

gender pay gap analysis

DSC 530 Project Summary



November 19, 2021

Subhashini Natarajan

Statistical Question:

The statistical question that I chose for the project is to see if gender pay gap exists in IT industry. I am curious to understand why there would be such a gap in IT industry, as companies look for experience and skillset to determine the compensation. Data source for the project is the salary survey conducted among IT employees in Europe. This dataset is obtained from Kaggle.

Outcome:

EDA performed on the data set includes CDF comparison, Pareto analysis, Simple linear regression, Multiple linear regression and T-Test. All the analysis consistently indicates that gender pay gap exists in IT industry and female employees receive lower salary than male employees.

Additional helpful dataset:

Additional variables to identify if career break was in place due to health/personal issues or break taken after child birth could have helped understand the results. Employees coming back to work after a break tend to settle down for lesser pay as compared to the market standard. Market standard pay is another variable that could have helped in the comparison.

Another observation in the dataset is the number of female employees in the data as compared to number of male employees. The number of female employees is only one fifth of the number of male employees. For the purpose of analysis, I assumed that the number is the true representation of the work force.

Missed part in analysis:

It would be even more insightful if the comparison was performed within the role as well. However, due the number of female employees represented in the survey is less and the data gets too fragmented for role wise comparison, I couldn’t perform that.

Incorrect assumptions:

I did not observe any assumptions becoming wrong in the analysis.

Challenges:

Gender being a categorical variable and not numeric, I wasn’t sure which analytic distribution or hypothesis will be appropriate. With some research found that T-Test can be used for a categorical predictor and continuous dependent variable.

Another challenge was in the kernel, that I was executing, the recursive length was minimal and continuously faced recursive error. It took some time to understand and resolve the issue. I set the recursion length to 20,000 and the error didn’t reoccur.