JSC370 Midterm Project: Modelling Compensation in the Software Industry

Ziwen Han

2022-03-10

Introduction

The worst kept secret in the current software industry is that it pays well. If you ever browse an online group focused on computer science careers, one of the most discussed topics is total compensation (TC) (and arguably becomes an unhealthy obsession at times). However, as with any information, as it spreads so does the noise in the data increase, and as a result hearsay can dominate over statistics. As students of data science, it would be ironic if we relied on such weak data points without statistics. My high level goal with this analysis is to quantify and summarize these compensation metrics for software professionals to make informed decisions and stay up-to-date on current trends of competitive compensation. In addition, we hope to discover and highlight trends in the industry compensation package that may not be observed at first glance.

While there are multiple sites which allow users to anonymously report their total compensation, one of the more heavily relied on ones in the software industry is levels.fyi [1]. Mainly, it allows users to submit responses on a company, their current experience, and exact offered compensation (this is different from sites such as Glassdoor, which often gives an unreliable range). However, while the site gives an easy interface for navigating individual reports, it lacks more in-depth statistical analysis which could reveal underlying trends beyond just a well-known name and big number.
I will be using a version of the dataset from levels.fyi today to help answer these general preliminary research questions.

RQ1: What are some of the main trends in reported compensation in the current software industry?
RQ2: What is a competitive compensation for software professionals at varying levels?

Methods

1. Getting The Data We first get the raw data from levels.fyi using a standard JSON request, which includes entries until August 2021 [1]. An API example, which has been adapted here, was provided in a blog I found [2]. Note: The site provides a curated and updated spreadsheet of more variables, but it costs three figures to obtain. Overkill for a class project, but given dynamic trends in industry anyone using the analysis to make business decisions should consider using the curated version.

A python embedding is used to extract the data (cleaner code representation), which is subsequently loaded into R.

```
import pandas as pd
import requests
pd.DataFrame(requests.get('https://www.levels.fyi/js/salaryData.json').json()).to_csv('salary_data.csv'
```

2. Cleaning and Transforming the Data We're mainly interested in metrics of compensation and other factors which could impact careers (eg. Company, Location, Years of Experience, Gender), which is what we will keep from the raw data.

However, due to the nature of self-reported data on a public site, there will be differences in how people choose to define certain variables. There are also mistakes in entering data. For example, compensation metrics are either reported in units of one US dollar or 1000 US dollars, and could be reported in different currencies. These are all considerations we have to make when cleaning the data.
Of course, since this is the internet, people could just report false data, so we must be cautious taking certain figures at face value.

All compensation metrics were adjusted to ensure they were reported per 1 US dollar (assuming US for everyone, since the vast majority of positions are in the US). Anyone who reports over 60 years at a particular company was removed, to avoid any misentered or false data. For the same reason, the top 99th and bottom 5th percentile of TC/base salary were removed. Any columns with missing/unreported TC or base salary were also removed, since it is the main variable of interest.

Analysis was carried out using the `tidyverse` and `lme4` packages in R.

