# Statistical Analysis

**Code**:

Text

Description automatically generated

All the required libraries are imported.

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.**

**Code:**

Text

Description automatically generated

Here after cleaning the data, we calculated the median of male and female employees.

Result :

A picture containing text

Description automatically generated

1. **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.**

Code :

Graphical user interface, text, application

Description automatically generated

Result :

Chart, histogram

Description automatically generated

1. 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

Code :

Text

Description automatically generated

• **Define the test statistic, the null hypothesis, and the alternative hypothesis**

**Test Statistic**: Generally, Test Statistic is some random value which is used to validate Null Hypothesis, this value is obtained from given sample data. In order to validate the data, it compares data from given sample to expected null hypothesis data.

**Null Hypothesis:** Null Hypothesis is a hypothesis that can be nullified. Here, sample and population mean are considered as same.

**Alternate Hypothesis:** This is opposite to Null Hypothesis, as it rejects the decision given by data.

Text

Description automatically generated

**• 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**

**Code:**

Text

Description automatically generated

**Result:**

Chart, histogram

Description automatically generated

• **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**

Text

Description automatically generated

**Result**:

Chart, histogram

Description automatically generated

• **Write a procedure to use bootstrap to produce at least 5000 samples**

**Code:**

Graphical user interface, text, application

Description automatically generated

Here, by taking sample size n with replacement along with repetition for 5000 times we get a sampled data, this sampled data is Bootstrap sample data.

Now, by calculating mean of each sample then we get 5000 estimates of resampled samples.

• **Draw the histogram of the test statistic of the bootstrap samples**

**Code:**



Chart, histogram

Description automatically generated

**• Define the confidence interval and P-value to validate the hypothesis you defined**

Text

Description automatically generated