Data Analysis on ‘Kaggle ML&DS Survey’ Responses

The objective of this analysis is to explore the impact of gender and education level on average yearly compensation, whether they are correlated or not. Given the dataset, I consider the salary as my interest variable, which is also the response variable. Then I look for some key features as explanatory variables which are expected cause of different salary level.

- Q1. Key Features Analysis

For the first plot (Figure1.1 in appendix), I chose the education level vs salary. Focus on three basic degrees, bachelor, master, and doctor degrees, we can clearly conclude that the mean salary increases with higher education level.

For the age vs salary (Figure1.2), there is no obvious trend showing the older the age, the higher the salary. But we can find the minimum average salary appeared in the age range of 18-21, the second minimum in range of 22-24. It stated that for those who just entered the workplace, salary would tend to be lower due to their less working experience.

Then the third plot I considered the professional experience and salary (Figure1.3). This plot has an explicit trend showing higher salary comes with longer professional experience.

- Q2. Women’s representation in ML&DS using t-test

Figure2.1 gives the descriptive statistics for women, the mean yearly salary on women is 34816.88 US dollars, with big deviation of 72017.35. The histogram of women’s salary shows significantly positive skewness, the mean is higher than median and large amount population focus on minimum salary at $1,000. Similar results in men’s descriptive statistics (Figure2.2), but the sample size of men group is about 5 times women group. The mean value of men is 51193.6 higher than mean of women group.

Then I conducted two sample t-test for these two groups. The hull hypothesis is that mean value of women group is same as mean value of men group. T statistic is 7.77, p-value is 8.09, which is much higher than alpha of 0.05. Thus, given confidence interval of 95%, we fail to reject the null hypothesis. There is no significant difference between mean values of two groups.

Next step is bootstrapping sample data. We use bootstrap to resample the dataset and ensure the samples are randomly selected from the population, which assigns measures of accuracy. By using 1000 replications, we can see bootstrapped distribution for both men and women groups tend to be normal distributed with a smooth bell curve (Figure2.3). The mean of men group is around 52000 whereas women group’s mean is

35000. Meanwhile, the difference of two groups is also normally distributed with mean of 16000(Figure2.4), which is consist with two groups distributions above.

I apply the t-test again for bootstrapped data. The null hypothesis is same as above, t statistic is -306.45 and p-value is 0.0 less than 0.05. Therefore, I reject the null hypothesis, that is the mean value of these two groups are not equal.

By all of results above, we can conclude that women's salary tends to be much lower than men who are working in data science and machine learning. This is consisted with the common bias that most of the world's data scientists are men. The small sample size of women group also proved this in dataset.

- Q3. Relationships between education level and salary using ANOVA

For the third part of this report, I focused on the effects of education level on salary. In this part I only considered three main education levels, bachelors, masters, and doctors. We have 4777 bachelors, 6799 masters and 2217 doctors as samples. Doctors have highest mean salary at $70,641, whereas $52,707 for masters and $35,578 for bachelors.

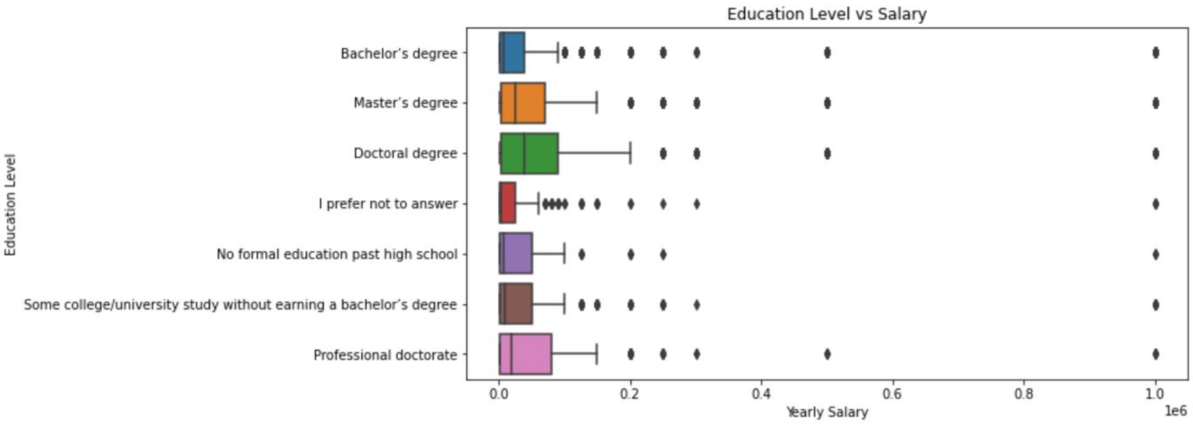
Due to three groups in consideration, t-test is no longer suitable for this hypothesis test. So, we create ANOVA test for these three groups. The null hypothesis is three groups' mean values are equal, mu1=mu2=mu3. Given F-statistic of 109.7578 and p-value of 5.1077, p-value is greater than 0.05. We fail to reject the null hypothesis. There is no significant difference between the means of three groups.

