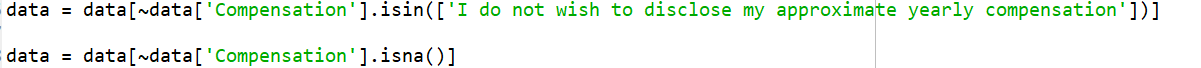
To work on this, I have used an open source dataset called “multipleChoiceResponses.csv” file from the Kaggle (<https://www.kaggle.com/supchanda08kol50/multiplechoiceresponses>). Attached is the Python code generated in Spyder, this code produces respective tasks mentioned below.

**TASK 1: Median Income of Male and Median Income of Female Employees.**

* As per the task requirement, I have read the CSV file and considered the Q1 and Q9 columns for values of “Gender” and “Compensation” respectively.
* The next step after considering required columns is “**Data Cleaning**”.

1. **Data Cleaning:**

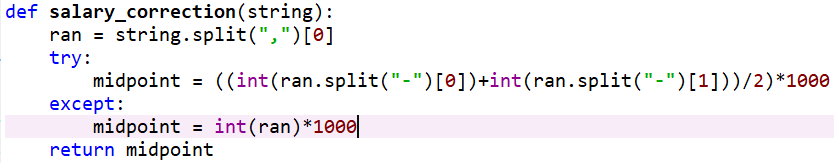
* Data cleaning should be done to get the accurate results by eliminating the unnecessary data which may impact the primary results like “Median of female/ Male” or “Ratio of Male and Female” and also to modify the existing data to make it readable (an example for this is getting the midpoints for the salary ranges).
* From Q9 column, I have deleted the missing rows/ observations and the rows with “I do not wish to disclose my approximate yearly compensation” using “Pandas .isin() and .isna()” to filter/ selection of particular rows from the data frame. Please find the below screenshot for reference.



* From Q1 column, I have deleted the rows with values as “Prefer not to say” and “Prefer to self-describe” using the same Pandas .isin().



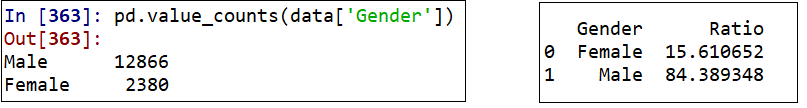
* And the last step is as we cannot process the range values to find the median, we should derive the mid points [(lower limit + upper limit)/2] from the salary ranges. To do this I have created a function to find the midpoints for the range salary values (ex: 10 – 20,000 will be mapped as 15,000) and to keep the integer values for the non-range values (ex: 500,000+ will be mapped as 500,000).



* The data is now clean and can be used for further analysis.

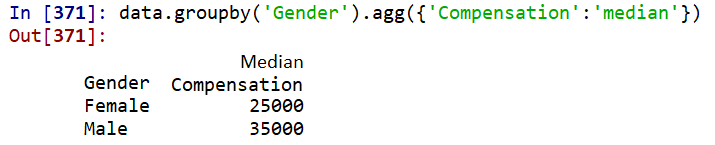
1. **Find the Female and Male ratio for the entire population which can be used further:**

* After the data is cleaned, there are total of 15246 observations. In which, Female to Male observations count and ratio can be found in below screenshot.
* The Male to Female ratio in the original population is nearly 84:16.



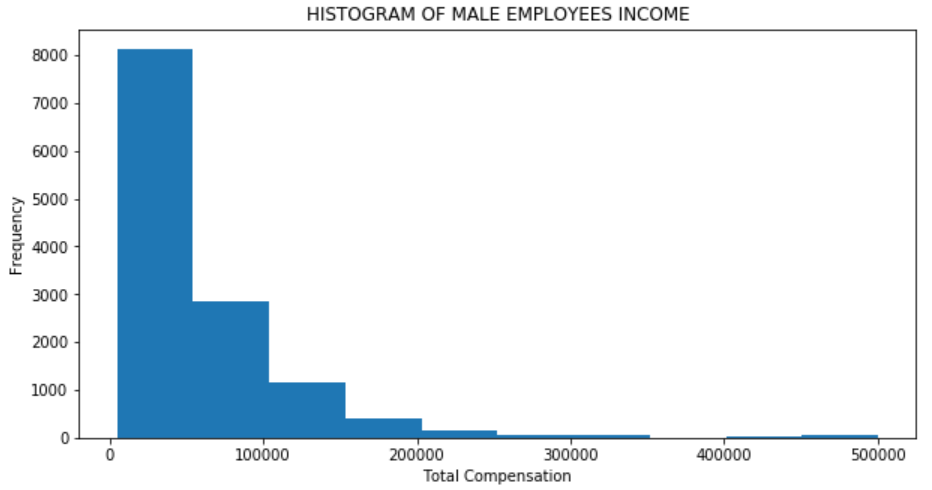
1. **Median Income of Male and Median Income of Female:**

