Highest Paying College Majors

Megan Eddy

2023-02-12

Introduction

Choosing a college major can be a daunting task. Incoming freshman have a wide variety of options to choose from. Advice and post-graduate employment information often varies from source to source. A key question college students may have concerns what their earning potential post-graduation will be with each major offered. It would be useful for information to be available to college students that would help them predict their chances of employment in their field as well as an approximation of the salary they can expect. Which major should college students pursue to increase their earning potential after receiving a bachelor's degree? Data science can provide the analysis of the current data to give this prediction.

Problem Statement

Which college major leads to the highest earning potential after graduating with a bachelor's degree?

Research questions

Draft 5-10 Research questions that focus on the problem statement/topic.

1. What are the college majors available?
2. What are the chances of employment after graduating with said major?
3. Do graduates tend to get jobs in fields related to their major?
4. What time frame should be considered?
5. How much does location affect salaries?
6. How does continuing education affect the salaries of graduates from said majors?
7. Do some majors need to be left out? (i.e. pre-med, pre-law)
8. Is there a difference in salary depending on the gender of the graduate?

Approach

Only majors with available data in the selected datasets will be considered. They will be compared to each other using mean salary information. Other predictors will be compared to determine if they could be influencing any differences in pay. Outliers will be identified and handled as is deemed appropriate to maintain the integrity of the results. The data will be further cleaned based on relevance of potential predictors. Employment numbers will also be considered since it does not really matter how much salary a major yields if the graduate cannot find employment in the field.

This will be done using the model recommended in the *Future Machine Learning* section.

The approach will directly compare various majors to their average earnings. By controlling for other variables, the results should paint a fairly accurate picture of which majors lead to the highest earnings. The scope will be limited by the time frame the data was collected as it is fairly old. It will also be limited to only the majors appearing in these datasets.

Data

Summary of Steps Taken To address the problem statement, data was taken from a reputable source and investigated to better determine its validity and meaning. It was then trimmed to include information related only to the direct question of which college major yields the highest earning potential. Columns were removed from the original datasets and then they were combined. Several graphs were plotted to give an idea of what the combined data looked like.

In order to actually analyse the data. The following model is recommended:

I would check the data to ensure it is normally distributed. I would also check homoscedasticity and the independence of errors. Assuming the parameters are met, I would then continue using Pearson's correlation method. I would look at the regression coefficients of each potential predictor of salary available. This would help determine if higher salaries are actually correlated to the major itself and not a third variable. The R2 and Adjusted R2 stats would be looked at to determine how much variation each predictor accounts for. Confidence intervals would also be calculated to help determine whether the model is a good fit. The data would be divided into training and test sets to verify the accuracy of the predictors of salary. This information could then be extrapolated to predict salaries for other majors not included in the dataset.

Description The data was originally obtained by FiveThirtyEight for an article about the earnings of college graduates. It was compiled by them from information gathered in the American Community Survey 2010-2012 Public Use Microdata Series. The purpose of the article was to highlight what specific college majors made, whether there was a discrepency between genders and also gives information on popularity of majors. The data was published on dataworld.com in 2017 but the data itself was gathered from 2010 - 2012.

The number of variables and column explanations in each dataset are as follows: (Note: The Class of Data is reported by the authors of the data and does not necessarily translate into the class of data it falls under in R)

Table 1: recent-grads: 20 variables

Column Title	Column Description	Class of Data
Rank	The rank of the major in terms of popularity	Integer
Major_code	The code associated with the major	Integer
Major_category	The category of the major	String
Total	The total number of students in the major	Integer
Sample_size	The sample size of the major	Integer
Men	The number of male students in the major	Integer
Women	The number of female students in the major	Integer
ShareWomen	The percentage of female students in the major	Float
Employed	The number of employed graduates from the major	Integer
Full_time	The number of full-time employed graduates from the major	Integer
Part_time	The number of part-time employed graduates from the major	Integer
Full_time_year_round	The number of full-time year-round employed graduates from the major	Integer
Unemployed	The number of unemployed graduates from the major	Integer

