How to get a Bag from College

CECS 450

December 8, 2023

Jaelyn Garrard

Matthew Halderman

Sarah Santos

Bonnie White

Table of Contents

Data Selection

The first step of our project was selecting the type of data we wanted to work on. This was an extensive process as we tried to gather several data sets of various kinds of topics to choose from online. Initially, we thought this process would be simple and quick. However, it became a challenge when looking into the data sets and not finding enough information to form our own data sets for this project.

The final data sets we chose related to colleges, majors, and salaries. With these data sets, we wanted to visualize the projected salary one would statistically be given based on their education level and type. The four data sets we worked on were based on college type, degrees, majors, and region.

Salaries By College Region

Hypothesis

When thinking about college graduate salaries by region, I predicted that California would be the region with the highest earnings. I came to this guess because I learned that California alone is the 7th largest economy in the world, so naturally with all of the tech companies and large corporations out here, it would make sense for graduates from here to have the highest earnings.

Collection & Preparation & Cleaning

Upon first receiving the dataset my first goal was to look through the data and record what needed to be cleaned. To do so, I used Jupyter Notebook and Python libraries such as Matplotlib and pandas to process and visualize data with supportive tools. Upon plotting the data for the first time, I created a graph that clearly did not understand the numerical values in the dataset, so I cleaned the data by changing the numerical datatypes to floats by stripping any non-numerical characters with regex and clearing null values.

Exploration

After testing and confirming that the numerical values were being recognized, I first generated a visualization of all data points in the set to see if I could make any observations right off the bat. As you can see in my notebook, the 2nd generated graph is too information-dense, I can see that there is a top earner, but I can't see names to observe any trends. As a solution, I figured it would be best to parse the data into different sections by region.

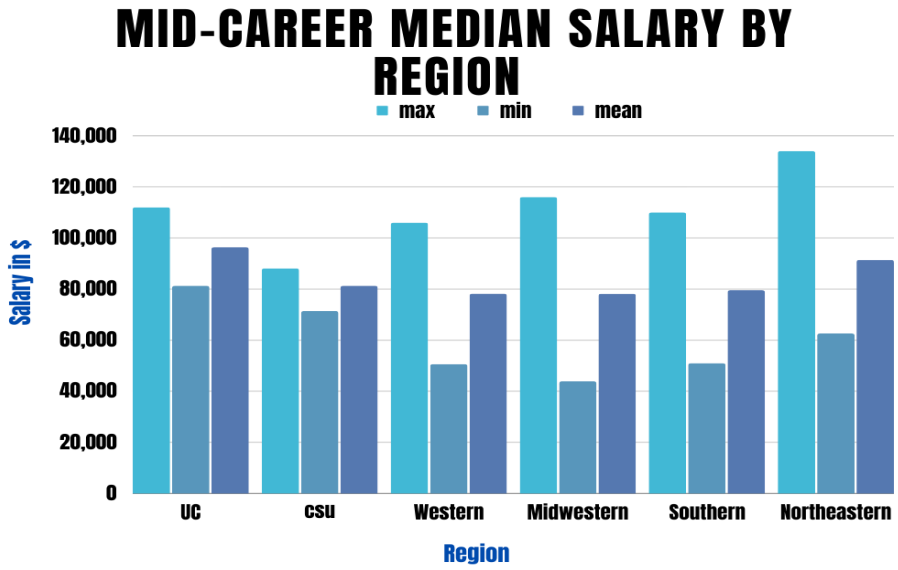
Graphing

Once the dataset had been successfully split between UCs, CSUs, Western, Midwestern, Southern, and Northeastern regions, I graphed each set on its own. Then I

decided it would be beneficial to graph the top 5 highest and lowest earning schools of each region for the presentation. I also figured it would be most beneficial to collect and graph summary statistics for overlaid data and clear calculated results.
Seeing everything in comparison side by side helped a lot with putting things in perspective.

