Analyzing Pathways for Career Success within the Field of Data Science

Ayaan Saifi and Antony Sikorski

0. Contribution Statement

Both Ayaan Saifi and Antony Sikorski contributed evenly to the project. The R code was evenly split, and each author contributed equally to the write-up of the report. The advanced analysis was additionally evenly split between the two. That said, specific primary contributions are as follows:
Ayaan Saifi was the primary contributor for Questions 2,4.
Antony Sikorski was the primary contributor for Questions 1,3.

1. Introduction

Within this report we examine a dataset of responses to the 2020 Kaggle Machine Learning & Data Science survey. Kaggle received 20,036 usable responses from members of their community in 171 countries, and all spam responses were filtered out to ensure a more accurate dataset. The survey consisted of 34 multiple choice and “select all” format questions, which asked the respondents basic questions about themselves, and then further asked a variety of questions about jobs, equipment, and expectations for the field of data science.

The field of data science and machine learning is a very new and dynamic field, with an extreme potential for growth. Analyzing the data set of responses from a community that primarily works within the field allows us to gain many insights about what it would be like to work within it, and what an aspiring data scientist should know before entering the field. We investigate compensation, tools utilized for jobs, differences in gender, and many more categories in order to answer a couple general questions about the field, along with approaching the primary goal of preparing a student for being maximally successful if they wish to enter the field. We assume that the reader of this report plans to work in the United States, and we cater to this by examining salaries only within the US. This allows for the most accurate results due to different costs of living and currencies.

Examining the data provided a few obvious conclusions, such as Python and it’s respective tools and libraries being exceptionally popular in the field, and advanced graphical hardware being very desired for data visualization. We found that data science is quite a profitable field, yet there are notable differences in the salaries of the different positions. In addition to this, we noticed that there is a distinct gap between the salaries of male and female workers in the field. This gap does not appear to have any strong correlation to experience within the field, which leads us to believe that there may be other, external factors that influence this decision. Overall, the data analyzed provides multiple excellent metrics for what a student looking to work in the United States in the field of data science should expect and strive for.

2. Basic Analysis

2.1. Data Processing

Methods:
The data was loaded with R. The cleaning of our data mainly pertained to identifying those who didn't respond to certain questions, although the data set was quite clean from Kaggle already. In addition to this, many interval ranges for observations such as salary, age, and experience were replaced with their numeric means for ease of computation and analysis.

Analysis:
The dataset "2020 Kaggle DS & ML Survey" contained 20,036 usable responses from the Kaggle community. We scanned for nonresponses, and for certain questions we filtered out categories of respondents who were asked not to respond to certain questions (ex. Students not responding to their salary). In addition to this, interval ranges for responses such as salary, experience, and age were converted into their numeric means, and the highest category was converted to its minimum number. While this may affect the exactness of our answers, our primary purposes in this report have to do with comparisons, and not point estimates. In addition to this, the data was filtered down to only respondents from the United States for questions having to do with anything financial, so we could more specifically analyze the job market our reader is most likely going to be working in.

Conclusion:
The filtering of our data set down to only United States respondents most likely increased our average salaries, yet it made our observations more accurate to our designated reader. Our conversion of interval ranges into numeric means lowered our exactness, yet it increased our ability to work with the data and give an effective approximation of multiple categories. The restriction of the maximum categories to their minimum number is similar to the removal of outliers, and if anything will only lower the skew and inaccuracy of our average calculations.

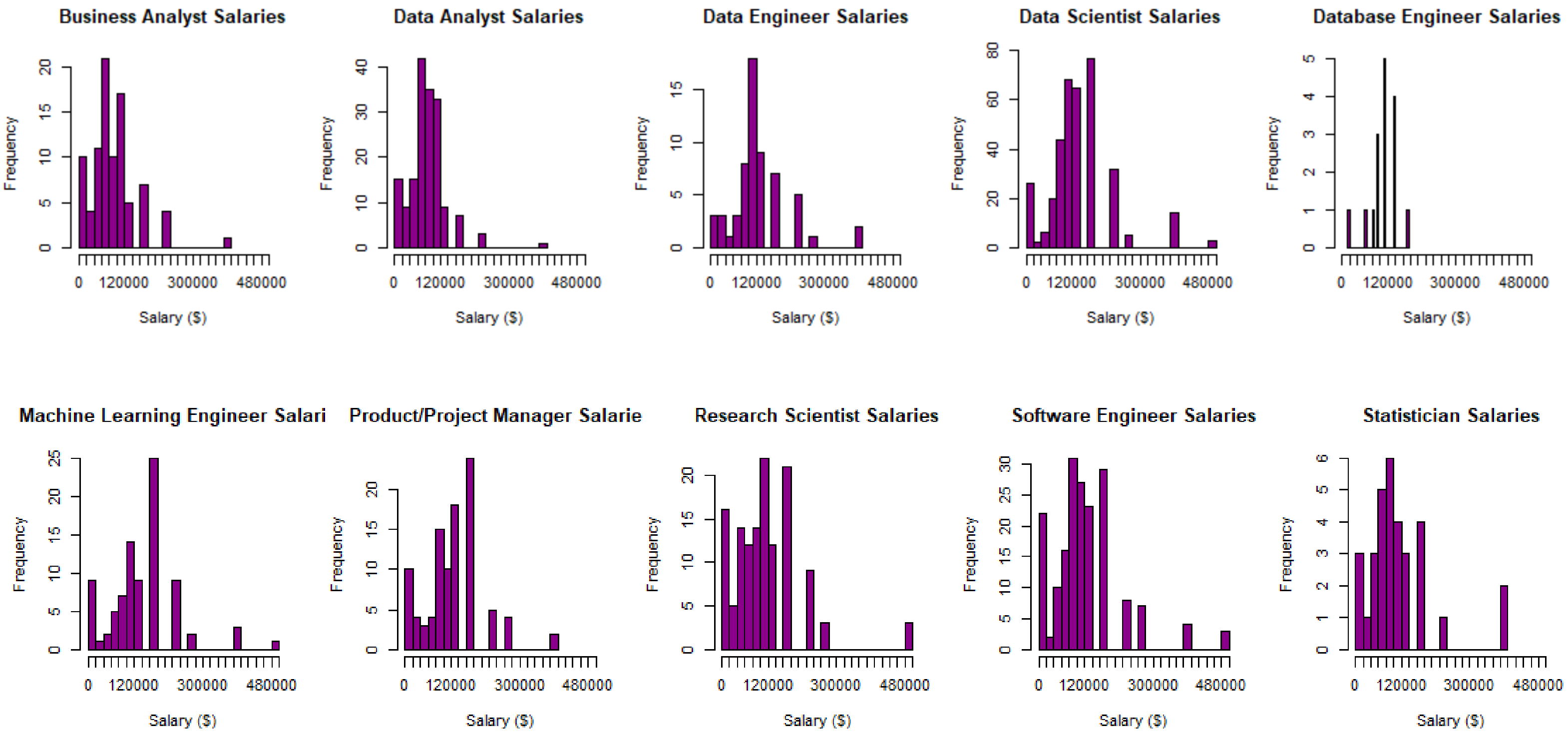
2.2. Salary Comparison of Different Data Science Jobs

