# Data Jobs on Glassdoor

Adam Bennett

April 28, 2023

STAT 441

## Section 1: Introduction

Data science is one of the fastest growing industries in the current job market. Companies are learning that having access to large amounts of well-structured data, and extracting valuable information from said data, directly results in greater profits. In a similar vein, as someone hoping to break into the data industry, I believe that gaining knowledge will be valuable toward my future career endeavors. My goal with this initial study is to use statistical tests to learn how the job market trends and what employers are looking for in a candidate.

Data jobs can be categorized into three primary job titles. Their definitions vary heavily from company to company, but as an overly simplified overview:

- **Data Engineers** create and maintain the pipelines in which data is collected and ensure the data is structured in a way that it can be useful.
- **Data Scientists** study and seek to understand the data and its characteristics to identify how it can benefit the company.
- **Data Analysts** explore, interpret, and visualize the data, and communicate their inferences to concerned parties within the company in order to influence strategic decision making.

On top of the three standard titles, the data industry also includes machine learning engineers, research associates, and business analysts, among other similar roles. There is overlap between each of their skillsets, but it is certain that each job title brings their own unique challenges to the table. As such, it is important to identify what skills would be most useful to each type in order to maximize the probability of finding a job within the desired field.

Since this is only an initial analysis of the subject matter, our research questions will remain simple and straightforward. The questions I will be exploring are as follows:

1. In the post-COVID world, remote work is as popular as ever. What percentage of the data job listings are remote?
2. Python is widely considered to be the premier programming language for use in data science. Is this statement supported by the data?
3. Between the three standard job titles, how do their average salaries compare to each other? Do our results match up with our expectations?

The data we will be using to perform this analysis was collected by Mohamed Sayed and published to Kaggle for public use. Sayed and his team used Python web scraping scripts alongside a browser automation program to quickly search Glassdoor for data job listings. For each corresponding URL, they extracted the HTML data from the website and used text searches to extract information from the page. We will observe the data structure in the next section.

In order for our tests to be useful, our data must meet specific requirements. Since we did not collect the data ourselves, we cannot be completely confident that these requirements are met, so we must make some assumptions about our data. Specifically, we want our data to be a random sample of the job listing population with independent observations. From what we know about their web scraping methodology, we can reasonably assume that the sample we are given is comprised of all the listings on Glassdoor at the time the script was run. Since each observation equates to one specific job opening, and we are not leaving anything out of our sample, we can treat our sample as a set of independent observations that comprise a strong representation of our entire population. Notably, there are many entries in the set that appear to be duplicates, which would present a problem, but we can assume that these are simply multiple job listings of the same title from the same company. As such, we can proceed without worry of violating our assumption of independence.

In this report, we will first explore and visualize our data to define the setting for our tests. We will then explain our methodology, perform our tests, and make our conclusions based on our test results. We will follow this with a short discussion on future plans and how we could improve on our process.

## Section 2: The Data

The original dataset consists of 2084 observations of job listings on Glassdoor, with each observation consisting of 23 variables. After cutting out the variables that are irrelevant to our research questions, we are left with a table of the following structure:

| location | salary estimate | python_yn | excel_yn | machine_learning_yn | job_simpl | seniority |
|---|---:|---:|---:|---:|---|---|
| San Diego, CA | 67800 | 1 | 1 | 0 | Data Analyst | Senior |
| Springfield, IL | 67882 | 0 | 1 | 0 | Data Analyst | Senior |
| Los Angeles, CA | 67918 | 1 | 1 | 0 | Data Analyst | Senior |
| Lexington, MA | 68229 | 0 | 0 | 1 | Other | Mid |
| Cold Spring Harbor, NY | 68500 | 1 | 0 | 1 | Machine Learning Engineer | Senior |

We have the following variables as descriptors of our observations:

- Location: The city and state of the job opening. If the role involves primarily remote work, the location is instead classified as remote.
- Salary Estimate: The estimated salary of the position. If the listing gives an exact salary, the column lists that number. If the listing gives a salary range, the column lists the average value of the range. If the listing does not state a salary, the column lists Glassdoor's salary estimate of the job. Unfortunately, the exact methodology behind Glassdoor's salary prediction is unknown, which may shroud our results. We will operate as if these predictions are accurate and revisit this idea later.
- Python_yn: A Boolean value corresponding to whether the job listing mentions Python in the description. If Python is mentioned, the value is 1. If not, the value is 0.
- Excel_yn: A Boolean value corresponding to whether the job listing mentions Excel in the description. If Excel is mentioned, the value is 1. If not, the value is 0.

- Machine_learning_yn: A Boolean value corresponding to whether the job listing mentions machine learning in the description. If machine learning is mentioned, the value is 1. If not, the value is 0.
- Job_simpl: The job title from the job listing classified into one of five categories: “Data Analyst,” “Data Scientist,” “Data Engineer,” “Machine Learning Engineer,” or “Other.”
- Seniority: A description of the seniority level of the role, classified into one of three categories: “Junior,” “Mid,” or “Senior.”

First, we will take a closer look at our salary estimate variable. Figures 2.1 through 2.4 show graphs of our salary distribution for the whole sample, and then for each of the respective job titles.