Analysis: Mid-Career Salaries by Region

After reviewing the summary statistics and the parsed data I concluded the following.



Based on this graph, I noticed that the average UC score is higher than Northeastern school salaries, I know that this average is displaying as a result of fewer UC schools than Northeastern schools, so the mean is not a fair representation. Considering this, I made conclusions based on the top 5 in each category to conclude the data fairly.

Conclusion: Mid-Career Salaries by Region: Top 5 schools per category

Average Earning Regions Ranked From Highest to Lowest:

1. Northeastern: $128,000

2. Southern: $107,200
3. Midwestern: $105,260
4. UCs: $102,060
5. Western: $94.340
6. CSUs: $85,300

CSUs Earned less than UCs and UCs earn less than the top Northeastern schools.

Results can be found in our GitHub repo Here:

- https://github.com/Immersor-git/CECS_450_Fall2023_Sem_Project/blob/main/Bonnie%20-%20Jupyter%20Notebook/bonnieWIP.ipynb

Work and code/graph comments can be found in the Bonnie-Jupyter Notebook Folder in the BonnieWIP.ipynb file)

- https://github.com/Immersor-git/CECS_450_Fall2023_Sem_Project/tree/main/Completed%20Graphs/Bonnie

Salaries by College Type

Hypothesis

Before computing data for salaries by college type, I predicted that a better college education (i.e., Ivy League and Engineering) would provide a higher salary compared to State and Party colleges. I would assume the education quality of a well-known university would hold up to its reputation and set up its students well for real-world careers.

Preparation & Cleaning

After selecting the data sets to work with, the next step was to prepare the data. The preparation for calculating the data was easy with the help of online resources such

as table converters websites. These websites automatically converted our data sets into CSV files allowing us to easily work with the data in RStudio.

In addition to converting the data sets into CVS files, preparing the numbers for calculation was another process. I had to research how to convert values in a CSV file into computable numerical values for the salaries. In addition, I had to clean up the data values by removing excess characters such as commas (,) and dollar signs ($). These computable values were then used to find the mean of starting and mid-career salaries. To find the mean, I first combined the values by the college type column, focused on the starting and mid-career salaries separately, and then calculated their means.

Graphing

Once the data was prepared and cleaned up, graphing the data became a simpler process. Initially, I attempted to graph the entire data set, however, there was an excess of information to digest. However, after grouping the data by college type, the computed graphs were simplified and readable. The first graph displays the mean starting salary by college type.* The statistics produced for the highest to lowest earnings were:

1. Ivy League - $60,475.00,
2. Engineering - $59,057.89,
3. Liberal Arts - $45,746.81,
4. Party - $45,715.00,
5. and States - $44,126.29.

The second graph displays the mean mid-career salary by college type.* The statistics produced for the highest to lowest earnings were:

1. Ivy League - $120,125,
2. Engineering - $103,842.10,
3. Liberal Arts - $89,378.72,
4. Party - $84,685.00,
5. and States - $78,567.43.

Analysis

When analyzing the data for salaries by college type I was pleased to see the results. My hypothesis was partially correct in the sense that Ivy League and specialized colleges produced the highest salaries for both starting and mid-career. However, I was surprised to see the Party colleges perform better than the State ones. For starting salaries, Party colleges made over $1,000 and over $6,000 more for mid-career salaries than State colleges.

*Results can be found in our GitHub repo in Completed Graphs/Sarah, work can be found in the “Sarah - R Studio Code” folder.

Salary by Major Type

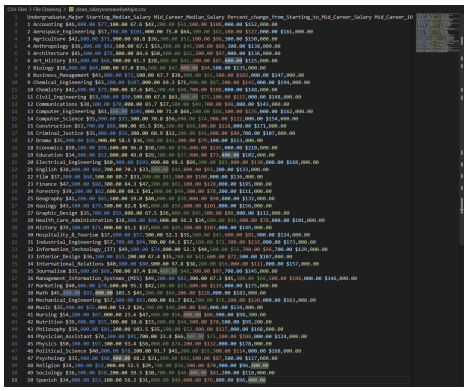
Hypothesis

The goal of the Salary by Major Type section was to see if categorizing groups of majors could provide useful insights for how salaries are distributed across broader fields - and to see how significant the opportunities for career progression are within these categories.

