Applictation of dplyr & ggplot2: College Majors and Income

terry min

10/30/2017

1.Intro

While there's no perfect correlation between one's major and his or her economic success, it is true that the average first-year salary for certain majors are higher than those for the others.According to GlassDoor, **computer science** majors received the highest median base salary, followed by **electrical engineering, mechanical engineering, chemical engineering, industrial engineering** and so on. According to Business Insider, **petroleum engineering, chemical engineering, and computer engineering, nuclear engineering, and electrical engineering** majors ranked in the top 5. Finally, according to PayScale, the five most promising majors in terms of early career median salary were **petroleum engineering, actualrial mathematics, actuarial science, nuclear engineering, chemical engineering**.

It seems that engineering majors are likely to start their career with higher salaries compared to non-engineering majors. I would first like to see if other data also reveal similar trend by using some of the functions of dplyr and ggplot2. For analysis, I would like to use the data and code from FiveThirtyEight's story on earnings of college majors. Since all data are from American Community Survey 2010-2012 Public Use Microdata Series, I would also include it in the reference citation.

Then a question arises: "Are there huge financial differences among different major categories?" In his article, Ben Casselmen, a senior editor and chief economics writer for FiveThirtyEight, states that "don't assume that all "STEM" — science, technology, engineering and math — majors are the same." Is this statement true? I would use some of the functions from the dpylr and ggplot2 in order to analyze and visualize the given data.

For the major categories, I would use the way how majors are categorized in Carnevale et al, "What's It Worth?: The Economic Value of College Majors." Georgetown University Center on Education and the Workforce, 2011.

Third, I would like to see what non-STEM majors earn the highest salary by using similar method to that I have used to address the first question and see if certain non-STEM major categories are dominant.

Finally, I would like to see if there is any relationship between STEM majors and unemployment rate.

2. Data Preparation

```
# packages
# I assumed that I've already installed the necessary packages.
# install.packages(c("dplyr", "ggplot2"))
library(readr)     # importing data
library(dplyr)     # data wrangling
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## 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)  # graphics
```

```
# download RData file into your working directory
github <- "https://github.com/fivethirtyeight/data/tree/master/college-majors/"
csv <- "all-ages.csv"
download.file(url = paste0(github, csv), destfile = 'all-ages.csv')
```

```
# import the data in R using the 'read.csv()' function
dat <- read.csv('../data/all-ages.csv', stringsAsFactors = FALSE)
```

3. Questions

(1) Does the data set I have also show that the engineering majors receive higher salaries?

As a reminder, some of the basic dplyr functions are listed in the following website. [Basic dplyr functions][https://www.r-bloggers.com/data-manipulation-with-dplyr/]

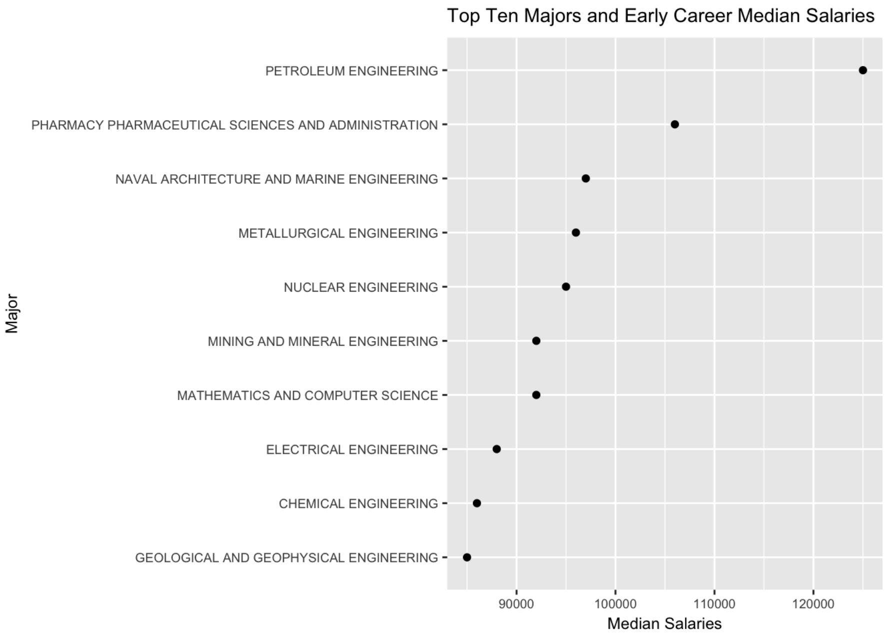
I would use the arrange function as well as the slice and select functions in package dplyr in order to see what are the top 10 majors that pay the graduates off the most.

```
# dplyr: slice and select
top_ten_dat <- slice(select(arrange(dat, desc(Median)), Major, Major_category, Median), 1:10)
top_ten_dat
```

```
## # A tibble: 10 x 3
##                                                Major
##                                                <chr>
##  1                              PETROLEUM ENGINEERING
##  2 PHARMACY PHARMACEUTICAL SCIENCES AND ADMINISTRATION
##  3           NAVAL ARCHITECTURE AND MARINE ENGINEERING
##  4                          METALLURGICAL ENGINEERING
##  5                                NUCLEAR ENGINEERING
##  6                     MINING AND MINERAL ENGINEERING
##  7                    MATHEMATICS AND COMPUTER SCIENCE
##  8                             ELECTRICAL ENGINEERING
##  9                               CHEMICAL ENGINEERING
## 10              GEOLOGICAL AND GEOPHYSICAL ENGINEERING
## # ... with 2 more variables: Major_category <chr>, Median <int>
```