Column Title	Column Description	Class of Data
Unemployment_rate	The unemployment rate of graduates from the major	Float
Median	The median salary of graduates from the major	Integer
P25th	The 25th percentile salary of graduates from the major	Integer
P75th	The 75th percentile salary of graduates from the major	Integer
College_jobs	The number of college jobs held by graduates from the major	Integer
Non_college_jobs	The number of non-college jobs held by graduates from the major	Integer
Low_wage_jobs	The number of low-wage jobs held by graduates from the major	Integer

Table 2: grad-students: 21 variables

Column Title	Column Description	Class of Data
Major	The specific major of the field of study	String
Major_category	The category of the major	String
Grad_total	The total number of graduates from the major	Integer
Grad_sample_size	The sample size of graduates from the major	Integer
Grad_employed	The number of graduates employed	Integer
Grad_full_time_year_round	The number of graduates employed full-time year-round	Integer
Grad_unemployed	The number of graduates unemployed	Integer
Grad_unemployment_rate	The unemployment rate of graduates	Float
Grad_median	The median salary of graduates	Integer
Grad_P25	The 25th percentile salary of graduates	Integer
Grad_P75	The 75th percentile salary of graduates	Integer
Nongrad_total	The total number of non-graduates from the major	Integer
Nongrad_employed	The number of non-graduates employed	Integer
Nongrad_full_time_year_round	The number of non-graduates employed full-time year-round	Integer
Nongrad_unemployed	The number of non-graduates unemployed	Integer
Nongrad_unemployment_rate	The unemployment rate of non-graduates	Float
Nongrad_median	The median salary of non-graduates	Integer
Nongrad_P25	The 25th percentile salary of non-graduates	Integer
Nongrad_P75	The 75th percentile salary of non-graduates	Integer
Grad_share	The share of graduates in the major	Float
Grad_premium	The difference between the median salary of graduates and non-graduates	Integer

Table 3: all-ages: 10 variables

Column Title	Column Description	Class of Data
Major	The specific major of the field of study	String
Major_category	The category of the major	String
Total	The total number of students in the major	Integer
Employed	The number of employed graduates from the major	Integer
Unemployed	The number of unemployed graduates from the major	Integer
Unemployment_rate	The unemployment rate of graduates from the major	Float
Median	The median salary of graduates from the major	Integer
P25th	The 25th percentile salary of graduates from the major	Integer
P75th	The 75th percentile salary of graduates from the major	Integer

Column Title	Column Description	Class of Data
Employed_full_time_year_round	The number of employed graduates from the major who are employed full-time year-round	Integer

There do not appear to be any discrepancies or N/A values present in the datasets. In order to combine them column names will likely have to be changed for consistency.

Importing the Data The data is saved in csv format in the same location in my working directory as my project file. I will use read.csv to import the datasets and save them as dataframes.

Required Packages Identify the packages that are needed for your project.

- ggplot2
- dplyr
- readxl
- car

Data Manipulation & Cleaning The plan is to summarize the relevant data into one data frame, taking only the variables that inform about the college major chosen, the number of students in the major, the number of graduates from the major, employment information and salary information. This will reduce the data to a more reasonable size. Each data frame will first be cut to represent only the relevant data and then cleaned before finally being combined.

