MIE 1624 Introduction to Data Science and Analytics – Winter 2023

Assignment 1

Name: Ayesha Patnaik, Student Number: 1008681696

The objective of this assignment is to explore the survey data to understand the nature of women's representation in Data Science and Machine Learning and the effects of education on income level. The following report contains analysis for each of the questions.

Question 1: In this part, an exploratory data analysis was performed to analyze the given data set. I chose to analyze the impact on salary by three categories : Age, Professional Experience and Job Role. The analyses are done first by comparing each of the three features with average salaries based on their gender and then solely on their categories. I cleaned the data further by removing unwanted columns and analyzed each combination of features through bar graphs.

First analysis is done between age vs average salary based on gender. We see from Figure-1 that the mean salary in general is higher for the age group of 40-44 and 45-49 across all genders which is on an average around 120,000 USD per year. However we do see some exceptions. In age groups of 30-34,35-39 and 70+, the non-binary population seem to have exceptionally higher mean salaries than the other genders. We do see a considerable proportion of people prefer to self describe their gender, especially the young population of 22-24 who seem to have really high yearly compensation and a good amount of people prefer not to reveal their gender. In general, we infer that irrespective of the age group, the mean salaries of men are always higher than the mean salaries of the women. And when we analyze average salaries solely on age group, we see in Figure-2 that with increase in age group the mean salaries get higher. Second analysis is done between professional experience and salaries based on gender. From Figure-3, we see that with increase in professional experience the mean salaries are increasing significantly for all gender categories. Even here we notice that men's mean salaries are higher than women's, and the gap increases significantly as experience increases. For non binary people, the case is different and it doesn't follow a trend but they have a significantly higher mean salary. When we analyze the relationship between average salaries with professional experience alone, we clearly see an increasing trend in salaries with more years of experience, as can be seen in Figure-4. Final analysis is done between job role and average salary based on gender and then solely on job role. From the graph Figure-5, we see that in general the high paying jobs are Developer Relations/Advocacy, Product Manager followed by Software Engineers, Data Scientists and Data Engineers. We also see a significant number of people, who are getting higher salaries, did not mention their job roles and labeled it as other which implies a lack of complete picture. There is no systematic trend observed for non-binary people and their average salaries are higher. Again in this case as well, we see the mean salaries for men is higher than women in almost all the job roles. Based on job roles alone, we see the highest paying job is of a Product Manager with almost 90000 USD. The next high paying jobs are Developer Relations/Advocacy, Project Manager, Data Scientist, Research Scientist and Data Engineer with an average of 50000 USD, as observed in Figure-6.

Question-2:

a. This part focuses on estimating the difference between the mean of males and females. The data for these two groups are isolated in separate dataframes and descriptive statistics are performed. The male participants(12642) are significantly higher than the female participants(2482). And according to statistics, the average salary of male participants is $51,193 (Figure-7) and it is much higher than the average salary of women $34,816 (Figure-8).

b. For performing t-test, an assumption of fitness, mainly normality, is validated. The normality assumption is checked visually where the distribution is right skewed (Figure-9 and Figure-10) and is also tested with *the scipy.stats.shapiro()* function. The null hypothesis that the data are normally distributed is rejected for both datasets. And hence a 2-sample t-test cannot be performed for these datasets.

c. In this part, the bootstrap method is understood to be a way of resampling the original sample and creating dummy samples. I have created a function to do this operation where both male and female datasets are sampled 1000 times. The *np.random.choice()* method is used. A *random seed* is provided inside the *for loop* to keep the

random sampling consistent for every run.The sample mean is recorded for each replicate. I experimented with different sample sizes and finally chose 50% size of both the datasets to bootstrap based on the normality of the distribution in graphs and *shapiro tes*t. The mean distribution for males and females and the distribution of their mean difference are shown in Figure-11 and Figure-12. From the figures it can be seen that both the male and female have a normal mean distribution and their mean difference also has a normal distribution.

d. To conduct a two-sample t-test on bootstrapped data we first check if all the assumptions are met. The method used is the same as before. The assumption of normality is not rejected, but the third assumption of equal variance is rejected using *levene* test. Since there is no homogeneity of variance, ideally a normal t-test can't be performed. However, we can proceed with Welch's t-test that is designed for unequal population variances but the assumption of normality is maintained. The t-test can is performed using *scipy.stats.ttest_ind()* function which does not assume equal variance. After running the test, the p-value is equal to 0, which is less than 0.05, the null hypothesis is rejected. The result of Welch's t-test shows that there is a statistically significant difference in the average salary between men and women.

e. Since p value is 0 , we can say that the mean salary of men is higher than mean salary of women.The bootstrapped mean distribution plots also show that the pay gap between men and women is quite large. The mean salary gap graph shows that the salary difference between the two groups is approximately $16,000. The average salary for women is around $35,000 (Figure-8), the difference between the two groups is almost 40% of the average salary for women. Also I infer that the bootstrap method to compare the mean really helps us find whether there exists a difference between two groups. It can make the distribution into a normal distribution which satisfies the assumption to perform t-test.

Question 3:

a. This final part focuses on the impact of education on income levels, three groups (bachelor, master and PhD) are selected for analysis.The data for these three groups are again isolated in separate dataframes and descriptive statistics are performed for each group. Master's degree participants were highest in number. Participants with a higher educational background tend to have a higher mean and median income. Results can be seen in Figure-13, Figure-14 and Figure-15 for exact values in USD.

b. Since there are now three groups, ANOVA is used instead of the two-sample t-test. Before running the test, the three assumptions are verified.. The same methods are used to test assumptions. It is based on the assumption of independence. All three datasets failed the normality assumption and the homogeneity of variance assumption because the p-value for both tests of the assumptions is less than the threshold of 0.05. Therefore, the ANOVA test is not performed in this case.The figures for distribution of salaries for each group are shown in Figure-16, Figure-17 and Figure-18 where we can see the right skewed distributions.

