In this assignment, we are going to explore the survey data to see the nature of women's representation in Data Science and Machine Learning, and also the effects of education on income level.

To perform exploratory data analysis on the survey dataset, I import the clean dataset first, and filtered the dataset with main characteristics country, age, professional experience and salary. Then, I get some statistics of the salary(Q25). Salary has mean at $49,116, and its standard deviation is around 98090. To present 3 graphical figures that represent different trends in data, I first want to see the relationship between country and salary. I extract the matching two columns, and then calculate the mean salary by grouping it by country. I also print out the 10 countries with highest average salary and 10 countries with lowest average salary. Based on the figure, we can see there's huge difference in mean salary between different countries. Switzerland is the country who has the highest mean salary with around $144,548, while Ethiopia has the lowest mean salary which is around $4,633. We can also discover that developing countries from South Asia and Africa tends to have lower mean salary than developed countries in North America and Europe. Secondly, I plot the relationship between age and mean salary. The plot shows there's a increasing trend in mean salary as getting older before age 60. There's a fall in mean salary at age 60-69, then rise after age 70. Age 18-21 has mean salary around $15,722 while age 70+ is around $100,649. The third plot compares the mean salary among different education levels in ascending order. The group "prefer not to answer" has the lowest mean salary, it may because their education is low, but we cannot make sure. bachelor's degree has the second lowest salary around $35,578. There's an obvious increase in mean salary after master's degree; professional doctorate and doctoral degree have the highest mean salary. This way of showing trends of data is simple and direct, but we are only grouping by one characteristic, it is hard to see the correlation between different variables.

To estimate the difference between mean salary of men and women, I first filtered the original dataset to get rid of other gender specification, and then extract gender and salary column. After checking there's no missing data, I grouped it by sex. There are 12642 men and 2482 women, the population of woman in data science and machine learning is far less than man. The mean salary for men is around $51,193 while for women is around $34,816. By applying a two-sample t-test with 0.05 threshold, with null hypothesis average salary of men and women are equal and alternative hypothesis that average salary of men and women are different. We get the result that t-statistic = 7.77406, and p-value = 8.08881e-15. Since p-value is around 8.09e-15, which is smaller than 0.05, so we have significant evidence to reject null hypothesis. We can conclude that the average salary of men and women are different, and from previous statistics, the average salary of men is much higher than women. T-Test is used to determine the significant difference between the means of two groups with similar features. It is essential for generalization and easy for interpret. The output tells you how different the mean of one sample is from the mean of another group. It also indicates the mean of each group and the average difference between the groups. A larger t-score indicate

the groups are different and a smaller t-score indicate the groups are similar. However, there may be carry-over effect and small amount of noise.

Next, I bootstrapped the data for comparing the mean salary of two groups with 1000 times replications and resample each group for 1000 times. Then, plot the bootstrap distribution for man's average, woman's average salary and sex difference in average salary. I also apply a two-sample t-test with 0.05 threshold on bootstrapped data, with the same null hypothesis and alternative hypothesis. We get the result that t-statistic = 80.7761 and p-value = 0. Since p-value is 0, which is smaller than 0.05, so we have significant evidence to reject null hypothesis. The average salary of men and women are different. From above bootstrap distributions figures, we can see the men's mean salary is much higher than women. With the bootstrapped data, most of the differences of mean salary for different sex locates between $10,000 and $20,000. Bootstrapping resamples a single dataset to create many simulated samples. It is also an appropriate way to control and check the stability of the results, and a convenient method that avoids the cost of repeating the experiment to get other groups of sample data. Limitations of bootstrap are if the bootstrap distribution is extremely skewed or sparse, the bootstrap interval might be quite unreliable, also a representative sample is required that still need a good representative sample from the population.

Following, we estimate the difference of average salary between highest level of formal education. I select "Bachelor's degree", "Master's degree", "Doctoral degree" from the original dataset. Here, I did not include "professional doctorate" since it may be outside the scope of formal education. After extracting the matching columns and checking there's no missing data, I grouped it by three education level groups. The population of having bachelor's degree, master's degree and doctoral degree is 4777, 6799 and 2217 respectively. The mean salary is higher when education level is higher. Here, we use ANOVA F-test (since there are three groups) with null hypothesis that average salary of bachelor's degree, master's degree and doctoral degree are equal, and alternative hypothesis average salary of three groups are not all equal. The result is statistic = 7332.13 and p-value = 0. Since p-value is around 5.11e-48, which is smaller than 0.05, so we have significant evidence to reject the null hypothesis. Therefore, there is at least one mean salary different from others. We also use bootstrapped distribution and repeat for 1000 times (same as the previous part), then produced three bootstrap distribution plots of average salary for these three education level groups and three distribution plots of pair difference in average salary. After that, I repeat the ANOVA F-test on bootstrapped data with same null hypothesis and alternative hypothesis. The result is statistic = 7332.13 and p-value = 0. Since p-value is 0, which is smaller than 0.05, so we have significant evidence to reject the null hypothesis. Therefore, there is at least one mean salary different from others. From bootstrapped distribution plots, we can see each group's mean salary is different from others. Also, distribution difference plots show that the mean salary difference between doctoral degree and bachelor's degree is the highest, mean salary difference between master and doctoral and difference between master and bachelor is similar.