DS-GA 1007 Group 3 Report:
Jean An, Clara Kim, Stephen Spivack, Jason Wang, Yiran Zou

Analysis of Full-Time STEM Compensation

1. Introduction

All students enrolled in this course would presumably wish to work as a full-time data scientist in the near future. One of the reasons for the popularity in data science -- and the STEM fields in general -- is the high compensation that comes with the title. However, it is not entirely clear which specific factors are driving the high compensations. Some of the questions that are often asked include: which STEM roles are the most well-paid? Are graduate degrees such as Master's degrees and Doctorates important driving factors of salary compensation? Do years of experience directly contribute to the amount of compensation? Lastly, which specific factors are the major driving forces of salary in the STEM fields? In this report, we will explore possible answers to these questions by analyzing data from an online site that offers services in salary negotiation.

2. Data

2.1 Data Source

We utilized data from the webpage levels.fyi, which contains self-reported information from full-time employees within the STEM fields. The information is verified by official offer letters and W-2s.[1] We found an article that examined its network calls using developer tools, and from June 7th, 2017 until August 17th, 2021, levels.fyi delivered its entire salary payload from a single endpoint.[2] Since the data is stored as a .json file, we imported the requests library and loaded the data as a Pandas DataFrame.

Unfortunately, it appears that levels.fyi started to monetize access to their data, so the .json file has not been updated since and only includes data up to Aug 17th, 2021. However, there were still 62,642 rows of data before we performed any cleaning or preprocessing of the data.

The inspiration for our dataset came from Kaggle[3], but obtaining our dataset and cleaning it were performed using our own methods.

2.2 Data Cleaning

First, we took a look at the 17 columns of the dataset: 'timestamp', 'company', 'level', 'title', 'total yearly compensation', 'location', 'years of experience', 'years at company', 'tag', 'base salary', 'stock grant value', 'bonus', 'gender', 'other details', 'cityid', 'dmaid', 'rowNumber'.

All the data types were strings, so we had to first change the 'timestamp' data type to datetime64 and all numerical columns' data types to float64. Also, missing string values were encoded as an empty string, so we replaced any empty string with NaN for easier analysis.

When we looked at the numerical columns 'total yearly compensation', 'base salary', 'stock grant value', and 'bonus', we noticed that there were inconsistencies in how users self-reported their

data, especially data with older timestamps. Total yearly compensation was always reported in thousands of dollars, while the other columns were sometimes reported in dollars. We first multiplied total yearly compensation by 1000, so that the column would be in dollars for ease of analysis. When we looked at the descriptive statistics of that column, the minimum total yearly compensation was $10,000 while the maximum was $5,000,000. Therefore, we could figure out among the other three columns 'base salary', 'stock grant value', and 'bonus' which rows were in thousands of dollars and which rows were in dollars. Given the known maximum and minimum, had the base salary value been less than $5,000, we would know that the given row was reported in thousands of dollars instead of dollars; we would then multiply the three columns by 1000.

Next, ‘other details’ sometimes included information about race, education, and other miscellaneous details in a single string. We wanted to extract this data, so we used the extract function and regular expressions to create two new columns, ‘Race’ and ‘Education’. Using regular expressions only obtained the first words of the race and education, so we also had to replace the word with its full text when appropriate.

Finally, we addressed logical errors in the numerical columns. The below equation,

Total yearly compensation = base salary + stock grant value + bonus,

should hold for all rows; but it wasn’t the case for some data points -- especially those with older timestamps -- because it either had missing values or just mistakes when users reported their data. Thus, for more accurate analysis, we removed all rows where the above equation did not hold, leaving us with 49,917 rows. This allowed us to set all missing values for ‘stock grant value’ and ‘bonus’ to 0, since if the above equation holds, that means that all NaN values are equal to 0.

3. Methodology and Results

3.1 Exploratory Data Analysis and Visualization

3.1.1 Distribution of numerical and categorical variables

We first conducted an exploratory data analysis of our cleaned dataset to gain an overall understanding of the structure of our data. To explore the distribution of our numerical variables, we generated a histogram for each of the columns as shown in Figure 1. Part of the color mapping code came from GitHub.[4]

All of our numerical variables were skewed to the right. ‘Years of experience’ ranged from 0 to 69 years with a median of 6 years. ‘Years at company’ also ranged from 0 to 69 years but with a median of 2 years. ‘Total yearly compensation’ ranged from $10,000 to $4,980,000 with a median of $190,000. ‘Base salary’ ranged from $1,000 to $900,000 with a median of $143,000. ‘Stock grant value’ ranged from $0 to $4,400,000 with a median of $25,000. Finally, ‘bonus’ ranged from $0 to $1,040,000 with a median of $15,000

We also printed tables that showed the tallies of each unique value in the categorical variables, displaying the top 10 counts for each, as shown by the tables in the Appendix.