Preparation and Cleaning

The raw data started out as a list organized by individual college majors. Each major was listed with its starting and mid-career salary, as well as some percentile metrics with them. There were a few problems with the data before it could be fully used. First, commas were used to separate values, but were also placed in all of the dollar figures in the data. Changing the CSV delimiter from commas to spaces fixed this - although the data required replacing some empty spaces with underscores to note which values should be separated.

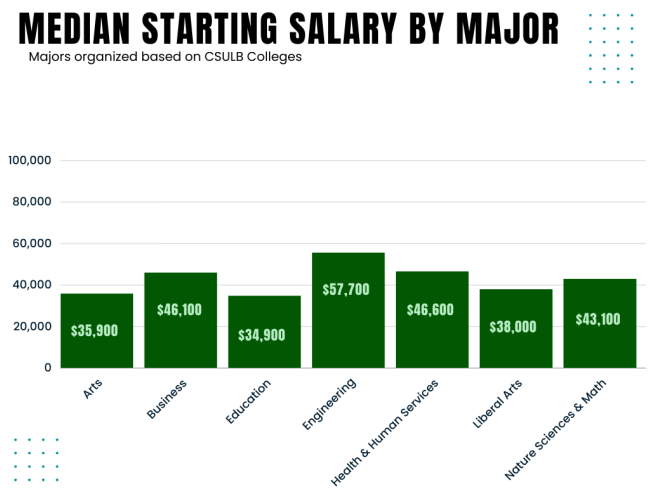
Once the data was ready to graph - it quickly became apparent that providing data for all college majors was difficult to show cleanly. To resolve this, we decided to make this section cover major categories. After moving the data to an excel sheet, we manually tagged each major with the associated CSULB college category (College of Engineering, College of Business, etc.) Finally, we were able to export the data as a CSV file to graph it in R and begin exploring the data.

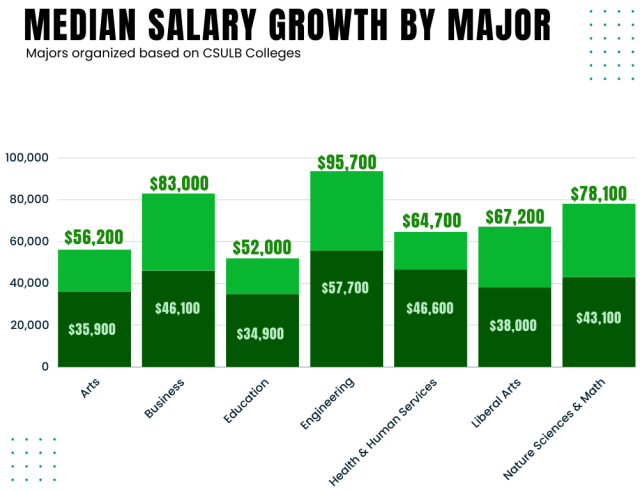


Graphing

Once the CSV files were ready for graphing we could get a sense of what the data conveyed. Initial graphs were messy, but those occurred before trimming the data by category. Those initial graphs are detailed in the next graphing section of this paper.

We split the category data into two matching graphs - with and without mid-career salaries. The idea was to introduce the initial salary first, and then dig into which fields pay off in the long run. These were the two final graphs.





Analysis

Results showed that Engineering related majors earn the most out of all categories for their entire career. Business, alongside Nature Sciences & Math showed the greatest mid-career growth, aside from Engineering. Education and Arts took up the bottom two slots across the board. The case of Health & Human Services proved unique, as it has the least career growth proportional to its starting salary. For this

project, we weren't able to explore the reasoning behind this more. Regardless, this form of graphing proved informative for guiding our project. Additionally, categorizing the data in this manner allowed us to hone in on more specific cases in the next section. First introducing salary by category before presenting the individual majors.

