Q1.
First of all, I am trying to get to introduce the data with 3 plots, to get to know some potential relationship between several variables and salary.
For the very first plot, I am trying to reveal the relationship between distribution (average and variance) of salary among age. From the resulted boxplot, we can observe that the span of salary range tends to spread as age increases, the salary range of age 18-21 is very tight, while the range grows larger when age increases, and the span tends to keep the same after age of 50. The average salary also grows with age from 18-45, after age of 50, the average salary flattens out. The distribution of salary is right skewed, as we can see a long tail among the high salary.
I plotted average salary vs country in the second figure. We can see that US, Switzerland and Israel has relatively high salary on average, countries like Ghana, Ethiopia and Bangladesh have relatively lower salary on average.
On the third plot, I try to see the relationship between Salary and Education. Here, I skipped 'prefer not to answer', 'no formal education past high school' and 'no bachelor's degree' because they don't have specified education type. For the other four degrees, we can see that Doctoral degree has highest mean salary and largest variance, bachelor's degree has lowest mean salary and lowest variance. Master's degree and professional doctoral has similar distribution.

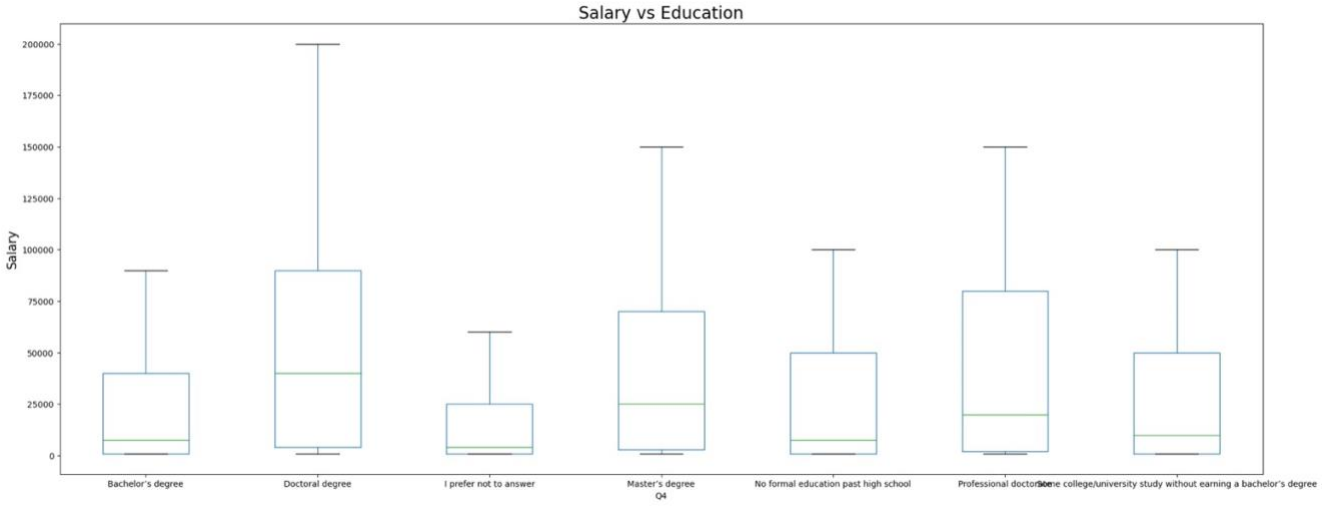
Q2.a.
In this step, we try to look at the distribution of woman and man salary, including mean, variance, and percentiles. To do this, I call a build in function *'df.describe()'* on man and woman dataset. The result is similar to what we can see in a boxplot, but it is more precise.
Here, we have much more samples for man (12642) than for woman (2482). It seems that man has a higher average salary than woman, and salary range of man spreads out more compared to woman (higher standard deviation). Next, we can see that man and woman all have same min and max salary, which is 1000 and 1000000, it is simply because we don't have a higher or lower salary option in our question. In addition, man has a higher salary value at 25^{th}, 50^{th} and 75^{th} percentile, which is consistent with the result that man has a relatively higher mean salary.
b.
We have seen above that the mean salary in the male and female populations were different. To test if this is significant, in this step, I used *'scipy.stats.ttest_ind()'* to do a 2-sample t-test.
c.
I extracted salary of man and woman as two lists, set up iteration for 1000 times, each time I bootstrapped list of salary respectively for man and woman, using *'random.choices()'*, and put the mean of bootstrapped sample into a new list. At the end, I get two lists of length 1000 of bootstrap mean in each iteration.
d.
In c, we can see that the sampling distribution of man and woman has different mean, a T-test here is to check if the difference is significant, the p value is exactly 0, which means that we are nearly 100% sure that bootstrapped distributions of mean salary man and woman has different mean.
e.
In question a, we can roughly see man salary has a higher mean and higher standard deviation, to see of the difference is significant, we did a T-test in b. The T-test statistic value is -7.77, result in an extremely small p value (= 8e-15 << 0.05). Therefore, we have strong evidence to