However, as we can see from above descriptive statistics, the mean value of these groups should not be equal. That is because the normality assumption of ANOVA test is not strictly followed, we can find the histogram of these groups' population are not normally distributed. Thus, we apply bootstrapping again on three groups. From the bootstrapped distribution plot of three groups (Figure3.1), the normality is now followed. At same time, the distributions of the differences between three groups are also normal, which proved the argument again that mean values should not be equal.

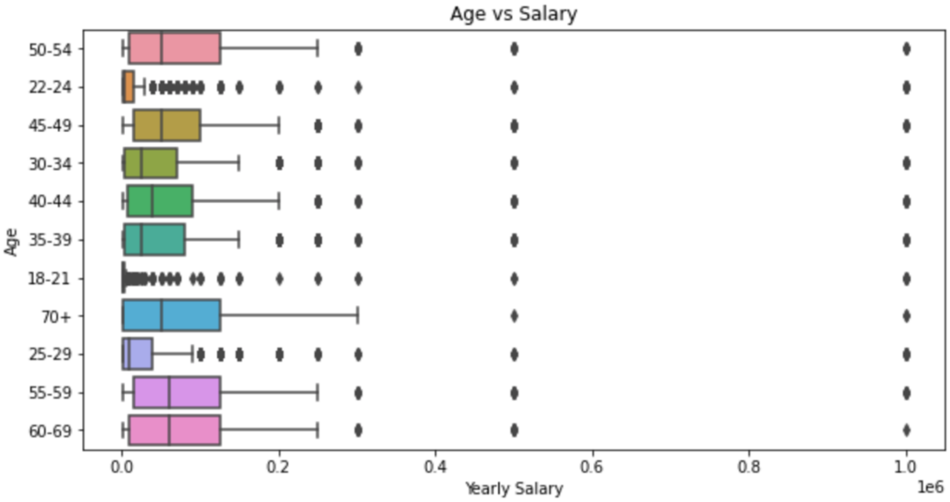
Then I tried ANOVA test again on bootstrapped data of three groups. The null hypothesis stays same, and F-statistic is 104628.1145, p-value is 0.0 less than 0.05. Thus, we reject the null hypothesis and concluded that the mean values of three groups are not equal.

In conclusion, people with a high level of education will get higher paid, which is consistent with our common sense. Because higher education talents have bigger knowledge reserves and can be promoted to more complex jobs and more important positions.