Methods:

To begin our analysis, we wish to investigate the differences between compensation in different jobs within the field of data science.Within the survey, participants were asked to select a category for their salary, such as “$30,000 - 39,999”. Since this is not a numeric value, we took the mean of each category and replaced the entries as such so we could work with our data more easily. We chose to only examine the salaries within the United States since this is the job market we are choosing to advise our reader on, and the economy we are most familiar with. We first perform visual analysis of the salary distributions of all of the different jobs options that were given within the survey. We then examine the means and IQR ranges of the salaries for these jobs to get a more concrete and numerical estimate of the compensation within the field.

Analysis:

First, we plot all of the distributions of salaries based on the job that was selected in the survey. While response frequencies differ, looking at the distribution is still very informative.



We then calculate the mean and IQR for each position in order to gain a point estimate of how much one can expect to get paid, along with how much variability there is to the salary.

Position	Business Analyst	Data Analyst	Data Engineer	Data Scientist	Database Engineer
Mean	$90,947	$85,491	$129,767	$142,610	$108,750
IQR	$57,500	$47,500	$51,875	$80,000	$42,500

Position	Machine Learning Engineer	Product/ Project Manager	Research Scientist	Software Engineer	Statistician
Mean	$145,095	$128,955	$115,794	$125,089	$115,016
IQR	$80,000	$85,000	$120,00	$100,000	$72,500

Conclusion:

It appears that the highest salaries are in the positions of machine learning engineer, data engineer, and project manager. These positions additionally have quite high IQR's, which indicates a large range of salaries, although this most likely points to high growth potential due to there being a low frequency of low paid positions in these areas. The lowest salaries were quite clearly in the Business and Data Analyst positions, which also had lower IQR, demonstrating that the mean is quite an accurate representation of that position. Machine learning and project manager were expected to be high due to machine learning being an extremely hyped up and new field, and executive positions reliably paying well. Business and Data Analyst both sound less technical in nature, and effectively the lowest salaries do not come as a surprise.

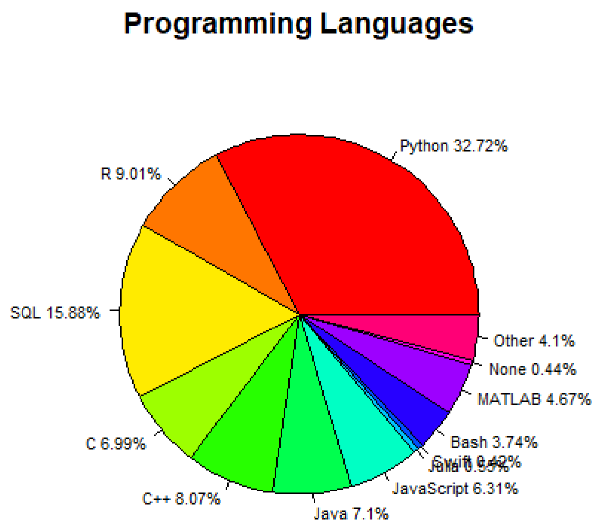
2.3. Examining the Most Common Programming Tools in the Workplace

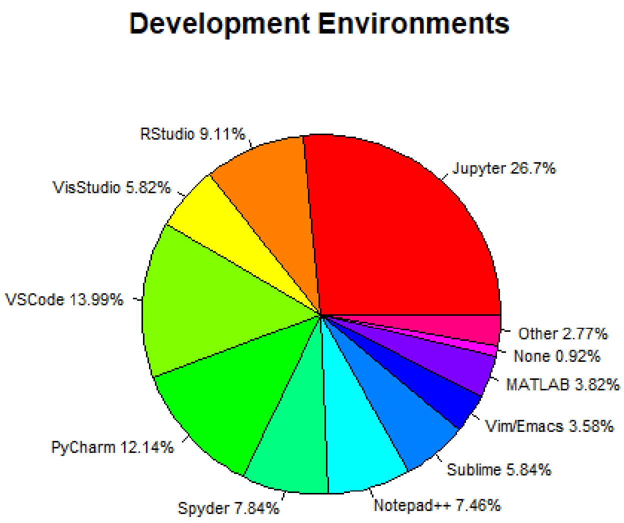
Methods:

The survey asked questions about which language, environment, hardware, and data visualization libraries workers used on an everyday basis in the workplace. Each of these questions allowed multiple responses in case people used more than one of these tools, which is quite understandable considering the fact that they all have different strengths, and often work cohesively. We will use data from the entire globe rather than just the United States, since we believe that the two are quite similar, and the frequency of programming language utilization is relatively uniform universally. We examine the most common languages, environments, hardware, and data visualization libraries based on frequency of response, and provide these percentages, along with visual demonstrations in order to get a sense for which tools are most commonly used in the field of data science.

