College Majors From an Economical Perspective

Sathwik Kesappragada
June 2nd, 2020

Importance of Analysis

The overarching question is to see

“which college majors are the most economically viable”

and get interesting Statistics on various degrees. We believe this is an interesting question because college is expensive and will continue to be more expensive in the coming years; more people accruing copious amounts of Student Loan debt. College has become a financial investment for most people and we want to see what degrees will give us the best return on investment.

Objectives and/or Questions of Interest

- ★ **Difference between STEM and Non-STEM majors**
- ★ **Difference within STEM majors**

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

https://www.census.gov/programs-surveys/acs/data/pums.html

Project Proposal Questions

- What are the best and worst degrees for men and women respectively?
- Salary difference between STEM majors and NON-STEM majors
- What majors attract the most by gender ~ major preference ?
- How much a salary gap is there between men and women for some majors?
- How do different categories of majors stack up against each other like engineering majors to science to math to arts ?
- What majors are the most popular college degrees?
- What is the salary gap between graduates and non-graduates?
- Unemployment rates respective to the major

Difference between STEM and Non-STEM major

- What falls under STEM and Non-STEM
- Top 5 median salaries by major
- Lowest Unemployment rates
- 5 Most popular majors
- Frequency Comparison
- Logistic Regression
- Nearest Centroid (NC) Classifier

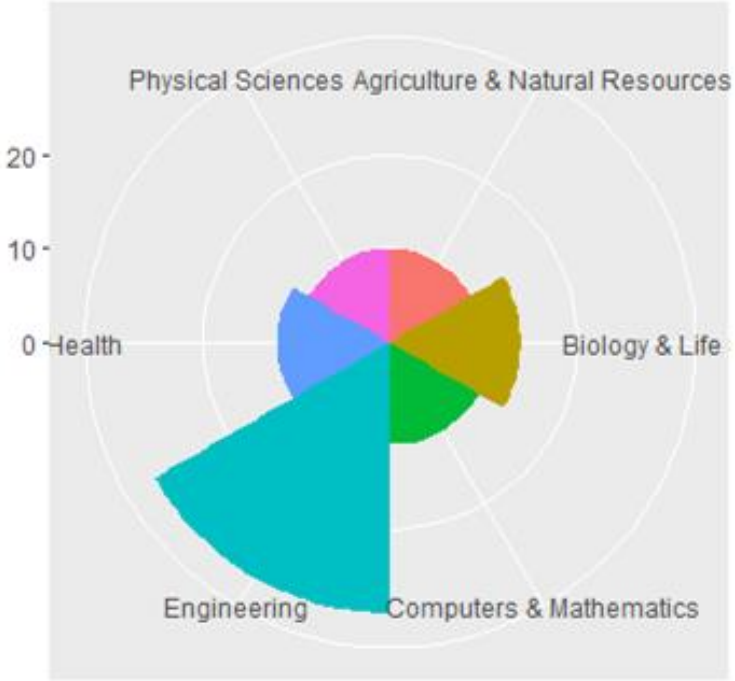
Subdivisions

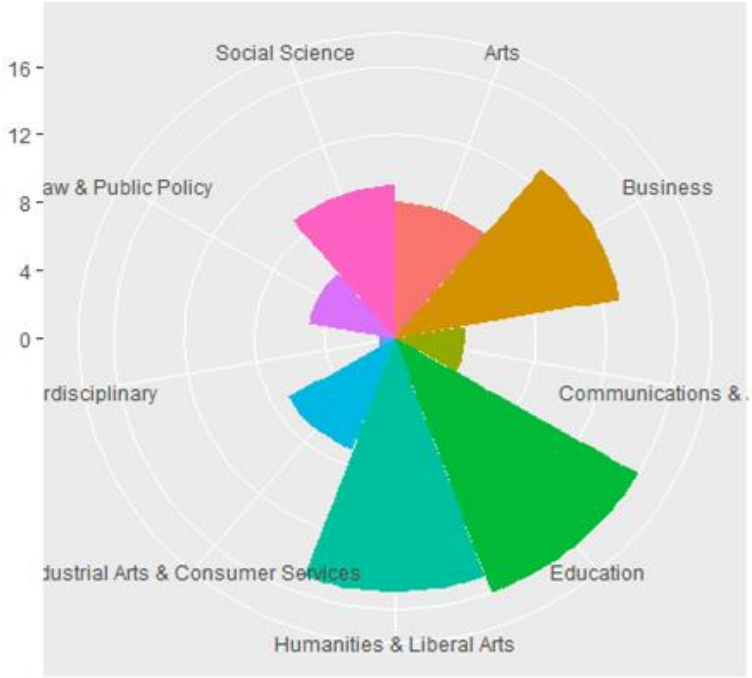
First step involves filtering the data to classify between STEM and non-STEM

STEM consists of Health, Physical Sciences, Agriculture & Resources, etc

Non-STEM consists of Business, Education, Arts, etc

STEM vs Non-STEM (pt.1)