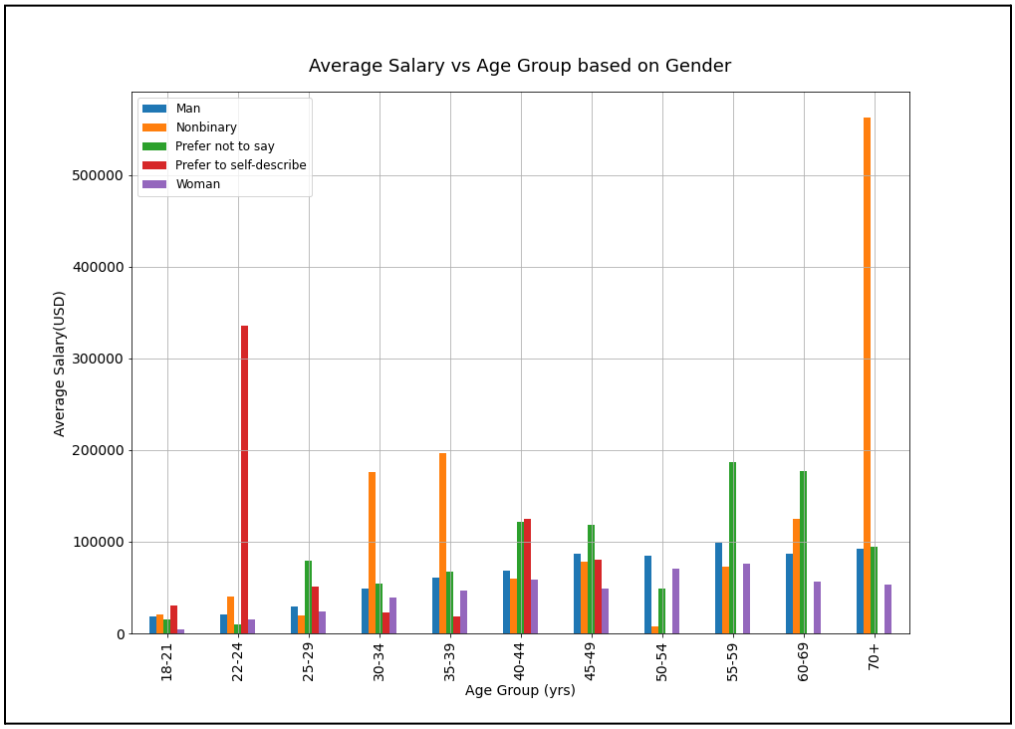
c. Similarly, three groups are selected and resampled using the defined function. Experiments are performed for each of the three datasets for bachelor's, master's, and doctoral earnings to find suitable sample size. Again based on visual judgement for normality and *shapiro* tests, the following sample sizes were selected: 1/10th of bachelor's, 1/8th of master's and 1/4th of doctoral. . In addition, the difference between each of the two groups is calculated. Three histograms are plotted for the bootstrap data in a single plot (Figure-19). The plots show a bell shape, suggesting that the data are now approximately normally distributed. A similar distribution also occurred with bootstrapped difference data (Figure-20, Figure-21 and Figure-22).

d. ANOVA is still not suitable here. We test the assumptions of normality and homogeneous variance. The assumption of independence is assumed. The normality assumption is not rejected, but the third equal variance assumption is rejected. Since the variances are in the same order, the homogeneity of the variances is retained. A simple ANOVA test is still performed to check its implementation. The *scipy.stats.f_oneway()* function is used to perform the one-way ANOVA test. The ANOVA result shows that there is a statistically significant difference in mean salary between the three groups

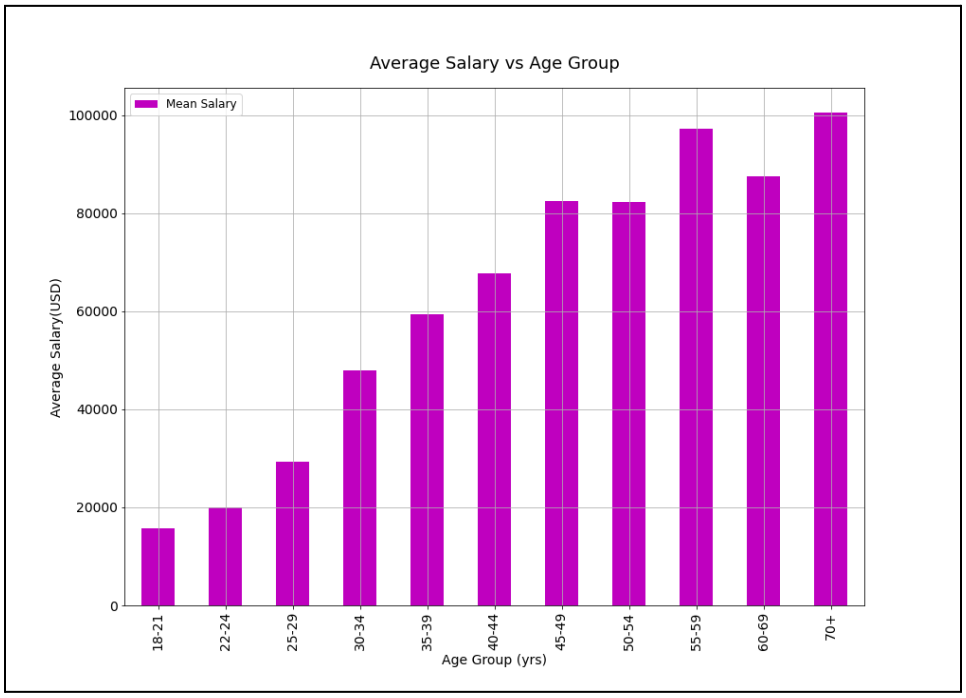
e. Due to p value being 0, it can be concluded that there is a significant difference between the salaries when two groups among the three are compared individually. From the plot Figure-19, we see that the mean salary of a doctoral degree is more than master's degree which in turn is more than bachelor's degree. In addition, I also infer from this process that bootstrapping data can only help us to achieve normal distribution of the mean of the salary data, but it does not help us to achieve equal variances.