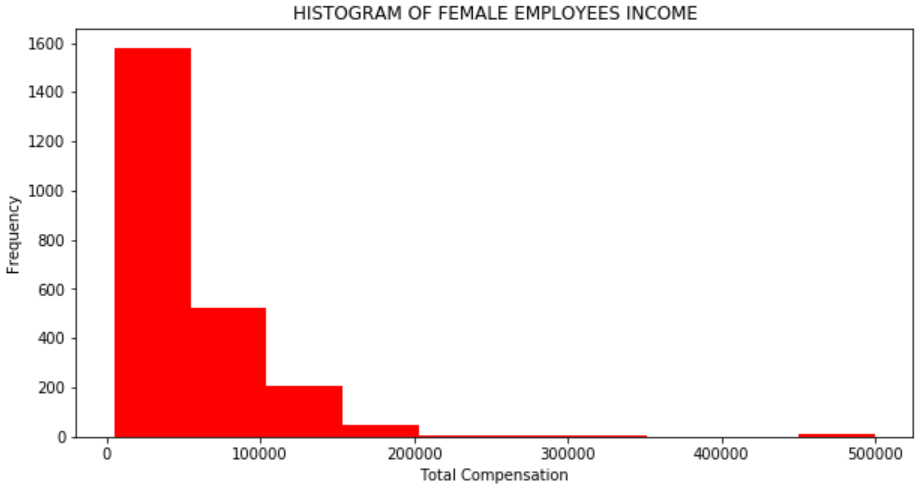
* From the cleaned data, the answer for the Task-1, the Medians for the gender group has been calculated and found as “25,000 $USD” for Female and “35,000 $USD” for Male.

