MIE1624 Assignment 1 Report

Exploratory Data Analysis

3 box plots have been created to show different trends in the data can be seen in the ipynb notebook.

Nature of Women's Representation in Data Science and Machine Learning

Descriptive statistics:

Salaries of men:	
count	8872.000000
mean	50750.619928
std	70347.974812

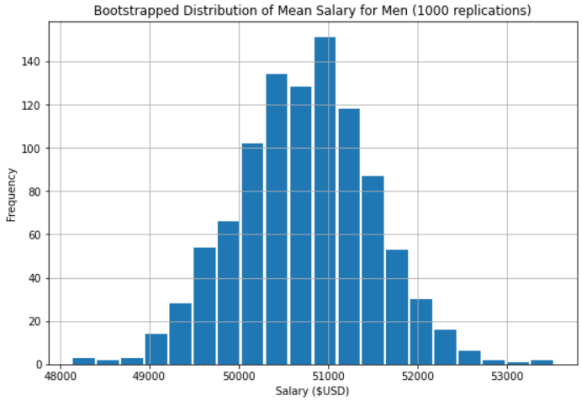
Salaries of women:	
count	1683.000000
mean	36417.112299
std	59442.716093

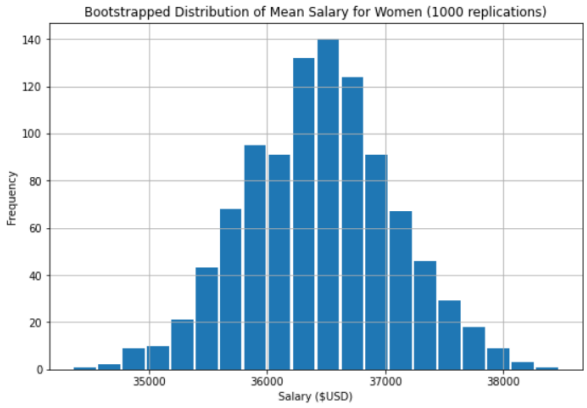
Looking at the above data, the mean salary of the male and female populations seems to be different. We will perform a t-test to verify this significance. The result of the t-test performed in Python are as follows:

```
t-test: t = -7.84433  p = 4.77315e-15
```

Since the p-value is much smaller than the 0.05 threshold we have set, this t-test proves that men generally have a higher salary compared to women.

Next, I will bootstrap the existing data for further analysis of the difference in salary between men and women. The bootstrapped distributions can be seen below. A graph of the difference in distribution between men and women can be seen in the ipynb notebook.





After performing the above bootstrapping for both gender groups, I performed another t-test, this time using the bootstrapped data. The results are as follows:

```
t-test: t = -458.916  p = 0
```

Since the p-value is much smaller than the 0.05 threshold, this t-test further reinforces the idea that men generally earn a higher salary compared to women.

<u>Comments on findings:</u>

Both t-tests (original data and bootstrapped data) resulted in very small p-values, meaning that the difference between the salaries of men and women are statistically significant.

Based on the results of the two t-tests performed, it is safe to assume that men tend to have a higher salary compared to women in the data science and machine learning industry. Additionally, from the descriptive statistics obtained it appears that women are very much underrepresented in this industry (8872 samples of men compared to just 1683 samples of women).

Effects of Education on Income Level

Descriptive statistics:

Bachelor degree:	
count	3013.000000
mean	35732.824427
std	60247.753546

Master degree:	
count	4879.000000
mean	52120.106579
std	67681.571528

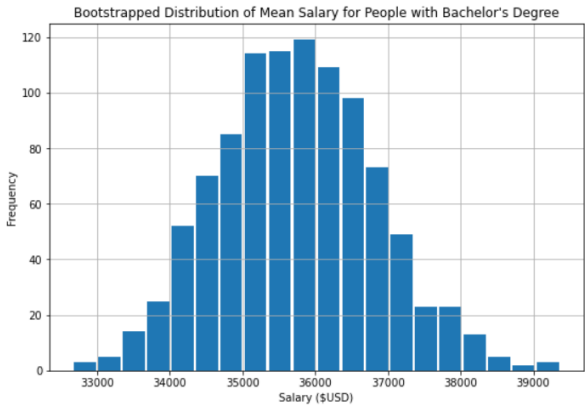
Doctoral degree:	
count	1718.000000
mean	68719.441211
std	85403.650394

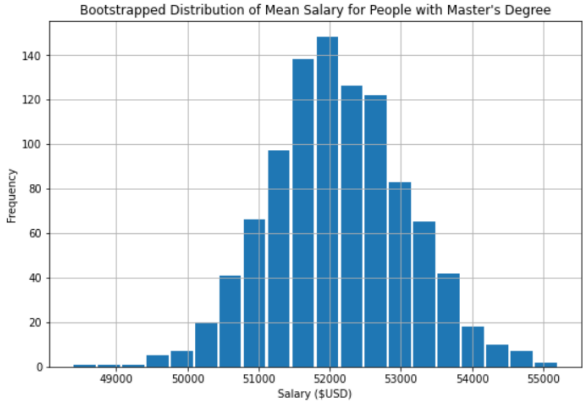
Looking at the above data, the mean salary of the populations of different degree holders seems to be different from each other. We will perform a one-way ANOVA test on these three groups to verify this significance. The result of the ANOVA test performed in Python is as follows:

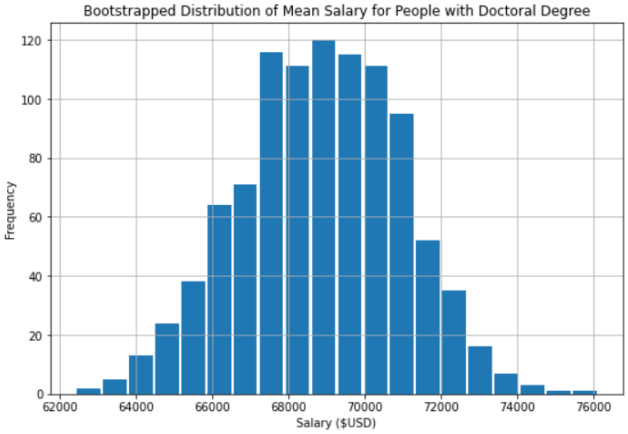
```
F-test: F = 129.756  p = 2.48521e-56
```

Since the p-value is much smaller than the 0.05 threshold we have set, the ANOVA test proves that the salary of different degree holders statistically all belong in different populations. Therefore, the test proves that higher degrees generally lead to better income.

Next, I will bootstrap the existing data for further analysis of the difference in salary among the three types of degree holders. The goal for bootstrapping is to obtain data that is more normally distributed compared to before. The bootstrapped distributions along with the distributions of the salary differences can be seen below.







Graphs of the difference in distribution among the three groups can be seen in the ipynb notebook.

After performing the above bootstrapping for the degree groups, I performed another ANOVA test, this time using the bootstrapped data. The results are as follows:

```
F-test: F = 124891  p = 0
```

Since the p-value is much smaller than the 0.05 threshold, this ANOVA test further reinforces the idea that higher degrees tend to lead to better financial gains.

<u>Comments on findings:</u>

Both ANOVA tests (original data and bootstrapped data) resulted in very small p-values, meaning that the difference among the three groups of degree holders are statistically significant.

Based on the results of these ANOVA tests, it is safe to assume that in the data science and machine learning field, those with a master degree will generally earn more than those with only a bachelor degree, and those with doctoral degrees will generally earn more than both bachelor and master degree holders.