3.1.2 Correlation Analysis

First, we created a heatmap to show the correlation matrix between the numerical variables (see Figure 2). Our target variable ‘total yearly compensation’ is most correlated with ‘stock grant value’ (r = 0.89), meaning that ‘stock grant value’ could be the leading indicator when measuring total salary. ‘Total yearly compensation’ is naturally also correlated with ‘base salary’ (r = 0.77) and ‘bonus’ (r = 0.51). It’s also worth noting that ‘total yearly compensation’ has a stronger correlation with ‘years of experience’ (r = 0.42) than with ’years at company’ (r = 0.16).

3.1.3 Salary Breakdown by Years of Experience, Title and Education

First, we grouped ‘years of experience’ into buckets of 5-year spans and plotted the average salary in each group (see Figure 3). Before reaching 40 years of experience, all three parts of ‘total yearly compensation’ -- ’base salary', ‘bonus’, and ‘stock grant value’ -- increases as ‘years of experience’ increase. The ‘total yearly compensation’ reaches its peak for those with 35 to 40 years of experience at around $380,000, which is more than twice for beginners with 0 to 5 years of experience at around $160,000.

Second, we created stacked bar plots of the average salary for each ‘title’ and sorted them by ascending order of ‘total yearly compensation’ (see Figure 4). Software Engineering Manager takes the lead by earning around $360,000, followed by Product Manager (~$260,000) and Technical Program Manager (~$240,000). The ‘base salary’ does not vary much for different positions, but the figure clearly shows that manager roles have higher ‘stock grant value’ than other roles, and that sales roles have higher ‘bonus’ than other roles.

Lastly, we created stacked bar plots of the average salary for each ‘education’ category and sorted them by ascending order of ‘total yearly compensation’ (Figure 5). In general, the higher the degree obtained, the higher for all three parts of ‘total yearly compensation’. The ‘total yearly compensation’ is around $290,000 for Doctorate degree holders, $220,000 for Master’s degree holders, and $180,000 for Bachelor’s degree holders. However, employees with a high school degree or equivalent and those who only completed some college coursework receive a higher ‘total yearly compensation’ of around $220,000 than Bachelor’s degree holders.

3.2 Prediction: Multiple Linear Regression

Since we are interested in knowing the most important factors that determine ‘total yearly compensation’, we decided to implement a multiple regression analysis. To do so, we first formatted our data so that we could train our model. For our predictors we constructed a Pandas DataFrame containing (1) years of experience, (2) years at company, (3) dummy variables for the top three companies in terms of popularity (here, operationalized as the total number of responses), (4) dummy variables for the top three job titles in terms of popularity (also operationalized as the total number of responses), (5) dummy variables for gender (male and female), (6) dummy variables for race, and (7) dummy variables for education. For our

outcome variable 'total yearly compensation' -- which is what we aim to predict using our factors -- we implemented SciPy's z-score function to normalize the data to a standard normal distribution. Next, we implemented scikit-learn's train-test-split function to randomly partition both our predictors and outcome variable using an 80/20 split. With the training subset we fit a multiple regression model, which we then used to make predictions using the testing subset of our predictors matrix. To compute the testing error and assess the score of our model we implemented scikit-learn's metrics module, which resulted in an RMSE of 0.799 and an r-squared of 0.279. Finally, using Matplotlib we plotted the beta coefficient for each of our predictors so that we could visually assess the importance of each factor (see bar plot in Appendix).

4. Conclusion

4.1 Summary

Our project aimed to investigate the factors that influence full-time STEM compensations. We conducted several analyses: exploratory data analysis, correlation analysis, visualization, salary decomposition by various factors, and multiple linear regression.