Preliminary Results

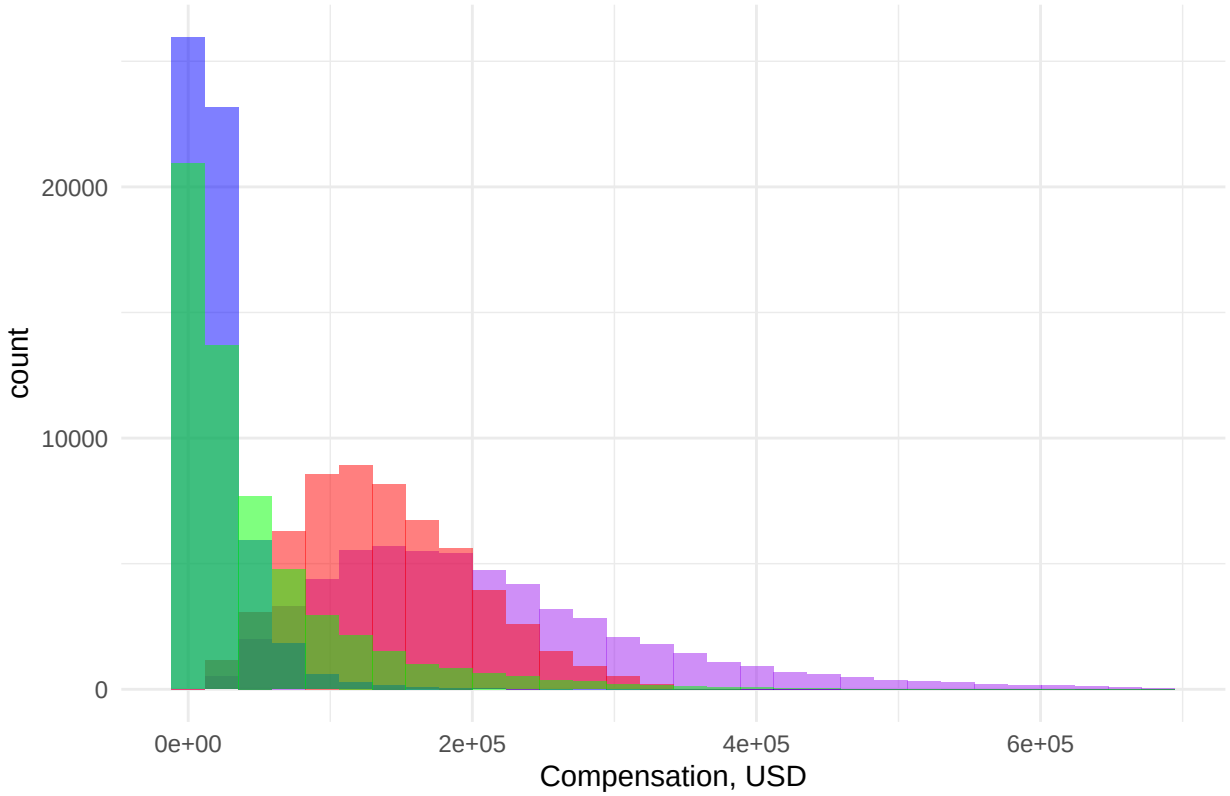
Exploratory Data Analysis We will first take an overview of the numeric variables in the dataset.

Table 1: A summary of the self-reported compensation and career variables in the software industry

Total Compensation (TC)	Years of Experience (YOE)	Years at Company	Base Salary	Stock Options	Bonuses
Min. : 25000	Min. : 0.000	Min. : 0.000	Min. : 22000	Min. : 0	Min. : 0
1st Qu.:125000	1st Qu.: 3.000	1st Qu.: 0.000	1st Qu.: 94000	1st Qu.: 1000	1st Qu.: 3000
Median :186000	Median : 5.000	Median : 1.000	Median :133000	Median : 25000	Median : 14000
Mean :207266	Mean : 7.093	Mean : 2.664	Mean :139499	Mean : 48477	Mean : 19515
3rd Qu.:264000	3rd Qu.:10.000	3rd Qu.: 4.000	3rd Qu.:179000	3rd Qu.: 64000	3rd Qu.: 27000
Max. :683000	Max. :58.000	Max. :47.000	Max. :336000	Max. :646000	Max. :472000

Of note, the median reporeted TC is 186k US, split base salary, stock options, and bonuses. The high amount here suggests a skewness of the reported data, as those who make less are less likely to self-report. Going on the site itself, you will notice that most reports are for companies with known higher compensations.
Another interesting observation is that although the median YOE is 5 years (with a mean of 7) in this self-reporeted data, the median number of years at a company is only 1 year (or a mean of 2.6 years). Although the YOE suggests that most observations are mid-career, the years at a company suggest that those in the profession often change companies during their career. However, our data lacks the reason for this change, whether it is simply due to higher compensation, work life balance, or other factors.

Figure 1: Breakdown of Compensation Distributon by Type



TC is represented in purple. Blue represents bonuses, green represents stock value, and red represents base salary.