APPENDIX

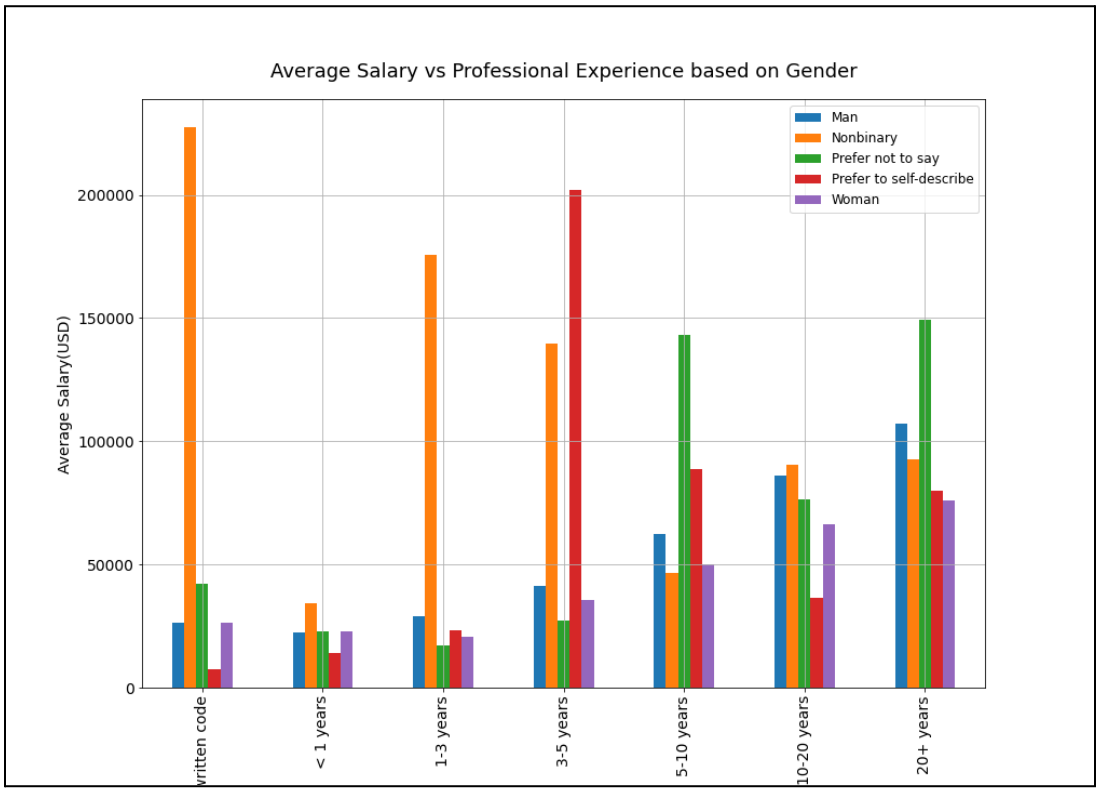
1. Figure-1: Average Salary vs Age Group based on Gender



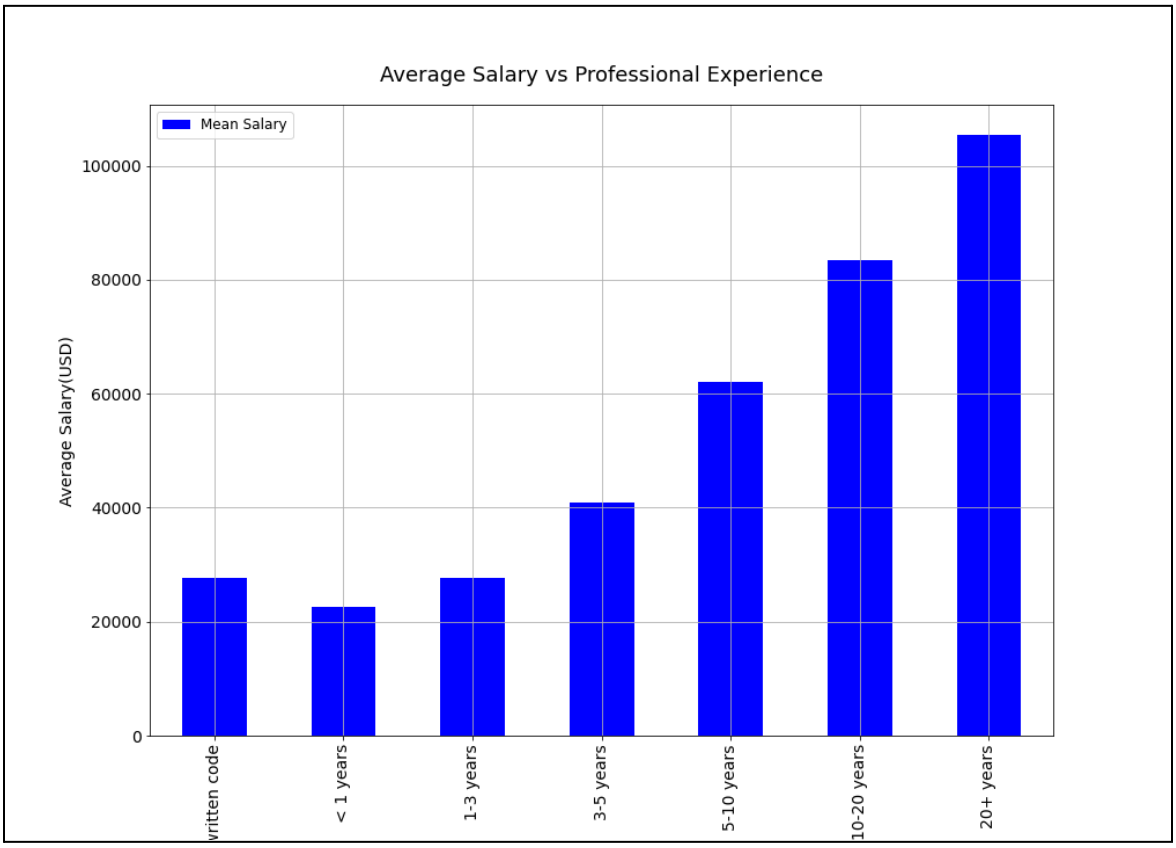
2. Figure-2: Average Salary vs Age Group