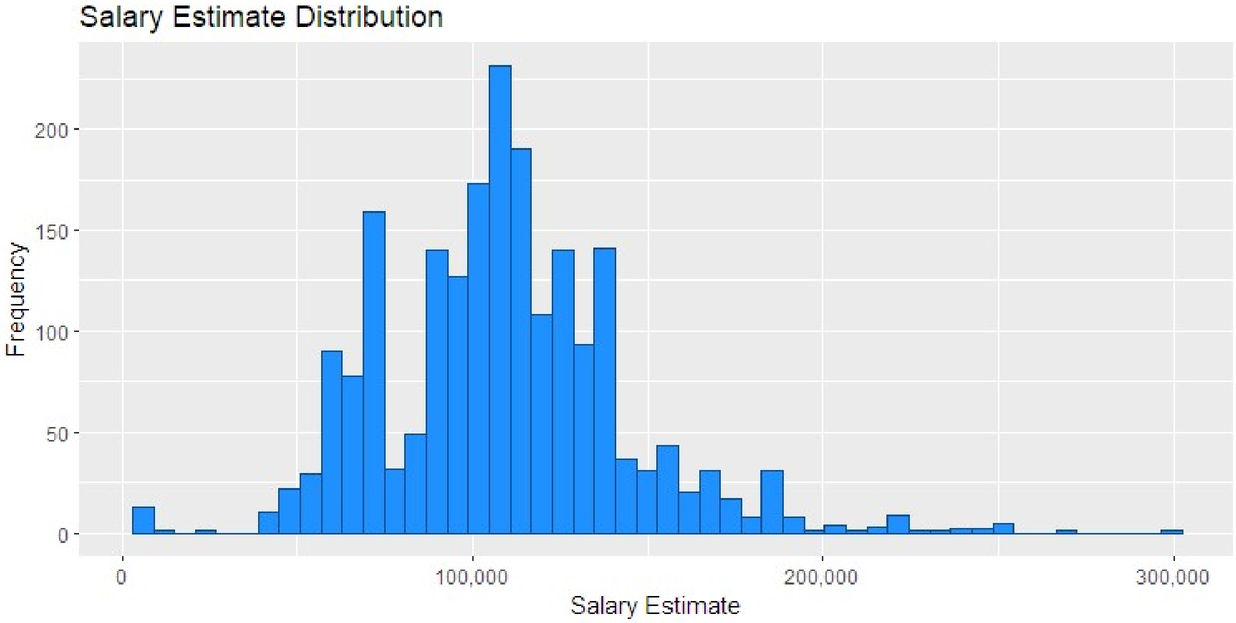


Figure 2.1

The overall mean salary estimate is $108768.64 with a sample size of 2084 and a standard deviation of $34434.63. We can observe that our distribution does have some characteristics of a normal bell curve, but has some outliers on the low end and is skewed slightly right. Since our sample size is very high, these departures from normality should not affect our test results with respect to means and proportions, but if we choose to perform any tests concerning our variance, we may not be able to fully trust our results.

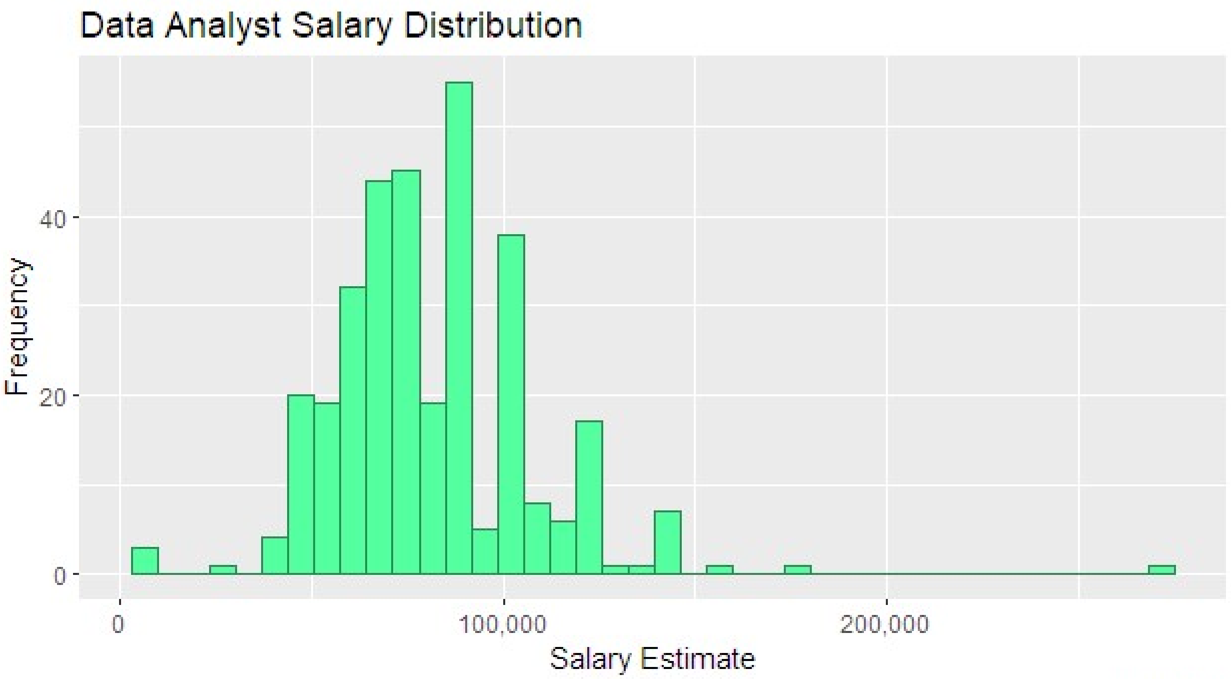


Figure 2.2

The mean salary estimate for data analysts is $81119.51 with a sample size of 328 and a standard deviation of $26491.46.