Results showed that total yearly compensation is affected more by stock grant value than by base salary or bonus, and the total number of years worked matters more than the number of years worked at a specific company. Salaries increased with years of experience and peaked at around 20-40 years. Manager roles tended to have higher total compensation, mainly due to greater stock grant value. Interestingly, besides the Doctorate degree, higher education did not necessarily predict higher compensation. In fact, the compensation of those with a Bachelor's degree was lower than those with just a high school or equivalent degrees or with some college coursework. Unluckily, for us Master's students, the compensation for those with a Master's degree was equivalent to that for those with only some college coursework. This may be a result of our limited data on certain rare categories. It may also be due to the fact that most people who are hired into STEM roles without even completing undergraduate degrees are likely the elite talents that do not require higher education, and thus are more likely to have had greater career success. Finally, our multiple regression analysis showed that the most important predictors of total yearly compensation are individuals who 1) have a PhD, 2) live in San Francisco, 3) work as a software engineering manager, and 4) work at Google.

4.2 Limitations and Future Studies

One possible limitation is that our dataset could be biased given that it was self-reported, not randomly selected. This means that our sample may not be representative of the total population, hence yielding biased results.

Furthermore, our dataset had limited information in particular STEM roles, such as Data Analysts and Data Engineers. Additional scraping of data from websites other than levels.fyi to incorporate such roles could produce a more representative analysis of full-time STEM role compensations.

5. References:

1) https://www.levels.fyi/verified/
2) https://towardsdatascience.com/a-beginners-guide-to-grabbing-and-analyzing-salary-data-in-python-e8c60eab186e
3) https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries/version/1
4) https://github.com/arseniyturin/Matplotlib-Histogram
5) https://numpy.org/
6) https://pandas.pydata.org/
7) https://scipy.org/
8) https://scikit-learn.org/stable/
9) https://matplotlib.org/
10) https://seaborn.pydata.org/

