MIE 1624 Introduction to Data Science and Analytics – Winter 2023

Assignment 1

Anbumanivel Mohan Suganthi - 1008696653

February 12, 2023

The aim of this assignment is to explore the survey to understand the nature of women's representation in Data Science and Machine Learning and the effects of education on income level.

Question 1

To analyze the dataset, exploratory data analysis was conducted. There are 15391 entries and 369 features in the dataset. Many of the features have missing values, and there are many parts to some of the features. For this question, I have chosen three features that do not have any missing values. These features are the highest level of formal education (Q4), coding experience (Q6), and current role (Q5). These three factors are examined with yearly compensation (Q25).

i). Based on the table of mean salaries for different educational levels, a bar graph is created between education level and salary. From this bar graph, people with doctoral degrees have the highest mean yearly compensation among others. It is also noted that people with professional degrees have a higher income level.

ii). The effect of programming experience on salary is explored by plotting a bar graph on the mean salary of different experience levels. The salary increases as the level of experience increases. However, people with no coding experience have a higher average income than those with less than one year of experience.

iii). A box plot is plotted between the current role and salary. Product managers have the highest 25 percentile, median, and 75 percentile salaries among all other people. Program/Project manager has the second highest 25 percentile, median, and 75 percentile salary.

Question 2

a. This part focuses on estimating the difference between the average salaries of men and women. There are other gender groups in the dataset other than men and women, which are discarded for this assignment. Based on the descriptive statistics, the dataset is unbalanced as the sample size of men is five times larger than women. From the statistics, the mean salary of men is $51,193 which is much higher than women $34,816.

b. The two-sample t-test requires certain assumptions to be met. They must be independent, the data must have a normal distribution, and the variance must be equal. The first assumption is tested during the survey. A histogram is plotted to test the normal distribution, it is seen that the distribution is right-skewed for both incomes of men and women. Additionally, the Shapiro-wilk approach utilizing the scipy.stats.shapiro() function is used to test the normality assumption. For both genders, the null hypothesis that the data is normally distributed is rejected. The assumption of equal variance can be

tested by the Levene method using scipy.stats.levene() function. As a result, the homogeneity of variance is rejected. Since both assumptions are not satisfied, the test cannot be performed.

C. Bootstrapping is done by resampling the data to produce many simulated samples. The male and female dataset is bootstrapped 1000 times. The mean of the sample is noted each time. The mean distribution for men and women and the distribution of the difference in means are plotted. The graph shows that both the distribution of men and women is normal. A normal distribution can be seen in both the mean graph and the difference graph.

d. The aforementioned conditions must be met in order to execute a t-test on the bootstrapped data. The normality assumption is tested in addition to the distribution plot using the same procedure as before. For both genders, the null hypothesis that the data is normally distributed is not rejected. But the homogeneity of variance is rejected. The t-test can still be performed by the scipy.stats.ttest_ind() function, as it does not assume equal variance. Welch's t-test results with a p-value equal to 0, which is less than the threshold of 0.05, and the null hypothesis is rejected. It shows that there is a significant difference in average salary between men and women.

e. The bootstrapped mean distribution plot shows that the average salary of women is around $35,000 and for men, it is around $51,000. The difference between them is around $16,000. From the t-test result, it is found that the male and female salaries are not equal.

Question 3

a. This part focuses on the effect of education on income levels. Three groups (Bachelor's, Master's, and Doctoral degree) are selected for this analysis. The descriptive statistics show that people with higher formal education tend to have higher salaries.

b. The assumption for the one-way ANOVA test is the same as the t-test. The same methods are used for validating the assumptions. All three datasets failed to satisfy the normality and homogeneity of variance assumption. Since the p-value for both assumptions is less than the threshold value of 0.05. Therefore, the ANOVA test is not performed.