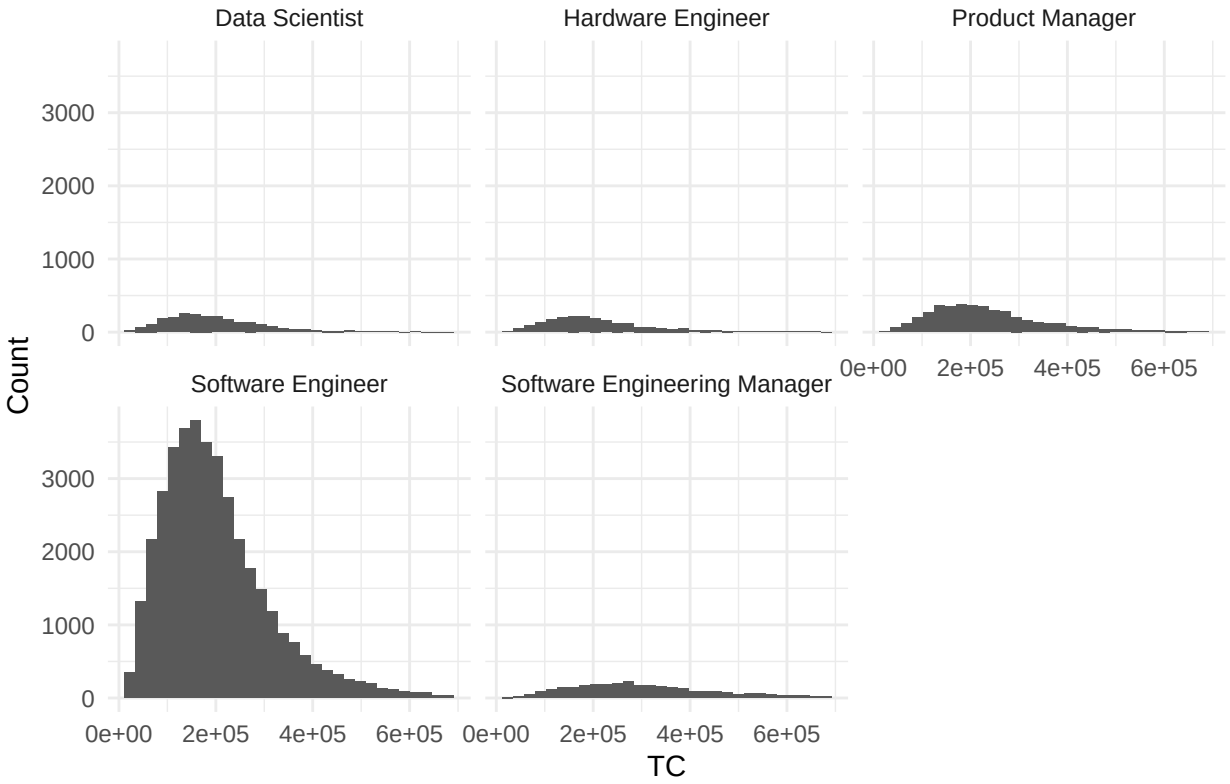
We observe that all types of compensation are skewed towards higher values. Of note, the majority of reported bonuses and stocks are close to 0, while base salary has a more symmetrical shape than TC. This indicates that the TC data is skewed by bonus and stock value.

Next, we will look at how the total compensation within certain categorical variables differs.

Table 2: Top 10 most commonly reported job titles with average compensation

Job Title	Count	Average TC	SD
Software Engineer	38314	200782.0	110843.63
Product Manager	4230	236320.2	119135.73
Software Engineering Manager	3020	301014.7	142808.63
Data Scientist	2401	198175.7	102857.70
Hardware Engineer	2124	211635.3	109859.03
Product Designer	1430	202327.6	104231.57
Technical Program Manager	1355	230341.7	104084.70
Solution Architect	1083	210733.5	98935.78
Management Consultant	955	158103.7	85548.07
Business Analyst	857	131486.6	73033.20

Figure 2: TC Distribution by Job Title (Top 5)