I would also use ggplot function in package ggplot2 in order to visualize the data by creating a bar chart of majors and median salaries.

```
# ggplot2: geom_bar, coord_flip(), theme(), and labs()
top_ten <- ggplot(top_ten_dat, aes(reorder(Major, Median), Median)) +
                labs(title = "Top Ten Majors and Early Career Median Salaries", x = "Major", y= "Median Salaries")
+
                geom_point() +
                coord_flip() +
                theme(text = element_text(size=9))
top_ten
```



Analysis

- A majority of top 10 majors (in terms of their average salaries) was under the major category of Engineering.
- Among the majors under Enginering, petroleum engineering is an outlier. According to US News, Texas A&M University–College Station, University of Texas–Austin, University of Oklahoma had the best undergraduate petroleum engineering programs in the States.

(2) Is it true that all STEM fields aren't the same?

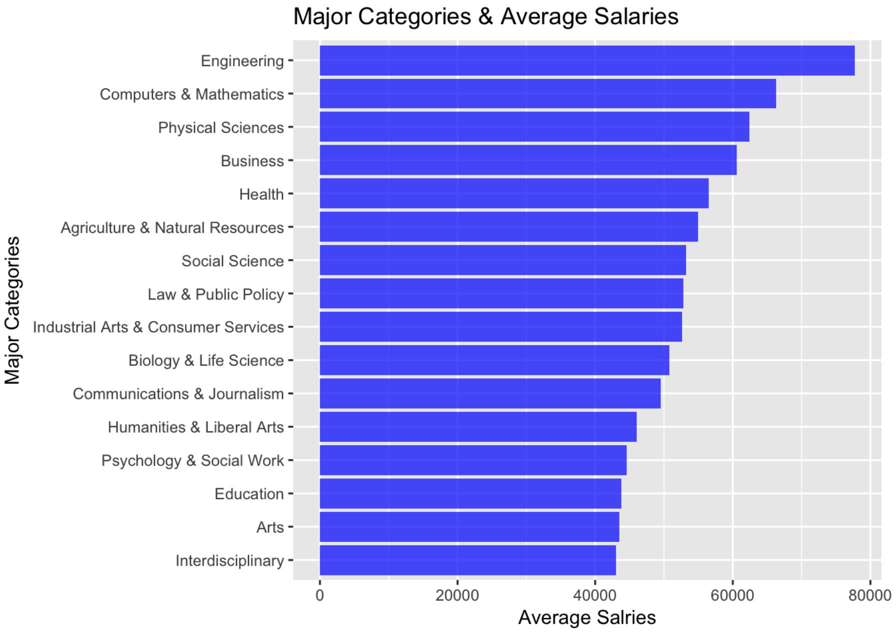
To see whehter Ben Casselmen's argument is true, I would use use the summarise function and group_by function in order to see whether the average salaries differ by major categories. I would then arrange them in order by using the arrange function to see if STEM majors receive higher salaries in general. I would consider Engineering, Computer & Mathematics, Physical Sciences and Biology & Life Science as STEM major categories.

```
# dplyr: group_by
dat_sum <- arrange(
            summarise(
              group_by(dat, Major_category),
              avg_mc_sal = mean(Median)),
          desc(avg_mc_sal)
          )
dat_sum
```

```
## # A tibble: 16 x 2
##                        Major_category avg_mc_sal
##                                 <chr>      <dbl>
##  1                        Engineering   77758.62
##  2              Computers & Mathematics   66272.73
##  3                   Physical Sciences   62400.00
##  4                            Business   60615.38
##  5                              Health   56458.33
##  6      Agriculture & Natural Resources   55000.00
##  7                      Social Science   53222.22
##  8                  Law & Public Policy   52800.00
##  9 Industrial Arts & Consumer Services   52642.86
## 10               Biology & Life Science   50821.43
## 11           Communications & Journalism   49500.00
## 12              Humanities & Liberal Arts   46080.00
## 13              Psychology & Social Work   44555.56
## 14                           Education   43831.25
## 15                                Arts   43525.00
## 16                   Interdisciplinary   43000.00
```