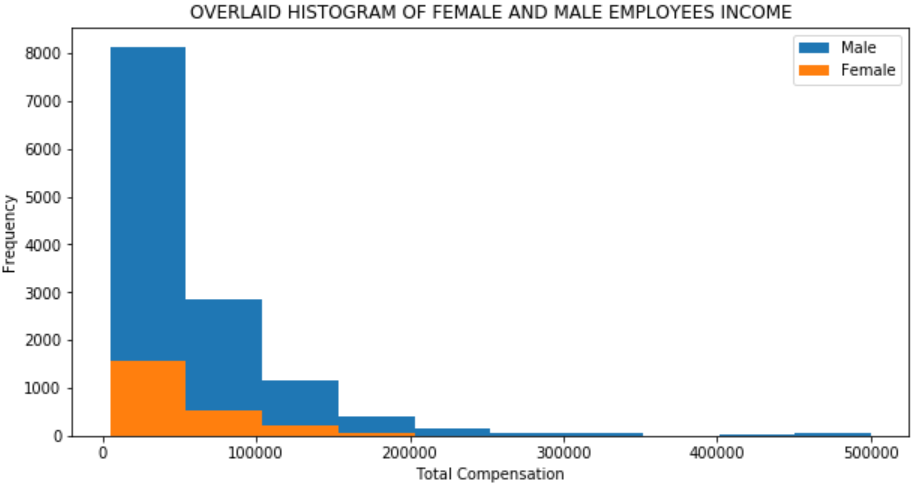


**TASK 2: Overlaid Histograms of the Incomes of Female and Male Employees in the Population.**

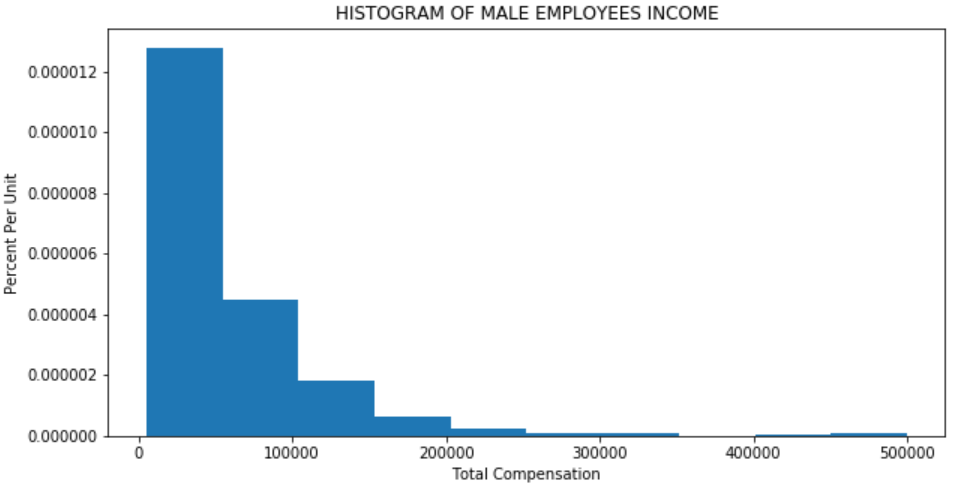
* I have generated two overlaid histograms for the female and male employees’ income from the population, one is using “Frequency” on the Y-axis and the second one is using “Percent per unit” on the Y-axis. Please find the reports below.

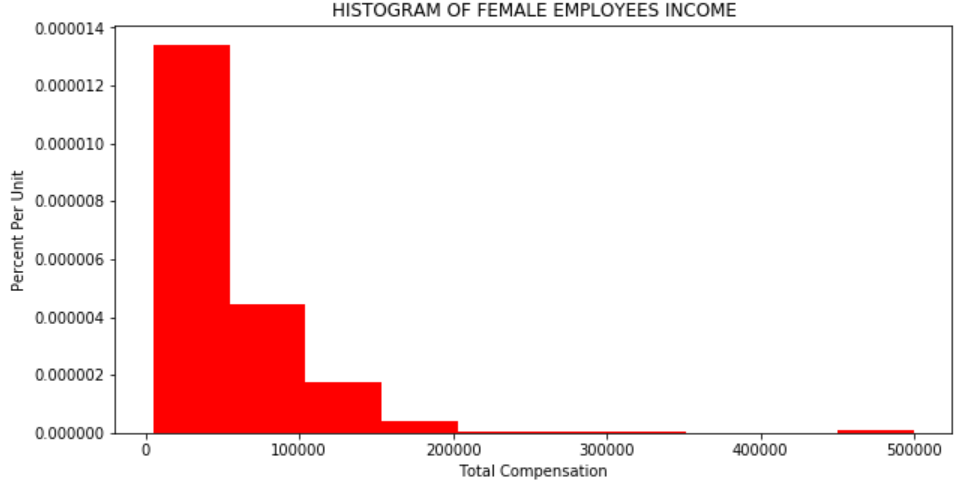
**Histograms with “Frequency” as Y-axis:** Below are the 3 histograms for Male, Female and Overlaid respectively.