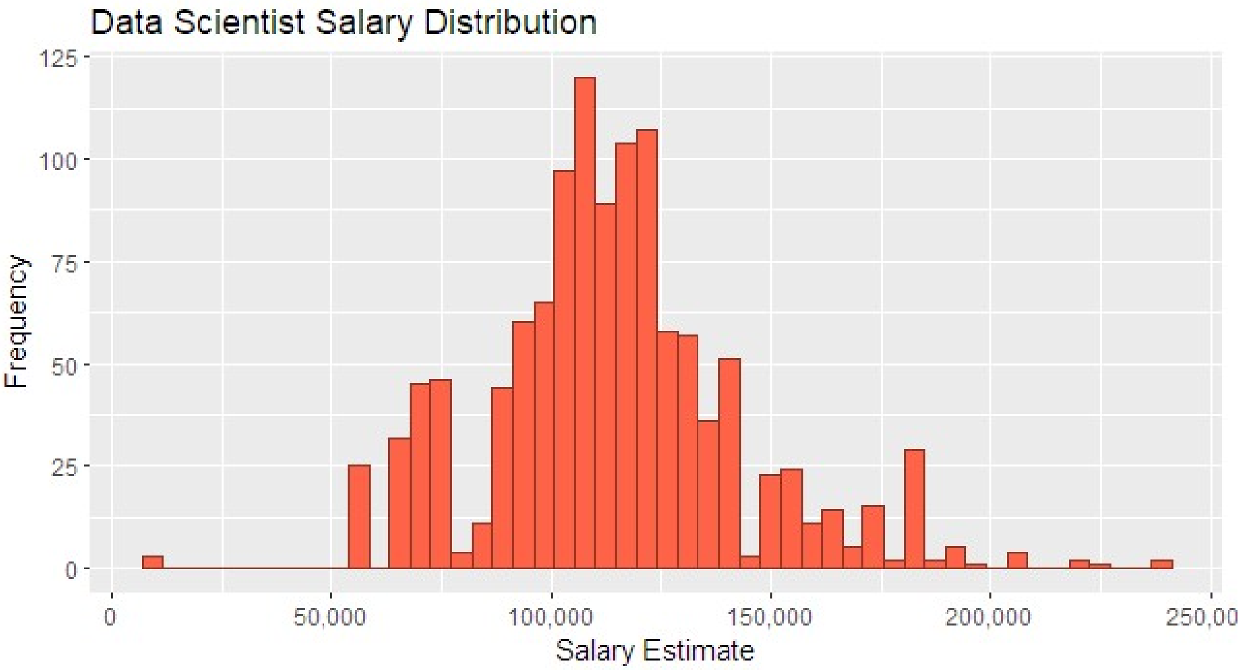


Figure 2.3

The mean salary estimate for data scientists is $114277.50 with a sample size of 1197 and a standard deviation of $29063.29.

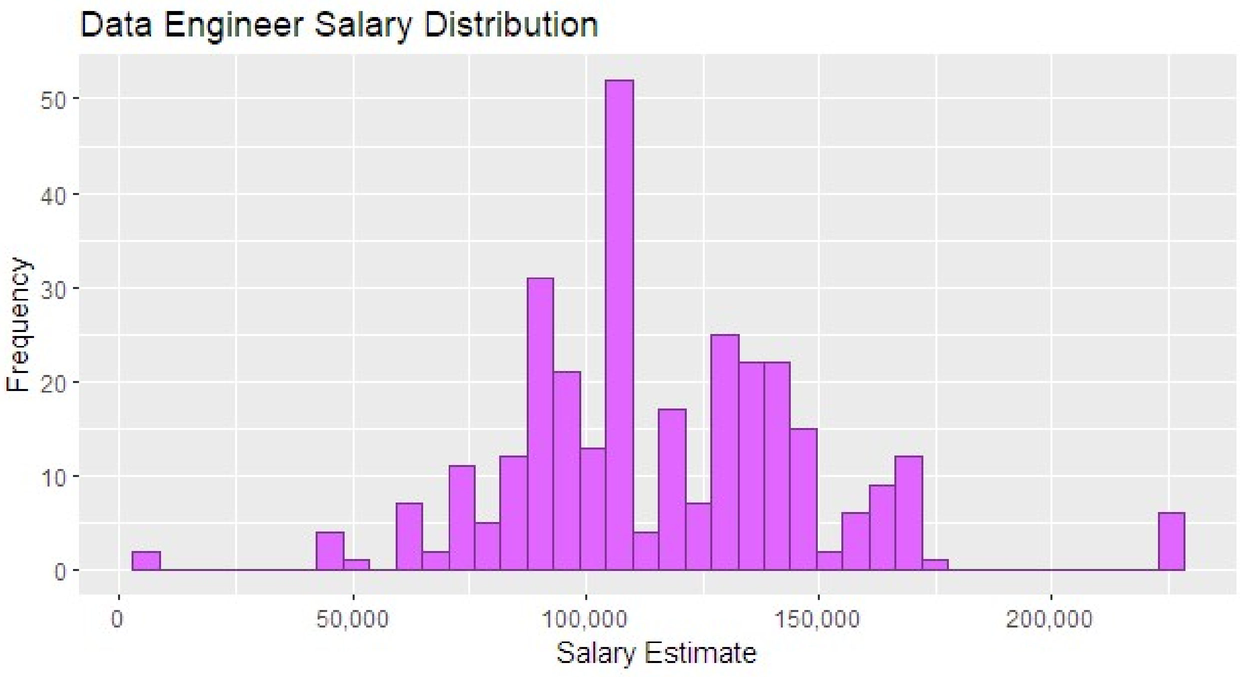


Figure 2.4

The mean salary estimate for data engineers is $116228.91 with a sample size of 309 and a standard deviation of $32773.12.

Figures 2.5 through 2.7 will show graphs of our categorical variables in relation to each job title.

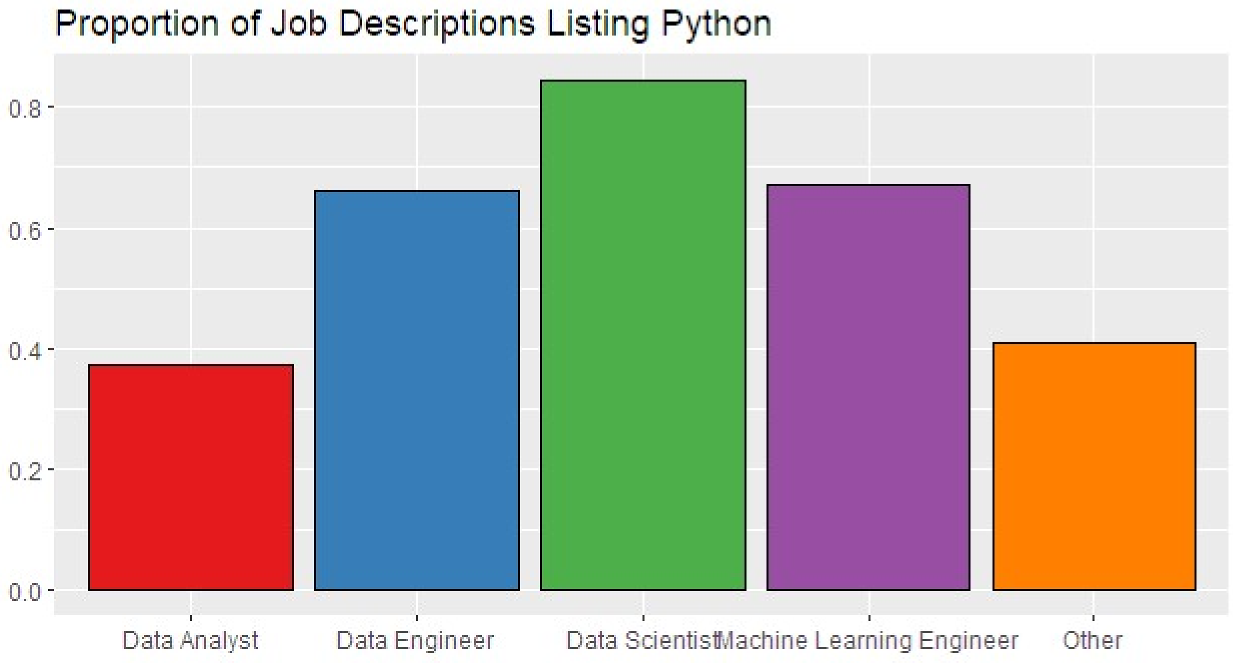


Figure 2.5

71.26% of our 2084 job listings contain Python in their description, with data scientists having the highest proportion of the different job titles at 84.46%.