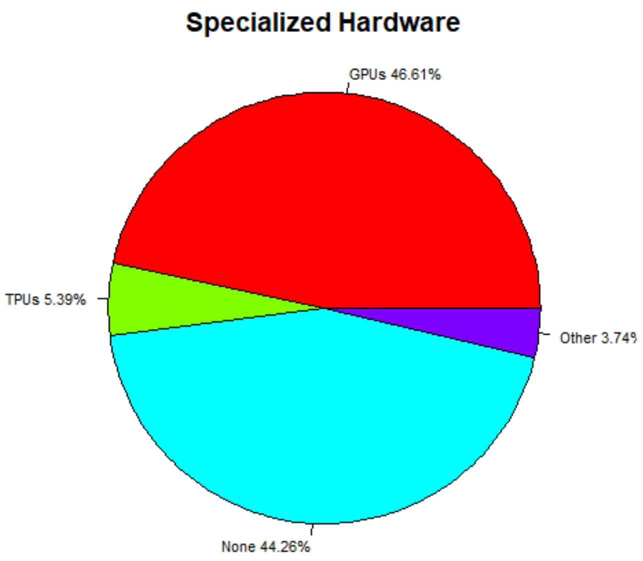
Analysis:

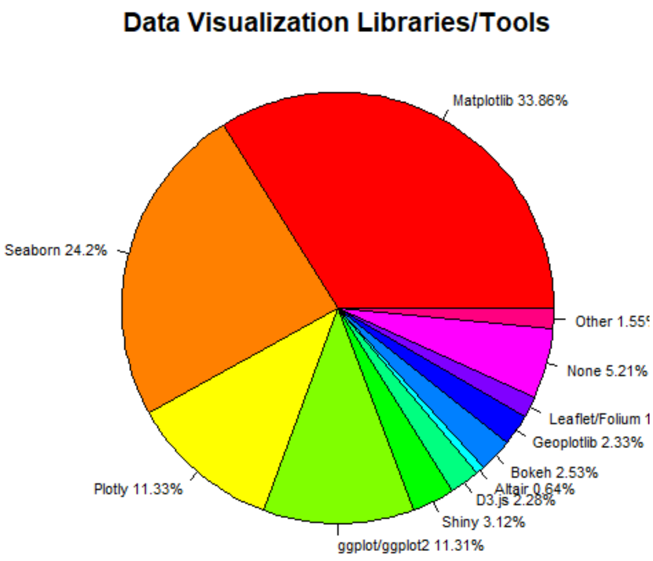
We first begin by finding the proportions of all of the programming languages and environments, and putting them into a pie chart to effectively demonstrate percentage.





We then follow with pie charts of percentage of use for the most common hardware types and data visualization libraries.





Conclusion:

After investigating the most common programming languages, it does not come as a surprise that Python (32.72%) dominates the field of data science. Other notable languages include R, which is very popular for statistical analysis, SQL, and common languages such as JavaScript. The development environments follow a very similar trend to the languages (ex. Jupyter Notebook for Python). The specialized hardware showed a surprisingly high percentage (44.26%) of workers in the field of data science who do not use special equipment, most likely because the processing power of their desktops is sufficient. Graphics cards were the most common equipment, which makes sense due to the necessity for complex data visualization. Finally, the libraries also showed a surprising amount of libraries that were designated specifically for Python, although this is countered by the fact that one of Python's strengths in the world of data science is the wide variety of powerful libraries compatible with it.

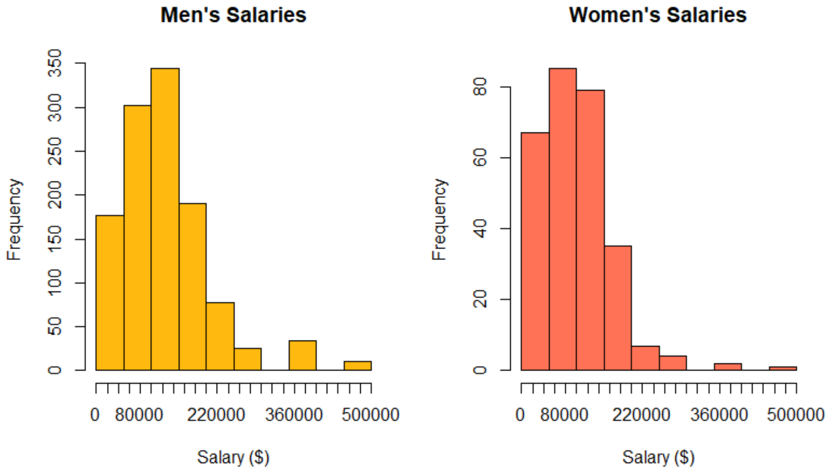
2.4. Differences in Pay by Gender in the United States Data Science Field

Methods:

Considering the magnitude of the pay gap within the United States and the presence of this social issue within our current society, we chose to examine the differences in pay between the two genders within the field of data science. We first visually examine the salary distributions and the average difference in salaries over the span of all the paid job options provided by the survey, and perform more formal statistical testing to determine whether there is a statistically significant difference in the two means. We also generate a 95% confidence interval for the difference in means. Finally, in order to more accurately observe whether the discrepancy is due to women taking lower paid positions, or if the pay is truly different between genders within the same position, we find the means of each salary for each position for comparison.

