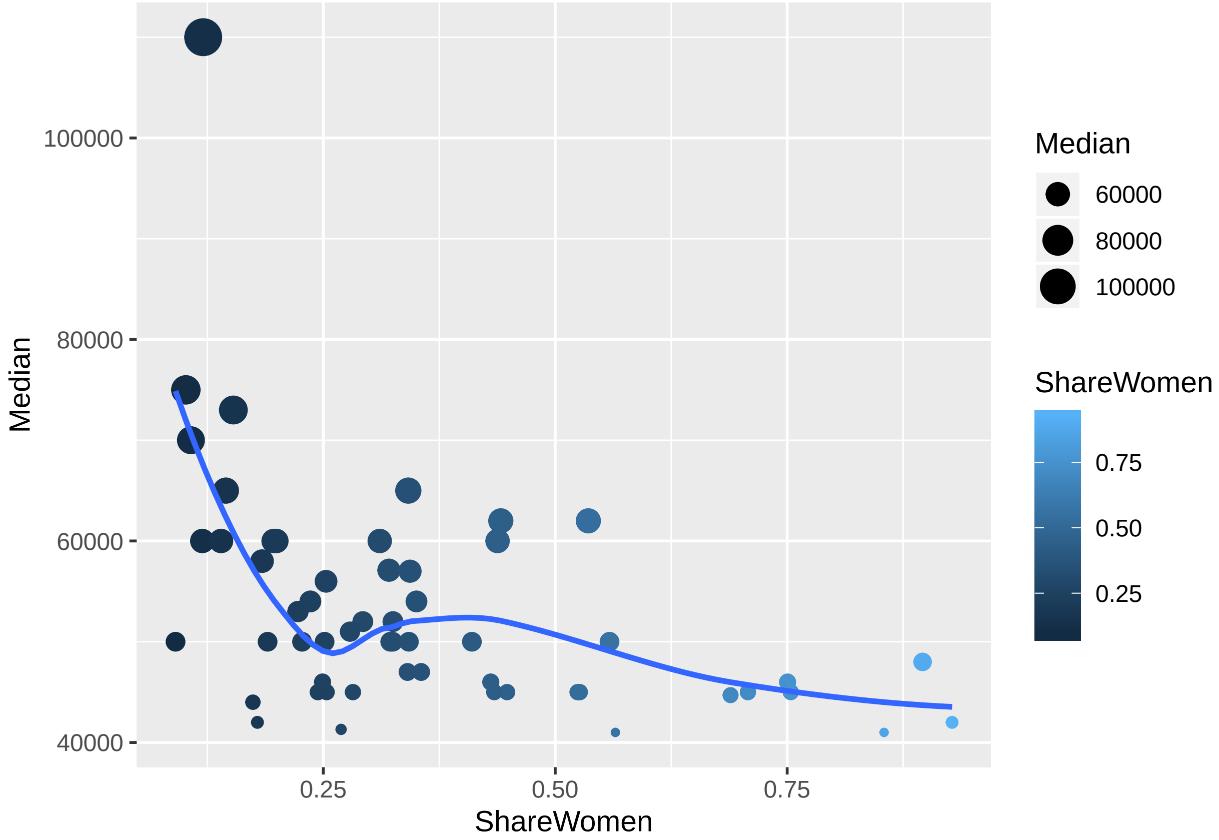
```
## Warning: Removed 1 rows containing missing values (geom_point).
```



```
plot(gg)
```