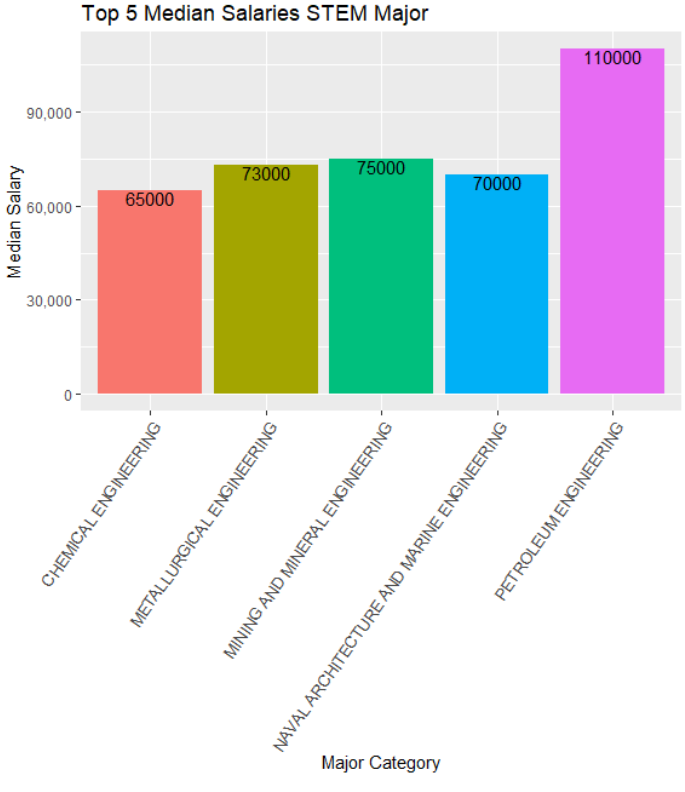
Median Salaries

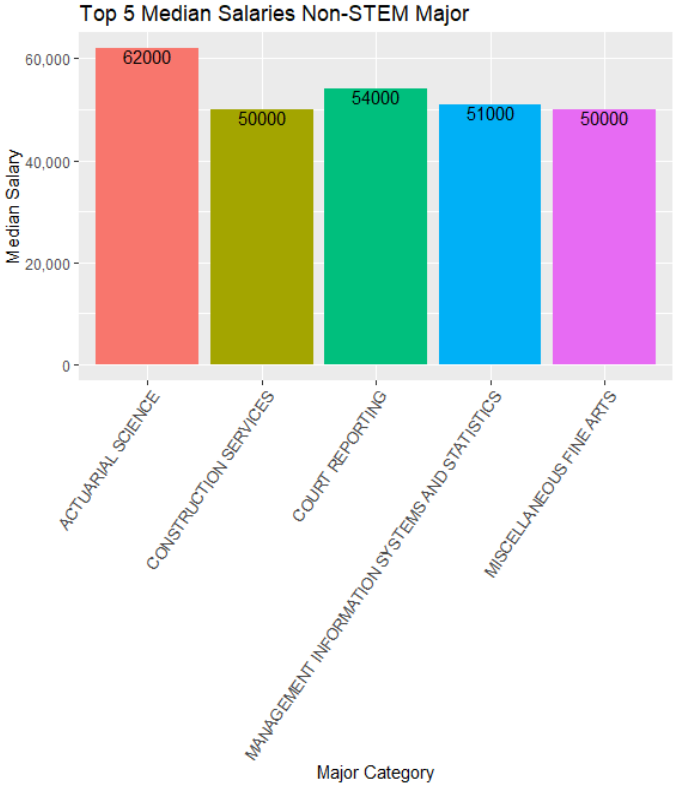
These are the top 5 highest paying majors

STEM: Petroleum Engineering

Non-STEM: Actuarial Science

STEM vs Non-STEM(pt.2)





Most Popular Majors
Lowest Unemployment Rates by Major

STEM vs Non-STEM pt.3

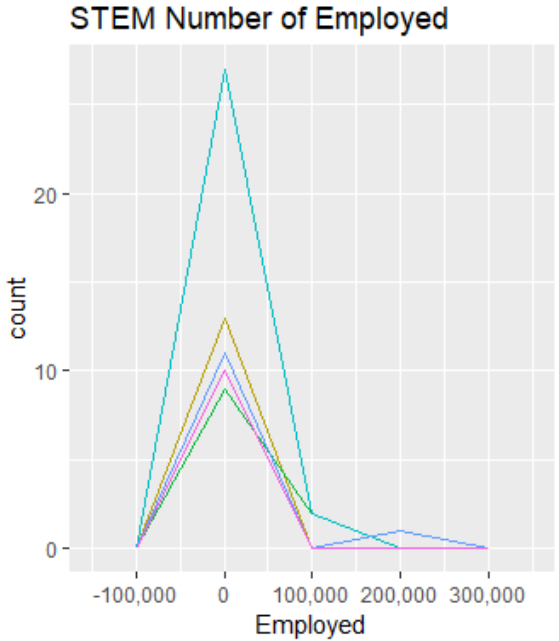
Biology	Business Management and Administration
Nursing	General Business
Computer Science	Communications
Mechanical Engineering	Marketing & Research
Electrical Engineering	Accounting