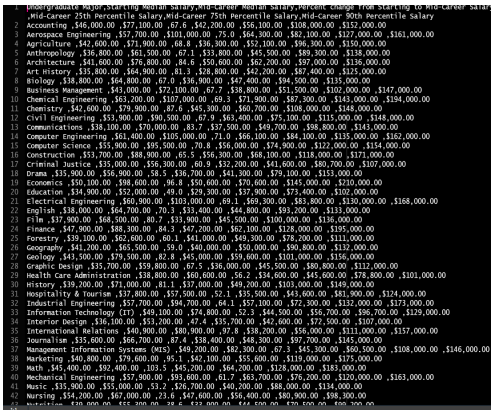
Salary by Major

Hypothesis

Starting with this assignment I believed that STEM majors would have the highest starting median salary. I thought they would occupy the whole top ten starting median salaries as well as the top ten for median mid-career salaries. As for the bottom ten, I expected that art and liberal arts majors, specifically majors such as Religion and Philosophy would have the lowest starting median salaries as well as the lowest mid-career median salaries. As for the percent change, I also thought that computer engineering, and computer science majors would have the highest jump from starting to mid-career salaries.

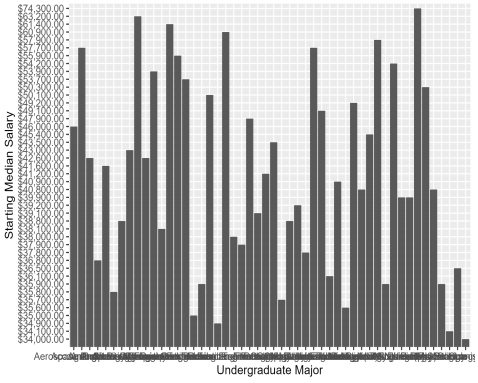
Preparation and Cleaning

Getting the data formatted in a way that would be easy to graph was a challenge in itself. The dataset that I was looking at had to be formatted to a CSV file. At first, I tried to format the data, I tried to do it programmatically. This was very difficult for me, but thankfully we found an online resource that allowed us to to convert our dataset to a CSV file easily. After this, there were still some modifications that needed to be made. A lot of our data had missing values and the numbers were formatted as strings or characters so we had to change the data to be integers so we could perform calculations on them.