Recent Grads

```
##                                    Major Total Sample_size Employed
## 1                    PETROLEUM ENGINEERING  2339          36     1976
## 2            MINING AND MINERAL ENGINEERING   756           7      640
## 3                METALLURGICAL ENGINEERING   856           3      648
## 4 NAVAL ARCHITECTURE AND MARINE ENGINEERING  1258          16      758
## 5                     CHEMICAL ENGINEERING 32260         289    25694
## 6                      NUCLEAR ENGINEERING  2573          17     1857
##   Unemployed Unemployment_rate Median P25th  P75th College_jobs
## 1         37        0.01838053 110000 95000 125000         1534
## 2         85        0.11724138  75000 55000  90000          350
## 3         16        0.02409639  73000 50000 105000          456
## 4         40        0.05012531  70000 43000  80000          529
## 5       1672        0.06109771  65000 50000  75000        18314
## 6        400        0.17722641  65000 50000 102000         1142
##   Non_college_jobs
## 1              364
## 2              257
## 3              176
## 4              102
## 5             4440
## 6              657
```

```
##                                     Major Total Employed Unemployed
## 1                     PETROLEUM ENGINEERING  2339     1976         37
## 2            MINING AND MINERAL ENGINEERING   756      640         85
## 3                 METALLURGICAL ENGINEERING   856      648         16
## 4 NAVAL ARCHITECTURE AND MARINE ENGINEERING  1258      758         40
## 5                      CHEMICAL ENGINEERING 32260    25694       1672
## 6                       NUCLEAR ENGINEERING  2573     1857        400
##   Unemployment_rate Median P25th  P75th
## 1        0.01838053 110000 95000 125000
## 2        0.11724138  75000 55000  90000
## 3        0.02409639  73000 50000 105000
## 4        0.05012531  70000 43000  80000
## 5        0.06109771  65000 50000  75000
## 6        0.17722641  65000 50000 102000
```

There are 173 observations across 11 variables. Data classes are: character, integer and numeric.

The column names all make sense and none are mislabeled so none will be changed at this time.

None of the character values in either the Major or Major_category columns are misspelled nor erroneous.

The Sample Size was checked against the Total to ensure none of the values exceeded those of the total. They did not

None of the Unemployment rate values exceeded 1, however there is an N/A which indicates the unemployment rate is 0% for that observation

The Median should have a value between the P25th and P75th. The median is less than all P75th values.

There are 3 instances where the P25th value exceeds or is equal to that of the Median.

- Value 57: The P25th value and Median value are equal.
- Value 75: The P25th, Median and P75th values are all equal.
- Value 172: The P25th and Median values are equal.

This may mean that the salary range for grads from these majors is not large which could be an indicator of small sample size.

The number of employed graduates should be equal to or greater than the sum of college jobs and non-college jobs. This holds true for all observations.

Grad Students

```
##                                  Major Grad_total Grad_sample_size
## 1                CONSTRUCTION SERVICES       9173              200
## 2     COMMERCIAL ART AND GRAPHIC DESIGN      53864              882
## 3                HOSPITALITY MANAGEMENT      24417              437
## 4 COSMETOLOGY SERVICES AND CULINARY ARTS       5411               72
## 5             COMMUNICATION TECHNOLOGIES       9109              171
## 6                       COURT REPORTING       1542               22
##   Grad_employed Grad_unemployed Grad_unemployment_rate Grad_median Grad_P25
## 1          7098             681             0.08754339       75000    53000
## 2         40492            2482             0.05775585       60000    40000
## 3         18368            1465             0.07386679       65000    45000
## 4          3590             316             0.08090118       47000    24500
## 5          7512             466             0.05841063       57000    40600
```

```
## 6          1008               0             0.00000000       75000    55000
##   Grad_P75
## 1   110000
## 2    89000
## 3   100000
## 4    85000
## 5    83700
## 6   120000
```

```
##                                    Major Total Employed Unemployed
## 1                  CONSTRUCTION SERVICES  9173     7098        681
## 2      COMMERCIAL ART AND GRAPHIC DESIGN 53864    40492       2482
## 3                 HOSPITALITY MANAGEMENT 24417    18368       1465
## 4 COSMETOLOGY SERVICES AND CULINARY ARTS  5411     3590        316
## 5             COMMUNICATION TECHNOLOGIES  9109     7512        466
## 6                         COURT REPORTING  1542     1008          0
##   Unemployment_rate Median P25th  P75th
## 1        0.08754339  75000 53000 110000
## 2        0.05775585  60000 40000  89000
## 3        0.07386679  65000 45000 100000
## 4        0.08090118  47000 24500  85000
## 5        0.05841063  57000 40600  83700
## 6        0.00000000  75000 55000 120000
```

There are 173 observations across 9 variables. Data classes are: character, integer and numeric.