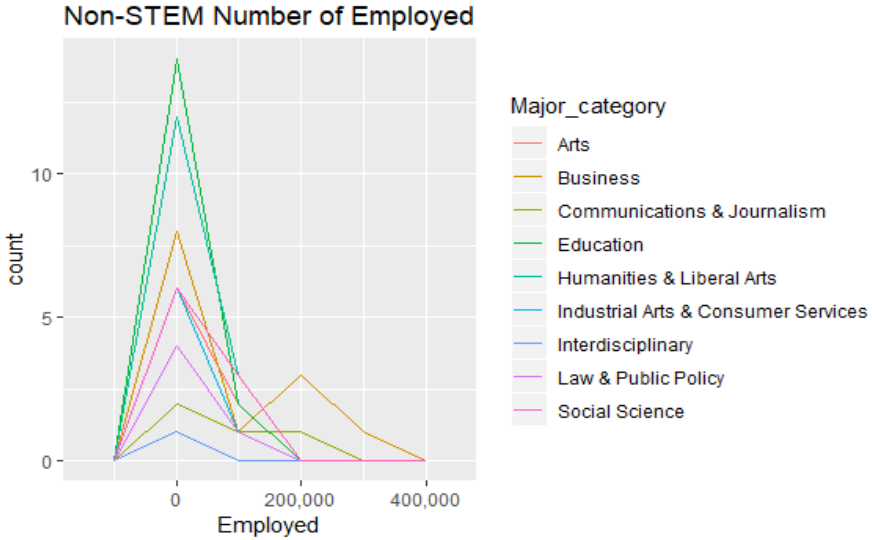
Botany (0.00)	Educational Administration and Supervision (0.00)
Mathematics & Computer Science (0.00)	Military Technologies (0.00)
Soil Science (0.00)	Court Reporting (0.011)
Engineering Mechanics Physics and Science (0.006)	Mathematics Teacher Education (0.016)
Petroleum Engineering (0.018)	Electrical, Mechanical, and Precision Technologies (0.029)

Number Employed

STEM: Engineering, Biology, Health
Non-STEM: Business, Social Science, Education

STEM vs Non-STEM pt.4





Logistic Regression & Model

```
model_glm_all = glm(SorN ~ Total+Unemployment_rate+Median+College_jobs ,
                    data = recentgrads_SorN_trn,
                    family = "binomial")
summary(model_glm_all)
```

```
Call:
glm(formula = SorN ~ Total + Unemployment_rate + Median + College_jobs,
    family = "binomial", data = recentgrads_SorN_trn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1990  -0.8799  -0.4259   0.9935   2.1391

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -4.964e+00  1.729e+00  -2.872  0.00408 **
Total            -1.594e-05  8.836e-06  -1.804  0.07130 .
Unemployment_rate  6.261e-01  8.627e+00   0.073  0.94214
Median            1.349e-04  4.187e-05   3.223  0.00127 **
College_jobs      3.065e-05  2.268e-05   1.352  0.17643
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 117.823  on 84  degrees of freedom
Residual deviance:  92.102  on 80  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 102.1

Number of Fisher Scoring iterations: 5
```

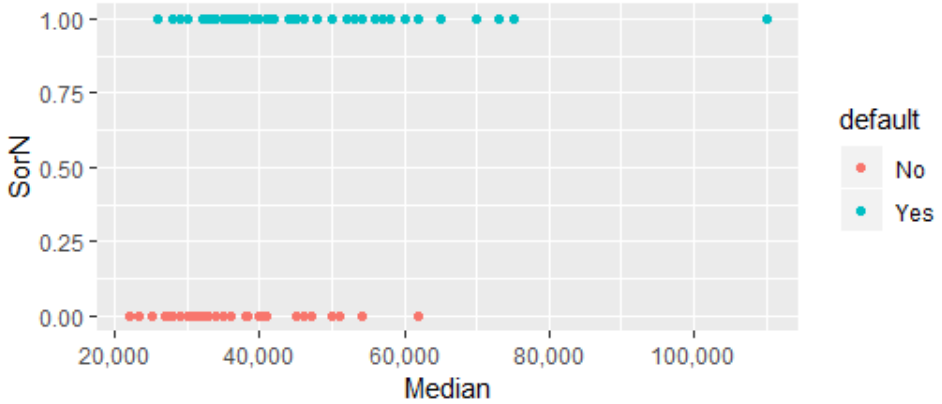
- **For College_jobs and Unemployment_rate**:
  - p-values are **0.176** and **0.942** respectively. As a result, the number with jobs requiring a college degree and unemployment rate are **NOT** significant predictors for STEM vs NONSTEM preference,