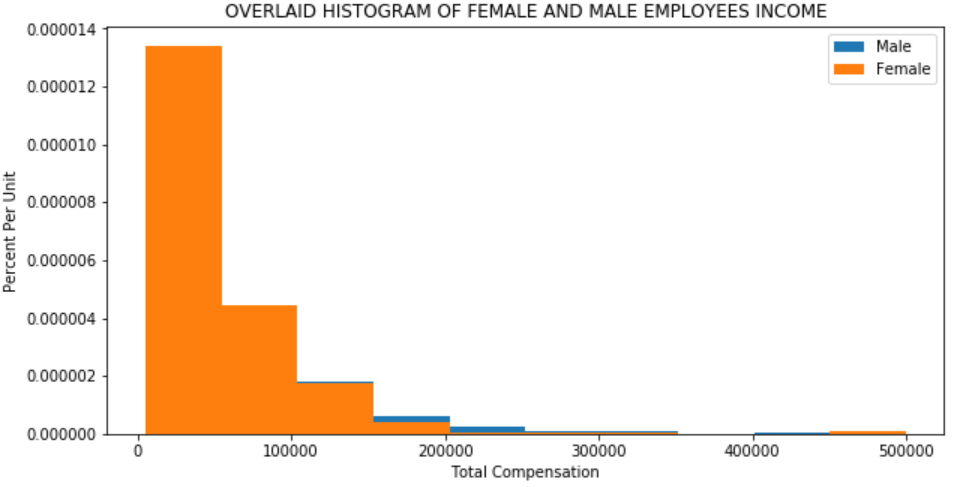




**Histogram with “Percent Per Unit” as Y-axis:** Below are the 3 histograms for Male, Female and Overlaid respectively.



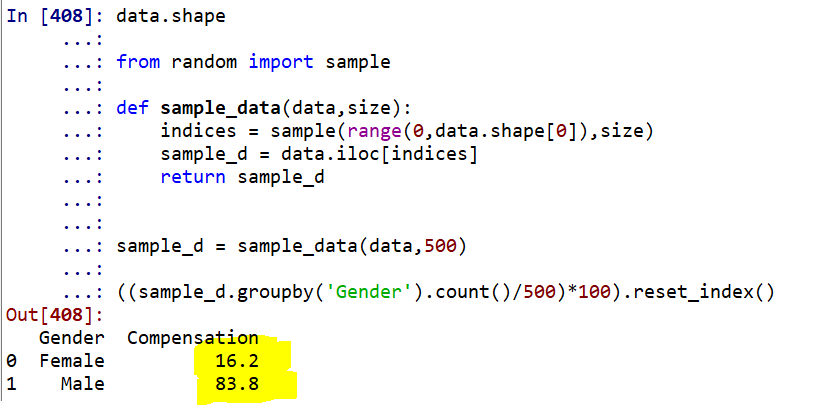




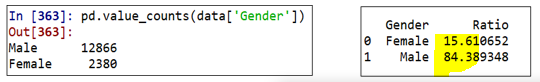
**TASK 3.1: Select a sample of 500 employees from the population and ensure the sampling strategy is fair since the datasets include overwhelmed male employees than female employees:**

* RANDOM SAMPLING: An important part of data science consists of making conclusions based on the data in random samples. **Random sampling is based on the probability, more chances will be given to the set which occurs more on the original population. Hence, the random samples are almost matches the original samples size/ ratio.**
* I have generated a sample of 500 employees without replacement using the sample(), an inbuilt function of random module and also using the shuffling technique.
* Once the 500 employees’ sample has generated, I have calculated the ratio, just to ensure if the selected 500 sample has the same ratio as we had for the original population.
* The ratio calculated for the 500 sample is, nearly 84 for Male and 16 for Female (please see yellow highlighted in screenshot below). This shows that the ratio of 500 sample is matching with the original population gender ratio [See Task1, point 2 screenshot (or) the second screenshot below].

**Gender ratio from sample of 500 employees:**



**Gender ratio from original population:**



**TASK 3.2: Define Test Statistic, Null Hypothesis and Alternative Hypothesis.**

1. **Null Hypothesis:**

* In the population, the distributions of the total compensation of the Female and Male employees are the same. Or the difference is low which can be happen just by chance.

1. **Alternative Hypothesis:**

* In the population, the total compensation of female employees is less than the male employees or the difference can be huge which is beyond the confidence interval levels.

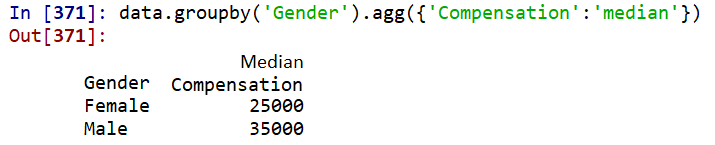
1. **Test Statistic:**

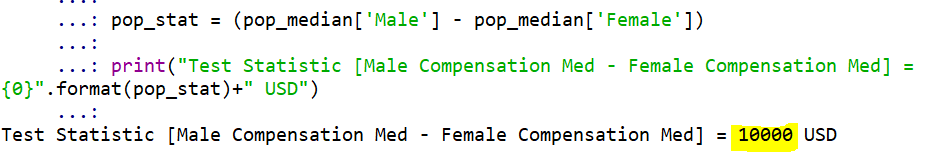
* The alternative hypothesis compares the median total compensation of the two groups (Female and Male) and says that the median for the female employees is smaller. Therefore, it is reasonable for us to use the difference between the two group medians as our statistic.
* Consider, Group A as “Female” and Group B as “Male” employees.
* Test Statistic: Difference between total compensation medians

**Group B Median - Group A Median**

* Large the values of the test statistic favor the alternative hypothesis.