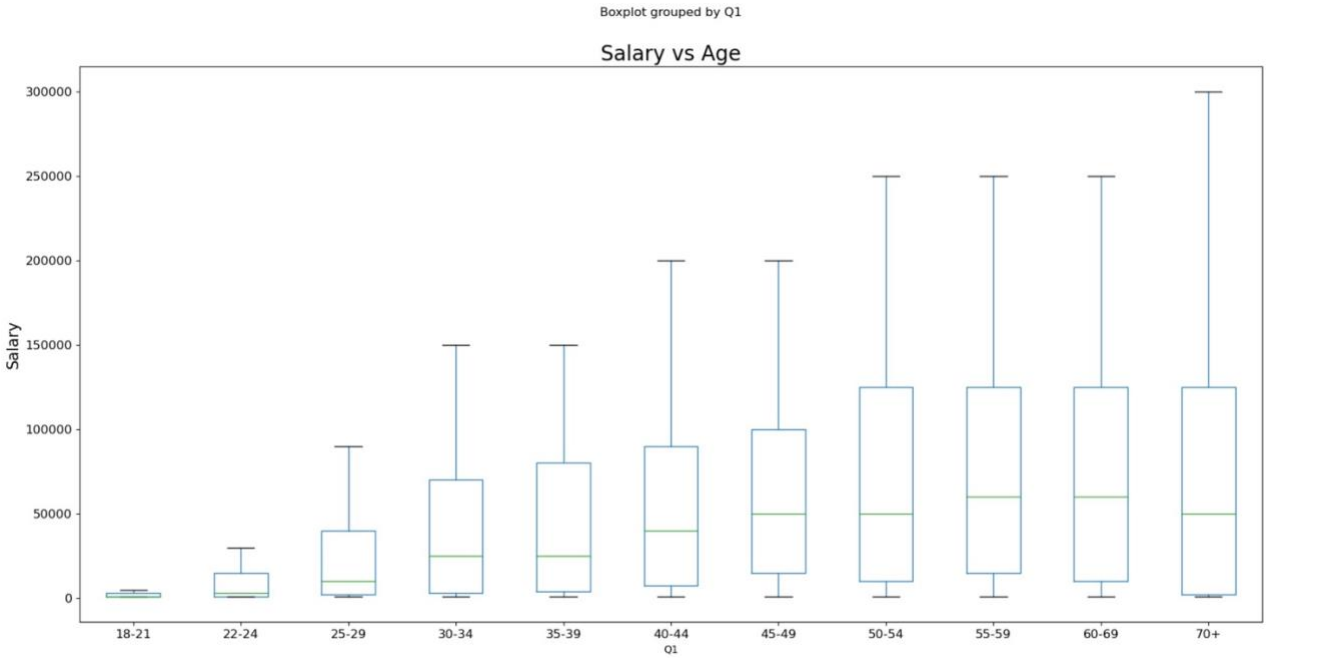
conclude that man and woman has a different average salary, and man tend to be paid more than woman.
In question c, we produce 2 plots to see our bootstrap result. The first plot in c shows distribution of mean salary for man and woman respectively, here, we can observe the distribution for man and woman mean salary is closed to normal. This can be explained by central limit theorem that sampling distribution of mean roughly follows a normal distribution. Distribution of mean salary for man and woman seems not to be overlapped. In addition, the second plot of difference in mean salary is also closed to normal, and the distribution shows a mean of about 16000, the average difference is much bigger than 0. This is consistent with the result from T-test that man and woman have different average salary. To verify our conjecture in c that bootstrapped distribution has different mean, we performed a T-test in part d. The p value is exactly 0, which means that we are nearly 100% sure that bootstrapped distributions of mean salary man and woman has different mean. The p value here is even smaller than the p value in question b, this can be explained by central limit theorem, because in b, we were exploring the distribution of individuals, but in d, we are exploring the distribution of sampling mean. Theoretically, if individual distribution follows normal (mean=a, variance=b), then the mean of this distribution would follow normal (mean=a, variance=b/n^2), since we have much smaller variance in the bootstrapped distribution, we will have more evidence to reject null hypothesis of equal mean.