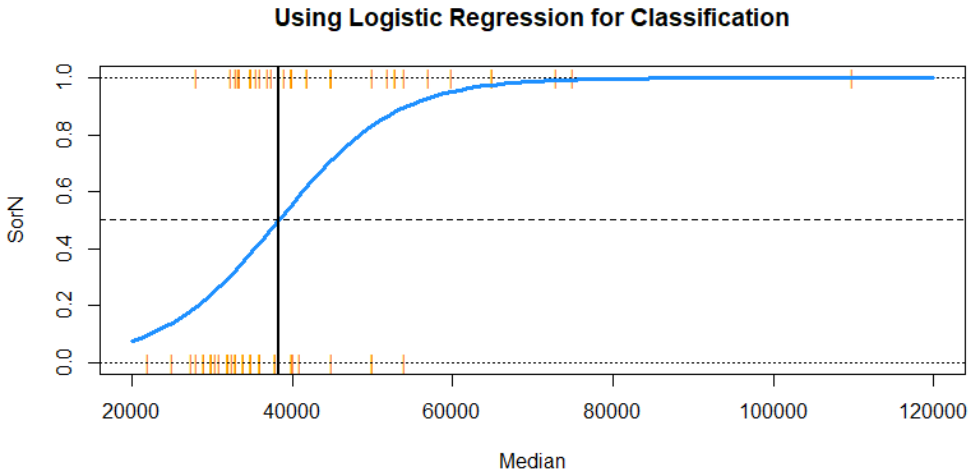
- **For Median Salary and Total**:
  - p-values are **0.001** and **0.071**, respectively, and are s**ignificantly less than 0.10**, which indicates that Median Salary and Total popularity of majors contributes to the prediction of joining STEM or NONSTEM.

Model = logOdds(Major_category -> STEM) = -4.964+ 0.0001349 *Median Salary

Logistic Regression : STEM v NONSTEM

- The blue “curve” is the predicted probabilities given by the fitted logistic regression.
- Orange lines represent Majors
- The solid vertical black line represents the decision boundary