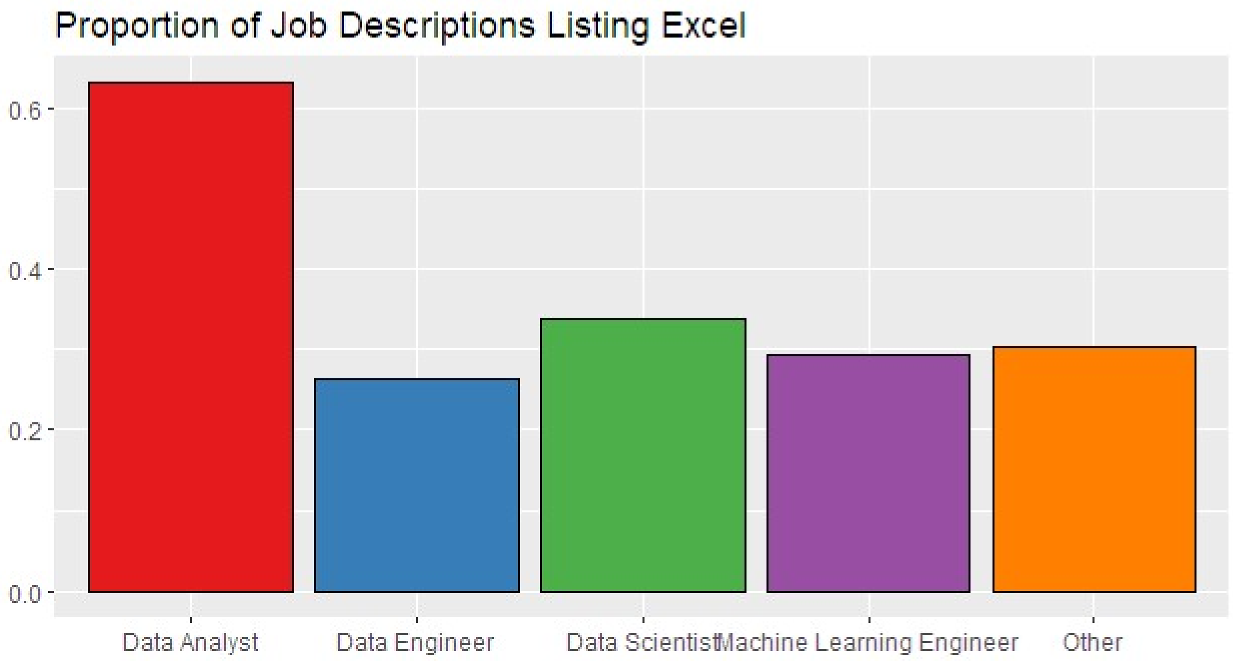


Figure 2.6

36.71% of our 2084 job listings contain Excel in their job description, with data analysts having by far the highest proportion at 63.11%. This can likely be explained by the fact that data analysts appear to have the highest representation of entry-level jobs, and it is generally expected that any more experienced worker in the industry will have plenty of experience with Excel.

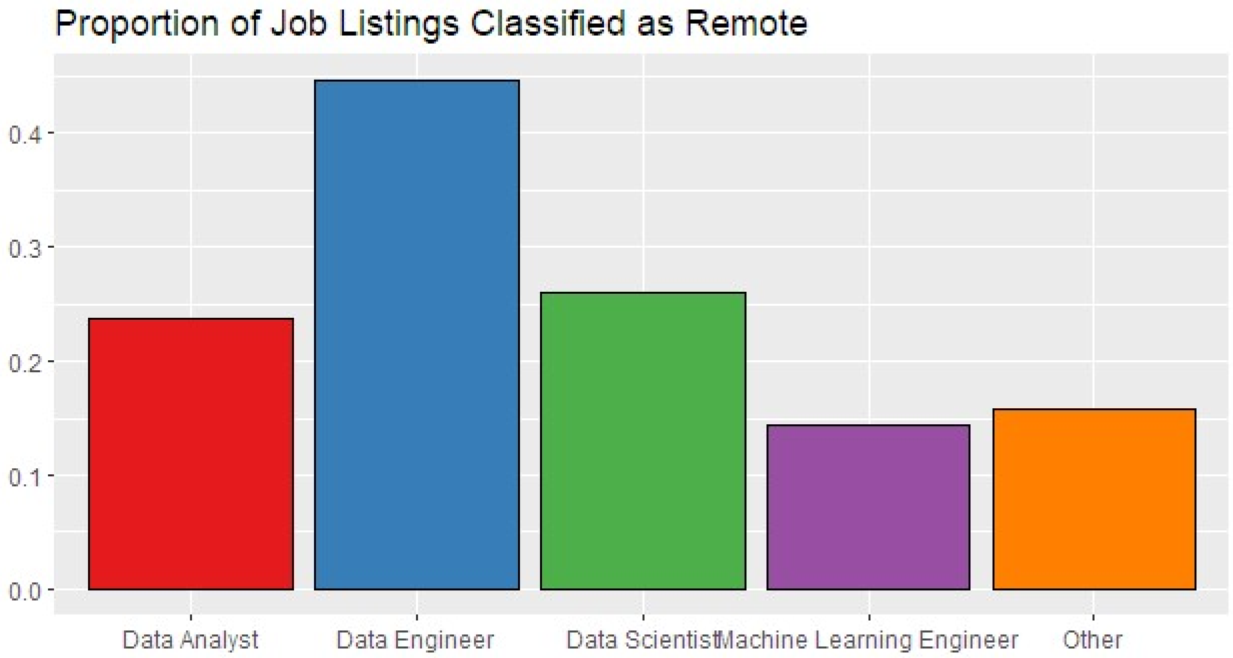


Figure 2.7

27.11% of our 2084 job listings are classified as remote, with data engineers having the highest proportion at 44.66%. This could be explained by data engineering commonly being considered a more independent role than a data analyst, for example, which may be more team oriented.