Analysis:

We plot the distributions of male and female salaries in the field to visually compare them.



The mean salary for men is $125,363, while for women it is $98,331. Using the Welch Two Sample t-test, we find that there is a significant difference in the means, with a p-value of $4.6 * 10^{-8}$, and the 95% confidence interval for the means is (17462.54, 36600.82).

We perform further analysis to find the mean salary for each position based on gender.

Position	Business Analyst	Data Analyst	Data Engineer	Data Scientist	Database Engineer
Mean (Men)	$95,773	$87,043	$121,078	$147,844	$115,000
Mean (Women)	$85,526	$81,663	$131,731	$119,480	NA
Position	Machine Learning Engineer	Product/ Project Manager	Research Scientist	Software Engineer	Statistician
Mean (Men)	$150,464	$128,994	$116,947	$124,648	$132,650
Mean (Women)	$111,300	$122,885	$102,471	$122,033	$85,625

The existence of the NA is due to the fact that our survey did not have any females who classified their current position most similar to that of a DBA/Database Engineer.

Conclusion:

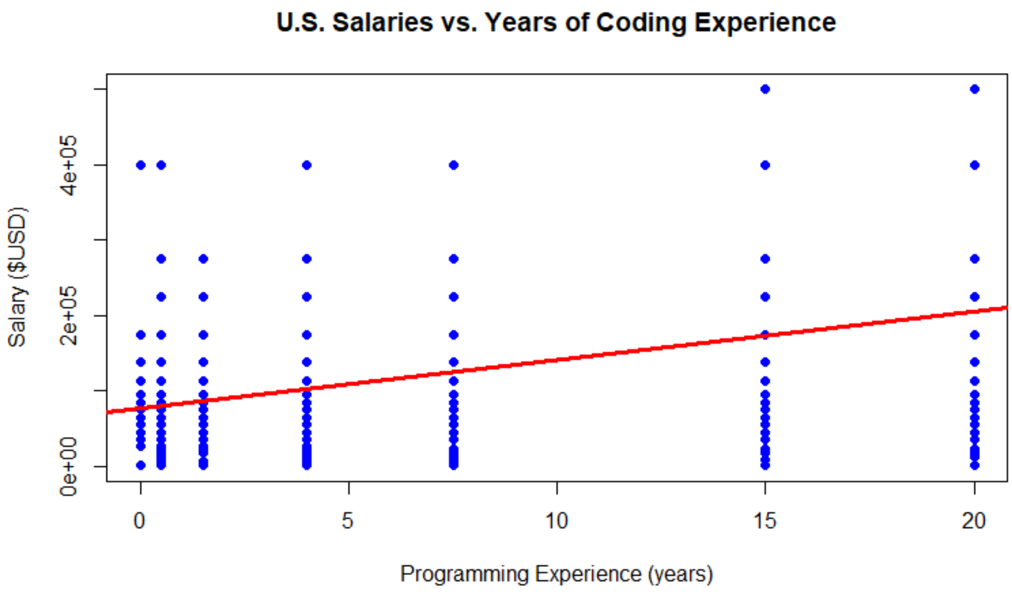
When comparing the male and female salaries within the field of data science, we found that on average males make nearly $27,000 more each year. We found the difference between the mean male and female salaries to be very statistically significant, with a 95% confidence interval for the difference between the means being (17462.54, 36600.82). Although there is a general societal pressure for women to often take lower paying jobs, our data does not support this conclusion. It rather supports the idea that women truly do get paid less for doing similar positions to their male counterparts. Only female data engineers had a reportedly higher salary than males, and on all other counts women appear to make less, with some of the differences, such as that of the machine learning or statistician positions, being extremely drastic. To conclude, the field of data science is not exempt from the unfortunate and general trend, and women appear to be paid a statistically significant amount less than their male counterparts.

2.5. Years of Coding vs Higher Education Effects on Salary

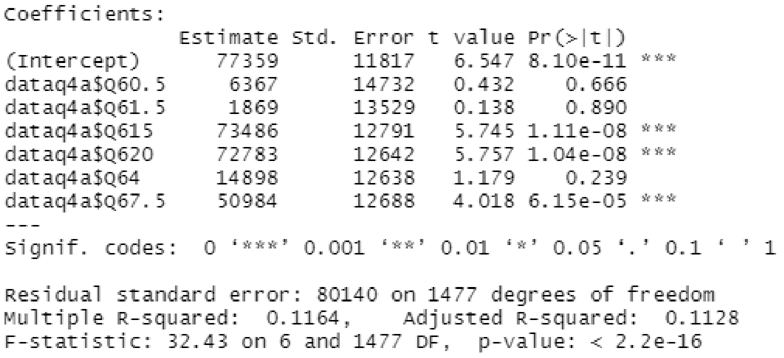
Methods:

A common question that is often asked in the world of data science and programming is: which is more important, your coding ability or your degree? We examine this question by performing a linear regression of salaries in the United States against both degrees and the number of years that one has reportedly been coding. The higher education degree options were converted into numerical values of their average time to completion. For example, a bachelor's degree was converted into 4 years of higher education. The number of years coding was provided from our data within ranges, so we used a similar method to our earlier questions and replaced each range of years coding with it's numerical average as an estimate. Both regressions were performed, and the residuals were inspected for normality, so we could accurately approximate which factor is more important to one's salary.