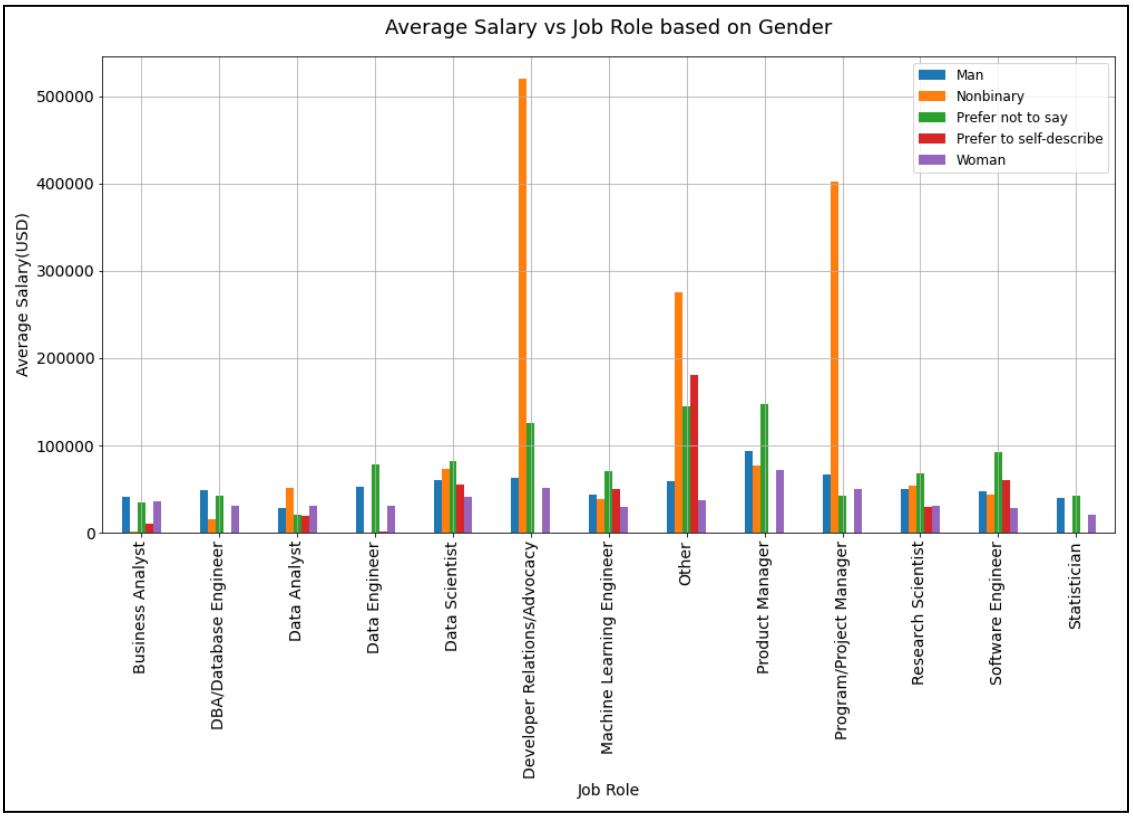
3. Figure-3: Average Salary vs Professional Experience based on Gender



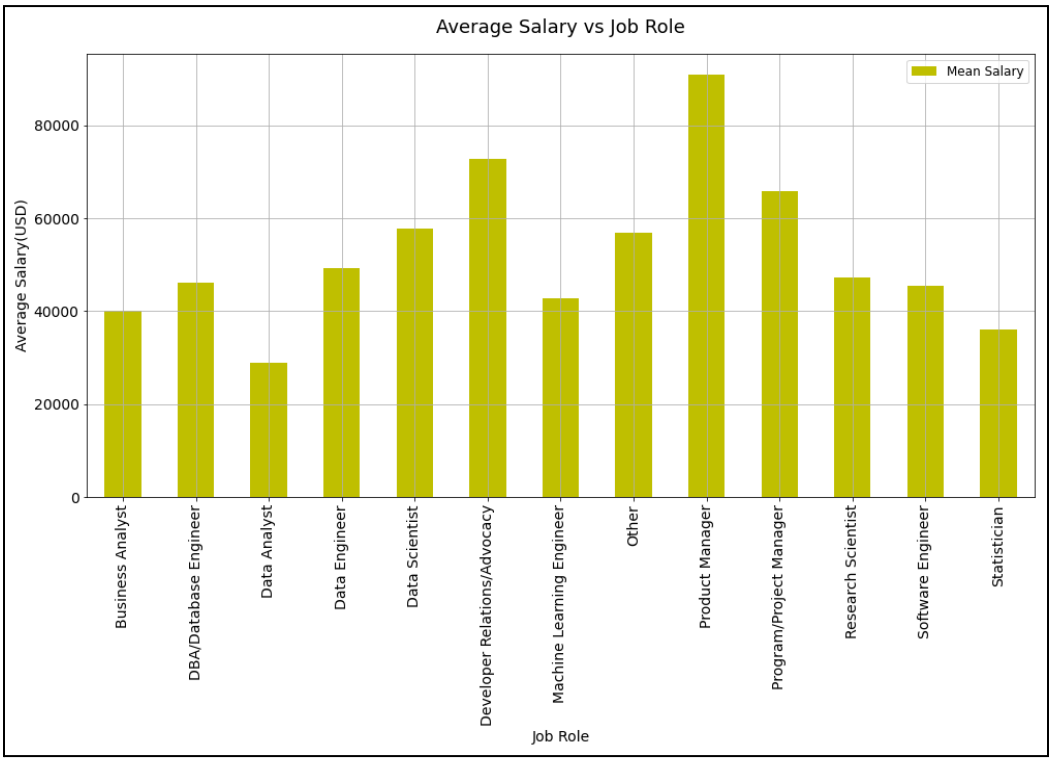
4. Figure-4:Average Salary vs Professional Experience