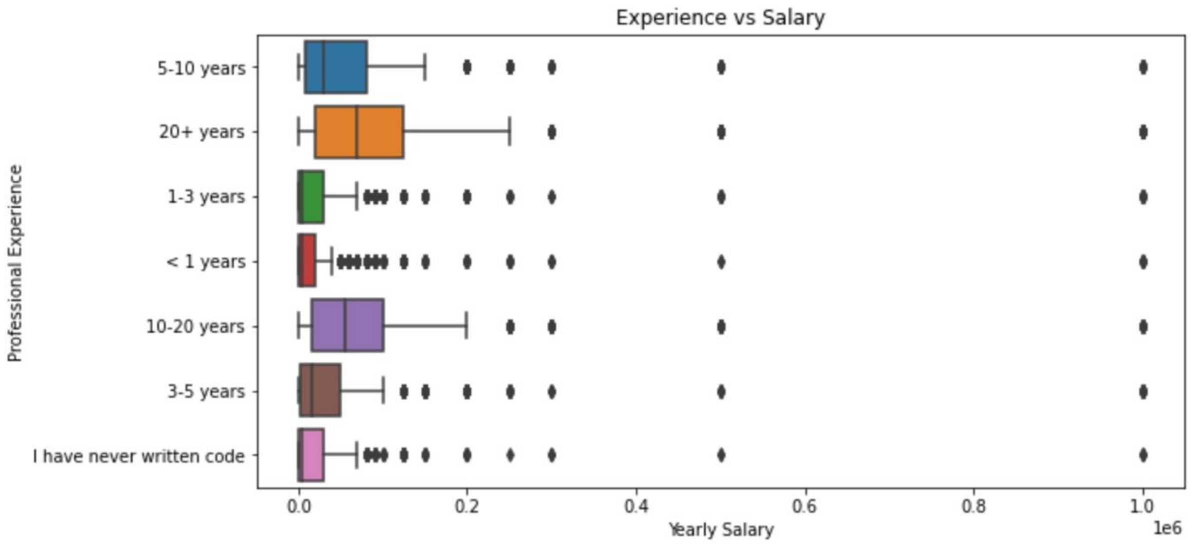
Appendix:



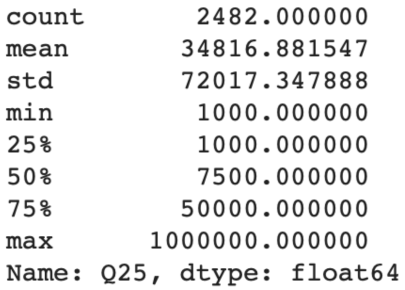
(Figure 1.1)



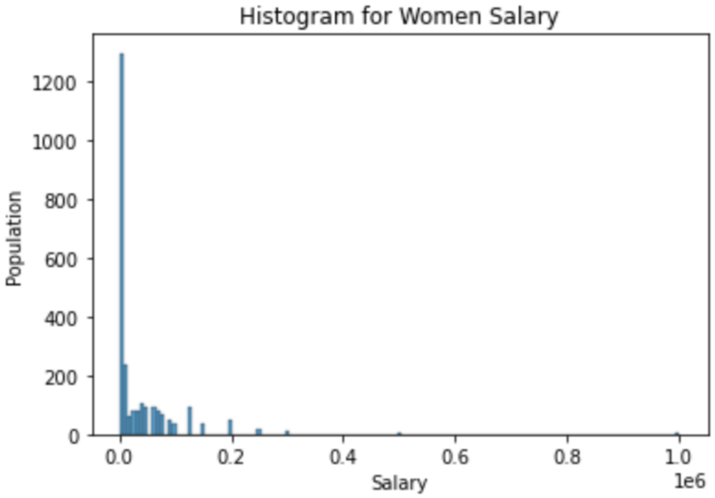
(Figure 1.2)



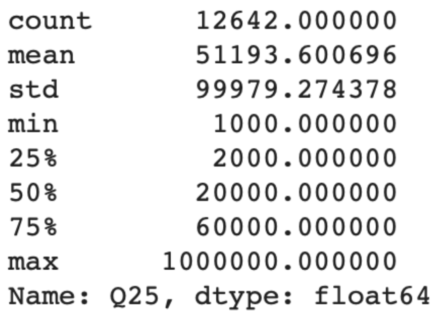
(Figure 1.3)



