JSC370 Final: Understanding Gender Bias in Self-Reported Software Compensation Data

2022-04-21

Introduction

It is no secret that the software industry pays lucratively, but are there trends beyond just the dollar signs? For example, traditionally, the workforce has been biased against women, but is tech any different due to its novelty as an industry, or is it all the same?

While my midterm project focused on the exploration of trends in the tech/software industry in an exploratory fashion to uncover patterns that the fancy user interfaces don't show, this final report will dive deeper into one suggestive trend: the potential bias against women in tech. However, keep in mind that while this analysis is statistically oriented, the results are only as good as the data. Unfortunately, software compensation data is not simply available as a population, much less to the public. To circumvent this, the following analysis uses the levels.fyi database, which is a website that allows users to anonymously self-report compensation data, along with company, location, gender, etc. data. Hence, the conclusions of this report is therefore not concrete due to the nature of self-reported data being potentially inaccurate, along with many hidden biases, and so dear reader take it with a grain of salt.

With that said, this investigation will focus mainly on two aspects of gender bias specified by the following reseach questions: - 1) How do gender ratios and roles differ in the tech industry based on self-reported data? - 2) How does compensation differ in the tech industry between genders based on self-reported data?

Methods

The data was extracted from levels.fyi [1] based on code from this blog [2]. A detailed guide on transformations applied to the data can be found under `data/data.Rmd` and `data/README.md` in the repo here:

(https://github.com/zw123han/JSC370_final/tree/master/data).

Dependencies Analysis, tables, and `ggplot2`/interactive visualizations was conducted using the `tidyverse`, `plotly`, `DT`, `knitr` packages.

Modelling was done using the `lme4` and `lmtest` packages, using a linear mixed approach. All analysis with the exception of retrieving the data was performed using `R` Version 4.1.3.

Removed Data With respect to the variable of interset (total compensation), the dataset is complete after cleaning. However, entries were removed two ensure stability of modelling and analysis. They are as follows:

- 5th and 99th percentile of reported total compensation removed. Due to the nature of self-reported data, it was manually verified that this removes most of the "extra zero" mistakes.
- Self-reported genders other than "Female" and "Male" were removed to ensure accuracy and to avoid any skewness due to small group size. Limitations of this decision are addressed at the end of this report. The number in each gender group can be found in this table below.

Table 1: Table 1: Average Reported Total Compensation and Counts by Gender

Reported Gender	Count	Average TC	SD
Female	6673	191239	97412
Male	33558	204864	114939
Other	361	217676	123910
Unreported	17585	217717	118162

We note that a large number of respondants do not report their gender, unfortunately due to the nature of this analyis they had to be removed. There is also suggestive evidence based on average total compensation that females make less than males, this will be explored later in the results.

- Interactive visualizations only show the 100 most commonly reported groups of interest (eg. Company, Location). This is to make them more user-friendly. Other analysis uses the full dataset, minus the alterations explained above.