5. Figure-5: Average Salary vs Job Role based on Gender



6. Figure-6: Average Salary vs Job Role



7. Figure-7: Descriptive Statistics for Men Group and their Salaries

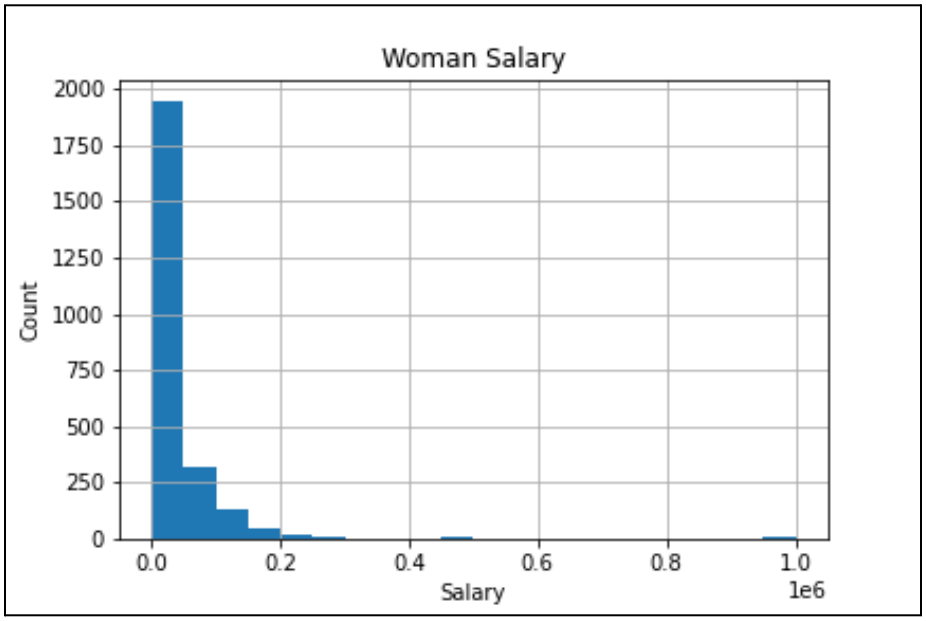
	Salary
count	12642.000000
mean	51193.600696
std	99979.274378
min	1000.000000
25%	2000.000000
50%	20000.000000
75%	60000.000000
max	1000000.000000

8. Figure-8: Descriptive Statistics for Women Group and their Salaries

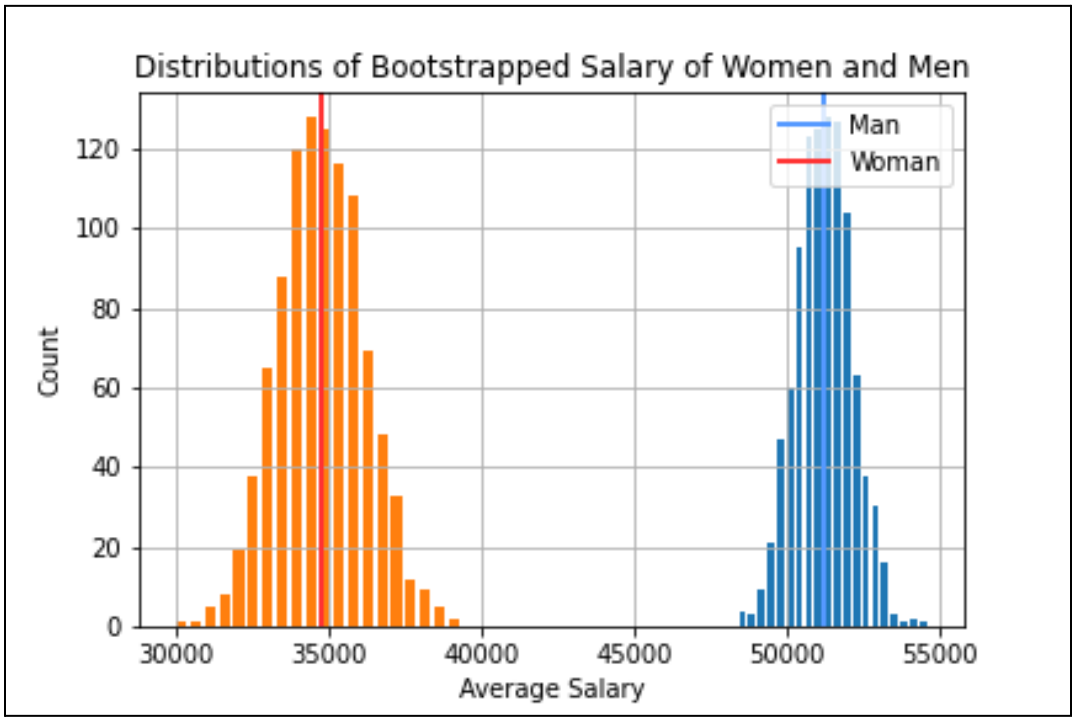
	Salary
count	2482.000000
mean	34816.881547
std	72017.347888
min	1000.000000
25%	1000.000000
50%	7500.000000
75%	50000.000000
max	1000000.000000