C. Bootstrapping is performed on the three groups of data. The dataset is resampled and bootstrapped 1000 times. The sample means are noted each time. The mean distribution and the distribution of the difference in means are plotted. From the graph, bachelor's, master's, and doctoral salaries have a normal distribution. The difference in the mean graph also shows a similar distribution.

d. To perform the ANOVA test on the bootstrapped data, above mentioned assumptions must be satisfied. In addition to the distribution plot, the normality and homogeneity of variance assumption are tested. The null hypothesis that the data is normally distributed is rejected. But the homogeneity of variance is not rejected. The ANOVA can still be performed by the scipy.stats.f_oneway () function, as it can tolerate non-normal data. The one-way ANOVA results with a p-value equal to 0, which is less than the 0.05 threshold value, and the null hypothesis is rejected. It shows that there is a significant difference in the average salary between the three groups.

e. According to the bootstrapped mean distribution plot, the average salaries for bachelor's, master's, and doctoral degrees are around $35,000, $52,000, and $70,000 respectively. From the distribution plots and the test results, the people with higher education levels have higher salaries.

Appendix

Table 1: Summary table of salary for different education level

	Salary							
	count	mean	std	min	25%	50%	75%	max
Education								
Bachelor's degree	4777.0	35578.291815	89382.060777	1000.0	1000.0	7500.0	40000.0	1000000.0
Doctoral degree	2217.0	70641.181777	117160.947589	1000.0	4000.0	40000.0	90000.0	1000000.0
I prefer not to answer	334.0	34191.616766	113660.692249	1000.0	1000.0	4000.0	25000.0	1000000.0
Master's degree	6799.0	52706.868657	90928.786678	1000.0	3000.0	25000.0	70000.0	1000000.0
No formal education past high school	228.0	38208.333333	100811.090707	1000.0	1000.0	7500.0	50000.0	1000000.0
Professional doctorate	290.0	67465.517241	136718.387541	1000.0	2000.0	20000.0	80000.0	1000000.0
Some college/university study without earning a bachelor's degree	746.0	41990.616622	110270.037272	1000.0	1000.0	10000.0	50000.0	1000000.0