I would visualize the data by using the ggplot functions. This time, I will use geom_bar see the relationship between the average salaries and major categories.

```
#ggpplot
dat_sum_plot <- ggplot(dat_sum, aes(reorder(Major_category, avg_mc_sal), avg_mc_sal)) +
                labs(title = "Major Categories & Average Salaries", x = "Major Categories", y= "Average Salries")
+
                geom_bar(stat = 'identity', fill = "blue", alpha = 0.7) +
                coord_flip() +
                theme(text = element_text(size=11))
dat_sum_plot
```



Analysis

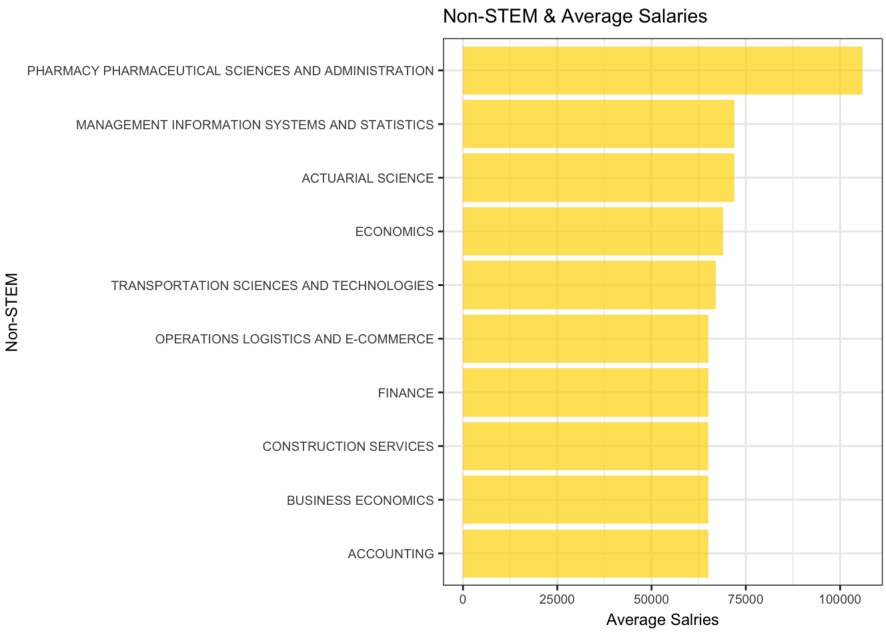
- In general, STEM majors receive higher early career median salaries.
- However, majors under the Biology & Life Science major category received less salaries than those under non-STEM majors such as Business or Health major category.
- Thus, it can be inferred that Ben Casselmen’s argument is true.

(3) Which non-STEM majors earn the most? And which Major Category earns the most?

```
#dplr
non_STEM <- filter(dat, (Major_category != "Engineering" & Major_category != "Computers & Mathematics" & Major_cat
egory != "Physical Sciences" & Major_category != "Biology & Life Science"))
ten_non_STEM <- slice(select(arrange(non_STEM, desc(Median)), Major, Major_category, Median), 1:10)
ten_non_STEM
```

```
## # A tibble: 10 x 3
##                                                  Major
##                                                  <chr>
##   1 PHARMACY PHARMACEUTICAL SCIENCES AND ADMINISTRATION
##   2                                  ACTUARIAL SCIENCE
##   3       MANAGEMENT INFORMATION SYSTEMS AND STATISTICS
##   4                                          ECONOMICS
##   5            TRANSPORTATION SCIENCES AND TECHNOLOGIES
##   6                              CONSTRUCTION SERVICES
##   7                                         ACCOUNTING
##   8                 OPERATIONS LOGISTICS AND E-COMMERCE
##   9                                 BUSINESS ECONOMICS
##  10                                            FINANCE
## # ... with 2 more variables: Major_category <chr>, Median <int>
```

```
#ggplot
dat_sum_plot <- ggplot(ten_non_STEM, aes(reorder(Major, Median), Median)) +
                labs(title = "Non-STEM & Average Salaries", x = "Non-STEM", y= "Average Salries") +
                geom_bar(stat = 'identity', fill = "gold", alpha = 0.7) +
                coord_flip() +
                theme_bw() +
                theme(text = element_text(size=9))
dat_sum_plot
```