The column names all make sense and none are mislabeled so none will be changed at this time.

None of the character values in either the Major column are misspelled nor erroneous.

The Sample size column was checked against the Total column and did not exceed its value at any observation.

The values for the Grad unemployment rate and Nongrad unemployment rate were all less than 1.

The Grad median, Grad P25th and Grad P75 values were all as expected such that the median value fell between the P25th and P75th values for all observations.

The sum of Nongrad employed and Nongrad unemployed does not equate to the value of the Nongrad total. The values in the Nongrad Total column are always higher than the values in the other two columns combined. This indicates that the Nongrad total column is not representative of sample size for the income information provided.

all-ages

```
##                                    Major  Total Employed Unemployed
## 1                    GENERAL AGRICULTURE 128148    90245       2423
## 2 AGRICULTURE PRODUCTION AND MANAGEMENT  95326    76865       2266
## 3                 AGRICULTURAL ECONOMICS  33955    26321        821
## 4                      ANIMAL SCIENCES 103549    81177       3619
## 5                         FOOD SCIENCE  24280    17281        894
## 6            PLANT SCIENCE AND AGRONOMY  79409    63043       2070
##   Unemployment_rate Median P25th P75th
## 1        0.02614711  50000 34000 80000
## 2        0.02863606  54000 36000 80000
## 3        0.03024832  63000 40000 98000
## 4        0.04267890  46000 30000 72000
## 5        0.04918845  62000 38500 90000
## 6        0.03179089  50000 35000 75000
```

There are 173 observations across 8 variables. Data classes are: character, integer and numeric.

The column names all make sense and none are mislabeled so none will be changed at this time.

None of the character values in either the Major column are misspelled nor erroneous.

The Total number of graduates exceeds the sum of Employed and Unemployed graduates which is to be expected.

The Employment rate is never above 1.

The Grad median, Grad P25th and Grad P75 values were all as expected such that the median value fell between the P25th and P75th values for all observations.

Slicing & Dicing I included only columns relevant to the problem statement. To create my graphs, I narrowed the data down further to only include the top 10 observations for each dataset. I also further reduced the columns included in order to combine all three datasets together.

Final Data Set To achieve the Final Data set, columns were renamed in each separate data frame so they matched one another. The datasets were also trimmed further so dataframes were available that only contained the information shared between the three.

```
##               Major   Total Employed Unemployed Unemployment_rate Median P25th
## 1        ACCOUNTING  569677   451610      19729        0.04185735  88000 59000
## 2        ACCOUNTING  198633   165527      12411        0.06974901  45000 34000
## 3        ACCOUNTING 1779219  1335825      75379        0.05341467  65000 42500
## 4 ACTUARIAL SCIENCE    2472     2020        162        0.07424381 110000 80000
## 5 ACTUARIAL SCIENCE    3777     2912        308        0.09565217  62000 53000
## 6 ACTUARIAL SCIENCE    9763     7846        466        0.05606352  72000 53000
##    P75th
## 1 131000
## 2  56000
## 3 100000
## 4 150000
## 5  72000
## 6 115000
```

Questions for Future Importing/Cleaning Steps I need to learn how to do the $<$, $>$, $==$ functions using a pipe. It would clean up my code nicely.

I need to learn how to represent each dataframe in a clean, concise manner to show how I selected only some of the columns from the original dataset.

I could use work on identifying outliers to clean up my data.