Figure 1: Barplot of Education vs Salary

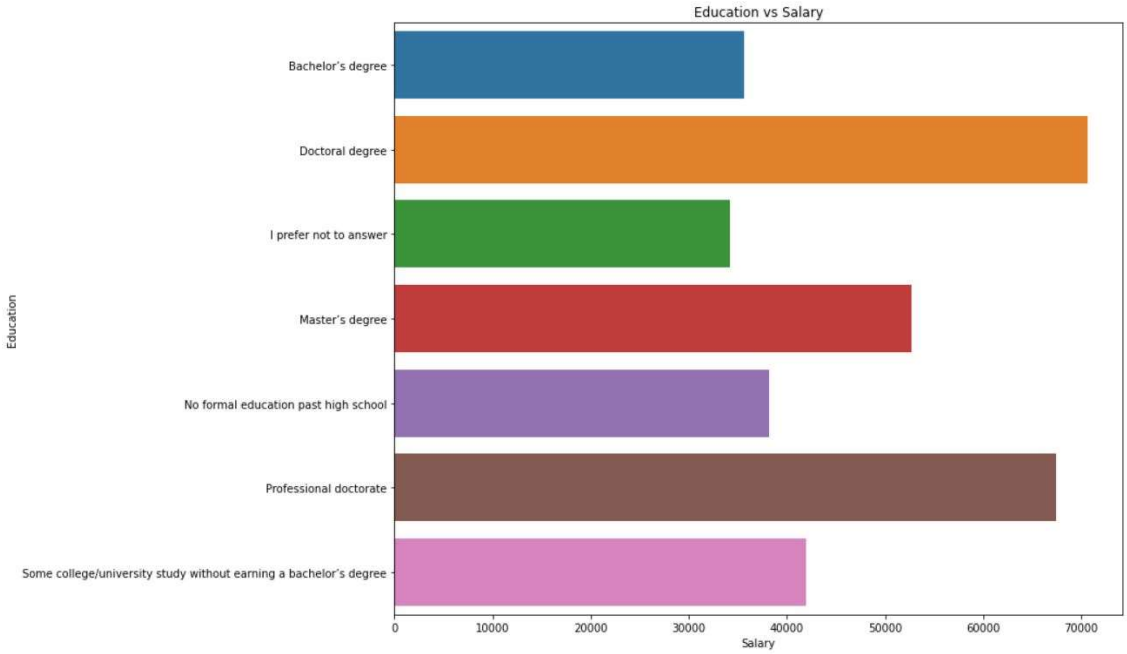

Table 2: Summary table of salary for different coding experience

	Salary							
	count	mean	std	min	25%	50%	75%	max
Experience								
1-3 years	3691.0	27763.885126	71454.037904	1000.0	1000.0	5000.0	30000.0	1000000.0
10-20 years	1846.0	83355.904659	119174.584809	1000.0	15000.0	55000.0	100000.0	1000000.0
20+ years	1624.0	105358.682266	144497.878806	1000.0	20000.0	70000.0	125000.0	1000000.0
3-5 years	2461.0	40831.369362	88867.577602	1000.0	2000.0	15000.0	50000.0	1000000.0
5-10 years	2345.0	62156.716418	98771.984542	1000.0	7500.0	30000.0	80000.0	1000000.0
< 1 years	2463.0	22604.141291	64817.238338	1000.0	1000.0	4000.0	20000.0	1000000.0
I have never written code	961.0	27651.404787	70575.009746	1000.0	1000.0	5000.0	30000.0	1000000.0