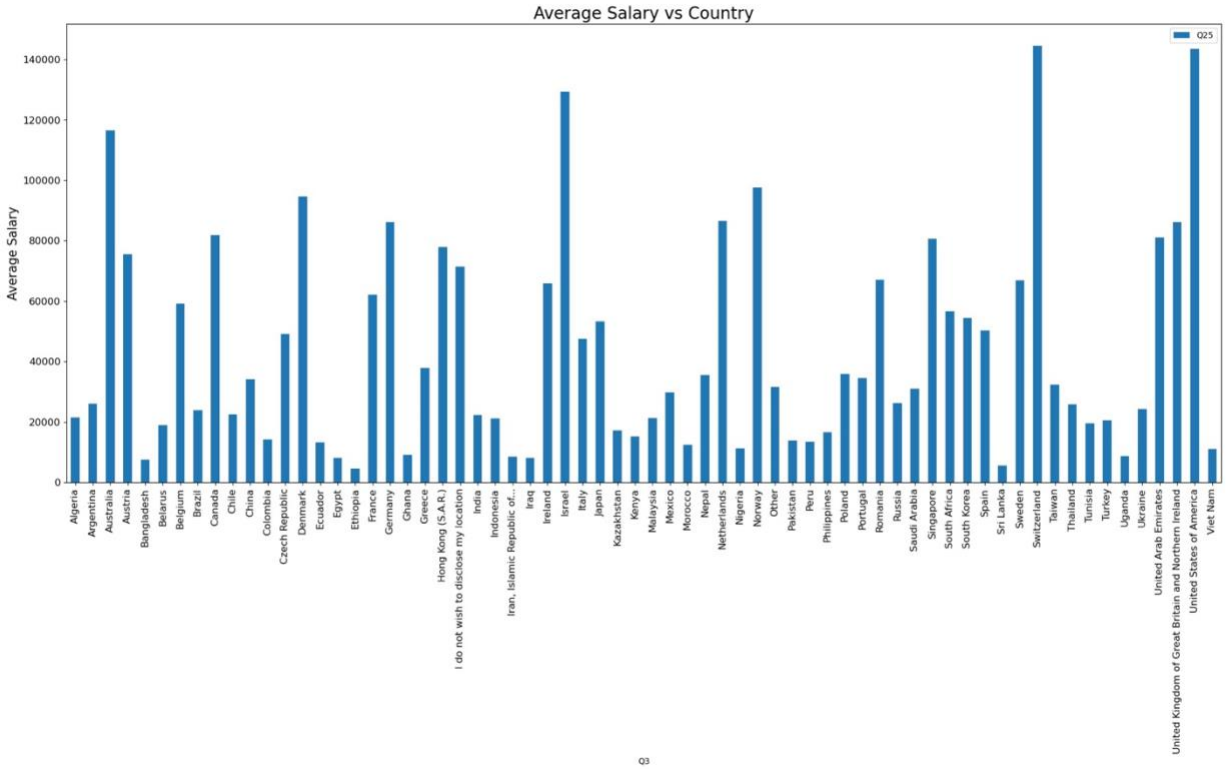
Q3.a.
In this part, I dropped several rows: 'prefer not to answer', 'no formal education past high school', 'professional doctoral' and 'college study without degree'. Therefore, we are only analyzing on bachelor, master, and doctoral degree in question 3. The code is very similar to question two, but we generated three descriptive results here.
b.
We have seen above that the mean salary for three of degree types were different. To test if this is significant, in this step, I used *'stats.f_oneway ()'* to do ANOVA.
c.
The programming part is similar to question 2, but we repeated bootstrapping 3 times, to get the mean salary for three groups.
d.
We want to see if the difference in mean of bootstrapping distributions for (master, bachelor, doctor) is significant, therefore we performed an ANOVA test.
e.
Looking at the plot of part a, we have a highest respondent on master's degree, followed by bachelor and doctor. People in doctoral degree has highest average salary, followed by master and bachelor. For analysis of variance, we have statistic value 109.8, and a very small p value, so we reject the null hypothesis that all three categories has same average salary, the evidence is significant that at least one of (master, bachelor, doctor) has different mean salary.
In part c, the first plot shows three distinct distributions for bootstrap result of these three groups, they do not overlap. This can also be shown in the following three plots: difference in mean salary is also closed to normal, the central of the normal distributions are far from 0. The relative rational is explained in question 2. To see if the difference is significant, we go to part d. The statistical value here is extremely big, resulted in a p value of 0, which means that we are nearly 100% sure that bootstrapped distributions of average salary of these three groups has different mean.