6. Appendix of Figures and Tables:

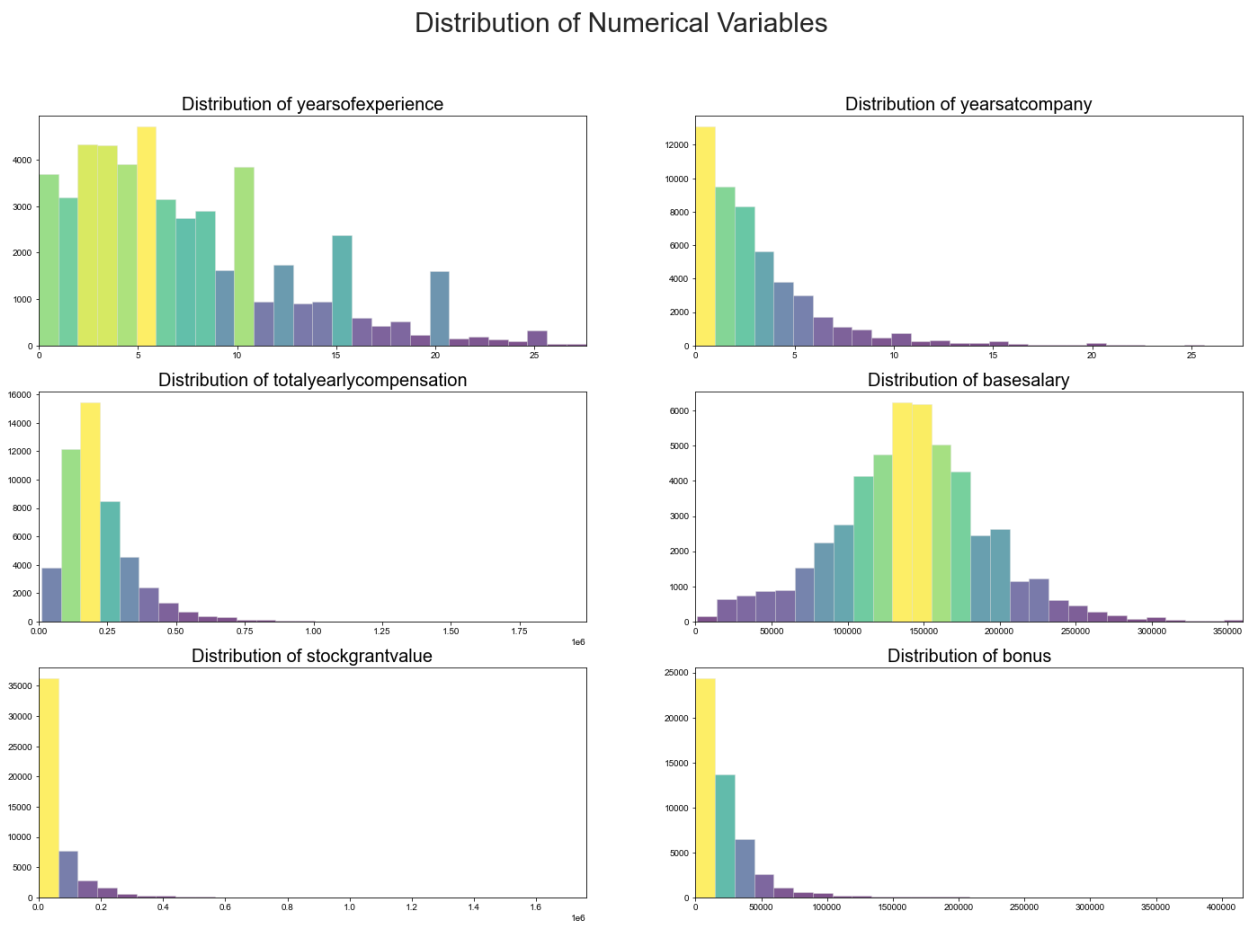

Figure 1: Histograms of the 6 numerical columns in our data type. Note that the counts may not be representative of all STEM salaries in the U.S, since levels.fyi does not randomly obtain data as compensation is self-reported.

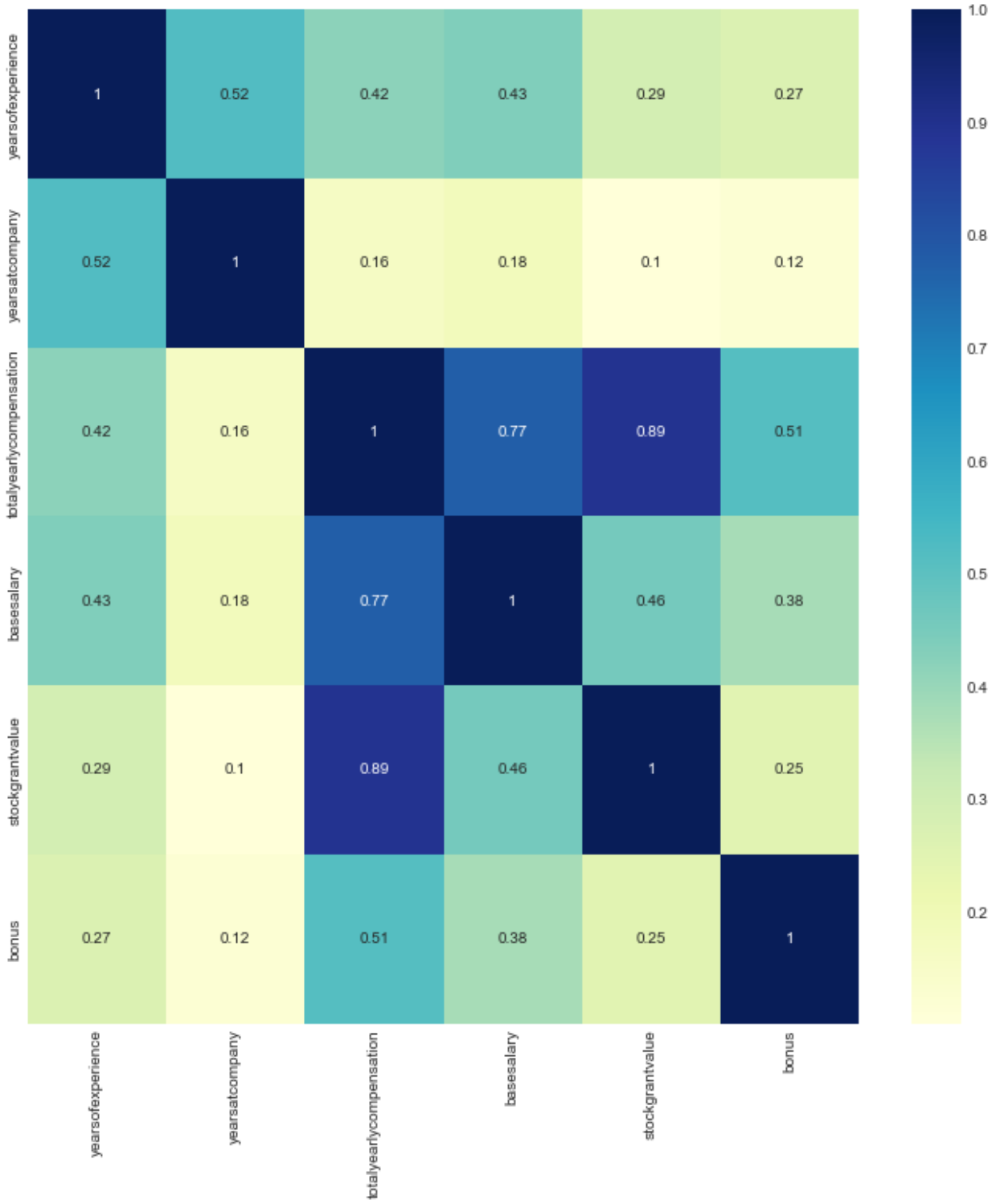

Figure 2: A correlation heatmap between the numerical variables.