Figure 2: Barplot of Programming Experience vs Salary

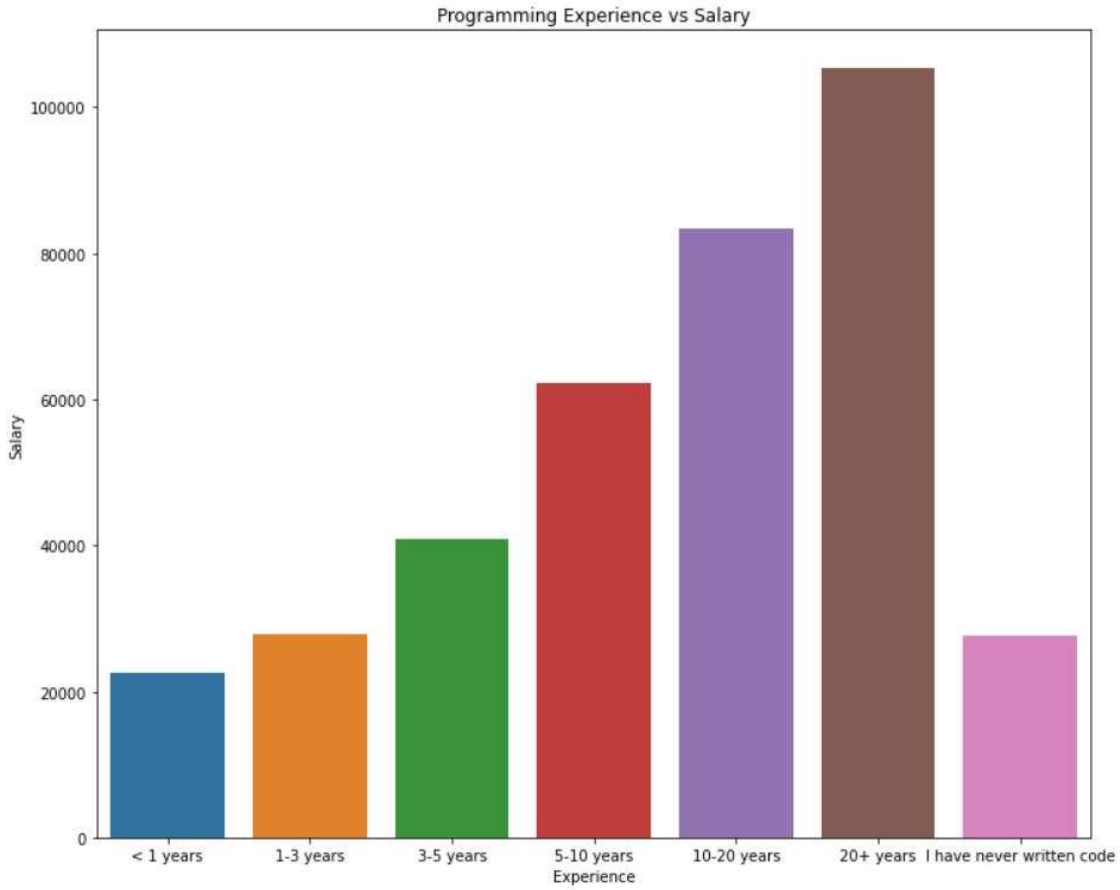

Table 3: Summary table of salary for different role

	Salary							
	count	mean	std	min	25%	50%	75%	max
Current Position								
Business Analyst	885.0	39983.050847	79124.960207	1000.0	3000.0	15000.0	50000.0	1000000.0
DBA/Database Engineer	151.0	46023.178808	59467.227326	1000.0	4000.0	20000.0	60000.0	250000.0
Data Analyst	2039.0	28827.856793	56755.093315	1000.0	1000.0	7500.0	40000.0	1000000.0
Data Engineer	597.0	49226.968174	87071.250580	1000.0	2000.0	20000.0	60000.0	1000000.0
Data Scientist	3240.0	57771.296296	106170.740600	1000.0	2000.0	20000.0	80000.0	1000000.0
Developer Relations/Advocacy	86.0	72656.976744	171850.127276	1000.0	1000.0	15000.0	60000.0	1000000.0
Machine Learning Engineer	1327.0	42787.490580	101921.189100	1000.0	1000.0	5000.0	50000.0	1000000.0
Other	2204.0	56941.923775	115403.732022	1000.0	3000.0	20000.0	70000.0	1000000.0
Product Manager	285.0	90877.192982	143369.772112	1000.0	10000.0	50000.0	125000.0	1000000.0
Program/Project Manager	784.0	65728.954082	102745.212731	1000.0	7500.0	40000.0	90000.0	1000000.0
Research Scientist	1404.0	47183.048433	80804.503293	1000.0	3000.0	20000.0	60000.0	1000000.0
Software Engineer	2110.0	45505.450237	101319.783478	1000.0	2000.0	15000.0	60000.0	1000000.0
Statistician	279.0	35992.831541	78035.974126	1000.0	1000.0	5000.0	40000.0	1000000.0