Analysis:
We first perform the regression of salaries in the United States and years of coding experience.



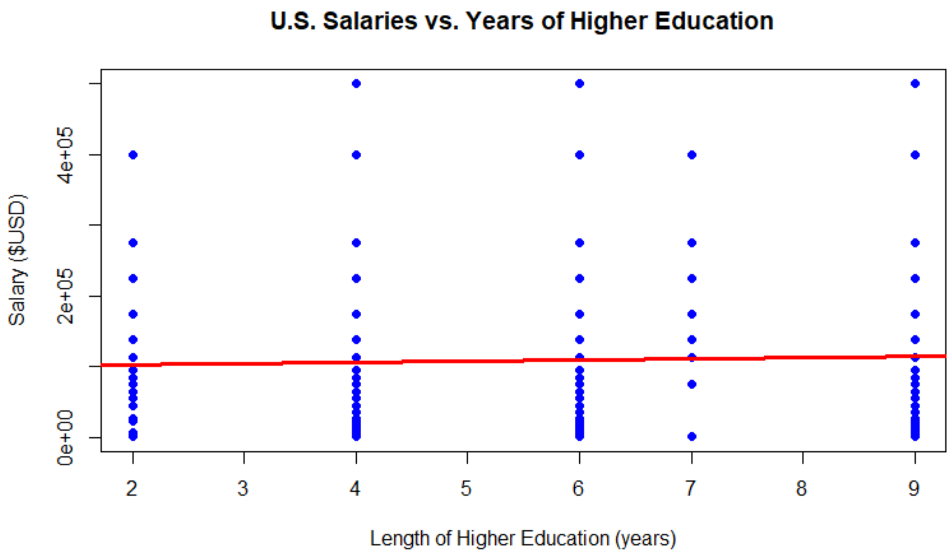
The fit provides the following results: most notably a low value of $R^2 = 0.1164$.



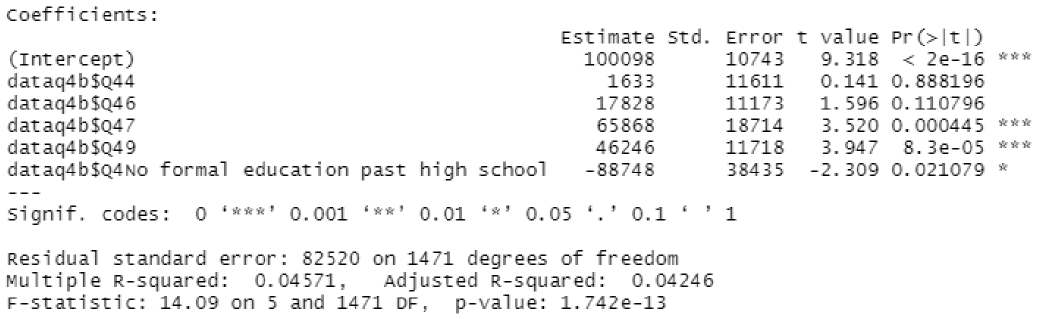
```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)        77359      11817   6.547 8.10e-11 ***
dataq4a$Q60.5       6367      14732   0.432    0.666    
dataq4a$Q61.5       1869      13529   0.138    0.890    
dataq4a$Q615       73486      12791   5.745 1.11e-08 ***
dataq4a$Q620       72783      12642   5.757 1.04e-08 ***
dataq4a$Q64        14898      12638   1.179    0.239    
dataq4a$Q67.5      50984      12688   4.018 6.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 80140 on 1477 degrees of freedom
Multiple R-squared:  0.1164,	Adjusted R-squared:  0.1128 
F-statistic: 32.43 on 6 and 1477 DF,  p-value: < 2.2e-16
```

The residuals were tested for normality using visual analysis from the plot and histogram, a Kolmogorov-Smirnov test for normality with an extremely low p-value of $2.2 * 10^{-16}$, and a QQ plot against the normal distribution. The residuals easily failed the test for normality. Figures can be found in the Appendix (Figures 1-3).

Afterwards, the same steps were taken to examine the effects of years of higher education on the current salary.



The linear regression showed the following results, with a $R^2 = 0.04571$value.



```
Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                    100098      10743   9.318  < 2e-16 ***
dataq4b$Q44                                      1633      11611   0.141 0.888196    
dataq4b$Q46                                     17828      11173   1.596 0.110796    
dataq4b$Q47                                     65868      18714   3.520 0.000445 ***
dataq4b$Q49                                     46246      11718   3.947  8.3e-05 ***
dataq4b$Q4No formal education past high school -88748      38435  -2.309 0.021079 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82520 on 1471 degrees of freedom
Multiple R-squared:  0.04571,	Adjusted R-squared:  0.04246 
F-statistic: 14.09 on 5 and 1471 DF,  p-value: 1.742e-13
```

The residuals were examined using plots, histograms, a KS test, and a QQ plot, and yet again failed the test for normality on all counts. Figures can be found in the Appendix (Figure 4-6).

Conclusion:

Unfortunately, the results from the linear regression did not provide statistically significant evidence of either factor. Whether it is number of years coding or number of years in higher education, neither had a strong enough R^2 value to claim there is a real correlation. Despite this, the two can still be compared relative to each other. The number of years coding regression demonstrates a higher slope per year, and also has a larger correlation coefficient than the number of years in higher education. This is not an argument against higher education, and students should still generally attempt to get at least a 4 year degree, but our dataset indicates that this industry places more emphasis on coding ability than education. Despite this relative analysis, it is extremely important to mention neither factor is statistically significant enough to demonstrate anything near a causal relationship.