Results

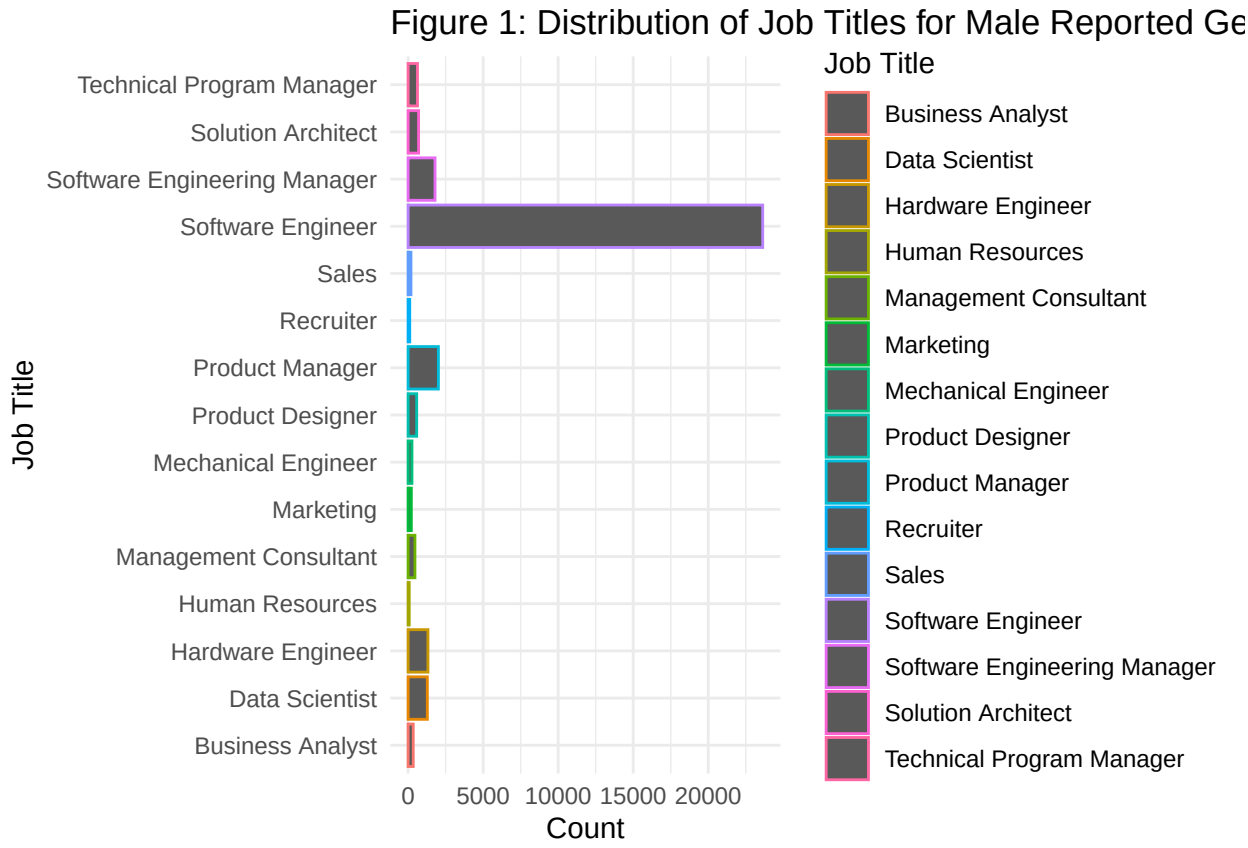
Figure 1: Distribution of Job Titles for Male Reported Ge
Job Title
Technical Program Manager
Solution Architect
Software Engineering Manager
Software Engineer
Sales
Recruiter
Product Manager
Product Designer
Mechanical Engineer
Marketing
Management Consultant
Human Resources
Hardware Engineer
Data Scientist
Business Analyst
0
5000
10000
15000
20000
Count
Job Title
Business Analyst
Data Scientist
Hardware Engineer
Human Resources
Management Consultant
Marketing
Mechanical Engineer
Product Designer
Product Manager
Recruiter
Sales
Software Engineer
Software Engineering Manager
Solution Architect
Technical Program Manager

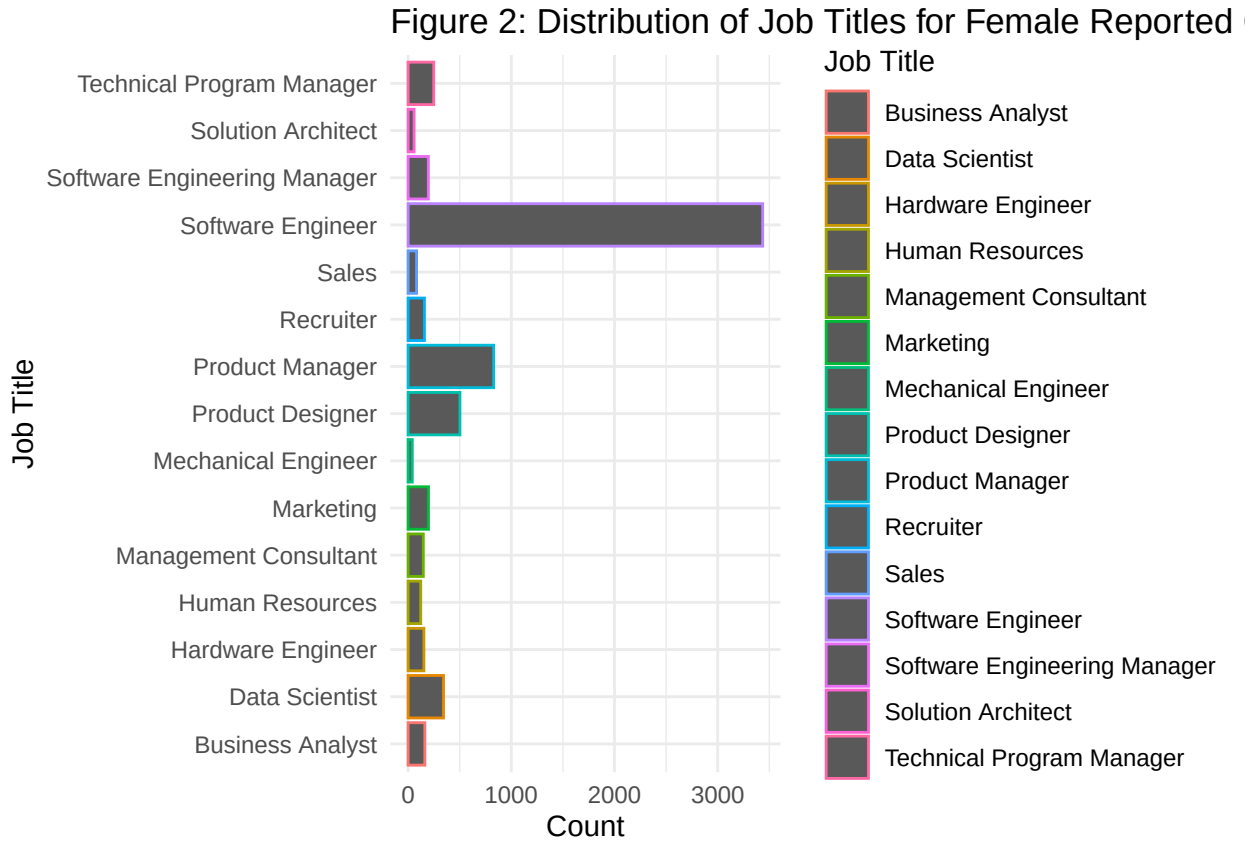
Figure 2: Distribution of Job Titles for Female Reported
Job Title
Technical Program Manager
Solution Architect
Software Engineering Manager
Software Engineer
Sales
Recruiter
Product Manager
Product Designer
Mechanical Engineer
Marketing
Management Consultant
Human Resources
Hardware Engineer
Data Scientist
Business Analyst
0
1000
2000
3000
Count
Job Title
Business Analyst
Data Scientist
Hardware Engineer
Human Resources
Management Consultant
Marketing
Mechanical Engineer
Product Designer
Product Manager
Recruiter
Sales
Software Engineer
Software Engineering Manager
Solution Architect
Technical Program Manager