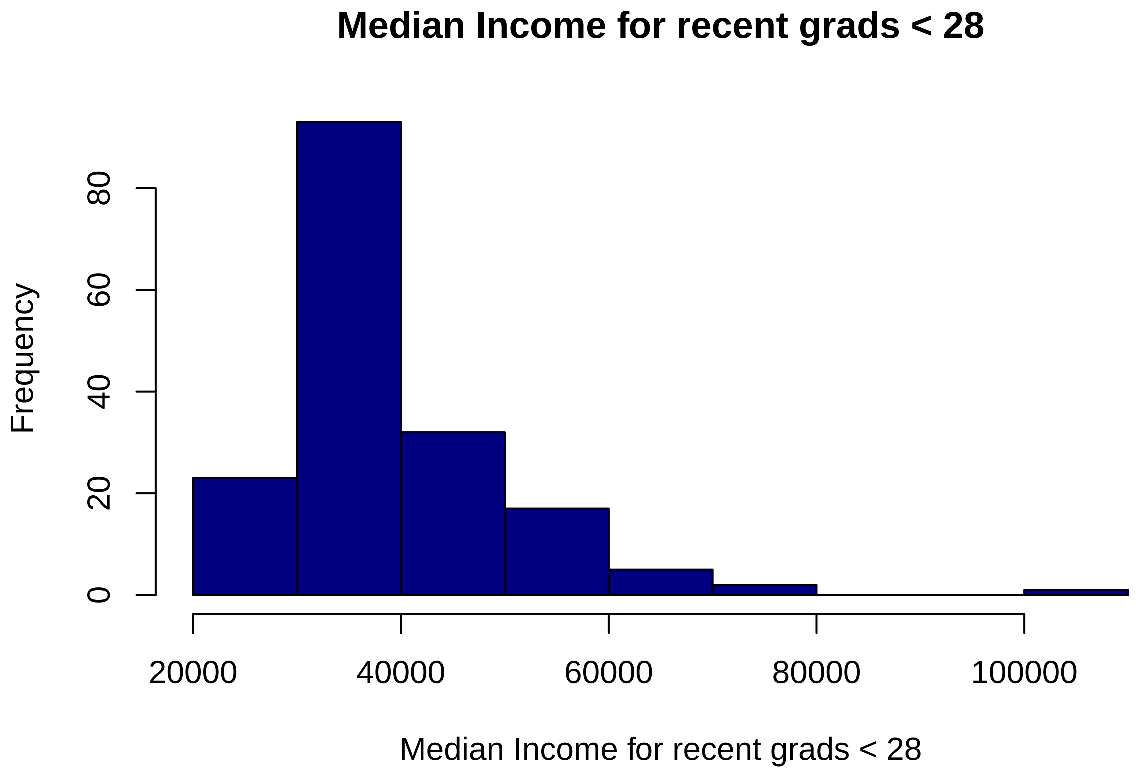
```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

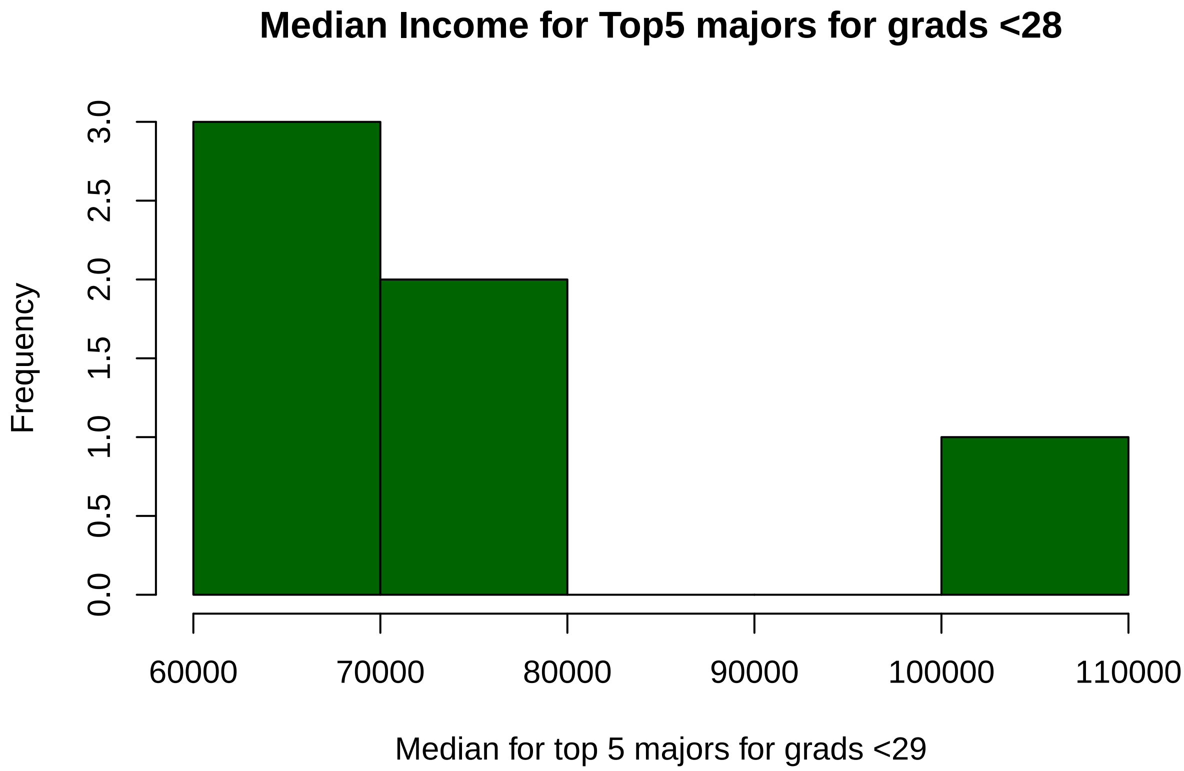


The histogram will show a normal distribution unimodal with some slight right skwe.