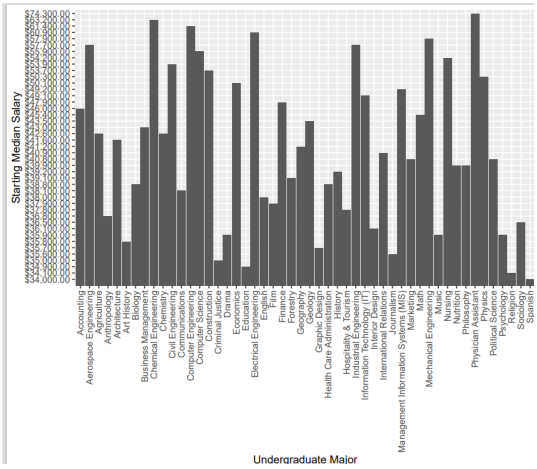


Undergraduate Major	Starting Median Salary	Mid-Career Median Salary	Percent change from Starting to Mid-Career Salary	Mid-Career 10th Percentile Salary	Mid-Career 25th Percentile Salary	Mid-Career 75th Percentile Salary	Mid-Career 90th Percentile Salary
Accounting	$46,000.00	$77,100.00	67.6	$42,200.00	$56,100.00	$108,000.00	$152,000.00
Aerospace Engineering	$57,700.00	$101,000.00	75.0	$64,300.00	$82,100.00	$127,000.00	$161,000.00
Agriculture	$42,600.00	$71,900.00	68.8	$36,300.00	$52,100.00	$96,300.00	$150,000.00
Anthropology	$36,800.00	$61,500.00	67.1	$33,800.00	$45,500.00	$89,300.00	$138,000.00
Architecture	$41,600.00	$76,800.00	84.6	$50,600.00	$62,200.00	$97,000.00	$136,000.00
Art History	$35,800.00	$64,900.00	81.3	$28,800.00	$42,200.00	$87,400.00	$125,000.00
Biology	$38,800.00	$64,800.00	67.0	$36,900.00	$47,400.00	$94,500.00	$135,000.00
Business Management	$43,000.00	$72,100.00	67.7	$38,800.00	$51,500.00	$102,000.00	$147,000.00
Chemical Engineering	$63,200.00	$107,000.00	69.3	$71,900.00	$87,300.00	$143,000.00	$194,000.00
Chemistry	$42,600.00	$79,900.00	87.6	$45,300.00	$60,700.00	$108,000.00	$148,000.00
Civil Engineering	$53,900.00	$90,500.00	67.9	$63,400.00	$75,100.00	$115,000.00	$148,000.00
Communications	$38,100.00	$70,000.00	83.7	$37,500.00	$49,700.00	$98,800.00	$143,000.00
Computer Engineering	$61,400.00	$105,000.00	71.0	$66,100.00	$84,100.00	$135,000.00	$162,000.00
Computer Science	$55,900.00	$95,500.00	70.8	$56,000.00	$74,900.00	$122,000.00	$154,000.00
Construction	$53,700.00	$88,900.00	65.5	$56,300.00	$68,100.00	$118,000.00	$171,000.00
Criminal Justice	$35,000.00	$56,300.00	60.9	$32,200.00	$41,600.00	$80,700.00	$107,000.00
Drama	$35,900.00	$56,900.00	58.5	$36,700.00	$41,300.00	$79,100.00	$153,000.00
Economics	$50,100.00	$98,600.00	96.8	$50,600.00	$70,600.00	$145,000.00	$210,000.00
Education	$34,900.00	$52,000.00	49.0	$29,300.00	$37,900.00	$73,400.00	$102,000.00
Electrical Engineering	$60,900.00	$103,000.00	69.1	$69,300.00	$83,800.00	$130,000.00	$168,000.00
English	$38,000.00	$64,700.00	70.3	$33,400.00	$44,800.00	$93,200.00	$133,000.00
Film	$37,900.00	$68,500.00	80.7	$33,900.00	$45,500.00	$100,000.00	$136,000.00
Finance	$47,900.00	$88,300.00	84.3	$47,200.00	$62,100.00	$128,000.00	$195,000.00
Forestry	$39,100.00	$62,600.00	60.1	$41,000.00	$49,300.00	$78,200.00	$111,000.00