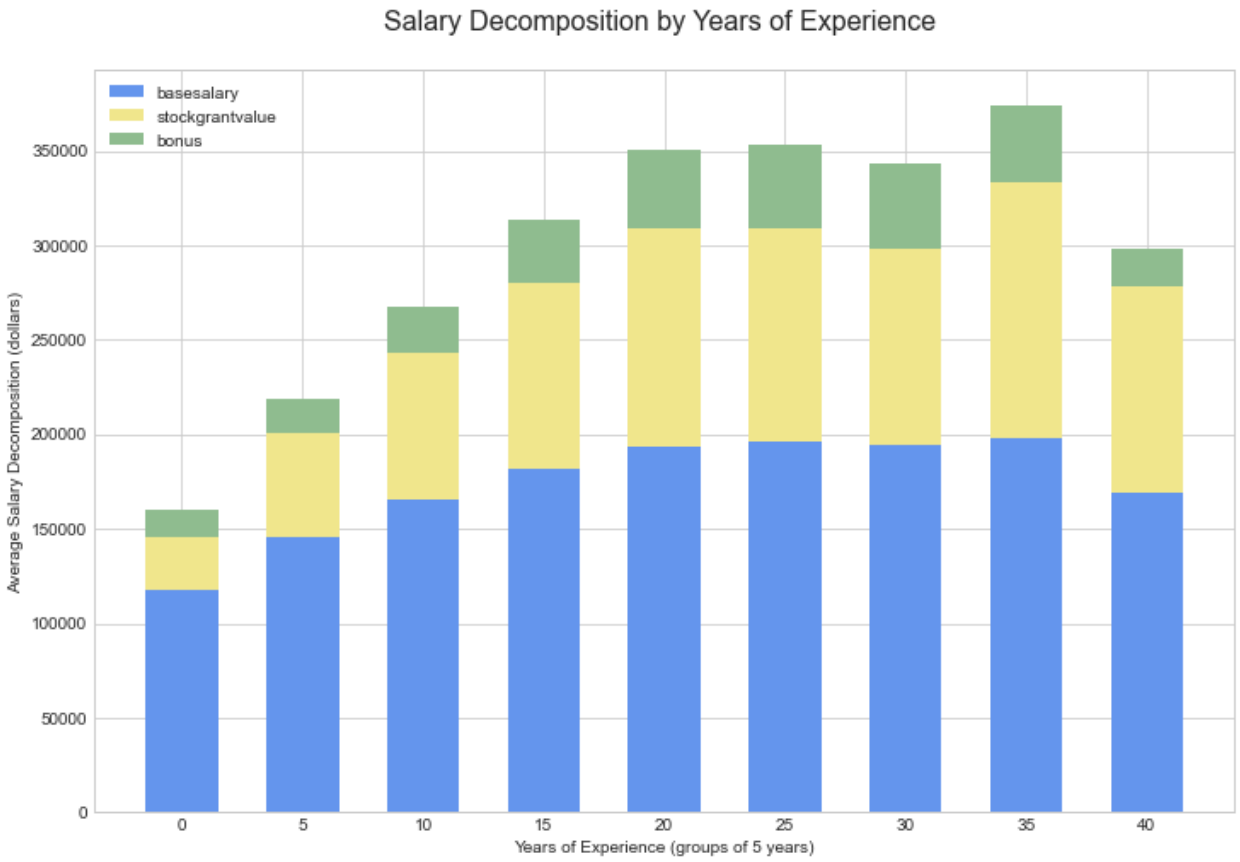

Figure 3: Breakdown of salary compensation by years of experience. Note that x labels refers to the lower bound of the bucket of every 5 years (eg. the first bar includes those with greater or equal to 0 years of experience but less than 5 years of experience).