## Section 3: Methods

Before we answer any of our research questions, we will be using resampling techniques to visually demonstrate the sampling distribution of the sample variance. Our goal is to confirm that, with a high sample size, the sample variance has a normal distribution and gives an unbiased estimation of the population variance. For this demonstration, we will be analyzing the salary estimate column.

To accomplish this, we will be treating our original dataset as the population and using Excel to take 5000 independent random samples from said population. By using the RANDBETWEEN function in Excel, we ensure that our sample is entirely random. By allowing for the same observation to be picked more than once, also known as replacement, we ensure that each observation in our sample is independent of the others. Finally, by having each sample have a sample size that matches our original dataset, we allow for our sampling variance to give similar values to the original population variance, removing potential bias.

To summarize, we will be taking a total of **5000** samples using resampling with replacement, and each sample will have a sample size of **2084**. We will then calculate the sample variance for each sample, plot them on a histogram, and report our results.

To answer our research questions with respect to the entire population of job listings, we will perform several statistical tests.

For question 1, we are interested in the proportion of all data jobs that are considered remote. We will first calculate the sample proportion with respect to our sample, then we will use this result to calculate a 95% confidence interval of the population proportion of data jobs that are classified as remote.

For question 2, we are interested in the proportion of data science jobs, specifically, that list Python in their job description. First, we want to know if there is evidence that, out of all data science jobs, more than 80% of them list Python. To answer this, we will perform a one-sided hypothesis test on the proportion of data science jobs listing Python. Afterwards, we will calculate a 95% confidence interval of the real number of data science jobs listing Python.

For question 3, we will be comparing the average salary of the data job titles to each other. First, we want to confirm our suspicions that data analysts have the lowest average salary. To do so, we will perform a one-sided hypothesis test on the difference in mean salary between data scientists and data analysts, and if applicable, we will find the 95% confidence interval of this difference. If we confirm our suspicions with that test, we will follow it up by analyzing whether either of the other two job titles make more than the other on average. For this, we will perform a two-sided hypothesis test on the difference in mean salary between data scientists and data engineers. If there is a significant difference, we will find the 95% confidence interval of the difference, and if not, we will perform a variance ratio test to see if there is any other information to be extracted from our sample.

# Section 4: Results

## Sampling Distribution Results

Figure 4.1 shows a histogram of the sample variances calculated from our 5000 resamples as detailed above.

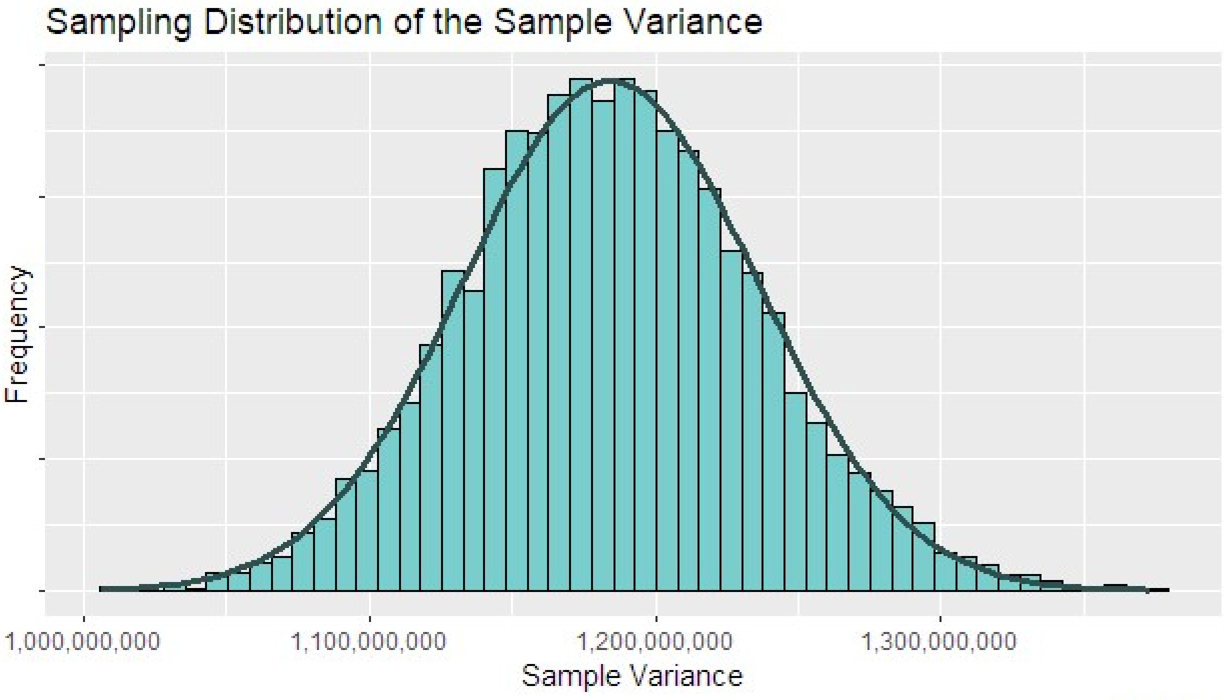


Figure 4.1

The mean of the sample variances from our experiment is **1,184,404,516**.
The sample variance from the salary estimate column in our original dataset is **1,185,744,070**.

These two values differ by only 0.11%. This serves as a strong demonstration that the sampling distribution of the sample variance has a mean equal to the variance of the population distribution.

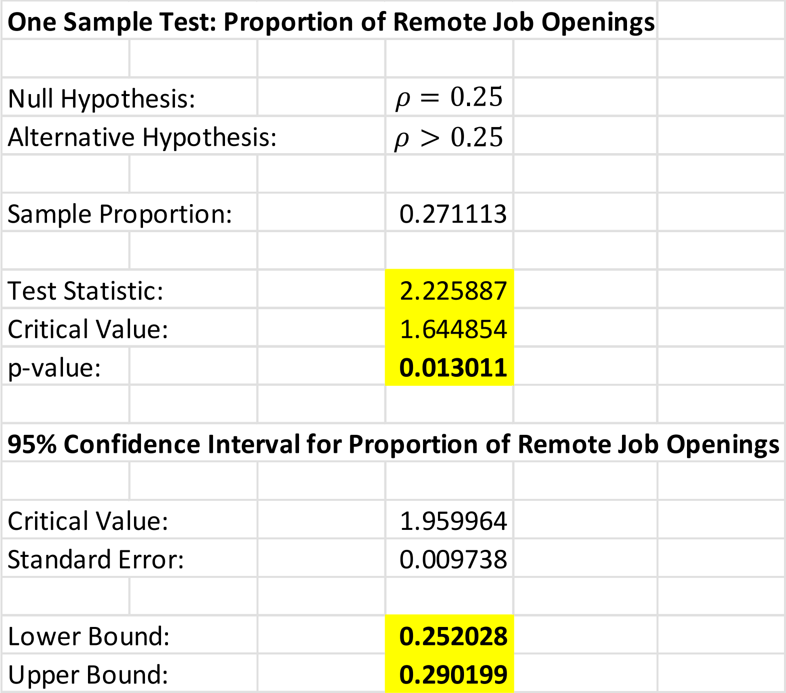
We can see by comparing the shape of the histogram to the shape of the overlaid normal distribution curve that our sampling distribution is very close to a normal distribution. This matches our expectation that the theoretical sampling distribution of the sample variance should be normal.