3. Advanced Analysis

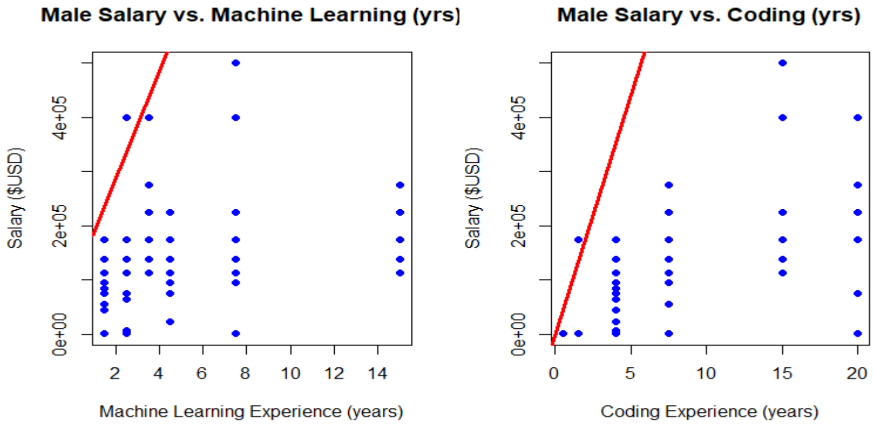
3.1 Further Examining the Gender Pay Gap

Methods:
We found the average gap between men and women to be extremely high, and wished to examine it further. We chose one of the positions with the highest gap between pay (Machine Learning Engineer), and decided to examine the reasoning behind this pay gap further. We first look at the average number of years that both men and women have been coding, and have been using machine learning methods. Both of these questions were present in our survey. We then perform regression analysis to see if either factor is truly significant towards changing the salary for either gender to see if we can validate the gender pay gap with some form of statistical evidence.

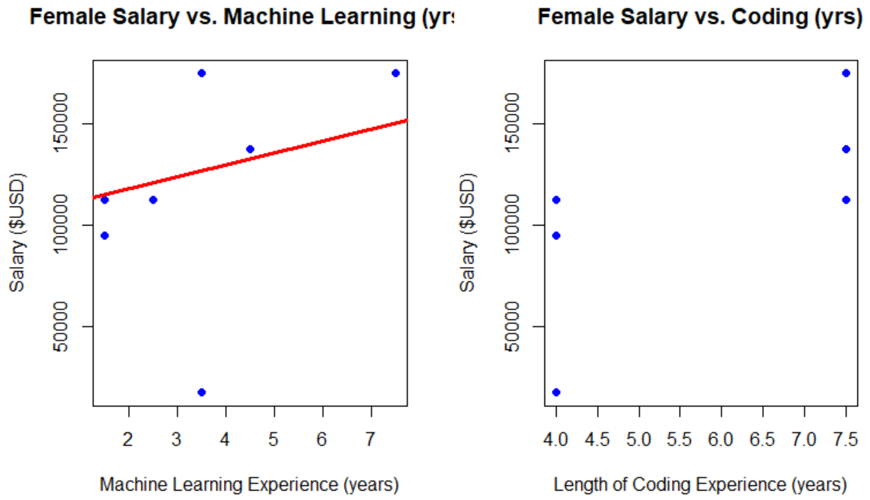
Analysis:
We first recall the mean salaries for men and women in the position of machine learning engineer, and then provide the mean number of years coding and mean number of years using machine learning methods for both genders.

Gender	Male	Female
Mean Salary	$150,464	$111,300
Coding Experience	10.13 years	6.333 years
Machine Learning Experience	4.845 years	3.722 years

We then perform linear regression analysis for both men and women to see if any of these factors are significant.



Neither years of machine learning, nor years of coding, appear to have a strong correlation with male salaries, with the trend lines appearing to be quite inaccurate and the respective R^2 values being 0.1632 (machine learning) and 0.2867 (coding).



The pattern follows for females, with the machine learning trend line seeming to be accurate, but the trendline for coding not even being present on the graph. The respective R^2 values being 0.3869 (machine learning) and 0.5171 (coding).

Conclusion:

When first observing the statistics, we are led to believe that men may be paid more due to the fact that they have significantly more years of both coding and machine learning experience than their female counterparts. Since both of these are huge parts of a machine learning engineer's job, this is easy to believe. Further investigation using regression analysis shows us that neither of those variables actually has a significant effect on the salaries of men or women. Interestingly enough, women do appear to have a stronger correlation between the number of years they have been using machine learning methods or coding, and their overall salary. Despite this, this analysis shows us that we cannot prove that men are paid more than women in the machine learning role due to the number of years of experience they have coding or using machine learning methods. This means there must be some form of external factors that influence the pay gap, which may be more in favor of the idea that discrimination against women is a larger factor than actual skill, although we do not have statistical evidence of this, so we will not draw this conclusion. All we can say is that neither machine learning nor coding experience are significant enough reasons for men to get paid more than women in the position.

4. Discussion and Conclusion

The objective of this study was to examine the survey responses from the Kaggle Data Science and Machine Learning survey in order to gain a better understanding of the field of data science itself, and advise a student wishing to enter the field how best to succeed. We examined the field on multiple fronts, one of those being the tools and programming languages one must be proficient in to succeed. We found Python to be the dominant programming language within the field, with its respective libraries and packages being the most commonly used within the field. The most common hardware that one must most likely have was a graphics processing unit (GPU), which is most likely necessary for computers to perform difficult data visualization demonstrations, although almost just as many jobs reported not using any special equipment. Another aspect the study focused heavily on was the financial aspects of the data science job market. Looking into the different positions, we found that machine learning engineers, data scientists, and project managers appear to have the highest salaries, and positions with "Analyst" in the title were paid much lower than the rest. We wished to observe the reasons that one could get paid more within their field, so we performed regression analysis on the years of coding and years of higher education to see whether one could guarantee a higher salary. We found the correlation between both of these factors to be extremely low, yet relatively speaking, coding experience was found to be a better predictor of higher salary than years of coding experience.