There are 15 possible job titles related to the software industry in the dataset. Though this is not all-encompassing, they capture the majority of roles. We once again note that the number of data points for males is much greater than females, and this skewness remains a limitation.

Visually, one can see that the distribution of job titles for males is much more skewed than for females. A large majority hold the "Software Engineer" title for males, and while also true for females, there is a wider distribution of job titles.

Statistics & Modelling

You, the reader, might wonder why there are not many obvious statistical tests in this report. Why not simply use Welch's t-test for unequal variance to check total compensation, or salary, or stock grant values, or bonses between males and females?

My reasoning is because they simply do not capture enough information. Factors such as job title, location, and company all play a role in affecting the compensation. This is why I instead chose to focus on visualizations and modelling, which both generalize (or in the case of linear models, equivalent) and capture more information than a t-or-other-statistical test might.

Linear Mixed Modelling As explored in the midterm report, the compensation data takes an approxmiately normal distribution, which allows us to use the powerful and easily interpretable linear regression model on it. However, there are many companies and locations within the dataset, which makes it difficult to explain each coefficient if we were just to naively use classical linear modelling.

To account for this, linear mixed effects model using both random and fixed effecs were used, where the influence of variables with many levels on the variance is taken into account, just not interpreted (random effect). Since our research focuses on gender, we will treat it as a fixed effect. Years of experience (YOE) are also taken into account as an interpretable fixed effect. Interactions are not used to avoid high collinearity/variance inflation factors.

Total Compensation An initial model was fit as follows.

Model 1: TC ~ Company (random) + Location (random) + Job Title (random) + YOE (fixed) + Gender (fixed)

A smaller model without the fixed effect of gender was also built, then compared using the likelihood ratio test (LRT) for nested models. A likelihood ratio test statistic of chisq = 194, $p < 0.001$ indicates very strong evidence against the fact that both models would explain the data equally well. Hence, we need to take into account the effect of gender.

This model estimates that holding comapny, location, job title, and year of experience all constant, being male increases the average total compensation receieved relative to baseline (females) by an average of 15014 US. The LRT test allows us to determine this is statistically significant. On a side note, every year of experience increases the average total compensation by 5688 US, all else being held constant. Not bad!

Next, we will try to understand the components of total compensation (base salary, stocks, bonuses) that gender affects the most.

Base Salary Model 2: Salary ~ Company (random) + Location (random) + Job Title (random) + YOE (fixed) + Gender (fixed)

This is similar to the first model, only using base salary instead of total compensation. A LRT test result of chisq = 68, $p < 0.001$ allows us to again conclude that gender plays a significant role in impacting base salary. In this case, being male increases the average base salary by 4941 US, all else being held constant. Aside, every year of experience will also increase average base salary by 2579 US, all else held constant.

Stock Value Model 3: Stocks ~ Company (random) + Location (random) + Job Title (random) + YOE (fixed) + Gender (fixed)

Same as previous, gender is significantly impactful on the amount of stock value receieved holding all else constant (chisq = 149, $p < 0.001$). With an even larger impact than base salary, being male adds an average of 8152 US relative to being female. A year of experience will only add 2359 US on average.