In the case of variance, it is essential that each of our resamples had the same sample size as our original to preserve its scale. This is what allows the final value we generated to be so close to the original, despite being such a large value. The number of samples chosen, 5000, is arbitrary, but chosen to be large enough to generate an approximately normal distribution graph. If we had used a higher number of samples, we may have an even more normal looking graph, while a lower number may have caused some additional deviation from normality.

## One Sample Results

First, we want to test whether the true proportion of job listings that are classified as remote is greater than 25% and calculate the corresponding confidence interval at the 95% confidence level.

Below are the results:



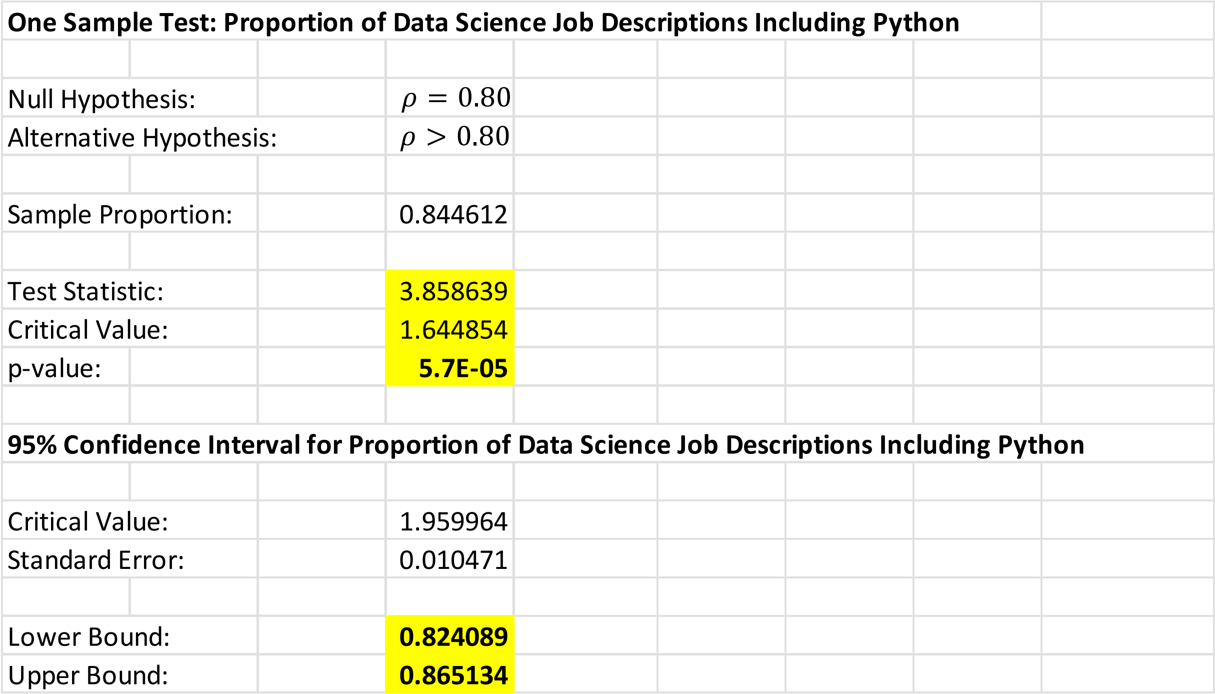
| **One Sample Test: Proportion of Remote Job Openings** | |
|---|---|
| | |
| Null Hypothesis: | $\rho = 0.25$ |
| Alternative Hypothesis: | $\rho > 0.25$ |
| | |
| Sample Proportion: | 0.271113 |
| | |
| Test Statistic: | 2.225887 |
| Critical Value: | 1.644854 |
| p-value: | **0.013011** |
| | |
| **95% Confidence Interval for Proportion of Remote Job Openings** | |
| | |
| Critical Value: | 1.959964 |
| Standard Error: | 0.009738 |
| | |
| Lower Bound: | **0.252028** |
| Upper Bound: | **0.290199** |

Based on this test, we **reject the null hypothesis**, hence we have statistical evidence that in the entire data job market, more than 25% of job listings are remote.

Furthermore, we conclude with 95% confidence that in the data job market, the percentage of job listings that are remote is between **25.2% and 29.0%**.

Next, we want to test whether the true proportion of data science job listings that list Python in the job description is greater than 80% and calculate the corresponding confidence interval at the 95% confidence level.

Below are the results:



| One Sample Test: Proportion of Data Science Job Descriptions Including Python | |
|---|---|
| Null Hypothesis: | $\rho = 0.80$ |
| Alternative Hypothesis: | $\rho > 0.80$ |
| Sample Proportion: | 0.844612 |
| Test Statistic: | 3.858639 |
| Critical Value: | 1.644854 |
| p-value: | **5.7E-05** |
| **95% Confidence Interval for Proportion of Data Science Job Descriptions Including Python** | |
| Critical Value: | 1.959964 |
| Standard Error: | 0.010471 |
| Lower Bound: | **0.824089** |
| Upper Bound: | **0.865134** |

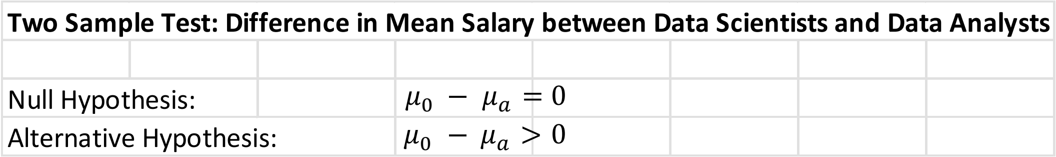
Based on this test, we **reject the null hypothesis**, hence we have statistical evidence that out of all data science job listings, more than 80% of them contain Python in the job description.