Furthermore, we found a statistically significant gap in salary between men and women in similar positions. Further examining this, we found that only female data engineers had a higher salary than their male counterparts, while in the rest of the positions the males were paid more. To further examine this, we examined the machine learning engineer position, which had one of the more drastic disparities in pay. Although men had both more years of experience in coding and machine learning on average, regression analysis showed that neither of these factors was statistically significant. This means we cannot prove that men were paid more due to these skills, so there must be some external factors such as discrimination that influence men's salaries being so much higher. This is not a new concept, and disparities in pay between genders has been found in many other studies, with an article published by Sabrina Baez (1) demonstrating that men were paid significantly more in the technological and data science fields.

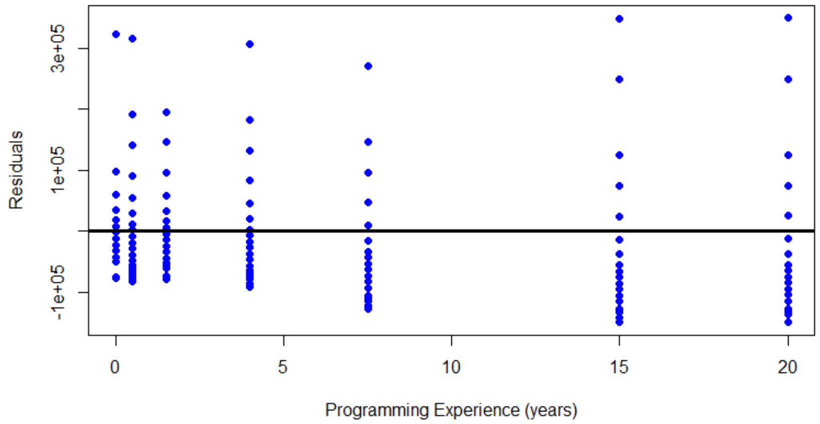
To conclude, it appears that an student aspiring in the field of data science should most likely learn Python and it's respective libraries, and aim for positions with more technical roles and titles. While we do not advocate for a lack of education, it appears that one should stress their coding experience more than their degree in order to obtain a higher salary. Further research should perform more thorough analysis of more positions to understand why women are paid less than men in the field. A survey asking for more information regarding the respondent's technical background would be useful for this purpose, so that more resume accomplishments can be cross referenced to men's and women's salaries to see if there is any correlation.

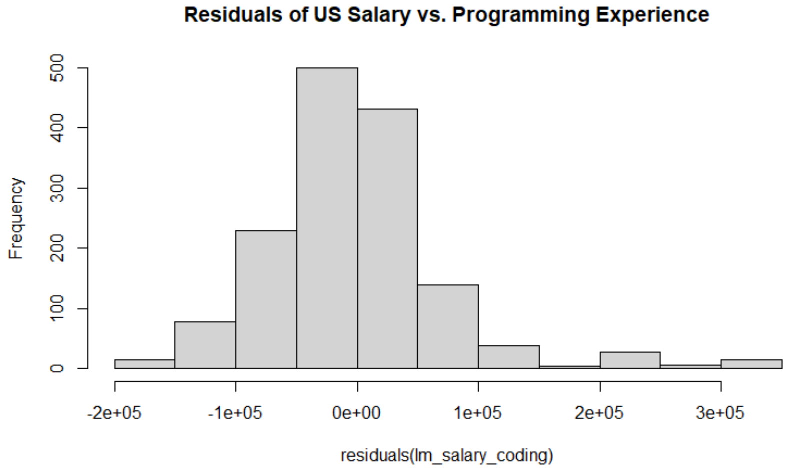
5. Work Cited

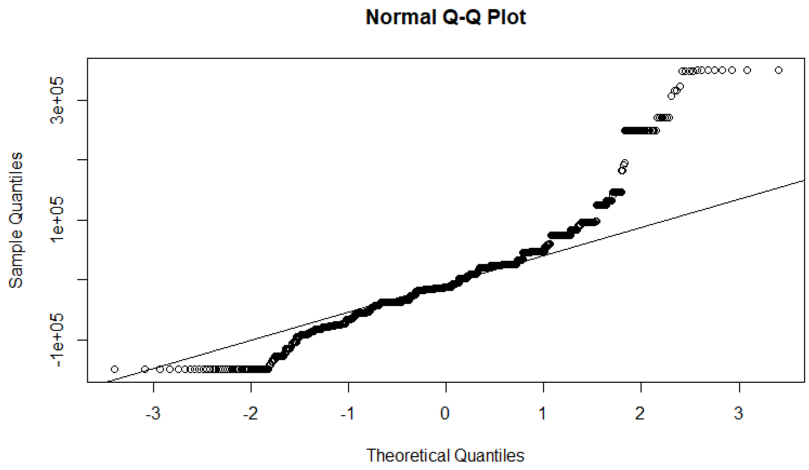
1. Baez, Sabrina. "Women, the Gender Gap in Data Science, and What You Can Do About It." *Dataquest*, 28 June 2019, www.dataquest.io/blog/women-data-science-gender-gap/.