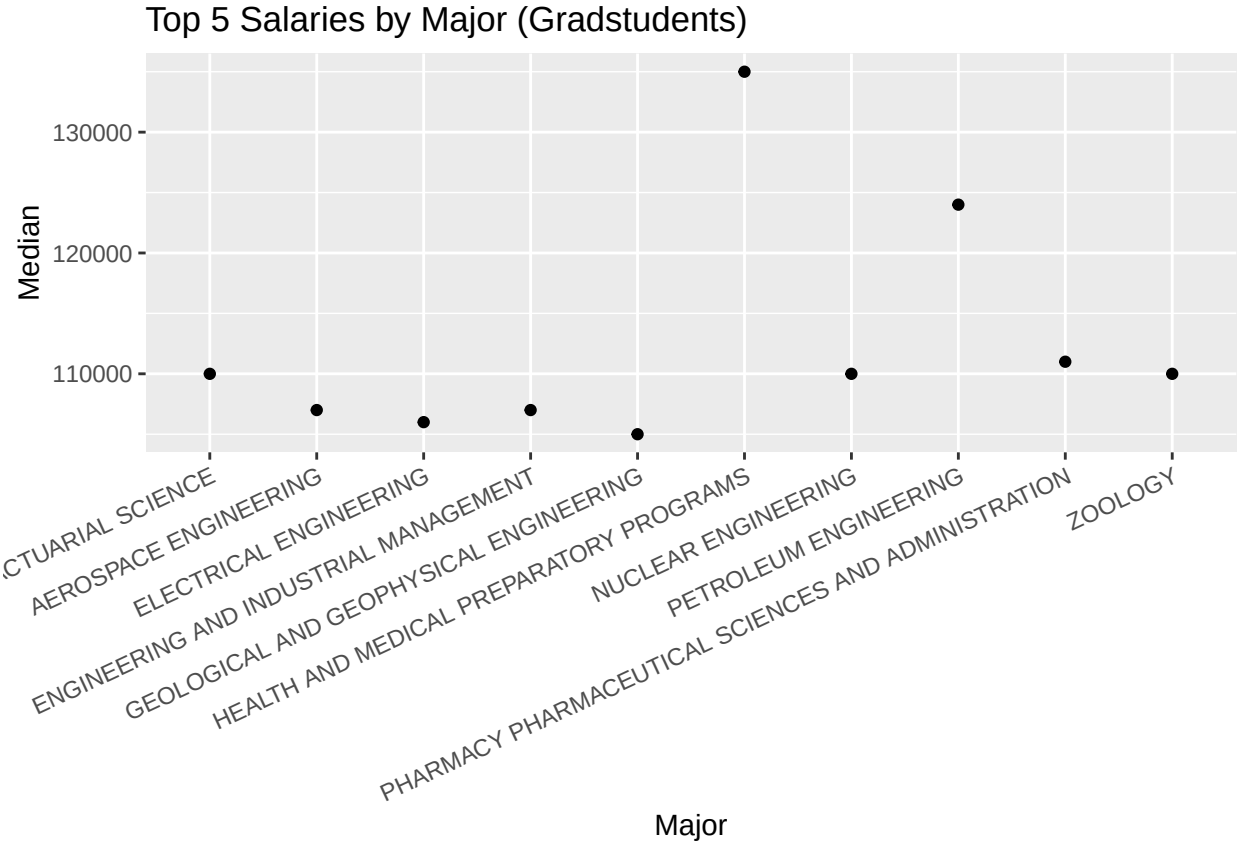
I need to figure out how to combine data that is representative of the same major. I would also need to know more about the data for each major to ensure it can be combined. The numbers are different for each Major observation but I do not currently know why.