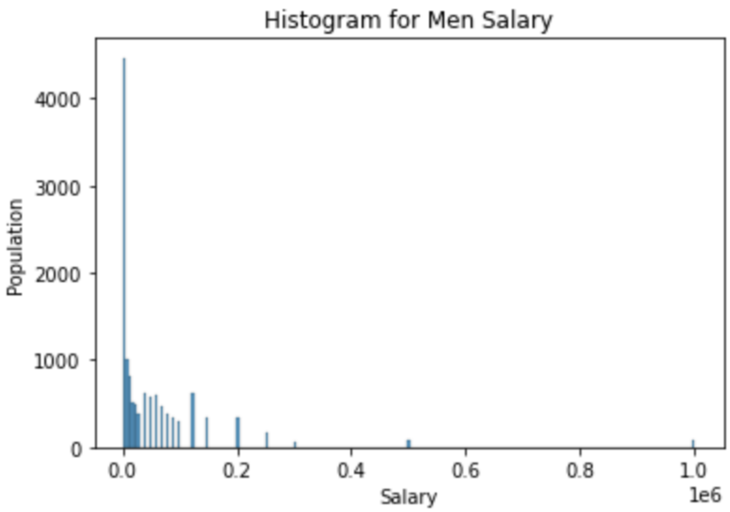
```
count        2482.000000
mean        34816.881547
std         72017.347888
min          1000.000000
25%          1000.000000
50%          7500.000000
75%         50000.000000
max       1000000.000000
Name: Q25, dtype: float64
```



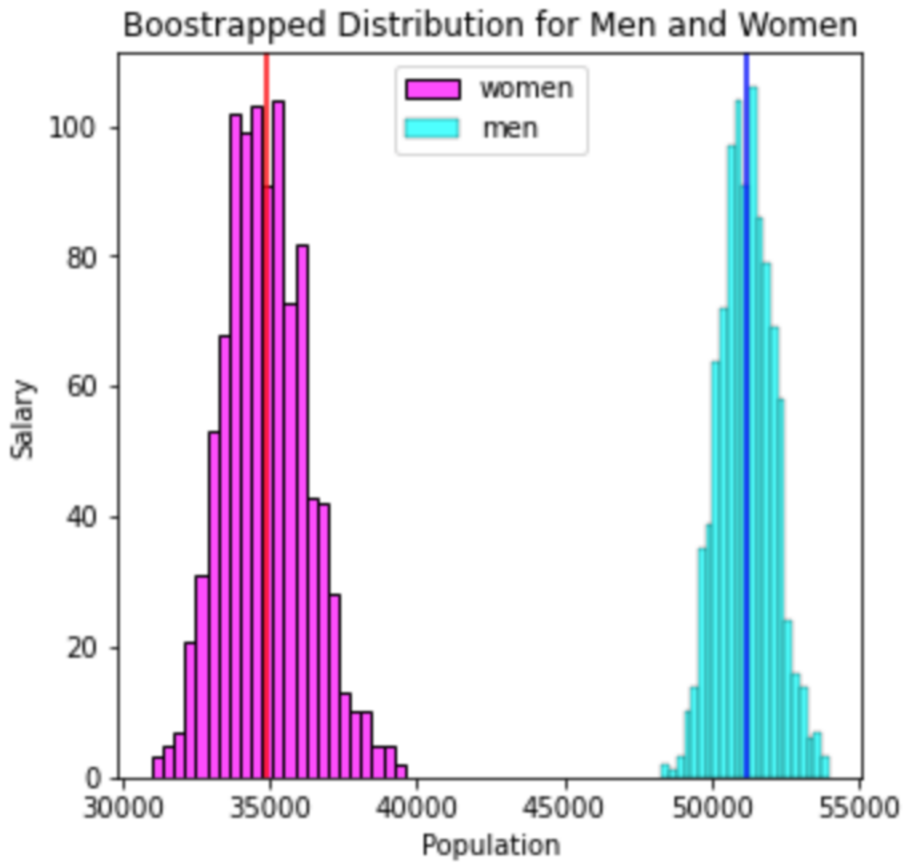
(Figure 2.1)