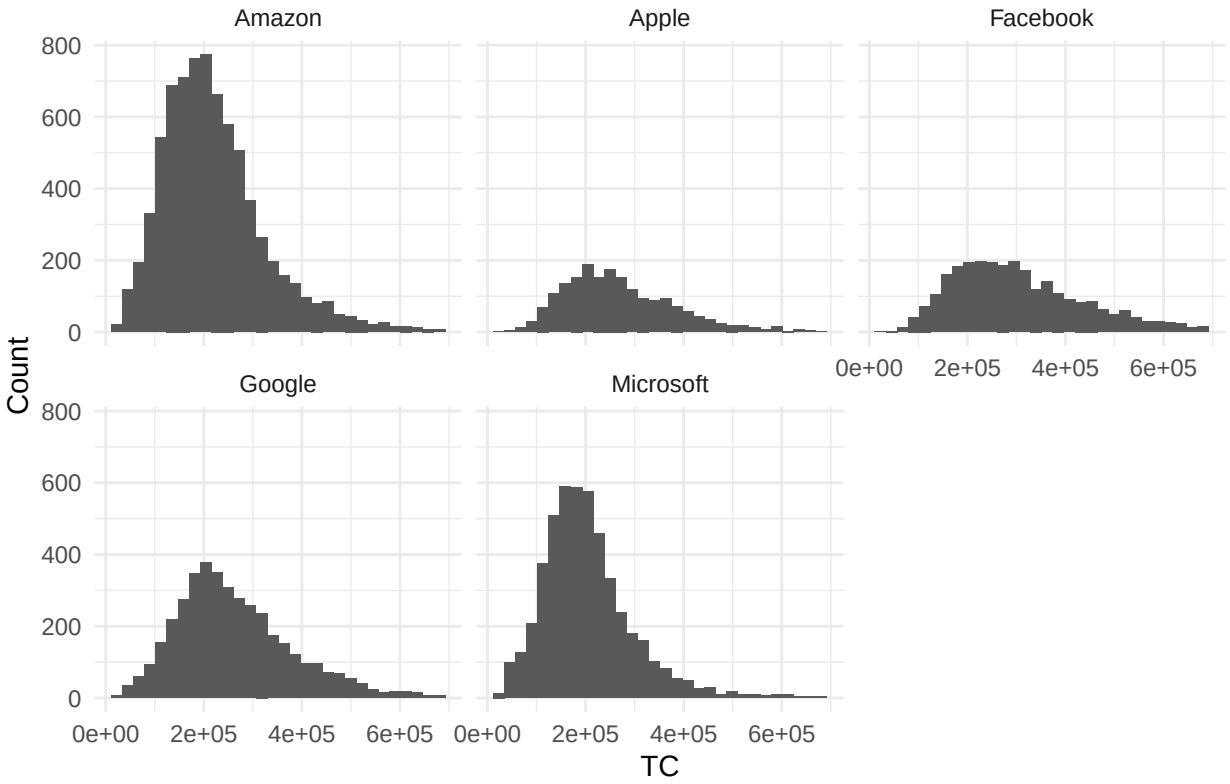
Unsurprisingly, majority of the reported data are from software engineers, with sparse data for even the other top 4 roles. This suggests that to avoid biases due to unevenly distributed data, future analysis may only want to focus on software engineers.

Table 3: Top 10 most commonly reported companies with average compensation

Company Name	Count	Average TC	SD
Amazon	7518	220620.3	104849.08
Microsoft	4886	203751.9	92844.10
Google	3982	263234.6	117330.44
Facebook	2707	304458.8	131392.72
Apple	1905	269217.5	112334.45
Oracle	1024	212715.7	114628.25
Salesforce	948	241167.5	108906.85
Intel	894	178475.9	87469.40
Cisco	845	197395.2	99216.99
IBM	818	137129.2	68331.63

Again, unsurprisingly, majority of the reported data comes from large established software companies. However, this distribution is not representative of the software/tech industry itself, but rather of those who choose to report their compensation.