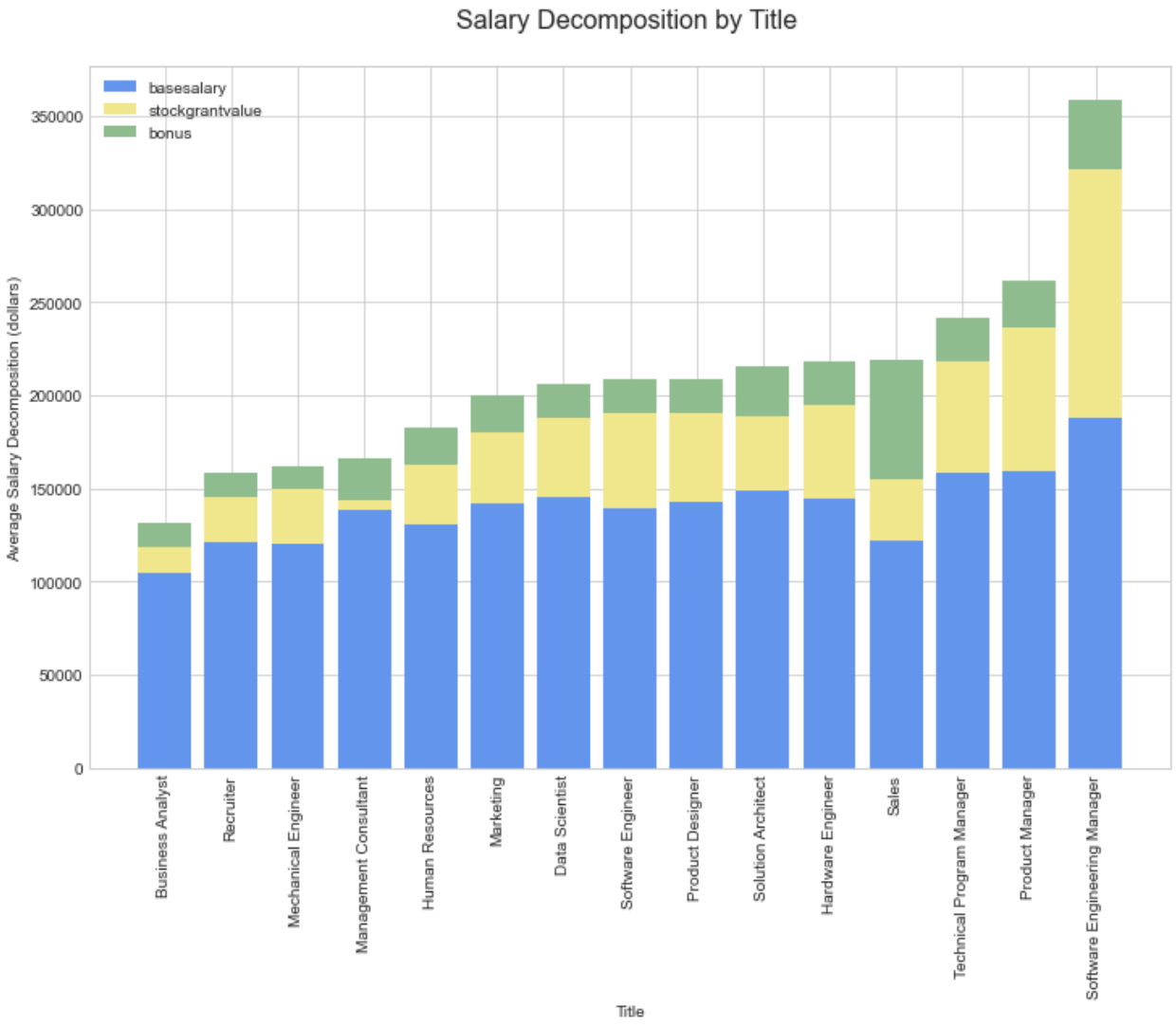

Figure 4: Breakdown of salary compensation by title. Note that levels.fyi only allows users to select from these 15 titles.