6. Appendix

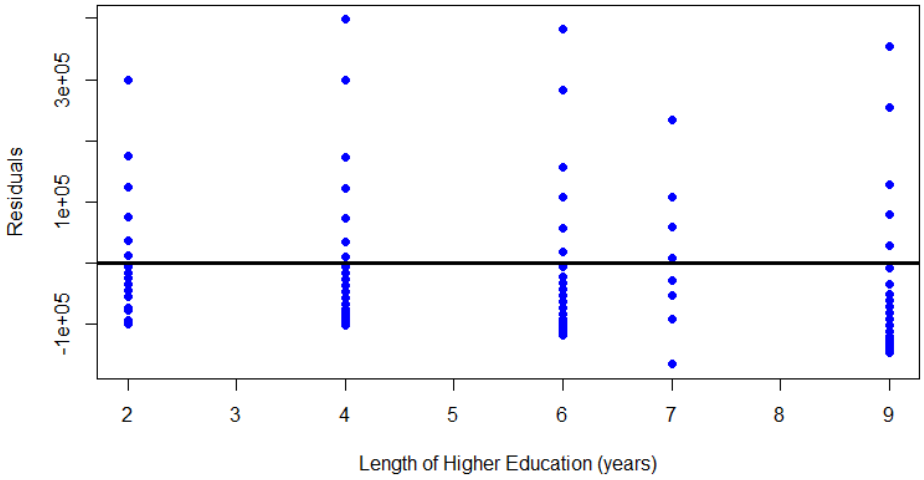
Figures 1-3 (Plot, Histogram, and QQ plot for residuals of years of coding experience vs salary):

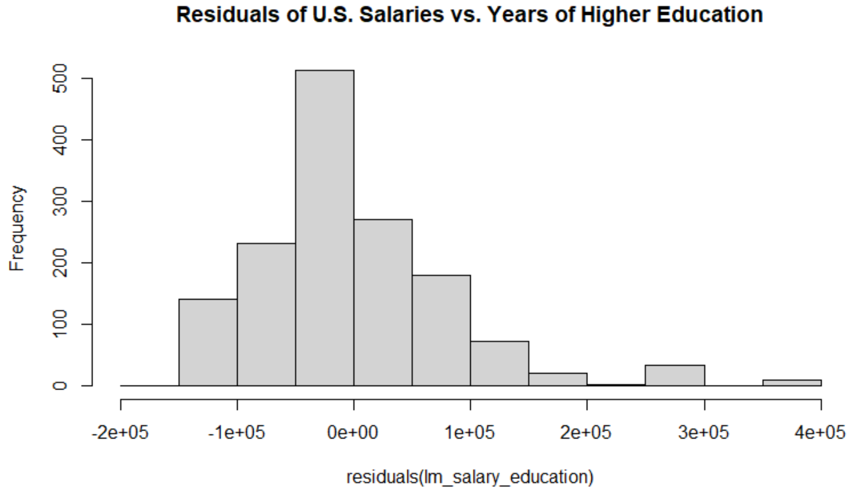


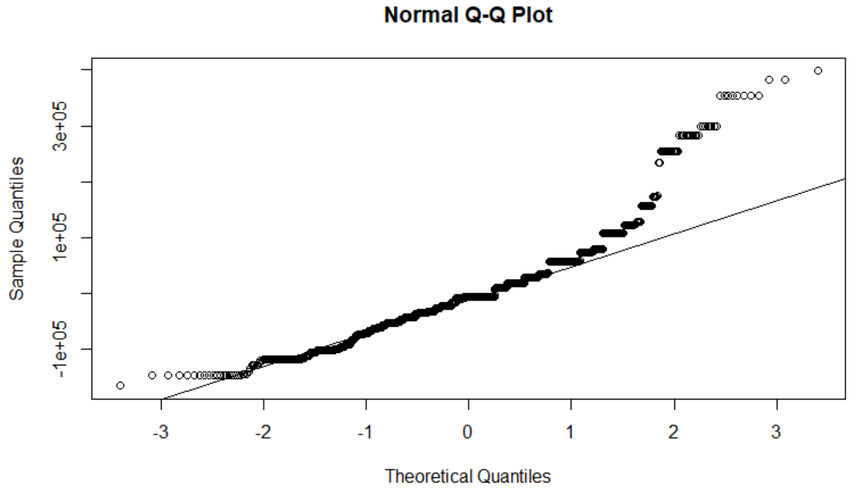




Figures 4-6: (Plot, Histogram, and QQ plot for residuals of years of higher education vs salary)







7. Questions

1. Examine the compensation responses of the survey participants and compare them to the jobs that they are currently working. What appear to be the highest paying jobs in the field of data science in the United States? Utilize visual displays and basic measures of center to answer this question.

2. What languages, environments, data visualization libraries, and hardware are the most common in the field of data science across the globe? Examine these responses to see how to see what the typical data scientist uses on a daily basis. How drastic are the differences in use? Use visual tools along with comparisons of percentages and tables to identify if certain aspects are much more common than others.

3. Compare the distribution of salaries for men and women in the field of data science in the United States. Compare the two means, and see if there is a statistically significant difference between the two means. Make sure to use more formal statistical testing such as the two sample t-test to answer this question. Where do you think this difference in means comes from?

4. What affects your pay more, the number of years that you have been coding, or what number of years of higher education (degree of education) you have been awarded? Perform a linear regression on both and observe the slope and the strength of the correlation to make a conclusion about which factor is more important in your success in this field. Make sure to pay attention to the visuals, slope, and fit.

5. **Advanced Analysis**: Further examine the Machine Learning Engineer position, which has one of the highest pay gaps between men and women. What factors could be the reason that women are paid less than men in this position? Particularly focus on the years of experience in coding and machine learning to see if the pay gap is warranted by these skills, or if there may be external factors at play.