Figure 3: TC Distribution by Company (Top 5)



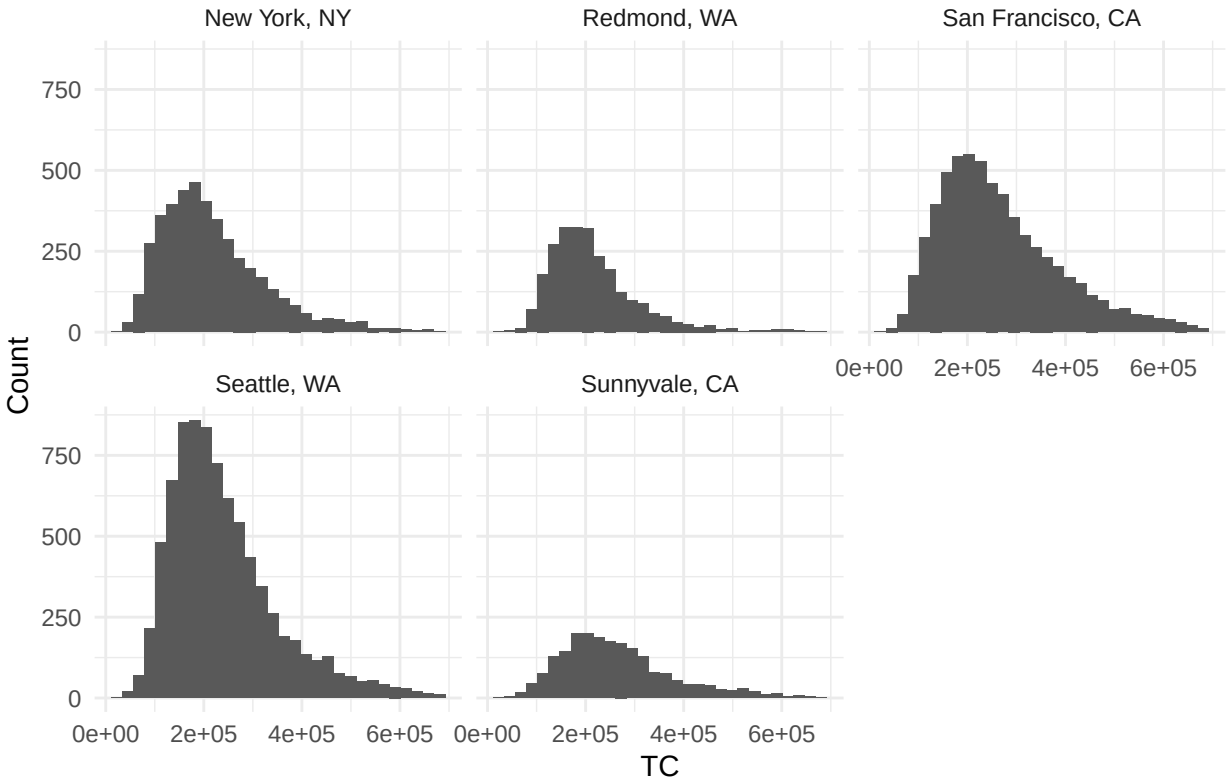
This highlights that each company has a different compensation policy and corporate structure, based on self-reported data. Companies such as Amazon and Microsoft have much more employees sitting at around the average compensation, while companies such as Apple and Google have fewer reported at this compensation. This could be due to the number of entry-level employees each company has, or simply because employees at certain companies are more motivated to share compensation packages.
The trend suggests that modelling compensation within a company could be of interest.

Table 4: Top 10 most commonly reported locations with average compensation

Location	Count	Average TC	SD
Seattle, WA	8073	240530.9	109906.97
San Francisco, CA	6209	266396.4	122704.64
New York, NY	4337	219613.1	107795.79
Redmond, WA	2513	216296.3	92758.21
Sunnyvale, CA	2103	264623.2	114769.50
Mountain View, CA	2079	275008.7	114508.39
San Jose, CA	1920	228440.2	108134.48
Austin, TX	1479	179294.8	87345.39
Cupertino, CA	1350	277325.1	111198.40
Menlo Park, CA	1282	317743.6	134975.11

There is an over-representation in the reported data from certain locations, with 8 of the top 10 being on the west coast. There is little data from elsewhere, which makes it difficult to generalize the results of this analysis for those who are not in one of the tech hubs.