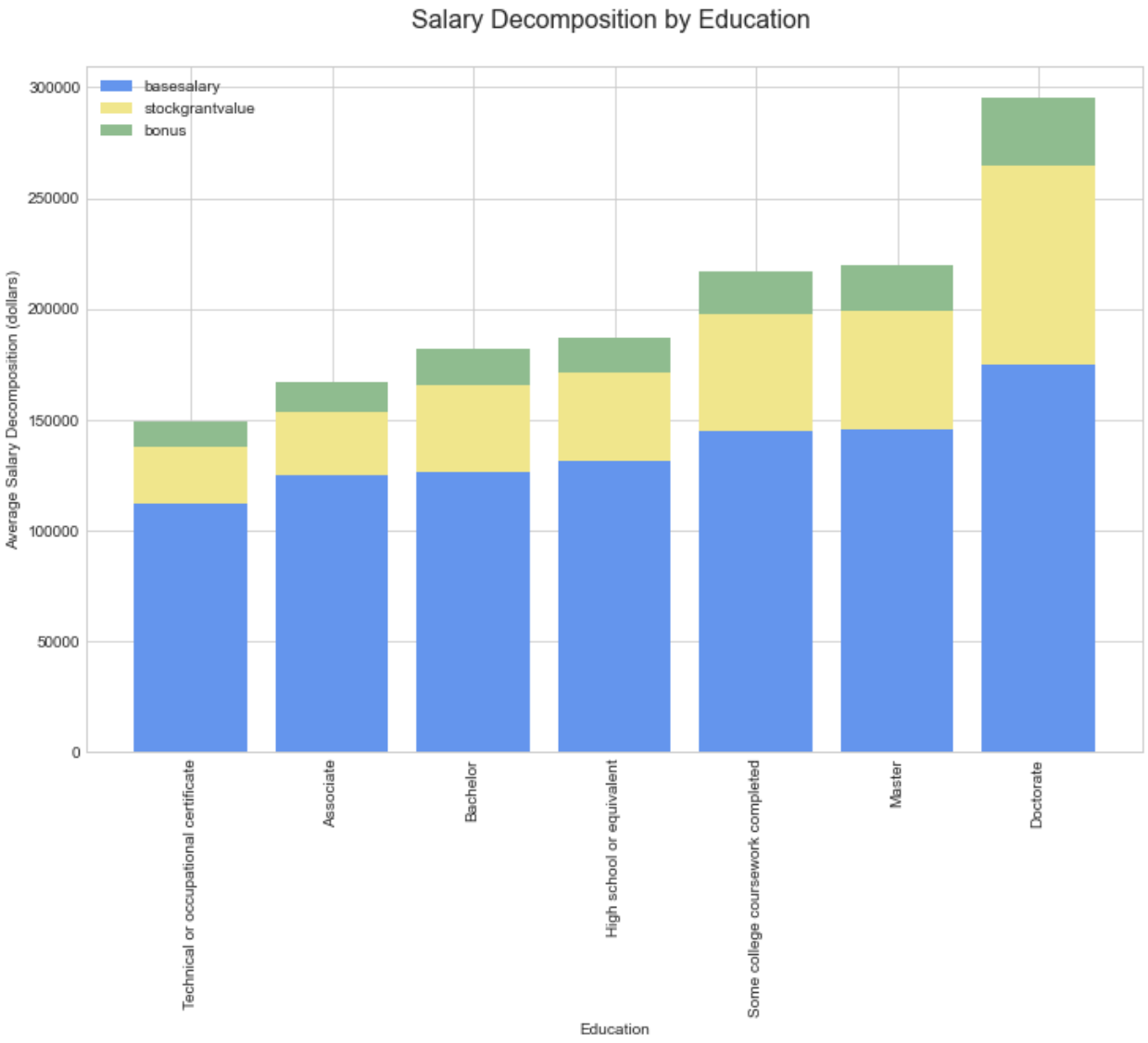

Figure 5: Breakdown of salary compensation by education. Although high school or equivalent and some college coursework completed data may be surprising, it is important to note that this is only data submitted by those already working in the STEM fields.