Figure 3: Boxplot of Programming Experience vs Salary

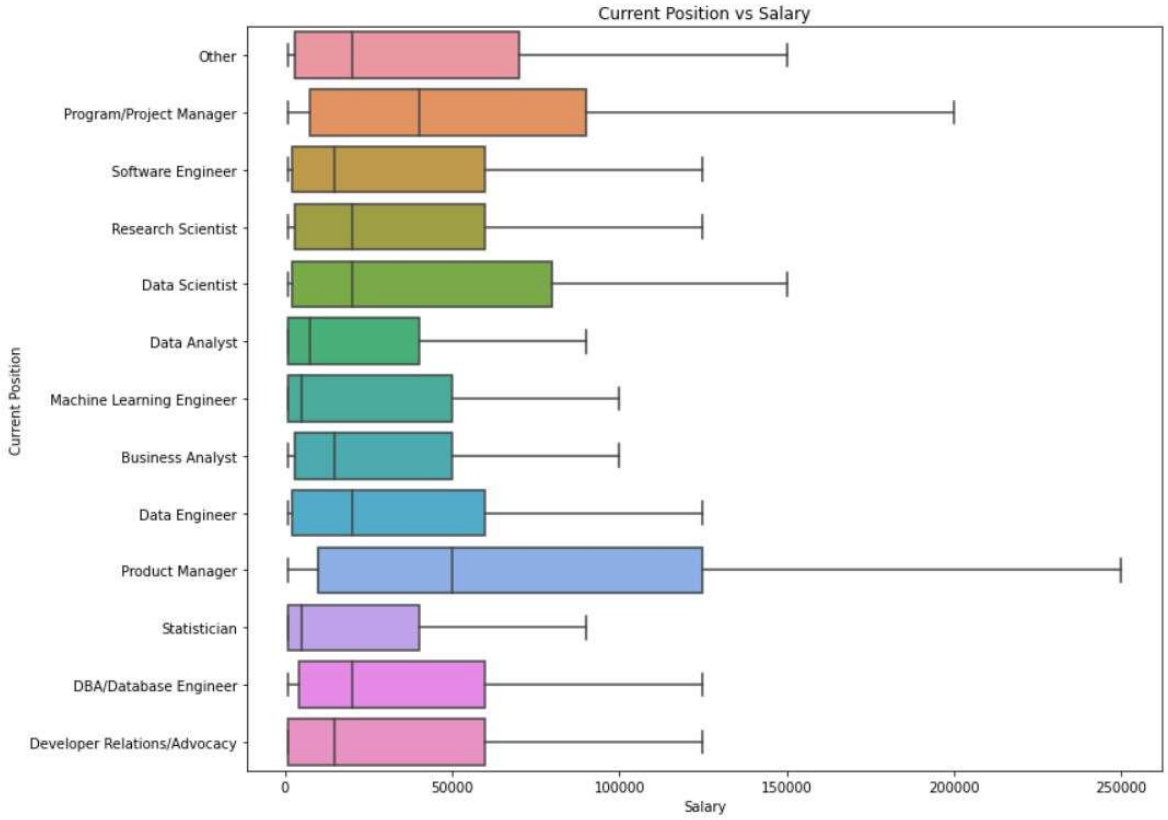

Table 4: Descriptive statistics for men

	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

Table 5: Descriptive statistics for women

	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