9. Figure-9: Distribution for Men's Salaries (Original data)

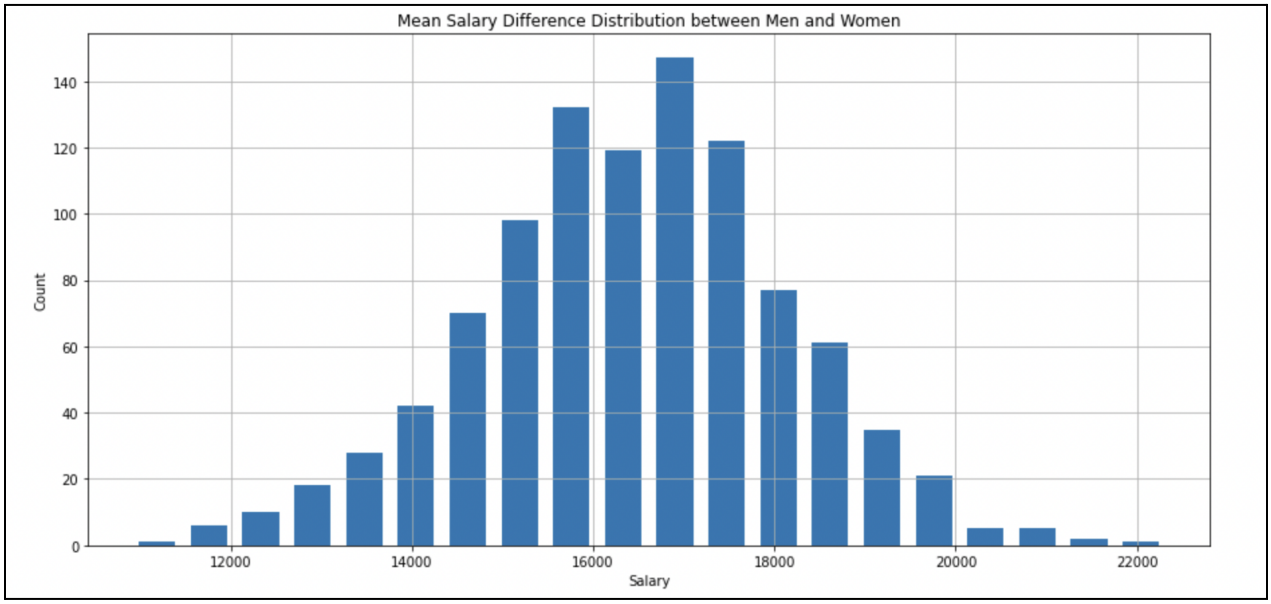
10. Figure-10: Distribution of Women's Salaries(Original Data)



11. Figure-11: Distributions of Bootstrapped Salary of Women and Men



12. Figure-12: Mean Salary Difference Distribution between Men and Women



13. Figure-13: Descriptive Statistics for Bachelor's Degree Salaries

	Salary
count	4777.000000
mean	35578.291815
std	89382.060777
min	1000.000000
25%	1000.000000
50%	7500.000000
75%	40000.000000
max	1000000.000000

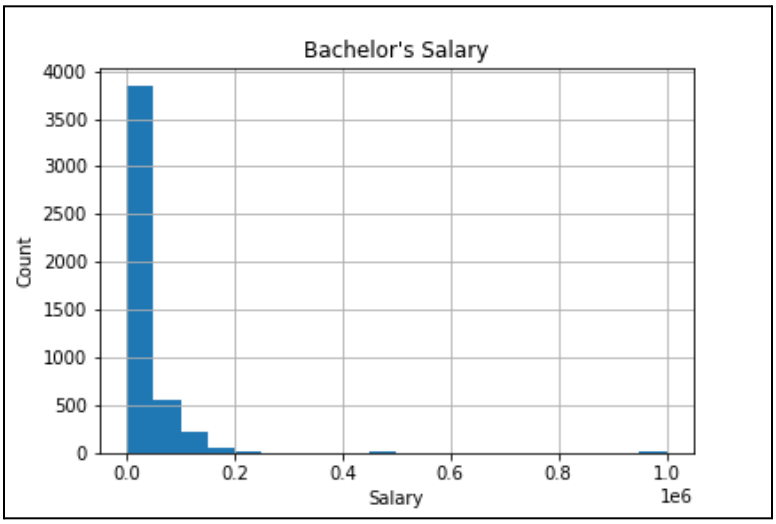
14. Figure-14: Descriptive Statistics for Master's Degree Salaries

	Salary
count	6799.000000
mean	52706.868657
std	90928.786678
min	1000.000000
25%	3000.000000
50%	25000.000000
75%	70000.000000
max	1000000.000000