Figure 4: TC Distribution by Location (Top 5)



There is likely substantial overlap between these distributions and TC distribution in the most commonly reported companies, since many of these cities are company headquarters (eg. Microsoft in Redmond, Amazon in Seattle). This partially explains similar patterns observed, for example the peak at around 200k in Seattle is similar to the peak observed in the TC distribution of Amazon. This highlights analysis in either the direction of company or location could yield similar results.

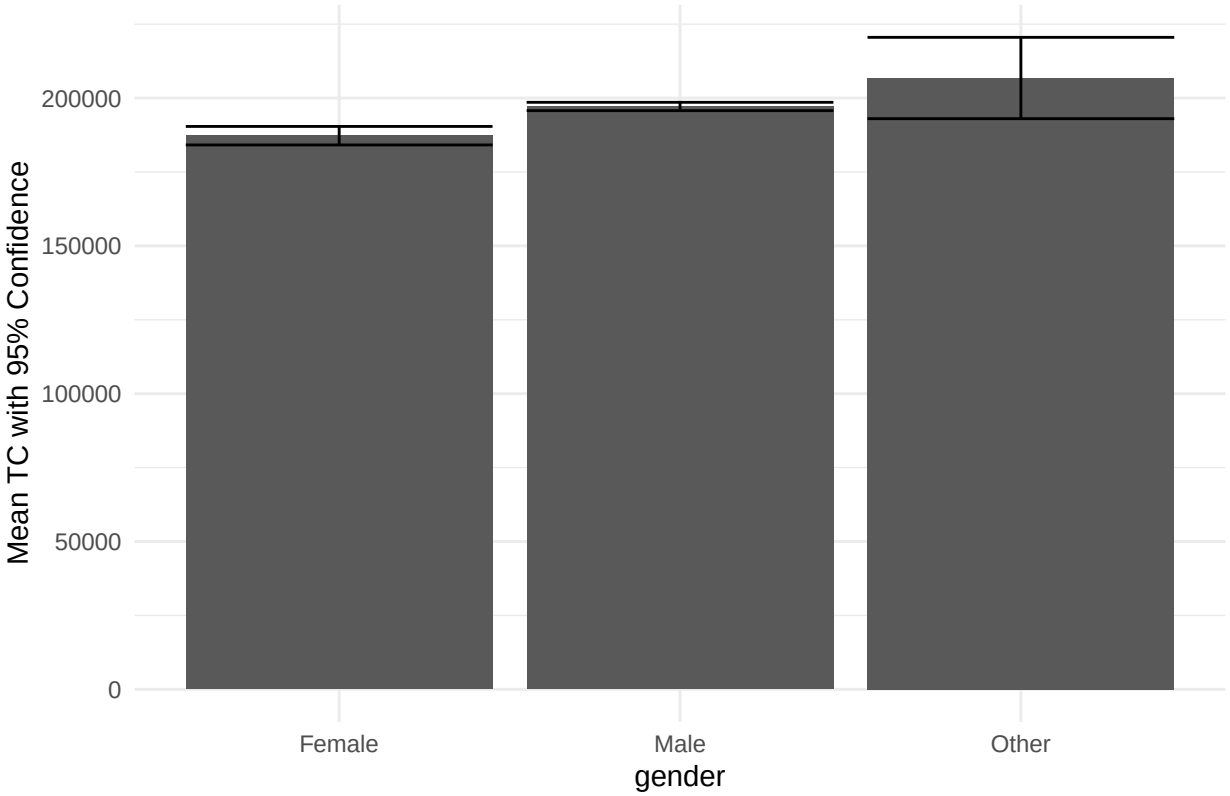
The Role of Gender in TC In many industries, an unfortunate trend is that equally qualified women make less than men. We are interested in comparing the total compensation between genders for Software Engineers (SWEs). This is a standard title in the industry, and as shown above accounts for the majority of reported job titles. This should also remove any skewnness induced by executive or other uncommon position titles.

Table 5: Table 5: Average TC and counts by reported gender for software engineers.