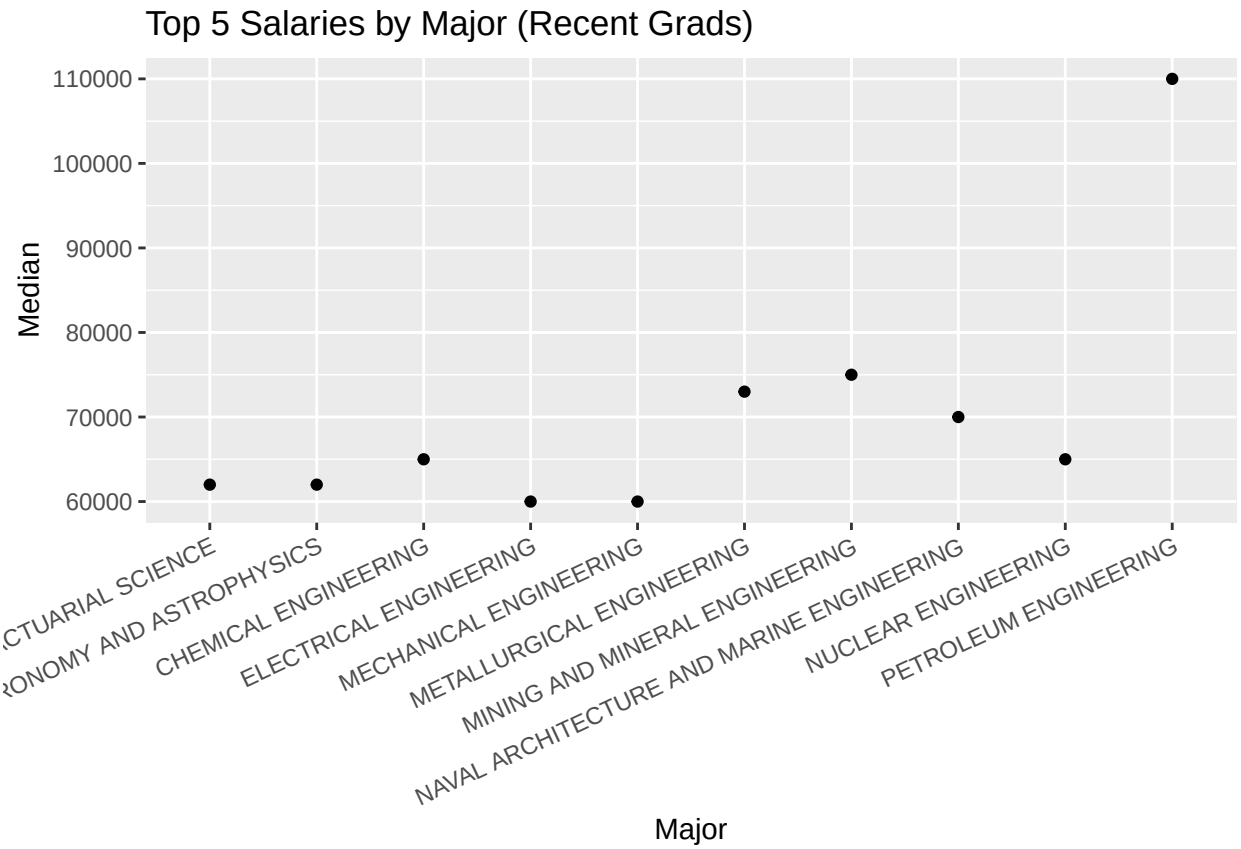
Plots and Table Needs What types of plots and tables will help you to illustrate the findings to your research questions?