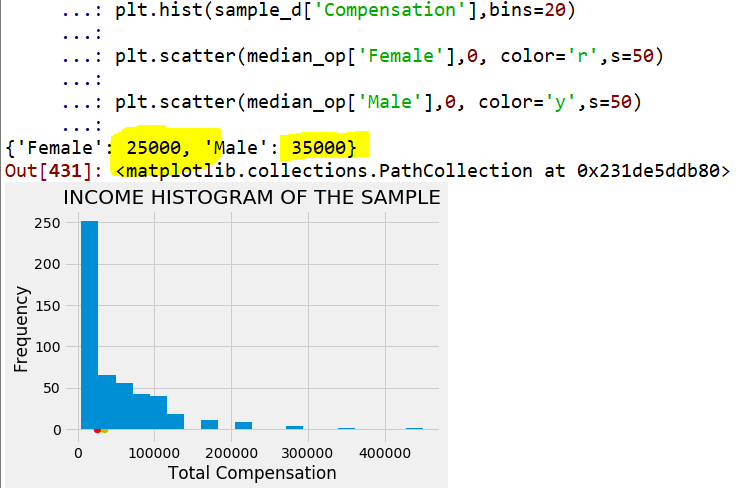
Test statistic can be calculated for our data as below. **As per our data, from the original population, there is a 10,000 $USD difference in the medians of Male and Female employees. Our data accepts the Null Hypothesis and Rejects the Alternative hypothesis, which is clearly explained in the last section.**





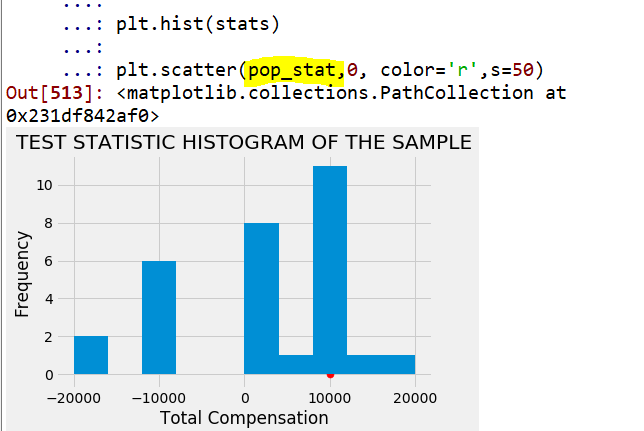
**TASK 3.3: Draw the income histogram for the sample, calculate the median income of the sample, and draw a red dot and a yellow dot of the female median income and male median income of the population, respectively, in the histogram.**

* The below income histogram is generated from the 500 sample that is selected above with the male and female ration as 83.8 and 16.2 respectively which is same as original population.
* You can identify the female income median (25,000) as red dot and male income median (35000) as yellow dot.



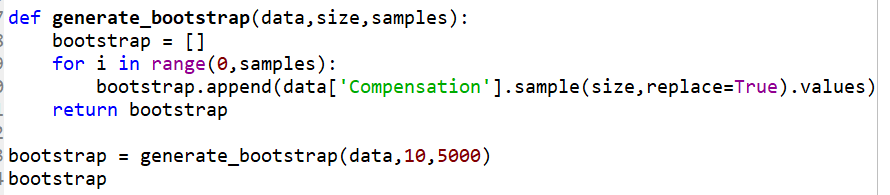
**TASK 3.4: Draw the histogram of the test statistic of the sample and draw a red dot to show the corresponding test statistic of the population (e.g. the diﬀerence of the median incomes between female and male employees) in the diagram.**

* Below histogram of the test statistic has been generated for the sample and plotted the red dot which represents the test statistic of the original population (pop\_stat: i.e., 10,000 $USD).



