Part1: Perform exploratory data

For this part, Age vs. Salary, Professional Experience vs. Salary and Gender vs. Salary is chosen. To do the analysis, first of all extract salary, age, gender, experience column is performed to create a new data frame, the advantage of doing that is to simplify the problem since only the useful column is kept and all the other data are filtered out. In addition, the `groupby` function is used here to group the same gender together (man and women). Since the figure plotted should describe the trends among those groups, so first of all, average salary of each gender is taken by using .mean() function, then the way of sorting the data is by using `.sort_values()`. Based on the generated plot, it can be observed that the older people usually have higher salary, people who has more years working experience will earn more salary. At last, man usually earns more than woman and people with Non binary gender earns the highest salary.

Part2: Estimating the difference between average salary (Q25) of men vs. women (Q2).

2.1 Descriptive Statistics

The descriptive statistics for men's salary and women's salary can be simply viewed by `df.describe()`, there are 12642 men in the sample space which have the mean salary 51193.600 and standard deviation is 99979.274. As for women, there are 2482 women take the survey and women have the mean salary of 34816.88 with standard deviation 72017.347

2.2 T-test

For performing the t-test on men and women salary, the null hypothesis H0 is Gender and salary are statistically not significant, when performing the t-test, the `ttest_ind` from `scipy.stats` is called and the returned p-value is 8.088e-15 which is way much lower than threshold 0.05 which means the null hypothesis is being rejected then gender and salary are statistically significant.

2.3 Bootstrapping

For bootstrapping the datasets, using the data frame built-in function df.sample(n = 6000), which will randomly draw out 6000 data from the original data , the reason why half (50%) of the original data is chosen is because if the number is chosen too large, some data will be represented many times, and the other datas might not be selected at all. and then take the mean of those drawn data and repeat this process for 1000 times and store the bootstrapped data back into the new data frame, by far, the bootstrapping process are completed. From the generated plot, it can be observed that both the man and woman salary become normally distributed. The Distribution Difference in Means is produced using bootstrapped man salary minus woman salary and it is also normally distributed.

2.4 Apply t-test on bootstrapped data

After use the ttest_ind function on bootstrapped data, the test statistic is now equal to 103.187 and p-value is now 0, since the p-value is smaller than 0.05, therefore, the null hypothesis is still rejected.

2.5 Comment On Findings

After performing t-test on original and bootstrapped data, it can be seen that the p-value is pretty low in both case, and p-value is even equal to zero when data is bootstrapped, which indicates that different gender earns different salary since p-value is lower than 0.05. The other finding is that when original data is bootstrapped, it creates many simulated data and the salary distribution becomes normally distributed.

Part3: Anova Test On Education VS Salary

3.1

Using the same function `.describe()`, it can be known that there are 4777 datas are collected from Bachelor's degree holders and the mean salary is 35578.29, standard deviation is 89382.06. There are 6799 Master Degree holder take the survey and their mean salary is 52706.86 and the standard deviation is 90928.78. As for the Doctoral's Degree, there are only 2217 people taken the survey, and they achieve the highest mean salary 70641.18 and the standard deviation is 117160

3.2

Anova test is used when the hypothesis test is conducted among more than two groups, using the `f_oneway()` function from `scipy` package, it returns test statistic =109.758 and p-value =5.10e-48. Based on this result, it can be concluded that they are statistically significant and the null hypothesis is rejected.

3.3

Using the same method mentioned in the section 2.3, and apply on to Bachelor's Degree, Master's Degree and Doctoral's Degree, and the number of data to be bootstrapped are 2400, 3400, 1110 respectively, then the same trend could be observed that the salary distribution on those three groups are all become normally distributed. The Distribution Difference in Means is produced using bootstrapped man salary minus woman salary.

3.4

After performing the Anova test on bootstrapped data, the new test statistic now becomes 15998.5 and p-value is equal to zero as well. Since p-value is smaller than 0.05, so it can be concluded that level of formal education is statistically significant with the salary and null hypothesis could be rejected.

3.5

If original data is used p-value is pretty close to zero, however when the using data is bootstrapped, then the p-value is exactly zero, which means the null hypothesis is rejected in both case, salary is statistically significant with the level of education. When the data is bootstrapped, more simulated new data will be created, and the distribution of salary will be more like normally distributed.