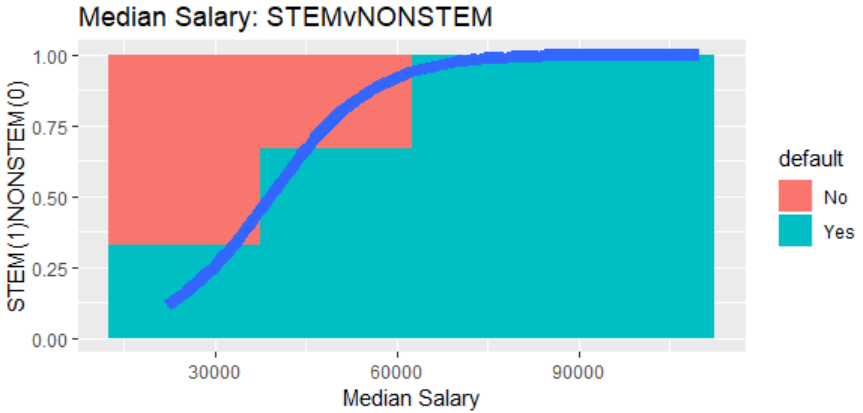


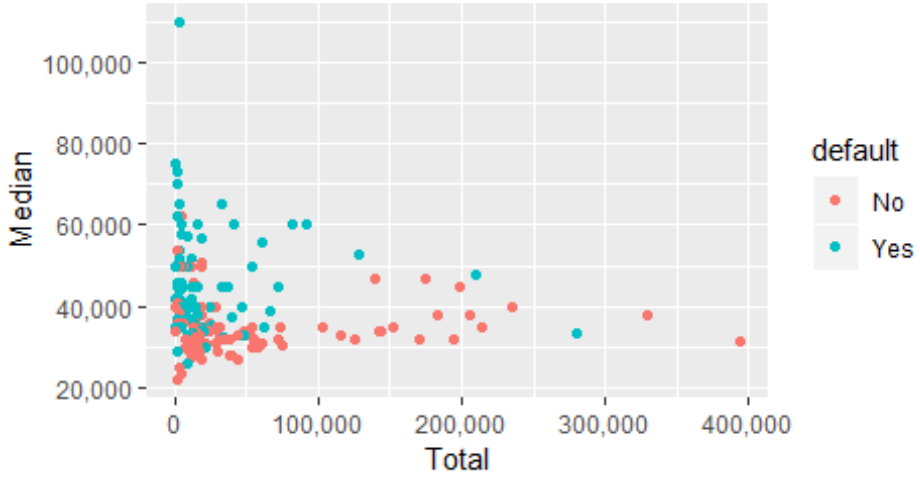


- Could spot all majors and range in salary for STEM and NONSTEM

STEM vs NonSTEM

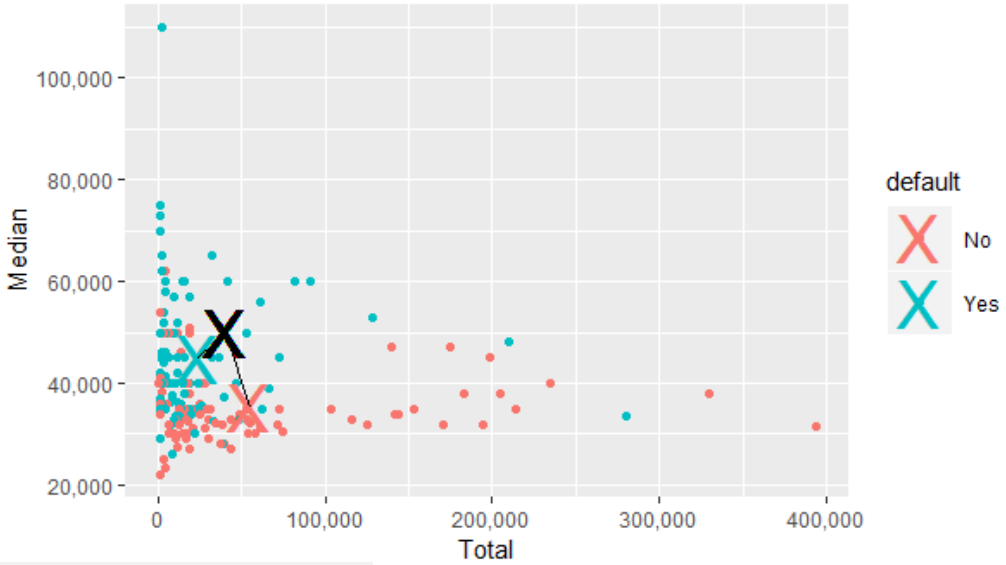
- No gaps
- Median Salary Range: 20,000-120,000
- STEM has a bigger range in salaries
- Estimates the conditional probability of STEM vs NonSTEM | Median Salary





Nearest Centroid (NC) classifier

- Provides mean difference classifier
- Computed the distance from the class mean to the test point
- Line Segment




```
# compute the observed class means
(obs.means <- recentgrads_sorN %>% select(default, Median, Total) %>%
    group_by(default) %>% summarise_all(mean, na.rm = TRUE))
(gg <- ggplot(recentgrads_sorN) +
    geom_point(aes(x = Total, y = Median, color = default), alpha = 1.0) +
    geom_point(x = x.test[1], y = x.test[2], shape = 'X', size = 10) +
    # plot the two class means
    geom_point(data = obs.means, shape = 'X', size = 10,
               aes(x = Total, y = Median, color = default)) +
  scale_y_continuous(labels = scales::comma) +
  scale_x_continuous(labels = scales::comma))
(dist <- obs.means %>% mutate(dist = sqrt((Total - x.test[1])^2 + (Median- x.test[2])^2)))
#add line segment to graph
gg + geom_segment(data = dist, aes(x = x.test[1], y = x.test[2], xend = Total, yend = Median))
```

default <chr>	Median <dbl>	Total <dbl>	dist <dbl>
No	35312.64	54676.69	20763.52
Yes	45046.51	23703.32	17032.88

Logistic Regression - Misclassification Rate

```
#logisticregression
logit.poly.fit <- glm(SorN ~ Median + Total, family = binomial(), data = recentgrads_SorN)
#predict the conditional prob
logit.fit.prob <- predict(logit.poly.fit, type = "response")
#Bayes rule
logit.fit.class <- as.factor(ifelse(logit.fit.prob > 0.5, "Yes", "No"))
#Misclassification error rate
mean(recentgrads_SorN$default != logit.fit.class)
```

Misclassification Rate = 0.2947977

The percentage of training and testing examples misclassified from a given data set.

R Code

```r
#graph for top5 stem median salaries
top5stem <-  head(med_sal_STEM, 5)
ggplot(data = top5stem, mapping= aes(x = Major, y = med_salary)) +
  geom_histogram(mapping = aes(fill= Major),  stat = "identity", show.legend = F) +
  scale_y_continuous(labels = scales::comma) +
  geom_text(mapping = aes(label = med_salary), vjust = 1) +
  labs(title = "Top 5 Median Salaries STEM Major", x = "Major Category", y= "Median Salary") +
  theme(axis.text.x = element_text(size = 10,angle = 55,hjust = 1,vjust = 1))
```

```r
## go through the each of the files and classify each record as stem and non-stem
majors <- recentgrads %>%
  select(Major_category) %>%
  distinct()
majors

STEM <- filter(recentgrads, recentgrads$Major_category == "Engineering"
               | recentgrads$Major_category == "Physical Sciences"
               | recentgrads$Major_category == "Computers & Mathematics"
               | recentgrads$Major_category == "Agriculture & Natural Resources"
               | recentgrads$Major_category == "Health"
               | recentgrads$Major_category == "Biology & Life Science")
STEM

NONSTEM <- filter(recentgrads, recentgrads$Major_category == "Business"
                  | recentgrads$Major_category == "Law & Public Policy"
                  | recentgrads$Major_category == "Industrial Arts & Consumer Services"
                  | recentgrads$Major_category == "Arts"
                  | recentgrads$Major_category == "Social Science"
                  | recentgrads$Major_category == "Education"
                  | recentgrads$Major_category == "Humanities & Liberal Arts"
                  | recentgrads$Major_category == "Pyschology & Social Work"
                  | recentgrads$Major_category == "Communications & Journalism"
                  | recentgrads$Major_category == "Interdisciplinary")

NONSTEM
```

```r
par(mar = c(7, 7, 7, 7))
#coxcomb for STEM by category
STEMbar <- ggplot(data = STEM) +
  geom_bar(mapping = aes(x = Major_category, fill = Major_category), show.legend = F, width = 1) +
  theme(aspect.ratio = 1) +
  labs(x = NULL,  y = NULL)
plot2 <- STEMbar + coord_polar()
grid.arrange(plot2, ncol = 2)
```