Figure 4: Distribution plot of men and women salaries

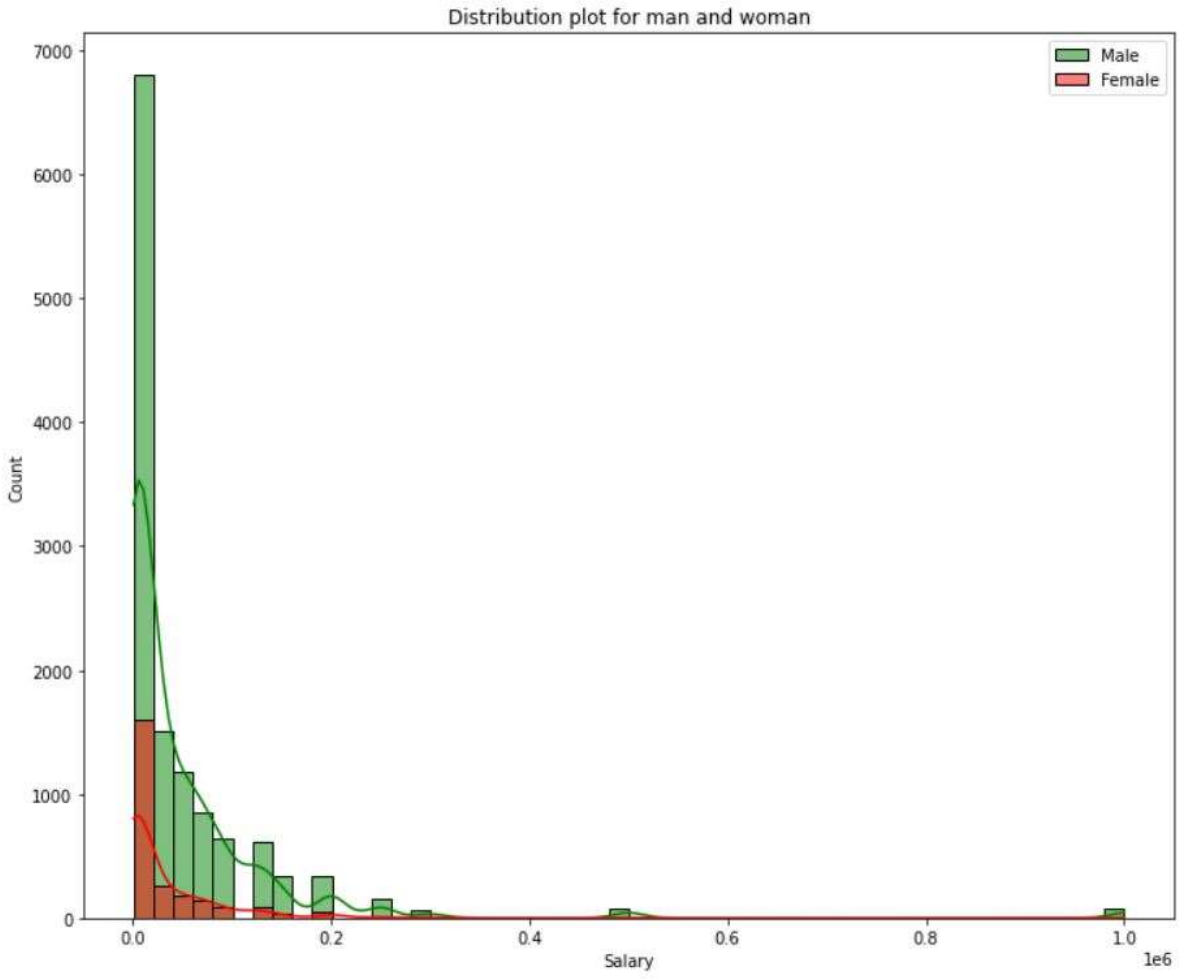

Figure 5: Distribution plot of bootstrapped men and women salaries

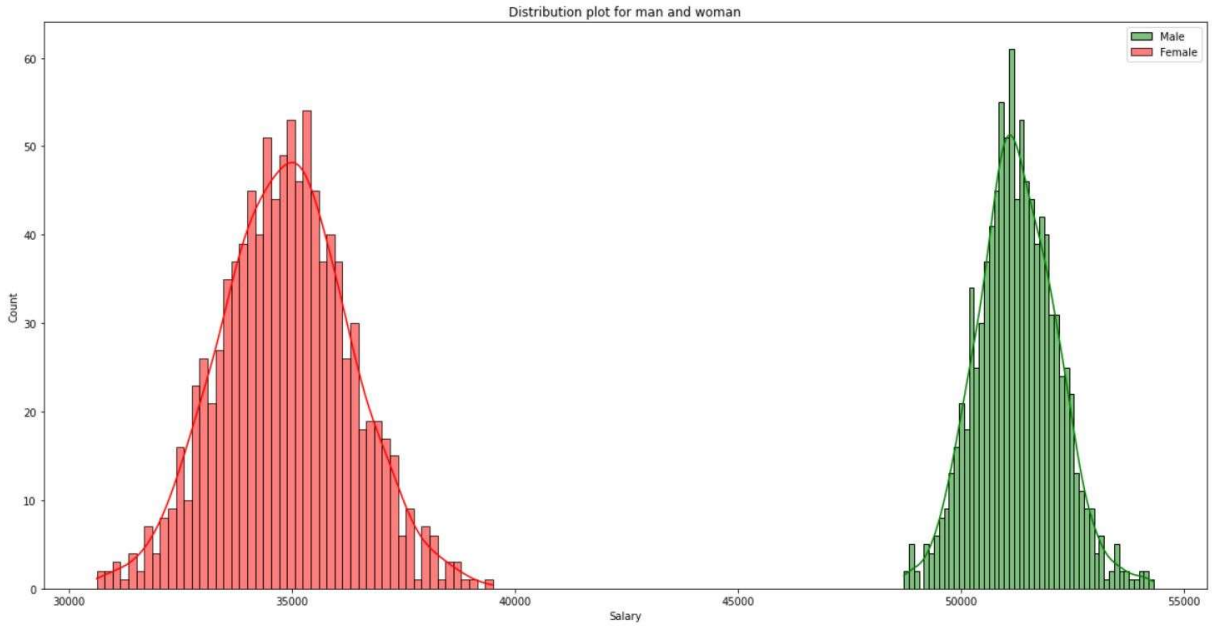

Figure 6: Distribution plot of difference in means of men and women salaries