Analysis

- Among non-STEM majors, Pharmacy Pharmaceutical Sciences and Administration under the Health major category is an outlier. The median salary for them is about 1.5 times higher than that of Management Information Systems and Statisitcs the which ranked second.
- Also, Among the top 10 non-STEM majors in terms of the average salary, more than half of the majors were under the Business major category. These majors include Actuarial Science, Management Information Systems and Statistics, Accounting, Operations Logistics and E-Commerce, Business Economics, and Finance.

(4) Unemployment Rate & STEM Major Categories

To see if there's a relationship between STEM major categories and unemployment rate, I would first create a new data frame of majors under STEM major categories.

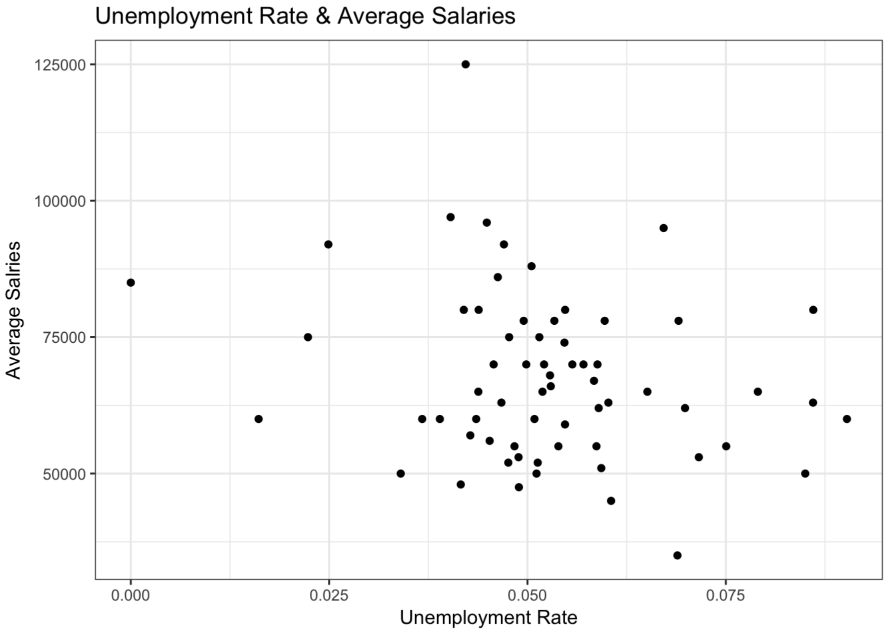
```
dat_STEM <- data.frame(filter(dat, (dat$Major_category == "Engineering" | dat$Major_category == "Computers & Mathe
matics" | dat$Major_category == "Physical Sciences" | dat$Major_category == "Biology & Life Science")))
```

Then I would export each data frame to a data file 'STEM.csv' in the 'data' folder.