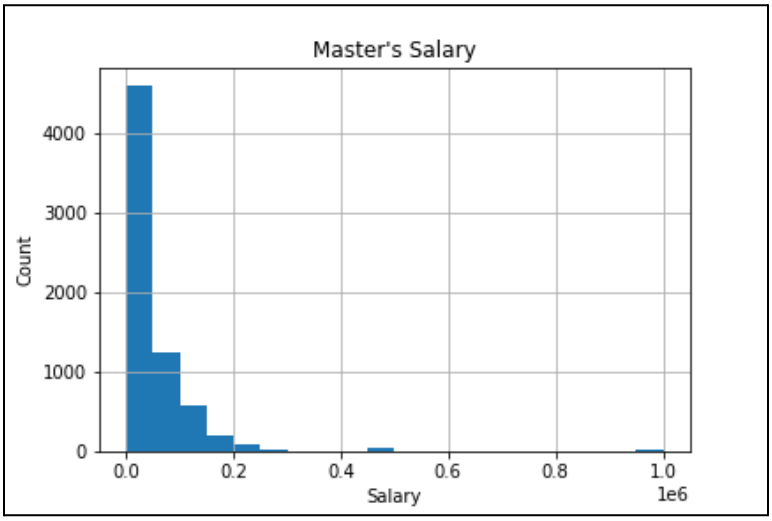
15. Figure-15: Descriptive Statistics for Doctoral Salaries

	Salary
count	2217.000000
mean	70641.181777
std	117160.947589
min	1000.000000
25%	4000.000000
50%	40000.000000
75%	90000.000000
max	1000000.000000

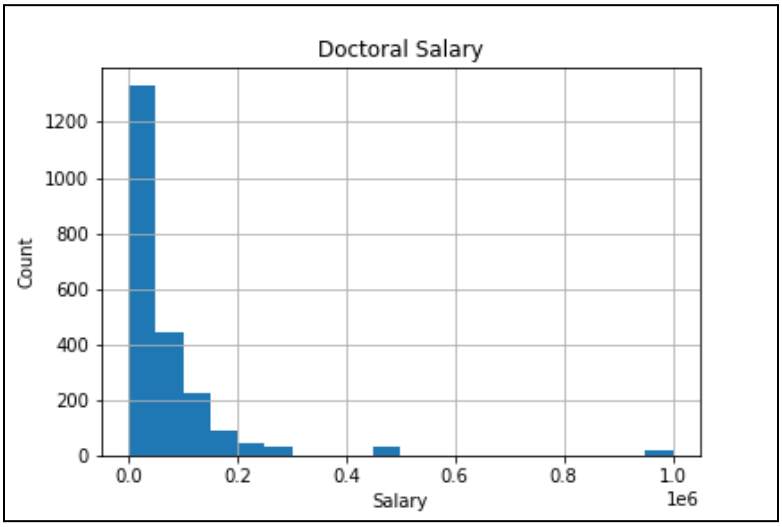
16. Figure-16: Distribution of Bachelor's Salaries



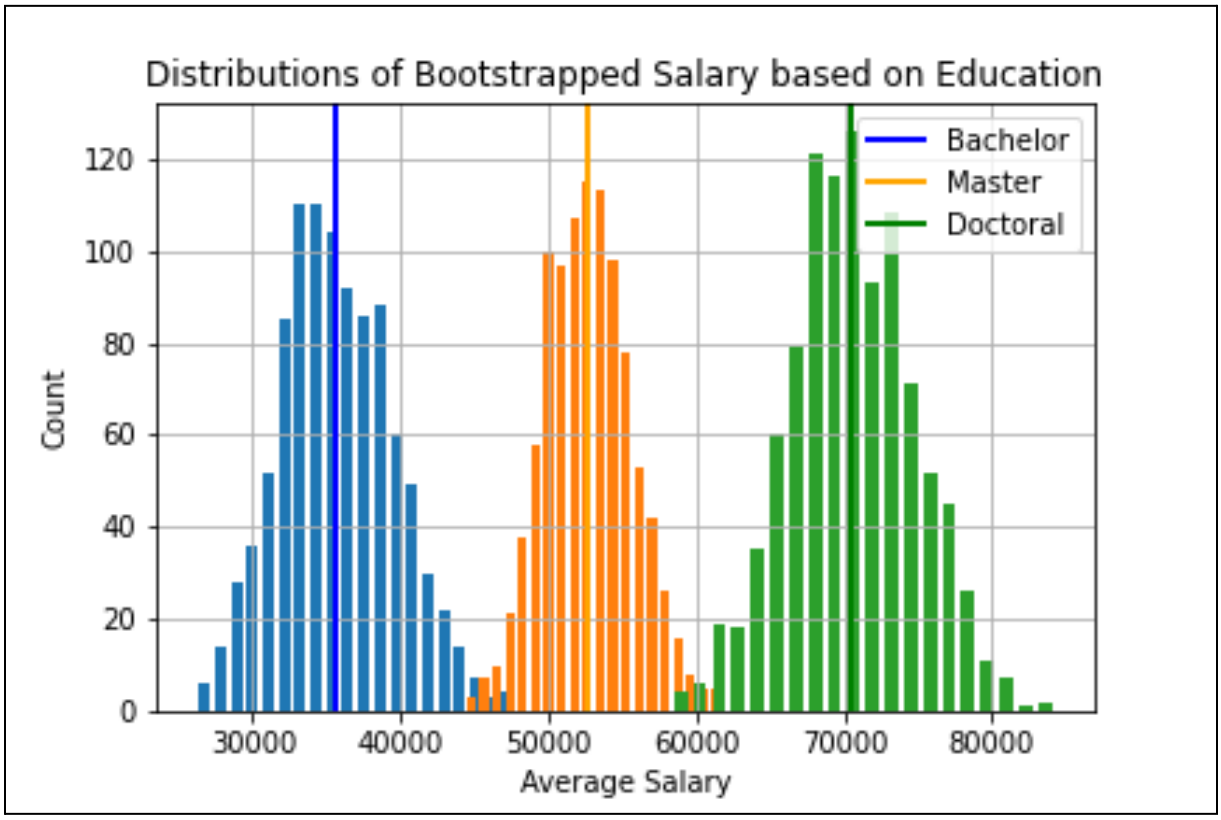
17. Figure-17: Distribution of Master's Salaries