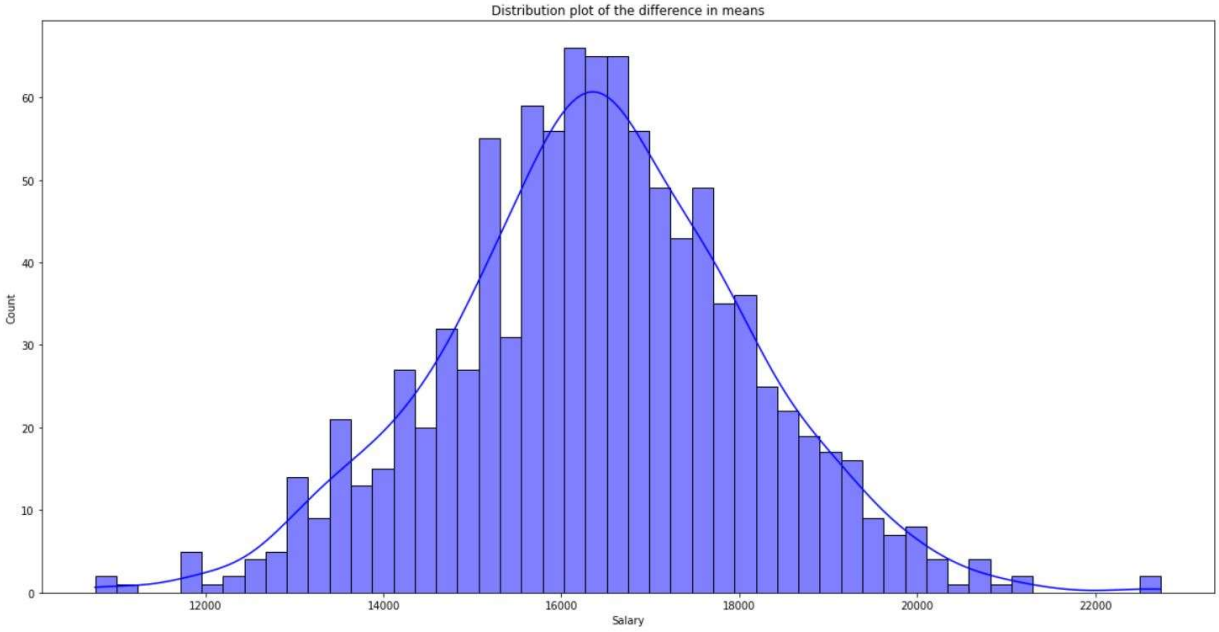

Table 6: Descriptive statistics for bachelor's degree

	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

Table 7: Descriptive statistics for master's degree

	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

Table 8: Descriptive statistics for doctoral degree

