INFO5502 - Assignment 4: Analyze the Income Gap between Female and Male Employees

Introduction

This assignment is based on the three datasets collected from Kaggle.com at: https://www.kaggle.com/kaggle/kaggle-survey-2018
An example notebook analyzing this data can be found at: https://www.kaggle.com/theoviel/kagglers-gender-pay-gap-salary-prediction You may refer to this website for basic explanation, cleaning, visualization, and analysis of the datasets. The three datasets are - **surveySchema.csv, freeFormResponses.csv,** and **multipleChoiceResponses.csv**.

Tasks

Complete the following tasks:

1. Calculate the median income of male employees and the median income of female employee in the population. Consider the set of all employees in the datasets as the population. (**1 point**)

2. Draw an overlaid graph to show the histograms of the incomes of female and male employees in the population. (You create one histogram for male, and another histogram for female, but the two histograms should be on display in the same graph with different colors. (**1 point**)

3. Use: random sampling, empirical distributions, sample comparisons, bootstrap, and hypothesis testing as well as A/B testing - that we discussed in the class - to analyze the income gap between female and male employees.

• Select a sample from the population. Make sure your sample include 500 employees selected from the population, and consider how to ensure the sampling strategy is fair since the datasets include an overwhelming number of male employees compared to female employees (**1 point**).

• Define the test statistic, the null hypothesis and the alternative hypothesis (**1 point**).

• Draw the income histogram for the sample; calculate the median income of the sample; and draw a red dot and a yellow dot for the female median income and male median income of the population respectively, in the histogram (**1 point**).

• Draw the histogram of the test statistic of the sample, and draw a red dot to show the corresponding test statistic of the population (e.g. the difference of the median incomes between female and male employees) in the diagram (**1 point**).

• Write a procedure to use bootstrap to produce at least 5000 samples (**1 point**).

• Draw the histogram of the test statistic of the bootstrap samples (**1 point**).

• Define the confidence interval and P-value to validate the hypothesis you defined (**2 points**).

4. Submit all your Python code; and in writing, explain the data cleaning procedure that explains how you defined the test statistic, the hypotheses, random sampling, bootstrap, confidential intervals, P-vales, as well as interpretation of your results, and all outputs described above.