Exploring

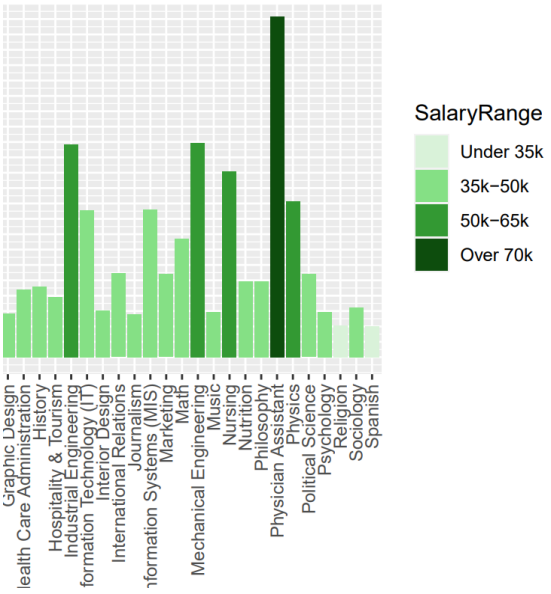


After cleaning and prepping the dataset, I started playing around to see how the data would look graphed. This created some very ugly graphs that weren't expressive or effective. This is when I had to play around and try to change how I wanted to graph the majors because it is very hard to convey fifty majors on a graph while maintaining expressiveness and effectiveness. What I did to combat this was change how I was graphing the data. First I changed the y-axis from starting at $0 to starting at $30,000. Since the salaries in the dataset ranged from $34,000 to about $75,000 the bars showed up bigger with the new y-axis. The next thing I did was change how words on the x-axis appeared. Instead of them being horizontal and writing over each other I changed the words to be vertical which allowed them to be read legibly. Even though it was now legible, it still wasn't good enough for me.

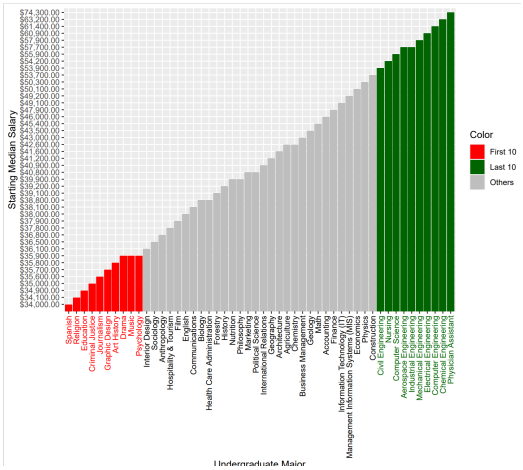


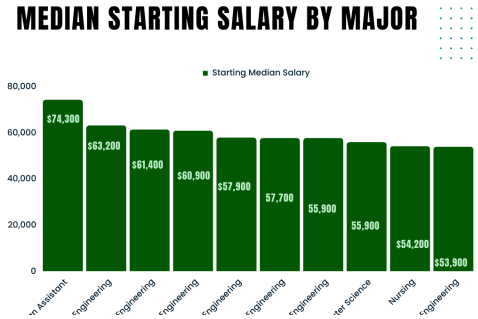
Graphing

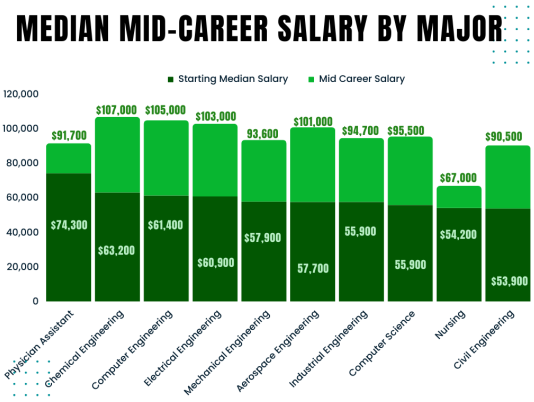
After exploring and seeing initial graphs I had to figure out how I was going to convey the data. First I kept the graph of all majors but had ranges in the color green depending on how much the starting median salary was. This still wasn't presenting the data in the way I wanted and this is where I realized I was not going to be able to keep all fifty majors in my graphs. This is where I decided to focus on the ten highest



and lowest starting median salaries as well as the ten highest and lowest mid-career salaries. I colored the top ten salaries in green, and also colored the text of the name in green and the opposite with the color red. This allows people to see the lowest and highest salaries. When I did this I got a graph that showed the information but because it was still so many different majors on the graph it was really difficult to get any real information from it. But while not expressing the salaries well, it did show the ranks well. This is when I decided on how I would do my graphs. I would only show the top ten and bottom ten starting salaries. From here I was able to clean up the data. Since there were only ten bars now, it was a lot clearer to tell what the graph was showing. It was now easy to view the names of the different majors and it was easy to tell the specific starting salary. To convey the mid-career salaries I did a stacked bar chart. I did it in a lighter color green so it would be easier for people to see the difference. I also showed the median career salary. Doing a stacked bar chart also allows the viewer to see how big the jump from starting salary to mid-career salary was. These graphs were done for both the top ten starting salaries as well as the bottom ten starting salaries.