Appendix

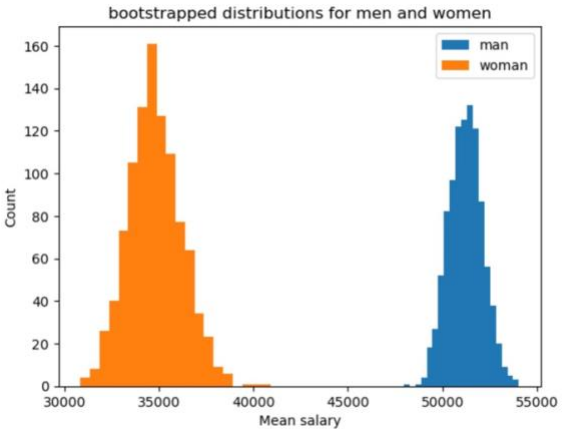
Figures for question 1

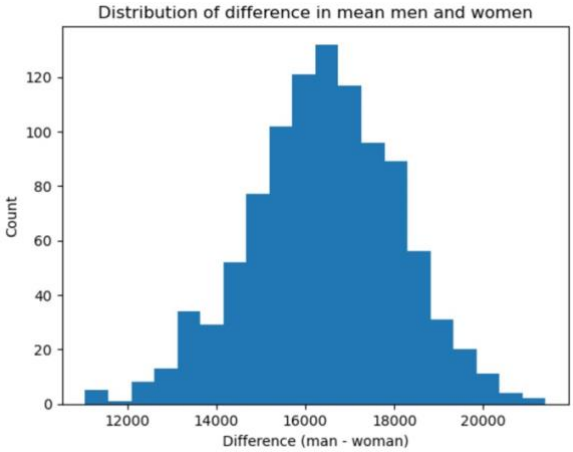






Figures for question 2





Figures for question 3