```
hist(recentgrads$Median, main = "Median Income for recent grads < 28", xlab = "Median Income for recent
```



```
hist(top5wMedian$Median, main = "Median Income for Top5 majors for grads <28", xlab = "Median for top 5
```



Part 4 - Inference

This is an observational study of the college majors and the income. From our textbook, Open Intro Statistics chapter 7.5 Comparing many means wiht ANOVA: "Generally we must check three conditions on the data before performing ANOVA:

- the observations are independent within and across groups,

- the data within each group are nearly normal, and

- the variability across the groups is about equal."

```
Majorcategory_mean_variance <- summarize(recentgrads %>% group_by(Major_category), mean = mean(Median))
Majorcategory_mean_variance
```

```
## # A tibble: 16 x 2
##    Major_category                      mean
##    <chr>                              <dbl>
##  1 Agriculture & Natural Resources    36900
##  2 Arts                               33062.
##  3 Biology & Life Science             36421.
##  4 Business                           43538.
##  5 Communications & Journalism        34500
##  6 Computers & Mathematics            42745.
##  7 Education                          32350
##  8 Engineering                        57383.
##  9 Health                             36825
## 10 Humanities & Liberal Arts           31913.
## 11 Industrial Arts & Consumer Services 36343.
## 12 Interdisciplinary                   35000
## 13 Law & Public Policy                 42200
## 14 Physical Sciences                   41890
## 15 Psychology & Social Work            30100
## 16 Social Science                      37344.
```

Looking at the analytis of the group of major category and it's mean, it appears that the variability across the groups is about equal

1) We see from the data that each major is indepedent
2) Reviewing the histograms we can see that the distribution of the median salary is a normal distribtion.
3) check the variability is about equal using and F-Test

Arts, Social Science & Humanities = 1 Arts = 2 Biology and Life Sciences = 3 business = 4 Communications & Journalism = 5 Computers and Mathematics = 6 Education = 7 Engineering = 8 Health = 9 Humanities and Liberal Arts = 10 Industrial Arts and Consumer Services = 11 Indterdisciplinary = 12 Law & Public Policy = 13 Physical Sciences = 14 Psychology and Social Work = 15 Social Science = 16

Hypothesis null and Hypothesis Alternative - H0: U Arts, So = u1 = u2 = u3 . . . = u16 HA: Median salary varies by major selected.

```
summary(aov(recentgrads$Median ~ recentgrads$Major_category))
```

```
##                              Df      Sum Sq   Mean Sq F value
## recentgrads$Major_category  15 12948122198 863208147      14
## Residuals                  157  9681069941  61662866
##                                          Pr(>F)
## recentgrads$Major_category <0.0000000000000002 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Per our textbook (pg 289), we will use the F statistic also called the F-Test to come to a conclusion that the alternative hypothesis is the correct choice as the p value is close to zero (0.0000000000000002). Median salary is not the same across majors.

Part 5 - Conclusion

1) Stick to Engineering and Technology - it will lead to higher paying jobs as an undergrad and have the least Share of women.
2) Looking at ratios of men to women to decide on picking a major may help to stand out as a women. Engineering majors tend to have a low percentage of men to woemen. there are culture trends related to woment diversity inititives like #girlpower and #equalpay. It may be beneficial to try to enter the workforce in a male dominted major as it could be a good way to stand out from other applicants.

References

1) https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/ by Ben Casselmen

2) All data is from American Community Survey 2010-2012 Public Use Microdata Series.

Download data here: http://www.census.gov/programs-surveys/acs/data/pums.html

Documentation here: http://www.census.gov/programs-surveys/acs/technical-documentation/pums.html

3) Open Intro Statistics 4th Ed. by David Diez, Mine Cetinkaya-Rundel and Christpher Bar Chapter 7 "Comparing many means wiht ANOVA

4) Shiny App: https://bencasselman.shinyapps.io/new-test/