Reported Gender	Count	Average TC	SD
Female	3436	187268.7	93586.96
Male	23630	197156.0	110634.55
Other	248	206786.3	109984.57

We note that women and other genders are underrepresented at least in the self-reported data, though most sources suggest that there is a vast underrepresentation in software industries as a whole. Curiously, the variation in TC is lower than other genders, which could be due to sparse data.

Figure 5: Average TC for SWEs by Gender



This bar plot shows some evidence that female SWEs make less than those who identify as male and others in self-reported compensation data.

Mixed Effect Modelling To quantify the role of gender in TC for SWEs, a mixed-effects model is used, setting gender as the fixed effects of interest, and controlling for company and locations. Any observations which did not self-report gender were removed.

Two models were constructed to test for the significance of gender in influencing TC. The formulas are stated in the format of the `lmer` functon from the `lme4` package.

Model 1: TC ~ Company + Location, both as random effects
`formula = totalyearlycompensation ~ (1|company) + (1|location)`

Model 2: TC ~ Gender + Company + Location, Gender is a fixed effect
`formula = totalyearlycompensation ~ as.factor(gender) + (1|company) + (1|location)`

A likelihood ratio test was used to test whether gender gives better explanability of the variation in TC compared to not including gender. H_0: Gender has no significant effect on TC H_1: Gender has a significant on TC
`anova(mixed_mod_no_gender, mixed_mod_gender, test = "LRT")`
Chi-sq statistic: 204.75, df=2, $p < .001$

```
summary(mixed_mod_gender)
```

The preliminary results show that it is highly unlikely ($p<.001$) gender does not play a role in the TC a Software Engineer receives, across 1296 companies and 739 locations. Setting female as the baseline, the estimated change in TC is for males is 22050 US and 18635 US for other.

Preliminary Discussion

There is suggestive evidence on average for Software Engineers, those who self-reported gender as Male make 22k US more, and those who self-report as other make 18.6k US more than females. However, we did not account for levels within a company, since there is no fixed definition between organizations. This means that we cannot differentiate whether there is a gender difference in total compensation in the same position, or whether those who identify as Male or Other hold higher positions. In addition, we do not differentiate between company and location interactions, which means we cannot account for gender influence within specific companies; the average difference is moreso reflected in the industry in general.
The analysis should be extended by accounting for non-independent assumptions in certain clusters. As demonstrated in earlier TC distributions, certain companies and locations will give out similar packages (eg. Amazon at 200k), which could potentially violate the independence assumption of any effect models.

The results from this preliminary analysis has demonstrated that there is substantial skewness in the reported total compensation data, including job titles, company, and location. Although the original goal was to model compensation in the software industry, it should instead be refined to focus on only tech centers or companies with many self-reported data points, since otherwise the data is unrepresentative of the overall industry.

Limitations As with any analysis, we are limited by the scope of our data, which will induce biases in our conclusions. Although levels.fyi has a substantial database in the industry often used to benchmark fair compensation, it is still limited by the voluntary nature of self-reporting, even if anonymized. This means that as we have observed, there is over representation of reported data points from certain locations, companies, titles, and levels of compensation in contrast with the all the positions across the US and Canada.

We are also limited by the nature of compensation in the industry, which is ever-increasing. Since this data spans over 4 years, including the pandemic, inflation and competition between corporations will undoubtedly change compensation between one time point and the next, even if for otherwise identical positions.

Conclusion and Summary

The goal of this preliminary analysis was to uncover trends in the compensation packages of professionals in the software industry, in order to shed light on industry patterns that may not be visible without aggregate. What was uncovered was under-representation of certain groups in self-reporting compensation data, in addition to evidence of a disparity between genders of Software Engineers. Future analysis will likely focus upon thsi gender difference in compensation, in addition to focusing more on the abundant clusters in the self-reported data.

References

[1] levels.fyi. (2022). Salary Data (updated 2021). https://www.levels.fyi/js/salaryData.json. Retrieved March 10, 2022.

[2] Grierson, M. (2020, January 8). https://towardsdatascience.com/a-beginners-guide-to-grabbing-and-analyzing-salary-data-in-python-e8c60eab186e. Retrieved March 10, 2022.