IDS 702: Final Project Proposal

Peining Yang

10/25/2021

Overview & Research Questions

In this project, I will analyze data regarding data science and STEM job salaries. As Data Science students, I think this project would generate interesting insights to what we can expect post graduation as we begin our careers. *Levels.fyi* is a website that allows you to compare career levels and compensations across different companies. We will focus on the technology related positions. The goal of this study is to determine the factors associated with a higher baseline salary in the technology field. In addition, I will use a multi-level model to account for potential clustering in location and explore the variation across state or regions. If time permits, I will also explore other response variables such as bonus or stock grand value.

Data

The data for this project contains entries of salary records from top tech companies. The dataset was scraped from the *levels.fyi* website then cleaned up and was uploaded to kaggle.
(The data can be found here: https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries)
The original dataset contains 62,642 observations and 29 variables. The variables describe the person's base salary, bonus, stock grant value, employment company, location, position title, amount of time working for the company, level of education, race, and gender. After loading the data, I've discovered that some observations contains missing values (e.g. 35.8% of the Race variable are missing). This means that before I even begin my analysis, I will need to investigate this issue more and determine the reason of missingness and whether they are viable ways to impute these values. If there isn't, I will need to decide whether removing these variables entirely would be better solution. The *level* variable represents the level of the position that are unique to a specific company. There is no way to generate useful results universally with this variable. Therefore, I will look into whether I should subset for a specific number of companies in order to better utilize this variable.

The dataset contains locations in the form of city and state as a single variable (e.g. Seattle, Washington). Since I plan on exploring the heterogeneity in salary by location, I will have to create separate variables for state and region, in addition to taking into account for international locations.

Project Plan

I anticipate using a hierarchical linear regression model for the analysis. The repsonse variable, baseline salary (and possibly also bonus), are continuous variables. Given that we want to explore variation across locations, we will also explore whether it is statistically significant to include varying intercepts or slopes by state or region. Since the video presentation is due November 22nd, I aim to have all the analysis completed by latest November 20th. Some milestones for this project would be data wrangling, exploratory data analysis, model fitting and evaluations, and report writing. In the next week, I plan on completing the data wrangling portion to have a cleaned dataset ready to work with. In the following week and a half, I will be mostly working on analysis and model fitting, with the remainder days set for completing the report.