DSC 530

Term Project

Brian Mann

This research project focused on assessing the wage gap between men and women in the USA. The data was sourced from the Economic Policy Institute’s State of Working America Data Library, which provided the average hourly wage for workers from 1973-2022. Every wage represents the average hourly wage for workers in that year converted to the value of 2022 US Dollars. We looked at how the wage gap has changed over time, differences by education level and other trends. Our main hypothesis was that there has been a statistically significant decline in the gap between the wages of men and women over time.

In our analysis, we assessed five target variables, each representing the highest level of education gained by that respective group - less than high school, high school, some college, a bachelor’s degree and an advanced degree. For each of these five variables, we also categorized them by gender, as well as subtracting the data sets of men and women to assess the wage gap itself. We then used a Jupyter notebook and python code to perform an EDA on the data by creating histograms, PMFs, CDFs, line charts, scatter plots and regression lines to help assess the wage gap.

After performing our analysis, we reached the conclusion that although there has been a significant decline in the wage gap for those with a highest education of some college or less, there has been a significant increase in the wage gap for advanced degree holders. For advanced degree holders, the average gap between the wages of men and women has grown from only $7 to nearly $20 in the past 50 years. No significant change was noted in the gap for those with just a bachelor’s degree.

During the analysis, we noticed that there was a sharp decline in the wage gap during the 1980’s that then seemed to level out in the 1990’s. We felt that we may have missed an important shift that happened to US wages that would need to be explored further. Additionally, there were several other variables in the data set that could have helped in this analysis that were ultimately left out, namely race/ethnicity. Type of job and years of experience would also have helped considerably. The assumption was also made that converting to 2022 USD would be the easiest choice to visualize the data, although there may be discrepancies in how this would be calculated by different entities.

Much of the challenge for this project was not actually in coding or generating the right plots but rather in interpreting the results. Knowing when to use which analytical distribution, which statistical hypothesis test, which correlation test and which regression model was typically quite challenging. Actual data from the real world is not always so cut and dry, and this data does not always easily conform to a standard distribution.