Who Makes the Big Bucks in the United States? A linear regression analysis on salaries of Stack Overflow Users in 2021

Anna Dai

December 12, 2021

Summary

While many organizations in the United States have actively implemented diversity, equity, and inclusion (DEI) programs in the past decade thanks to the spotlight that several social movements, such as #metoo and Black Lives Matter, have shone on the embedded injustices in modern America, the programs' effects appear to not always penetrate through to leadership positions of these organizations. In this report, we leverage results from the 2021 Stack Overflow Annual Developer Survey and use a hierarchical linear regression model to investigate primarily how different demographic variables amongst other characteristics are associated with receiving higher compensation in the U.S. technological sector. TODO: INCLUDE FINDINGS/CONCLUSIONS

This study explores factors associated with the prices of methadone on the black market with a multilevel linear regression model. The response variable in our study is the price per milligram (ppm) variable. In combination, the linear regression models and the stepwise regression process suggested that `mgstr`*, source, and bulk_purchase as predictor variables. USA_regions and state are suitable as hierarchical predictor variables. Our final model contains mgstr, source, and bulk_purchase as fixed effects, and USA_regions and state as random intercept variables. The fixed effects suggested that there is a tendency for lower prices with higher dosage strength as well as with bulk purchase. The hierarchical variables suggested that methadone prices could vary by region within the United States. For example, the price per milligram (ppm) in the South tends to be higher than in other regions. There are three significant states: California, Arizona, and Tennessee, in which California and Arizona seem to have the cheapest methadone prices, while Tennessee is paying significantly more compared to all the other states.*

Introduction

Diversity, equity, and inclusion efforts have gained traction within the last decade, but the results of such efforts are not always reflected in or visible from leadership roles of Corporate America. Unfortunately, personnel data are well guarded secrets of each company, so such information can be challenging to obtain for analysis. Stack Overflow is by far the most widely-used, community-based platforms where the software programmers from around the world post their coding questions and answers with over 16 million registered users in 2021 (https://stackexchange.com/sites?view=list#users). Each year, Stack Overflow surveys its users to improve the community and platform and in turn provides valuable insights into the technology sector. In this report, we intend to leverage the self-reported survey responses in 2021 to identify which factors are associated with receiving higher compensation in the technology sector of the United States in an attempt to shed light on whether demographic characteristics, such as ethnicity and gender, influence one's annual earnings.

Data

Data Preprocessing

The raw survey data contains XX variables and responses from XX users, spanning from all around the world. To start, we imported only the data that are interesting to us. We then filter for "I am a developer by profession" and "Employed full-time" and location to "United States" in order to focus the study on employed professionals in the country. which elimates the data set to only 10,000.

Then we evaluate the compensation variables, which span over four columns in the data set: (TODO: NAME THEM). Upon evaluation, we consoidated them into one response variable: compyearly, which Stack Overflow had calculated for us using `pay per period* periods * currency`. I aslo found from [SRC] that the reasonable salaries for professional deveoper to be range(). However, looking at the distribution of the variable, we can see that extremely biased. We notice some who had chosen a reasonable US salary in UGX, who we then corrected inthe data. In addition Ievaluated the distirbution of the variable [insert plot] and carefully examined the bottom and top tails. For the bottom 0.25% of the data, the salaries were under [insert lower bound], which appears to be too low for full-time employed developers inthe united States, so we remove them, since it could not be a calculation error when any calculation error (i.e. pay per year inserted as per period times period would result in only lower salary) so we remove them. Regarding the extreme left skew of the data with a max of $21,822,250, which is equivalent to the Apple C-Suite's salaries (SRC), it is evident that something is off. So, I take a close examination of the top 2.5% of the dat, which anyone earning more than [insert upper bound]. This alone is already 2x the higher bound of the reaosnable range per [SRC]. of the highest 5%, I looked at those who selected a non-yearly pay period and I'm convinced that it is more likely for them to have made a mistake in interpreting the question and inputting their total yearly comp as the per period comp. So we correct it and recalculat thier comp. This helped remove some majoe outliwers but still very right skrwed. This we try a log transform on the response variable.

Next we move to fix the multitude of issues on the predictor variables. We encounter primarily 3 types of issues: single select, multi select, etc.

Missing Data Imputations

With cleaned data, we proceed with missing data imputations. I have the following distribution of missing data [insert missing data percentages?] or ranging from 0.1 to 13% for my response variable. I employ multiple imputations using the norm and _ ? in order to perform the imputaitons. Due to the fact that States are unable to be imputed because it has too many levels for the number of observations we have, we manually replace N.As with missing state? and remove the variable from the imputation process. I try pmm as well as norm methods and decided that norm made the most sense as it employs Beysian methodology, when the represtation of real data of imputed is roughly the same. [insert imputation evaluation plots] [PLOOOOOTS] With my imputations complete, I evaluate such imputed results.

Exploratory Data Analysis

I want to log transform my response varialble We take state as our hierarchical method to help...

Even after data cleaning the distributions of rental and sales prices are still skewed. This might lead to problems when fitting linear models, but log transforming could be a potential remedy to this issue in the modeling phase. Moreover, the univariate distribution of the number of rooms (Appendix A) reveals that properties for sale tend to have more rooms and that rental properties have more missing data in this variable than properties for sale.

The original factor variable source has high cardinality with few cases in certain factor levels. Therefore, we decided to group some levels to have a clearer picture in the exploratory data analysis. All internetbased

sources, such as the different URLs, “Internet Pharmacy”, and “Google” are grouped into a single level named “Internet”, and values such as “None” and “N/A” are grouped into the “No Input” category. Moreover, all entries with missing ppm are removed, which also eliminated rows with missing mgstr values. The variable mgstr has six unique values and numeric data types. After an initial inspection, the mgstr variable is transformed to a factor variable, and 1mg, 2.5mg, and 15mg cases are filtered as they have only one or two entries per value, leaving only 5mg, 10mg, and 40mg. Usually Methadone pills come in 5mg or 10mg doses, while the 40mg pills are not FDA-approved and thus only appear on the black market.

Model

Conclusion

References

Link to GitHub Repo