Analysis

Before the data can be analyzed I needed to summarize the statistics. The summary statistics are as follows:

Starting Salaries:

Top 10 Best Starting Median Salaries:

1. Physician Assistant ($74,300)
2. Chemical Engineering ($63,200)
3. Computer Engineering ($61,400)
4. Electrical Engineering ($60,900)
5. Mechanical Engineering ($57,900)
6. Industrial Engineering ($57,700)
7. Aerospace Engineering ($57,700)
8. Computer Science ($55,900)
9. Nursing ($54,200)
10. Civil Engineering ($53,900)

Top 10 Worst Starting Median Salaries:

1. Spanish ($34,000)
2. Religion ($34,100)
3. Education ($34,900)
4. Criminal Justice ($35,000)
5. Journalism ($35,600)
6. Graphic Design ($35,700)
7. Art History ($35,800)
8. Drama ($35,900)
9. Music ($35,900)
10. Psychology ($35,900)

Highest Paid Early Career Major
Physical Assistant $74,300

Lowest Paid Early Career Major
Spanish $34,000

Median Salaries:

Top 10 Best Mid-Career Median Salaries:

1. Chemical Engineering ($107,000)
2. Computer Engineering ($105,000)
3. Electrical Engineering ($103,000)
4. Aerospace Engineering ($101,000)
5. Economics ($98,600)
6. Physics ($97,300)
7. Computer Science ($95,500)
8. Industrial Engineering ($94,700)
9. Mechanical Engineering ($93,600)
10. Math ($92,400)

Top 10 Worst Mid-Career Median Salaries:

1. Education ($52,000)
2. Religion ($52,000)
3. Spanish ($53,100)
4. Interior Design ($53,200)
5. Music ($55,000)
6. Nutrition ($55,300)
7. Criminal Justice ($56,300)
8. Drama ($56,900)
9. Hospitality & Tourism ($57,500)
10. Sociology ($58,200

Highest Paid Mid-Career Major
Chemical Engineering $107,000

Lowest Paid Mid-Career Major
Education & Religion $52,000

Takeaways:

Now that we have the numbers we can do some analysis. The first thing I noticed was that my hypothesis was pretty spot on. I expected that the top ten starting salaries would all be occupied by stem majors and that was the case.

- Engineering Consistency: Engineering fields offer strong earning potential, both early & mid-career.

- Passion vs. Paycheck: Pursuing passion may lead to lower initial salaries. For example, majors like Spanish and Religion have lower starting median salaries, showing the trade-off between pursuing personal interests and financial gain.

- Psychology's Paradox: Psychology majors, despite exploring human behavior, often face lower earnings, showcasing the disparity between societal importance and financial reward.

- Physician Assistants, despite starting with a high salary, experience a small salary increase in mid-career compared to other fields, which shows peak earnings.

- Civil Engineers, while maintaining competitive mid-career salaries, leave the top ten. Because civil engineers work in various sectors, including construction, transportation, infrastructure, etc. their mid-career salaries vary based on their specific area of expertise.

- Analytical Skills are valued. Several majors with the highest earning mid-career salaries (economics, computer science, and math) emphasize the significance of strong analytical and problem-solving skills in the job market.

- Chemical engineering having a high early career salary as well as the highest mid-career salary showcases the long-term financial benefits of this field
-
- Economics is the only non-STEM major in the top 10 of the mid-career salaries.

Conclusion

Northeastern Colleges average the highest starting and mid-career salaries. Ivy League and engineering schools have the highest starting and mid-career salaries. State schools are last. Engineering majors have the highest median salary across their entire careers. Chemical engineering and Computer Engineering take the top 1 and 2 spots. Education, Arts, and Liberal Arts majors statistically make the lowest starting and mid-career salaries regardless of what college attended.