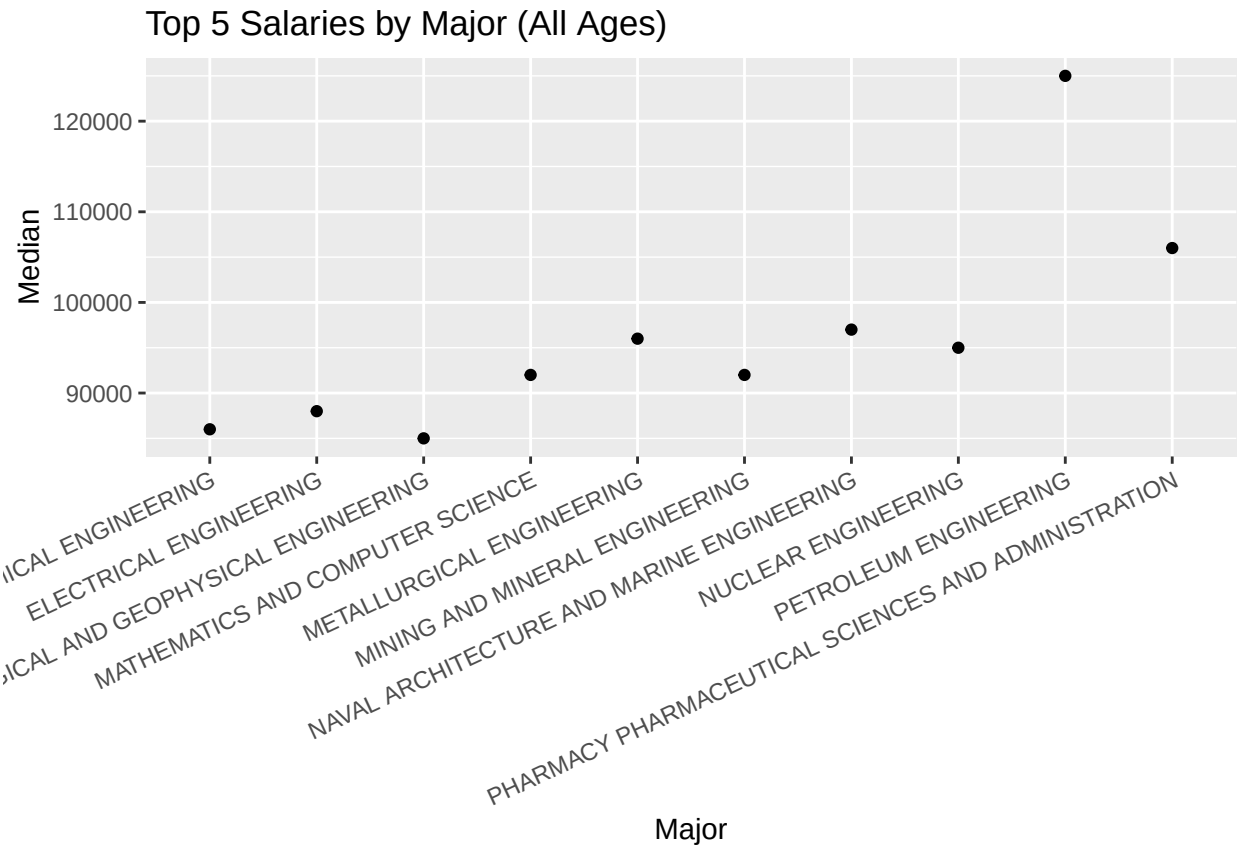
Scatterplot: Will help visualize the data distributions. I will use a scatterplot to show an example of the majors and their respective median salaries. There are too many majors to include them all on a single scatterplot so only the top 10 from each data set will be plotted.

Histogram: Will help check for normal distribution and create a visual of findings.

Q-Q Plot: Will help check the distribution for potential outliers influencing the results. This would need to be done after the model was built.




Top 5 Salaries by Major (Recent Grads)
Median
110000
100000
90000
80000
70000
60000
CTUARIAL SCIENCE
RONOMY AND ASTROPHYSICS
CHEMICAL ENGINEERING
ELECTRICAL ENGINEERING
MECHANICAL ENGINEERING
METALLURGICAL ENGINEERING
MINING AND MINERAL ENGINEERING
NAVAL ARCHITECTURE AND MARINE ENGINEERING
NUCLEAR ENGINEERING
PETROLEUM ENGINEERING
Major



Information That is Not Self-Evident The ranges of the salaries by major is not self evident. A column could be added to include them. It might be important to know the range because it might affect a student's choice of major.

The total range for all salaries could be shown to provide a better idea of the entire range of the dataset.

It might be worth comparing the total number of students enrolled in a major to the total number of graduates.

Different Ways to Look at the Data The majors associated with the highest salary and lowest unemployment rate might be worth investigating.

It also might be worth further exploring the dataset that compares graduate salaries to non-graduate salaries.

Summarizing the Data to Answer Key Questions

The top 5 salaries and their majors could be shown using the median value.

The top 5 salaries and their majors could be shown using the P75th value.

The top 5 majors with the lowest unemployment rates could be shown to highlight the majors that offer the best chances of obtaining employment. Their salaries could also be shown

Questions for future steps

I need to gain a better grasp on how to clean the data from the beginning of the analysis. This includes checking for and removing outliers.