Difference within STEM major

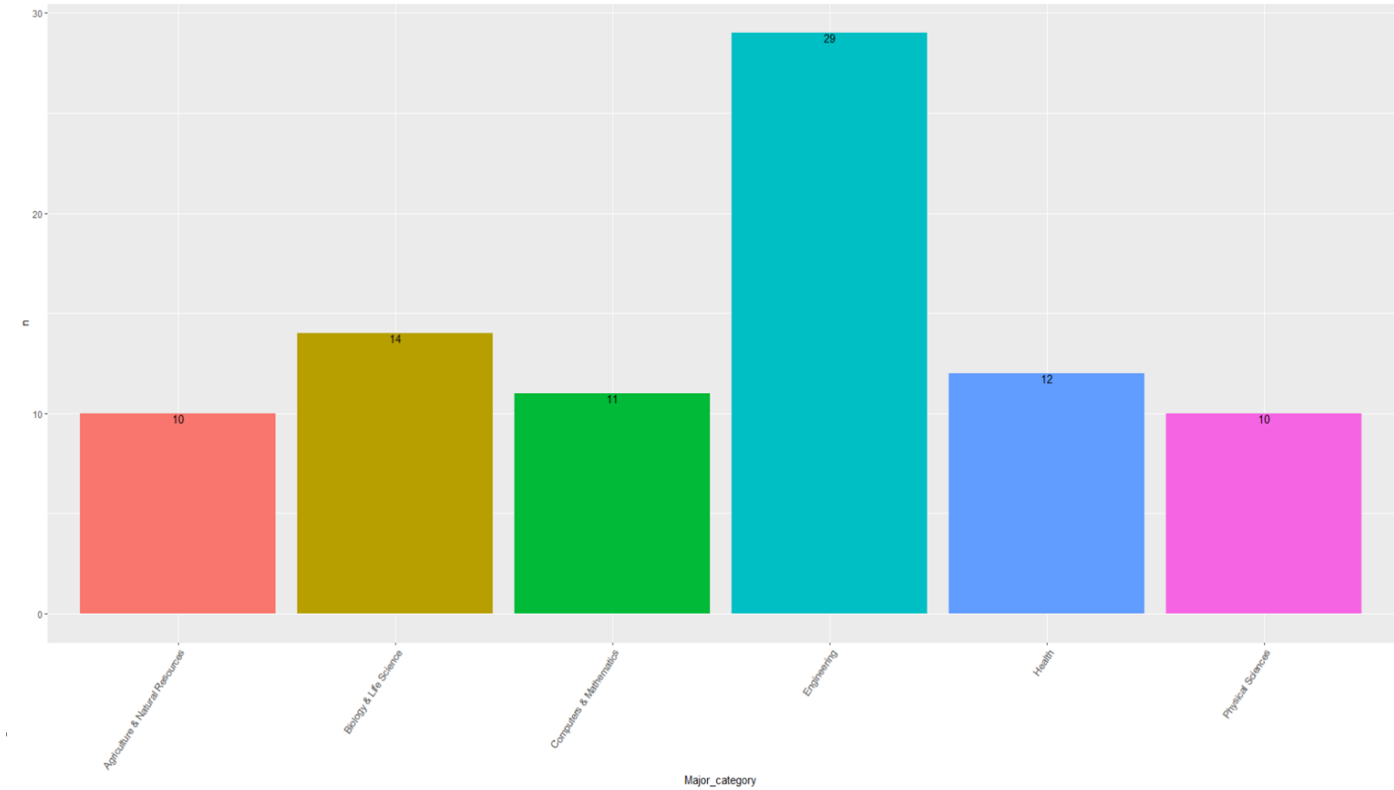
- STEM vs STEM

Differences Within Stem Majors (pt.1)

Amount of Stem Majors by each Category

Out of the 6 major stem categories, the one with the most diversity in majors is engineering.

The major categories with the least diversity in majors is Agriculture & Natural Resources and Physical Sciences.