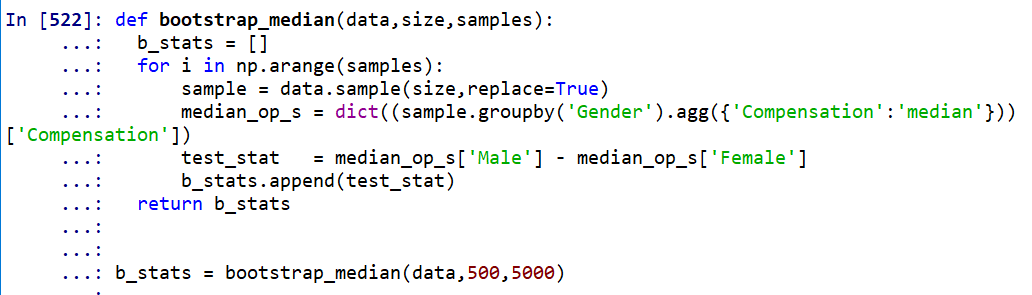
**TASK 3.5: Write a procedure to use bootstrap to produce at least 5000 samples**

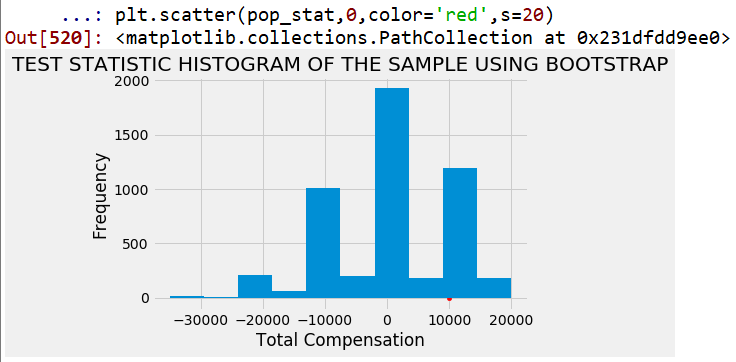
* DEFINITION OF BOOTSTRAP: The bootstrap generates new random samples by a method called resampling: the new samples are drawn at random from the original sample with replacement.
* I have defined a function to create 5000 samples using bootstrap.
* Here, I have used replace=True, to produce the **random sample with replacement.**
* Below is the screenshot of the function that I have defined to generate the bootstrap of 5000 sample each of 10 size.



**TASK 3.6: Draw the histogram of the test statistic of the bootstrap samples**

* Using the bootstrap sampling (random sampling with replacement), I have calculated the test statistic of the samples and populated the below histogram.
* I have also plotted the test statistic of the original population as a red dot just to co-relate the data.
* To calculate the test statistic, I have defined a function as below with replace=True, option.





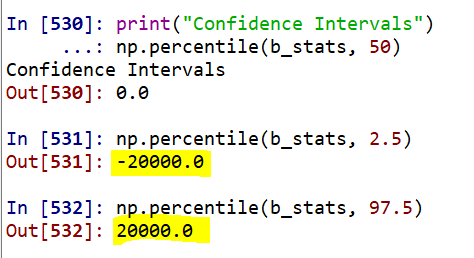
**TASK 3.7: Deﬁne conﬁdence interval and P-value to validate the hypothesis you deﬁned**

1. **Confidence Interval**:

* Confidence interval has a single purpose, i.e., to estimate an unknown parameter based on data in a random sample.

**Using Confidence Interval to Test Hypotheses:**

Our approximate 95% confidence interval for the Median Income in the population goes from -20000 $USD to 20000 $USD (see below screenshot). Test the following hypotheses:



* **Null hypothesis.** The difference in median income between male and female in the population is 10,000 $USD or less as it should be by a chance.
* **Alternative hypothesis.** The difference in median income between male and female in the population should be huge, i.e., beyond the confidence interval range.

CONCLUSION:

Hence, for the 5% cutoff of the P-value, I would **accept the Null Hypothesis and reject the Alternative Hypothesis**. This is because, 10,000 $USD is in the 95% confidence interval range for the population median income.

1. **P-Value (observed significance level):**

* DEFINITION:
* The *P*-value of a test is the chance, based on the model in the null hypothesis, that the test statistic is equal to the value that was observed in the data or is even further in the direction of the alternative.
* If a P-value is small, that means the tail beyond the observed statistic is small and so the observed statistic is far away from what the null predicts. This implies that the data support the alternative hypothesis better than they support the null.
* If the P-value is less than 5%, it is considered small and the result is called "statistically significant."

CONCLUSION: As per our data, the P-value is 0.28 which is greater than the 5%, hence the difference of the test statistic is not significant. Hence, **I accept the NULL HYPOTHESIS.**