I also need a better grasp on how to combine data across various datasets after determining which are important to answer the question.

I need to review pulling references from a .bib file as I could not get it to work for this part of the project.

I need to incorporate more visuals such as a bar graph comparing the differences in salaries for the same major that are shown across the three datasets.

Future Machine Learning

I would incorporate the machine learning technique of linear regression to the dataset. This would allow for future predictions about salary as opposed to only the current snapshot.

I would check the data to ensure it is normally distributed. I would also check homoscedasticity and the independence of errors. Assuming the parameters are met, I would then continue using Pearson's correlation method. I would look at the regression coefficients of each potential predictor of salary available. This would help determine if higher salaries are actually correlated to the major itself and not a third variable. The R2 and Adjusted R2 stats would be looked at to determine how much variation each predictor accounts for. Confidence intervals would also be calculated to help determine whether the model is a good fit. The data would be divided into training and test sets to verify the accuracy of the predictors of salary. This information could then be extrapolated to predict salaries for other majors not included in the dataset.

Analysis

It would be very interesting to see what other predictors lead to an increase in future earning potential. Its very possible that the major of choice does not necessarily influence the earning potential as much as I might think.

I found the cleaning of the data to be interesting because I didn't end up understanding the original data as much as I thought I did. I believed all the datasets were from the same original data with extra variables included in some of them. As I compared them to each other, though, I discovered that this is likely not the case. The numbers for the same variables did not match up across datasets as one would expect if it was the same data.

Implications

The implications of this research could be wide reaching. It could be provided to counselors both at colleges and high schools to assist students in deciding what major they would like to choose. It would offer guidance that may not be readily available at the current time. This would also help students later because they would not be as likely to be surprised by the salaries they receive once in the work force. It could also assist them in determining whether a major is worth taking out large student loan amounts. It would lead to better informed students who would later become better informed new employees.

Limitations

At this time, this analysis is limited in several ways. It is unknown which school the reported students graduated from. There might be a large difference in salaries between states, universities, cities, etc. This data does not take those differences into account.

The data is also relatively old. It would be best to augment the model with data from the last several years to ensure salaries are more accurately represented.

It is also limited to the majors available, however, predictions could be made about other majors if predictors were found to be significant that are shared between majors. It would be best if a larger variety of majors could be incorporated.

Sample size of graduates could always be increased.

More data could be collected on where (i.e. Industry, Private) graduates from each major are employed. There are potentially large differences in salary based on where a graduate works.

Concluding Remarks

After selecting, cleaning and suggesting a model for the analyses of the three datasets, it became clear that more work needs to be performed to determine what college majors lead to the highest earning potential in the job market. The data chosen was already clean in its original state, but columns had to be selected carefully to combine the three into one intelligible dataset. It would be worth the effort to analyse each dataset on its own to glean more information and assess the influence of predictors other than major on future earning potential.

Using the linear model in this paper, it would be possible to gain insight into which college majors lead to larger incomes later. This could be a helpful tool for college students to use when deciding which career path they would like to pursue.

Bibliography

- datasetage{fivethirtyeight2017,
  title={all-ages.csv},
  author={FiveThirtyEight},
  series={College Majors and Their Graduates},
  url={https://www.kaggle.com/datasets/thedevastator/uncovering-insights-to-college-majors-and-their?resource=download&select=all-ages.csv},
  year={2017},
  publisher={ABC News} }
- datasetgrad{fivethirtyeight2017,
  title={grad-students.csv},
  author={FiveThirtyEight},
  series={College Majors and Their Graduates},
  url={https://www.kaggle.com/datasets/thedevastator/uncovering-insights-to-college-majors-and-their?select=grad-students.csv},
  year={2017},
  publisher={ABC News} }
- datasetrecent{fivethirtyeight2017,
  title={recent-grads.csv},
  author={FiveThirtyEight},
  series={College Majors and Their Graduates},
  url={https://www.kaggle.com/datasets/thedevastator/uncovering-insights-to-college-majors-and-their?select=recent-grads.csv},
  year={2017},
  publisher={ABC News} }