**Exploratory Data Analysis: Gender Pay Gap**

Jahedur Rahman

Bellevue University

DSC 530

Dr. Shankar Parajulee

March 5, 2022

This EDA was done to find out if other variables impact income more than gender. The data set that was used was a decent size. Only six out of the nine variables were used to this analysis. There variables are Gender, Age, Education, Dept, Seniority, and BasePay. The hypothesis being tested was if the variable BasePay had some sort of relationship with any of the other variables. The results of the analysis showed relationship between BasePay and Age. In addition, the regression analysis showed a potential relationship between BasePay and Seniority as well.

The sample was split based on gender to compare differences in Education. The PMFs showed that males are more likely to go on and complete Post graduate degrees like master’s and PhD, while females are more likely to graduate with a high school and college degree. When generating the scatterplots, the sample was also split based on gender, however, this time it compares relationships between BasePay and Age for each gender. The scatterplots showed that there is a positive relationship between BasePay and Age. In addition, this separation based on gender helped in visualizing and comparing the range of BasePay for each gender. The CDF showed that most of the people earned between 60,000 and 120,000, which is about 70% of the sample. If this were to be analyzed further, it can show whether most of these were male or female. Additionally, the results of Pearson’s correlation showed that there is a moderate relationship between Age and BasePay. Also, the results after testing the difference in means shows that there is a significant difference between male BasePay and female BasePay.

The analysis did feel a little incomplete. I felt as if there were somethings that could’ve been done to improve on it even more. Generating more scatterplots that compare some of the other variables would be one improvement. In addition, some variables that could help in the analysis was a person’s race, and a person’s residing location. Additionally, an assumption that was made that I felt was incorrect was the Education Histogram. The education counts were almost equal. Usually, this wouldn’t be the case, however, I believe the person or group who made the data set did this on purpose so there will be minimum differences when comparing the income between each gender. Also, a challenge I face occurred when generating the histogram for BasePay. When using thinkplot.Hist it kept coming out wrong. After doing research I was able to find an alternative way for the histogram and it generated correctly.