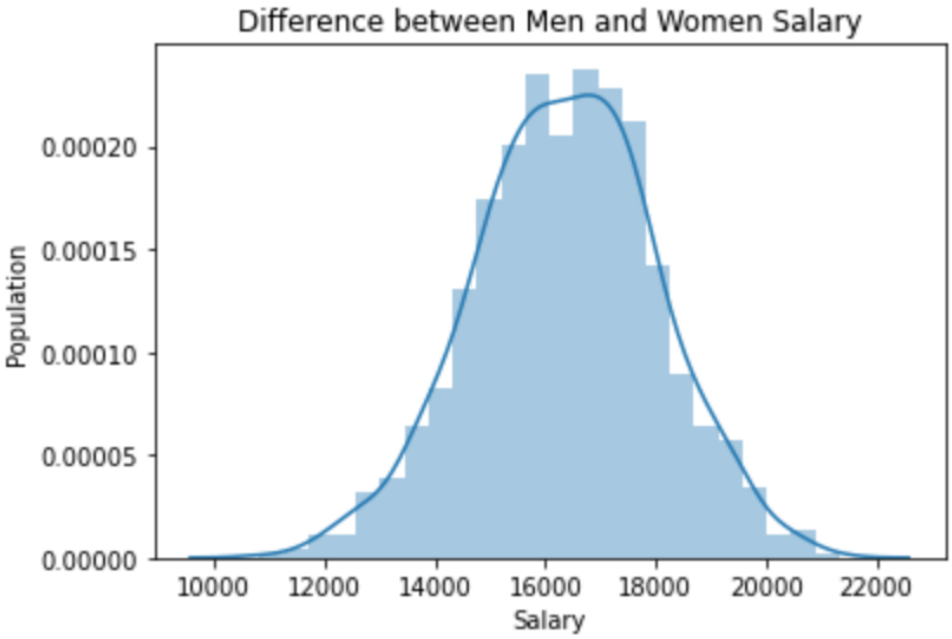
```
count        12642.000000
mean         51193.600696
std          99979.274378
min           1000.000000
25%           2000.000000
50%          20000.000000
75%          60000.000000
max        1000000.000000
Name: Q25, dtype: float64
```



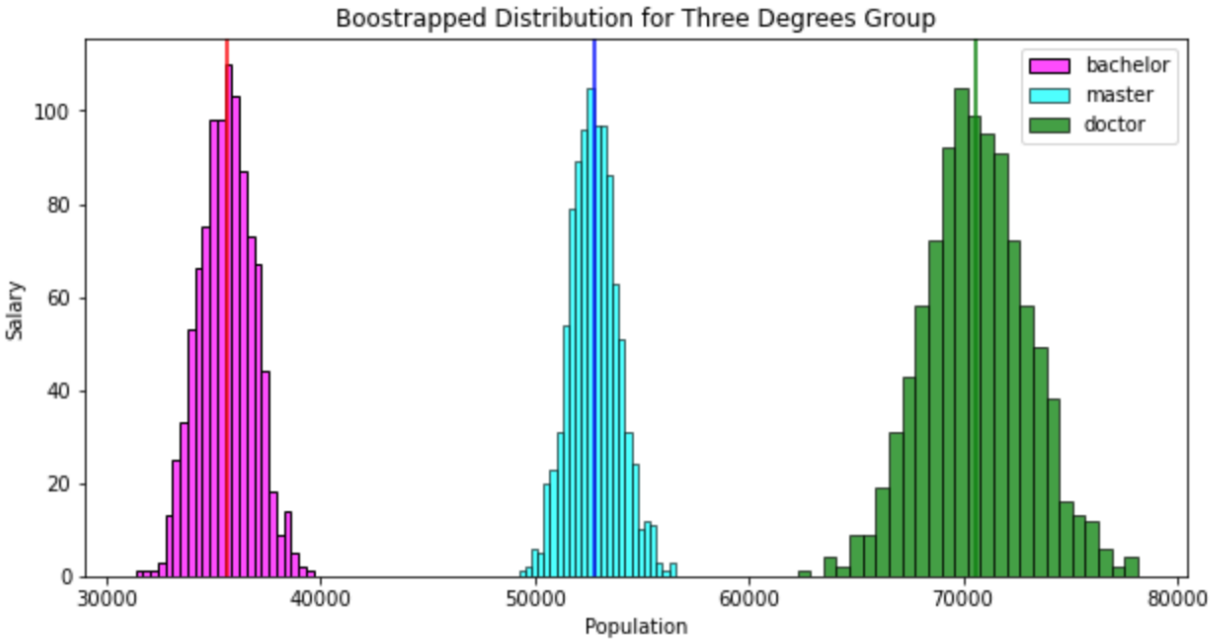
(Figure 2.2)



(Figure 2.3)



(Figure 2.4)



(Figure 3.1)