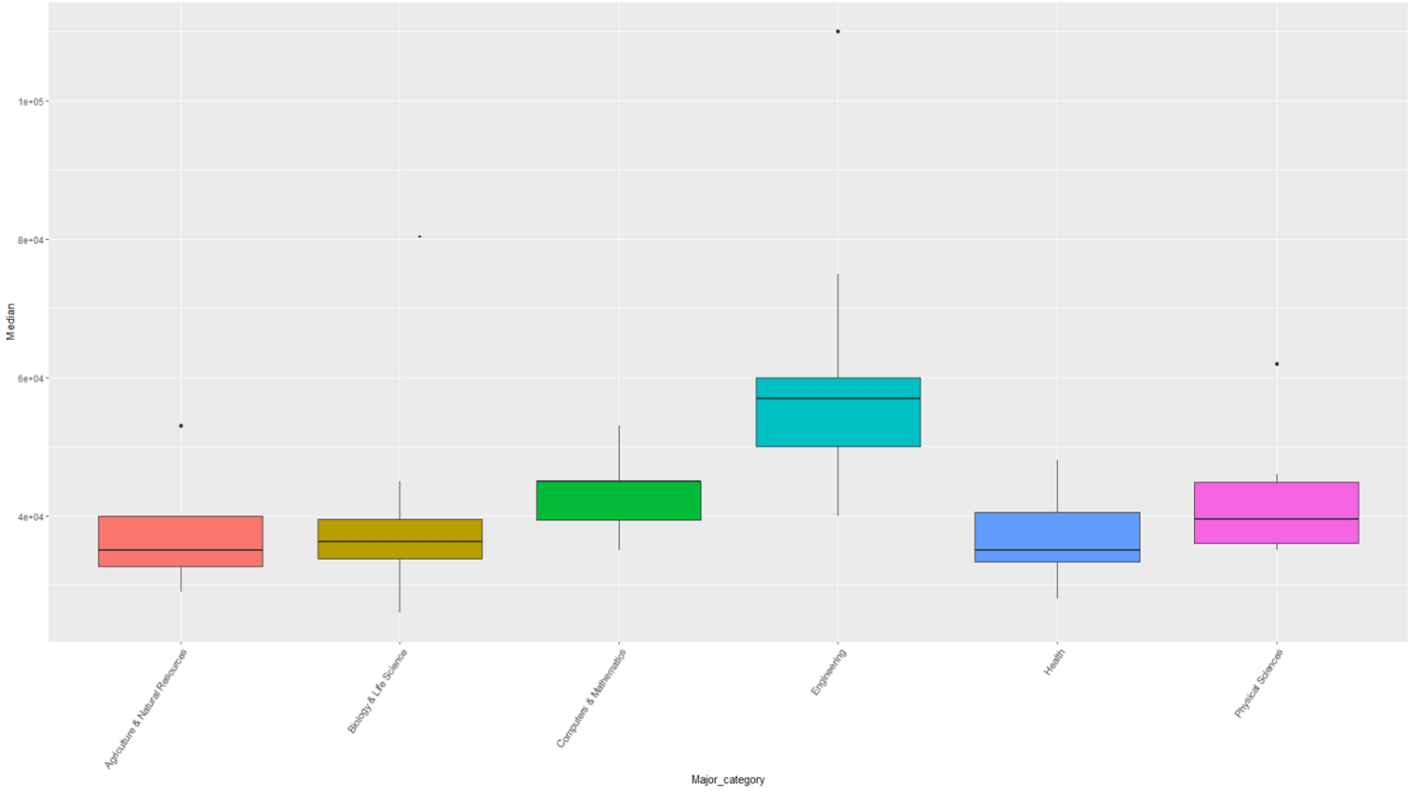


Differences Within Stem Majors (pt.2)

Boxplot of Median Salary of 6 Major Categories

From the boxplot, the highest median salary among the 6 major categories is engineering.

The lowest median salary is Agriculture and Natural Sciences.