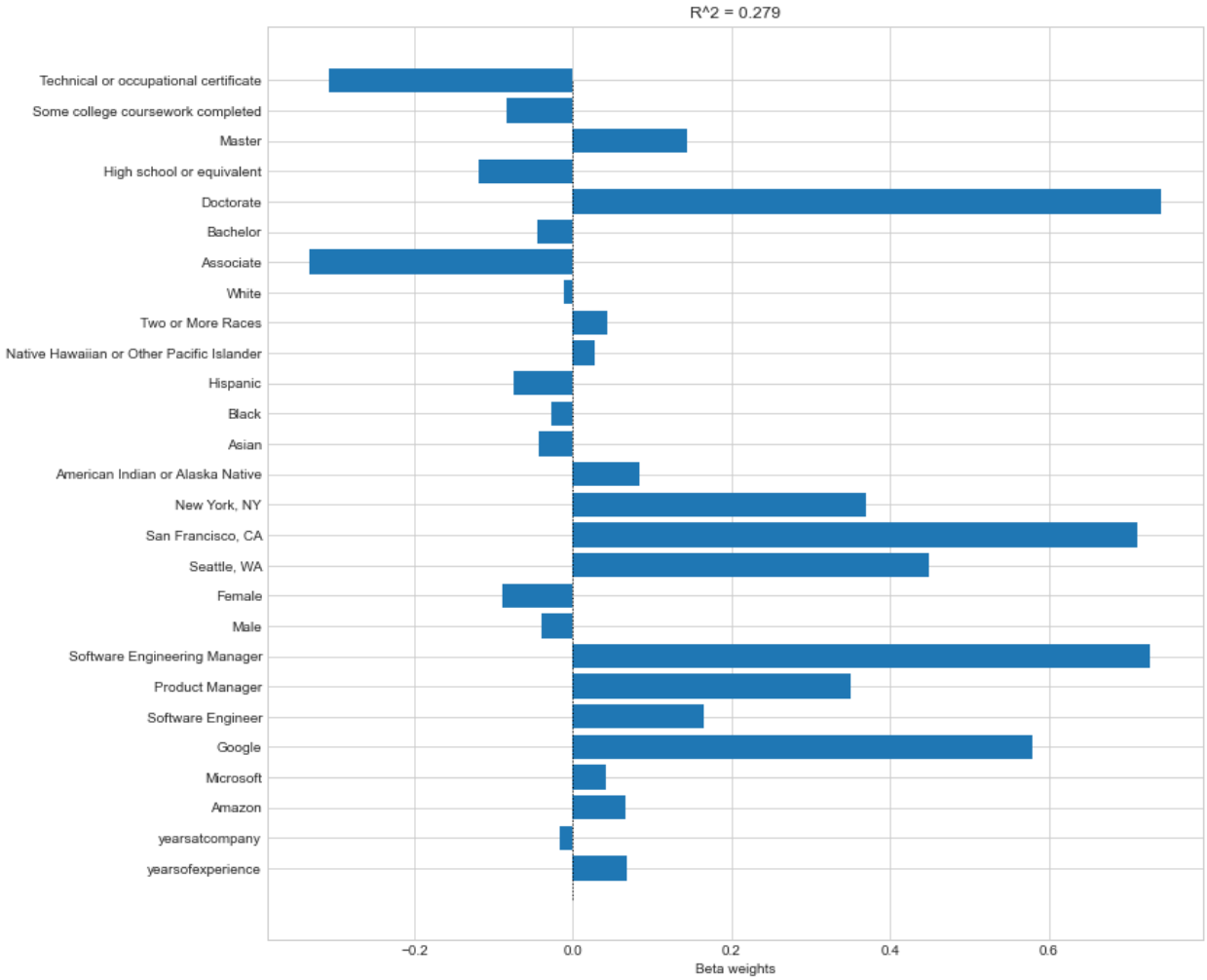

Figure 6: The beta coefficients after fitting a multiple regression to the data. Note that because of the existence of categorical predictors such as job title and location, we only selected the top three in counts within each category.

location	
Seattle, WA	6654
San Francisco, CA	5427
New York, NY	3860
Redmond, WA	1989
Mountain View, CA	1773
Sunnyvale, CA	1744
San Jose, CA	1643
Austin, TX	1253
Cupertino, CA	1173
Menlo Park, CA	1151

company	
Amazon	5978
Microsoft	3827
Google	3356
Facebook	2368
Apple	1639
Oracle	947
Salesforce	789
IBM	724
Intel	714
Cisco	697

title	
Software Engineer	32637
Product Manager	3592
Software Engineering Manager	2855
Data Scientist	2098
Hardware Engineer	1838
Product Designer	1213
Technical Program Manager	1161
Solution Architect	921
Management Consultant	846
Business Analyst	721

Table 1: Top 10 locations, top 10 companies, top 10 titles with highest counts

Education	
Bachelor	10438
Master	9102
Doctorate	1038
Some college coursework completed	287
High school or equivalent	268
Associate	119
Technical or occupational certificate	90

Race	
Asian	9582
White	6825
Hispanic	941
Two or More Races	680
Black	577
American Indian or Alaska Native	55
Native Hawaiian or Other Pacific Islander	27

gender	
Male	28876
Female	5549
Other	330

Table 2: Distributions of education, race and gender with counts