Bonus Model 4: Bonus ~ Company (random) + Location (random) + Job Title (random) + YOE (fixed) + Gender (fixed)

As expected, being male is statistically significant (chisq = 54, $p < 0.001$), with males making on average 1874 US more than females all else being held constant. Surprisingly, an year of experience only adds an average 733 US to the bonus.

Interactive Visualizations

The following are interactive visualizations best viewed on the website. They do not render in PDF.

Though we have suggestive evidence the industry trends show a gender bias in terms of compensation and roles taken, purely using aggregate loses information. The following interactive visualizations allow one to check gender ratios and total compensation by company or location. Outliers could highlight that not every company or location follows the same trend.

Interactive Visualizations of Gender Ratio

By Company

By company, we note the ratio of employees being male (relative to female only) is significantly skewed, averaging 0.8. Out of the top 100 most reported companies, no company is female majority.

For PDF Report: Refer to Figure 3 on the interactive website for the above visualization.

By Location

By location, we note the ratio of employees being male (relative to female only) is significantly skewed, averaging 0.8 again. Out of the top 100 most reported locations, no locations are female majority.

For PDF Report: Refer to Figure 4 on the interactive website for the above visualization.

Interactive Visualization of Total Compensation by Gender

By Company

This plot shows that not every company follows the gender trend in terms of compensation.

For PDF Report: Refer to Figure 6 on the interactive website for the above visualization.

By Location

This plot shows that not every location follows the gender trend in terms of compensation.

For PDF Report: Refer to Figure 6 on the interactive website for the above visualization.

Discussion and Conclusions

The goal of this investigation was to (1) determine gender ratios and roles and (2) determine any gender bias in self-reported compensation data. The main concrete finding is that there is significant gender imbalance between men and women in the tech industry. Whether this is due to women being less likely to report their compensation data publicly (though anonymously) or due to a suggestive likeliness of gender imblaance in the industry as a whole, more comprehensive data is required. The pattern persists when we looked at job titles: women were more likely to report job titles other than "Software Engineer" compared to men, even if marginally. This does not necessarily mean women are taking on less senior roles than men, as we see an inflation in management type roles too.

To address the second research question. Based on this self-reported data, there is significant evidence to suggest that women are making less than men. Even taking into account company, location, job title, and years of experience, in all aspects of compensation, being a man means on average more compensation. In particular, stock grant value given to women are much less than their male equivalents, standing out even among base salary and bonus compensation.

Limitations and Future Work

1) As described multiple times through this report, the data used in this analysis is not very reliable, but it is the best that is publicly available to analyze. The levels.fyi database is quite comprehensive, though because it is self-reported there is significantly missing data in terms of gender, and often mistakes in data entry. There are also many confounding factors in self-reported data, and as a result we cannot conclude with very strong absolute evidence gender bias, though we can say there is gender bias in self-reported data at the very least. Future work, especially concerning this field of human resources, should instead choose to pay for the more details levels.fyi curated database, though it would still be limited by the nature of self-reported data.

2) There is an underreprsentation of non-binary genders in many aspects of society, and this is not an exception. Due to the very small numbers of those who did not report male/female genders (outside of those who chose to abstain from gender reporting entirely), this group had to be removed from modelling to ensure analysis stability. Future work can focus on underrepresented gender groups in a more meaningful fashion.

3) This data is collected over a period of at least two years, which indicates time dependence. With inflation and ever-increasing compensation, simply averaging over multiple years of data can lead to problematic results. On the other hand, the publicly available portion of these databases do not have enough data points to model the increase over time. Furthermore, this means the results of this data will generalize increasingly less as we move forward in time, due to the aforementioned factors. A more updated analysis to keep into account inflation (not just economical, but compensation increases due to industry demand) should be in order to make this analysis more robust.

Reproducibility

This document is fully reproducibile unless the hosted JSON data source is removed. I chose to upload the extracted CSV files publicly since they are not mine to distribute. Refer to the `index.Rmd` file (which generates this report!) in the repository for my code and comments.

Acknowledgements & References

Some content and figures in this report was adapted from the midterm report for this course, and from the visualization website created in assignment 5, both of which use lab materials by the JSC370 teaching team.

You can find the midterm report here (https://github.com/zw123han/JSC370_final) and the assignment 5 website here (https://zw123han.github.io/hw5_website/).

Refer to `data/README.md` in the repo (https://github.com/zw123han/JSC370_final/blob/master/data/README.md) for references and source of the data.

[1] levels.fyi. (2022). Salary Data (updated 2021). https://www.levels.fyi/js/salaryData.json. Retrieved April 21, 2022.

[2] Grierson, M. (2020, January 8). https://towardsdatascience.com/a-beginners-guide-to-grabbing-and-analyzing-salary-data-in-python-e8c60eab186e. Retrieved April 21, 2022.