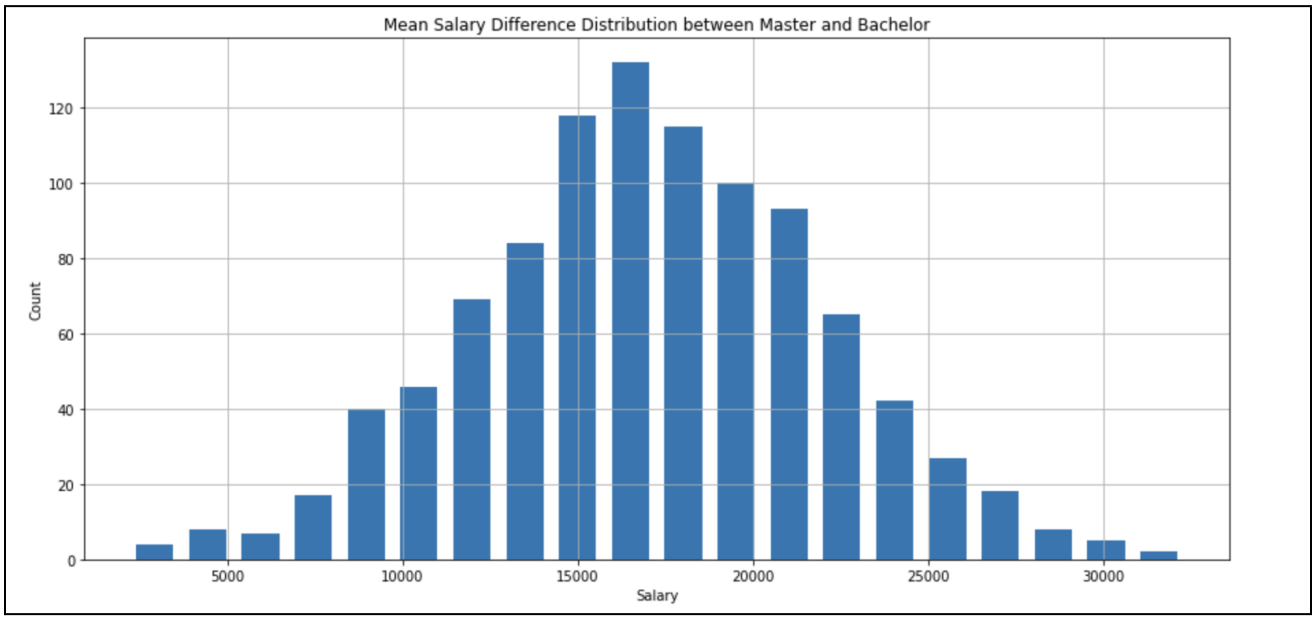
18. Figure-18: Distribution of Doctoral Salaries



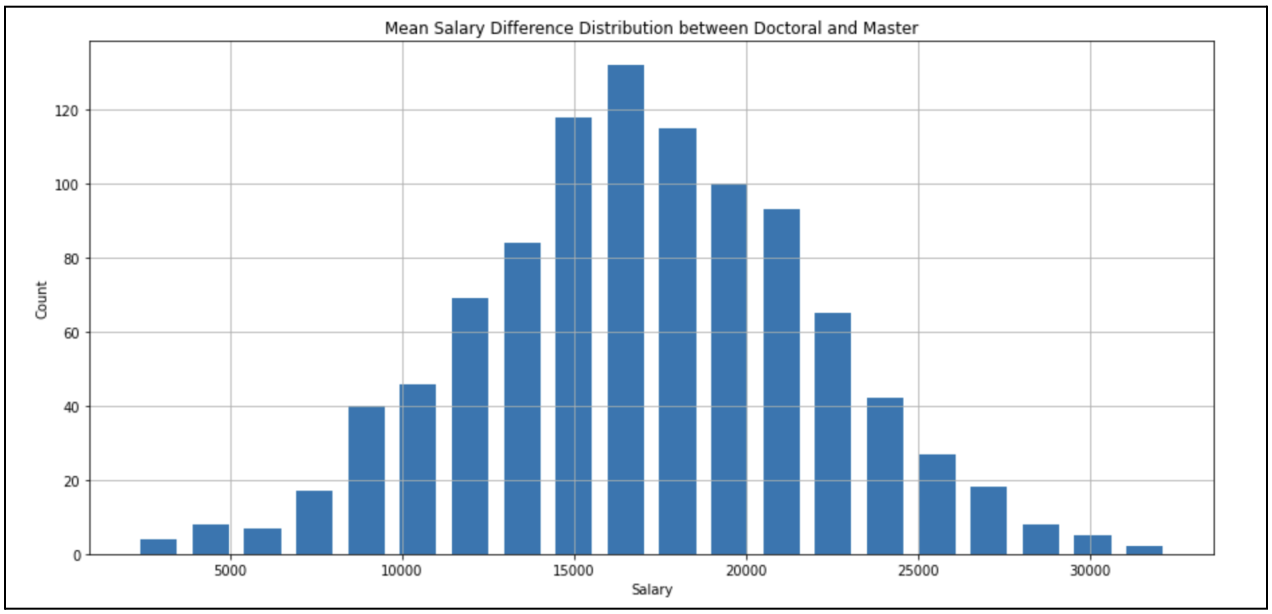
19. Figure-19: Distributions of Bootstrapped Salary based on Education



20. Figure-20: Mean Salary Difference Distribution between Master and Bachelor



21. Figure-21: Mean Salary Difference Distribution between Doctoral and Master



22. Figure-22: Mean Salary Difference Distribution between Doctoral and Bachelor