Furthermore, we conclude with 95% confidence that out of all data science job listings, between **82.4% and 86.5%** of them contain Python in the job description.

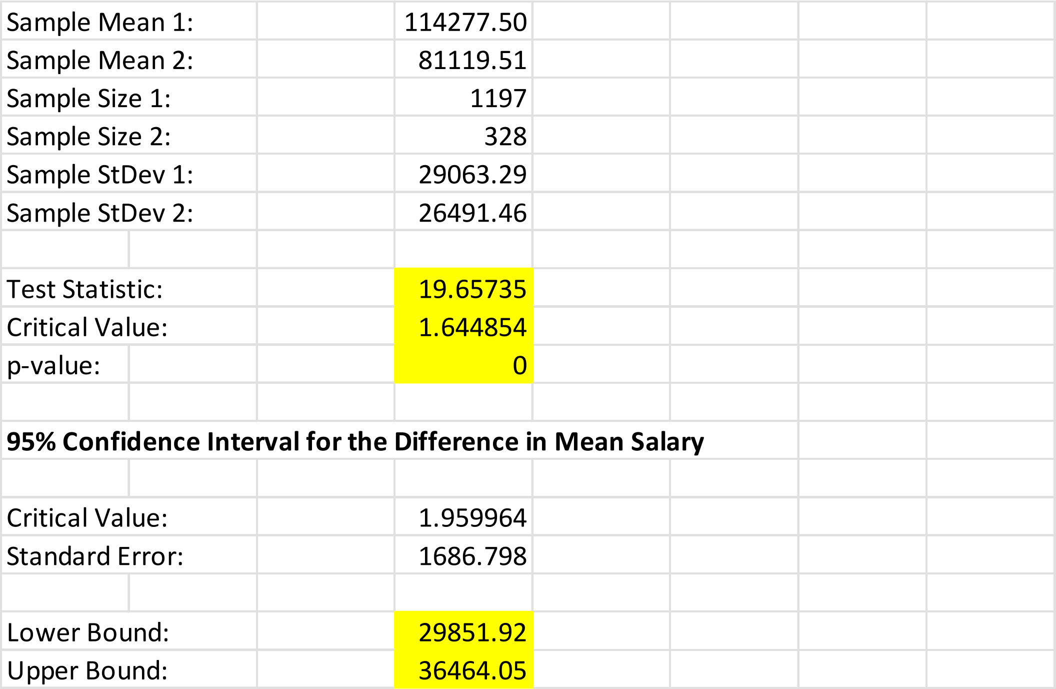
## Two Sample Results

First, we want to confirm our assumption that data scientists make more on average than data analysts. We want to test this suspicion and calculate a 95% confidence interval of the difference between mean salaries of data scientists and data analysts.

Below are the results:



| Two Sample Test: Difference in Mean Salary between Data Scientists and Data Analysts | |
|---|---|
| Null Hypothesis: | $\mu_0 - \mu_a = 0$ |
| Alternative Hypothesis: | $\mu_0 - \mu_a > 0$ |



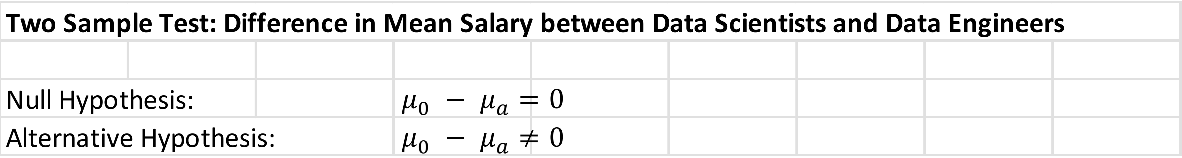
| | |
|---|---|
| Sample Mean 1: | 114277.50 |
| Sample Mean 2: | 81119.51 |
| Sample Size 1: | 1197 |
| Sample Size 2: | 328 |
| Sample StDev 1: | 29063.29 |
| Sample StDev 2: | 26491.46 |
| | |
| Test Statistic: | 19.65735 |
| Critical Value: | 1.644854 |
| p-value: | 0 |
| | |
| **95% Confidence Interval for the Difference in Mean Salary** | |
| | |
| Critical Value: | 1.959964 |
| Standard Error: | 1686.798 |
| | |
| Lower Bound: | 29851.92 |
| Upper Bound: | 36464.05 |

Based on this test, we **reject the null hypothesis**, confirming our suspicions that data scientist job listings have a greater average salary than data analyst job listings.

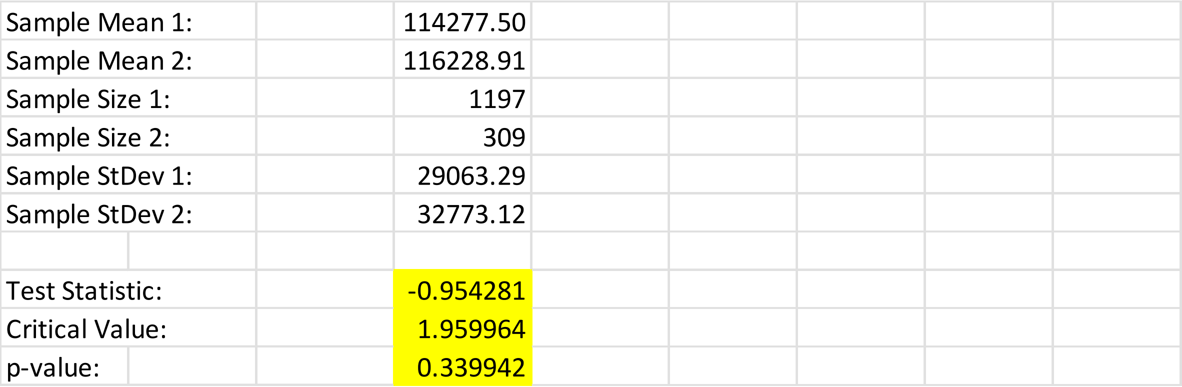
We conclude with 95% confidence that, in the job market, the real difference in average salary between data scientists and data analysts is between **$29,851.92 and $36464.05**.