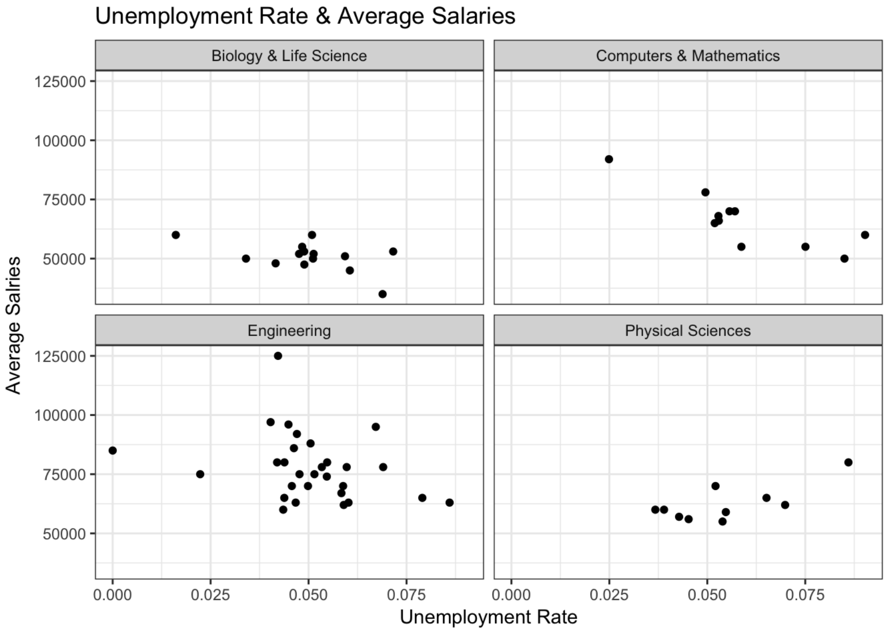
```
# Use the function `write.csv()` to export (or save) the data frame `dat_STEM` and `dat_non_STEM` to data files `S
TEM.csv` and `non_STEM.csv`, respectively, in the `folder/` directory.
write.csv(dat_STEM, file = "~/stat133/stat133-hws-fall17/post01/data/STEM.csv",row.names = F)
```

Finally, I would use ggplot function in order to visualize the data. I would also use facet_wrap in order to get scatterplots of unemployment rate and average salaries grouped by STEM major categories.

```
# Use ggplot and facet_wrap to see get scatterplots of 'unemployment rate' and 'average salaries' for STEM majors.
STEM_unem_median <- ggplot(dat_STEM, aes(Unemployment_rate, Median)) +
                labs(title = "Unemployment Rate & Average Salaries", x = "Unemployment Rate", y= "Average Salries"
) +
                geom_point() +
                theme_bw()
STEM_unem_median
```



```
# Use ggplot and facet_wrap to see get scatterplots of 'unemployment rate' and 'average salaries' grouped by 'STEM
major categories'.
STEM_unem_median_by_mc <- ggplot(dat_STEM, aes(Unemployment_rate, Median)) +
                labs(title = "Unemployment Rate & Average Salaries", x = "Unemployment Rate", y= "Average Salries"
) +
                geom_point() +
                theme_bw() +
                facet_wrap(~ Major_category)
STEM_unem_median_by_mc
```



Analysis

- It looks like there is no specific relationship between being a STEM major and average salaries. Even if I tried to group them by different major categories, the relationship was not clear.

4. Conclusion

To sum up, I got the following results.

- Engineering majors receive more early career median salaries.
- STEM majors received more payments in general, except for the majors under the Biology & Life Science major category. (However, we should not conclude that all majors under that category receive less payments: It's about the average!)

- Among non-STEM majors, Pharmacy Pharmaceutical Sciences and Administration under the Health major category received the most. However, in general, majors under the Business major category took the majority of the top 10 non-STEM majors in terms of early career median salaries.
- No relationship could be found between being a STEM major and average salaries.

Later, I would like to delve into the follwing questions

- Are there data that show changes in average wages for each marjor?
- If so, is there any noticeable trend?
- Are there any other factors people consider when deciding their majors in college?

5. References

1. [GlassDoor: 50 Highest Paying College Majors][https://www.glassdoor.com/blog/50-highest-paying-college-majors/]
2. [BusinessInsider: The college majors with the highest starting salaries][https://www.youtube.com/watch?v=_oRlrcoy4xw]
3. [PayScale: Highest Paying Bachelor Degrees by Salary Potential][https://www.payscale.com/college-salary-report/majors-that-pay-you-back/bachelors]
4. [Github Data][https://github.com/fivethirtyeight/data/tree/master/college-majors]
5. [American Community Survey 2010-2012 Public Use Microdata Series][http://www.census.gov/programs-surveys/acs/data/pums.html]
6. [The Economic Guide To Picking A College Major][https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/]
7. [Carnevale et al, “What’s It Worth?: The Economic Value of College Majors.”][http://cew.georgetown.edu/whatsitworth]
8. [US News: Best Undergraduate Petroleum Engineering Programs (Doctorate)][https://www.usnews.com/best-colleges/rankings/engineering-doctorate-petroleum]