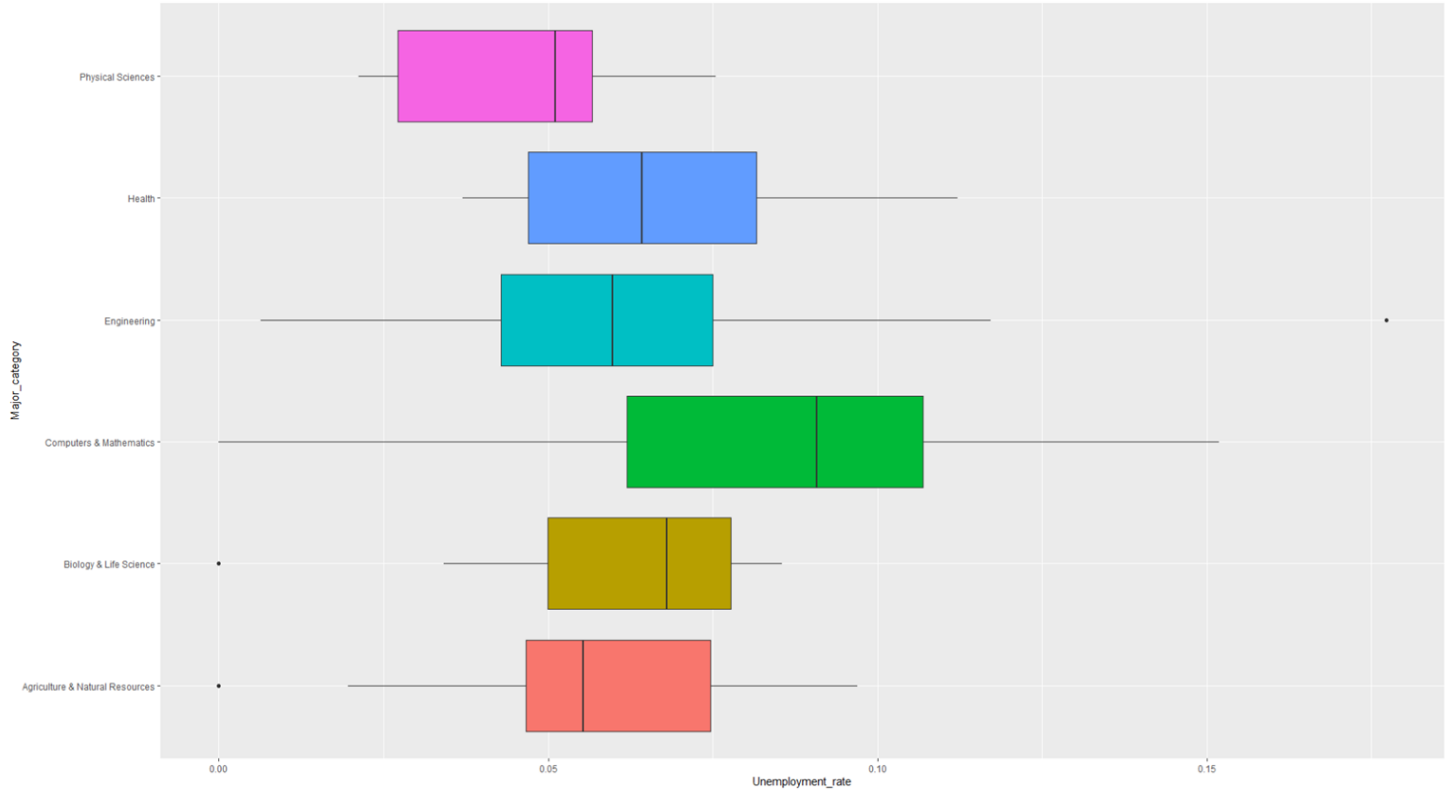


Differences Within Stem Majors (pt.3)

Boxplot of Unemployment Rate

Highest unemployment rate variance among 6 major categories is Computer & Mathematics.

Highest overall unemployment rate among 6 major categories is also Computers and Mathematics.



R Code: Within Stem Majors

```{r, collapse=T}
library(tidyverse) # for `ggplot2`, `dplyr`, and more
## data
library(readr)
library(dplyr)
recentgrads <- read_csv(file = "C:/Users/Huy Tran/Downloads/recent-grads.csv")
recentgrads
```

```{r,collapse = T}
## go through the each of the files and classify each record as stem and non-stem
majors <- recentgrads %>%
  select(Major_category) %>%
  distinct()
majors

STEM <- filter(recentgrads, recentgrads$Major_category == "Engineering"
               | recentgrads$Major_category == "Physical Sciences"
               | recentgrads$Major_category == "Computers & Mathematics"
               | recentgrads$Major_category == "Agriculture & Natural Resources"
               | recentgrads$Major_category == "Health"
               | recentgrads$Major_category == "Biology & Life Science")

STEM
df <- STEM %>% group_by(Major_category) %>%
  count(Major_category)
df

```

```{r}
ggplot(data = df) +
  geom_bar(mapping = aes(x = Major_category, y = n, fill = Major_category), stat = "identity") +
  theme(legend.position = "none", axis.text.x = element_text(size = 10,angle = 55,hjust = 1,vjust = 1)) +
 geom_text(mapping = aes(x = Major_category, y = n, label = n), vjust = 1)

```

```{r}
ggplot(data = STEM) +
  geom_boxplot(mapping = aes(x = Major_category, y = Median, fill = Major_category)) +
   theme(legend.position = "none", axis.text.x = element_text(size = 10,angle = 55,hjust = 1,vjust = 1))
```

```{r}
ggplot(data = STEM) +
  geom_boxplot(mapping = aes(x = Major_category, y = Unemployment_rate, fill = Major_category)) +
  coord_flip() +
 theme(legend.position = "none")
```

Conclusion

STEM vs. Non-STEM majors:

- ★ **Top 5 STEM major median salaries is much greater than that of Top 5 Non-STEM major**
- ★ **Median Salary and Popularity of majors are significant predictors for preference**

Within STEM majors:

- ★ Major category with the highest median salary: Engineering
- ★ Computer & Mathematics has the biggest diversity in unemployment rates