Next, we want to test if there is any significant difference in the mean salaries of data scientists and data engineers. For this test, we will be operating under the assumption that the two populations do not have equal variance, since we have no evidence to indicate otherwise. Since our samples are sufficiently large, we will be estimating the population variances using our sample variances.

Below are the results:



| | |
|---|---|
| **Two Sample Test: Difference in Mean Salary between Data Scientists and Data Engineers** | |
| | |
| Null Hypothesis: | $\mu_0 - \mu_a = 0$ |
| Alternative Hypothesis: | $\mu_0 - \mu_a \neq 0$ |

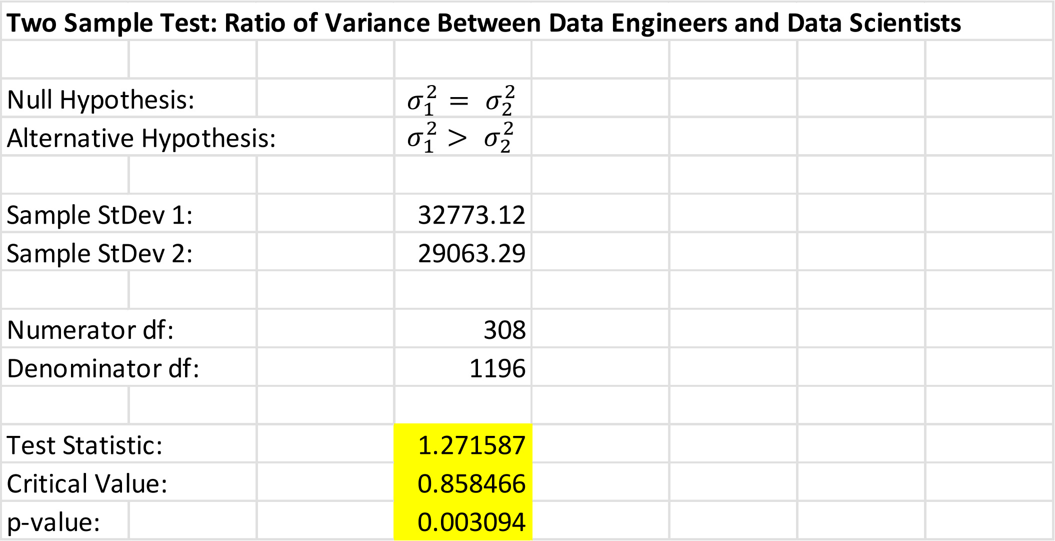


| | |
|---|---|
| Sample Mean 1: | 114277.50 |
| Sample Mean 2: | 116228.91 |
| Sample Size 1: | 1197 |
| Sample Size 2: | 309 |
| Sample StDev 1: | 29063.29 |
| Sample StDev 2: | 32773.12 |
| | |
| Test Statistic: | -0.954281 |
| Critical Value: | 1.959964 |
| p-value: | 0.339942 |

Based on this test, we **fail to reject the null**, thus we do **not** have evidence to conclude that the average salaries of data scientists and data engineers are different at the 95% confidence level.

To further investigate this relationship, we will perform a variance ratio test on the estimated salaries of data scientists and data engineers. In doing so, we acknowledge that the distributions of our variables do not appear visibly normal, so the results of this test may not be particularly trustworthy. This test is just to appease my curiosity.

Below are the results:



| **Two Sample Test: Ratio of Variance Between Data Engineers and Data Scientists** | |
|---|---|
| | |
| Null Hypothesis: | $\sigma_1^2 = \sigma_2^2$ |
| Alternative Hypothesis: | $\sigma_1^2 > \sigma_2^2$ |
| | |
| Sample StDev 1: | 32773.12 |
| Sample StDev 2: | 29063.29 |
| | |
| Numerator df: | 308 |
| Denominator df: | 1196 |
| | |
| Test Statistic: | 1.271587 |
| Critical Value: | 0.858466 |
| p-value: | 0.003094 |

Based on this test, we **reject the null hypothesis**, concluding that, at the 95% confidence level, the estimated salary of data engineers has a greater variance than the estimated salary of data scientists. Again, our underlying assumptions required for this test were not sufficiently met, so these results are inconsequential, albeit interesting.

## Section 5: Conclusions

In general, the results from these tests aligned with my suspicions about the job market. A significant portion of data jobs, more than 25%, are remote. In the post-pandemic world, remote positions are in high demand, so this is not a surprise. More than 80% of data science positions list Python as a skill in their job descriptions. As mentioned in the intro, Python is widely considered the premier coding language for data science, but 80% is such a high number that it suggests one may be at a severe disadvantage if they do not know Python.

We confirmed that data scientists make more on average than data analysts, which matches expectations, but requires further investigation. It could be the case that a higher proportion of data analyst jobs are entry-level, which would bring the average down. To be sure, we would need to further break down our data and analyze average salaries at each seniority level.

Finally, we observe that there is no significant difference between the average salary of data scientists and data analysts. To me, this is the most interesting result. A typical career trajectory involves starting in analytics and then transitioning into either data science, data engineering, or some other more specialized role, so it is reassuring to see that neither option appears to be at a clear disadvantage.

The implications of these results are clear: if one is aiming to break into the data industry, they should have their sights set ultimately on either data scientist or data engineering roles to maximize their potential earnings. If they are looking to land a remote position, they should have plenty of options to choose from. Most importantly, if they want to find a position in data science, it is very highly recommended that they start practicing with Python.

If I could change any aspect of my work here, I would start by ensuring that the dataset includes listings from multiple websites. The Glassdoor salary estimate for listings with no printed salary is very unclear and may be skewing my results away from reality. As such, I would love to revisit my tests after taking new samples that include listings from LinkedIn, Indeed, and other such websites. I would also like to scrape the data to find how many of the job listings include R, which is a programming language often referred to as an alternative to Python in the industry. It would be fascinating to see if either language has a higher representation than the other.

The original author of this dataset gathered the data with the intention of building a predictive model which would input job characteristics and output a predicted salary estimate. As someone looking to learn machine learning and predictive analytical techniques, I plan to iterate on his work to build predictive models of my own. For the reasons stated previously, I think new samples with improved diversity are necessary to create accurate models, so this will be a large endeavor, but it will be useful for me and provide valuable practice.

There is much work left to be done to fully understand the workings of the job market, but our results provide useful insights nonetheless.

# Section 6: References

Original data, authored by Mohamed Sayed and uploaded to Kaggle:
https://www.kaggle.com/datasets/mohamedsiika/data-related-jobs-in-us