	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

Figure 7: Distribution plot bachelor's, master's, and doctoral degree salaries

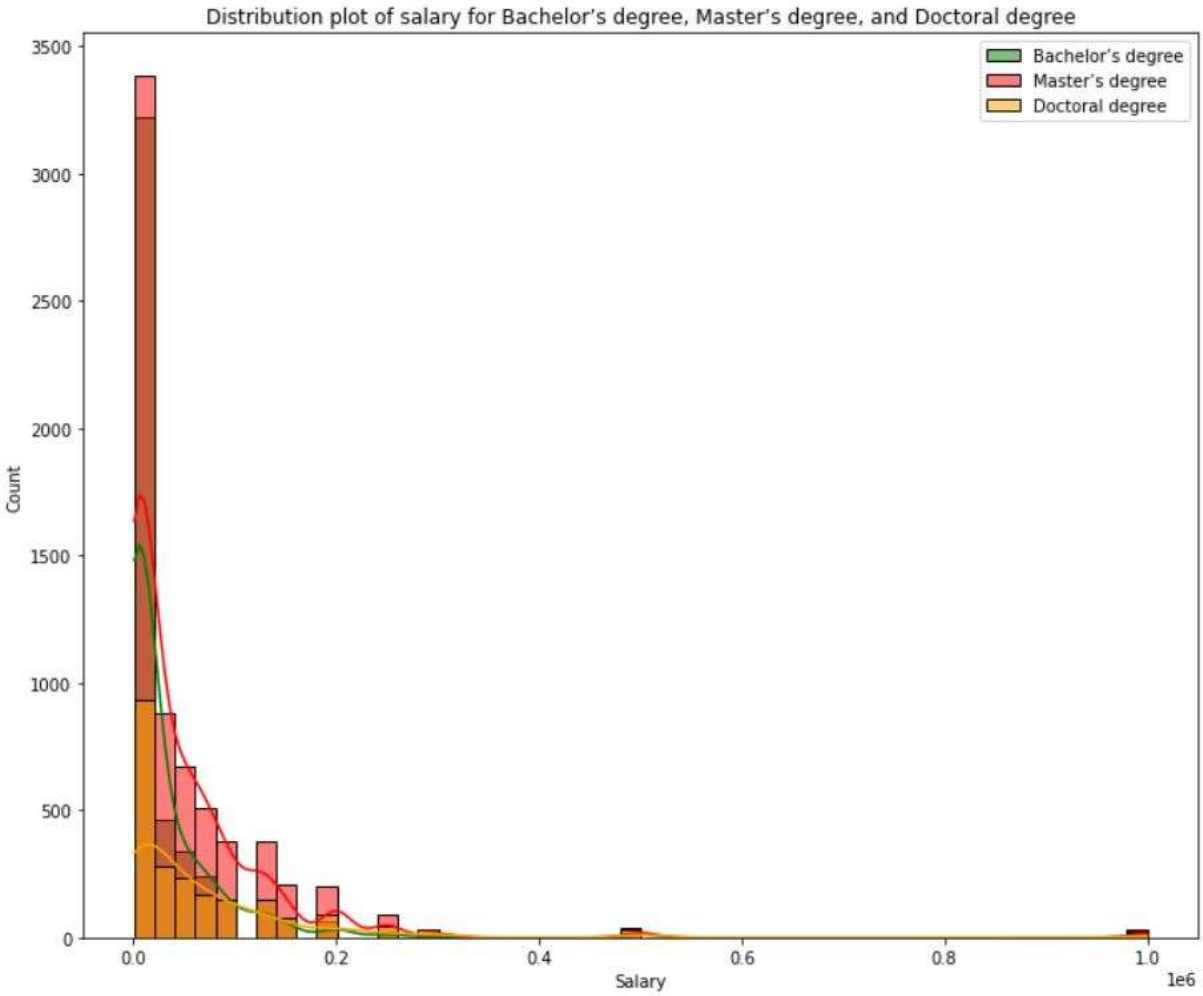

Figure 8: Distribution plot of bootstrapped bachelor's, master's, and doctoral degree salaries

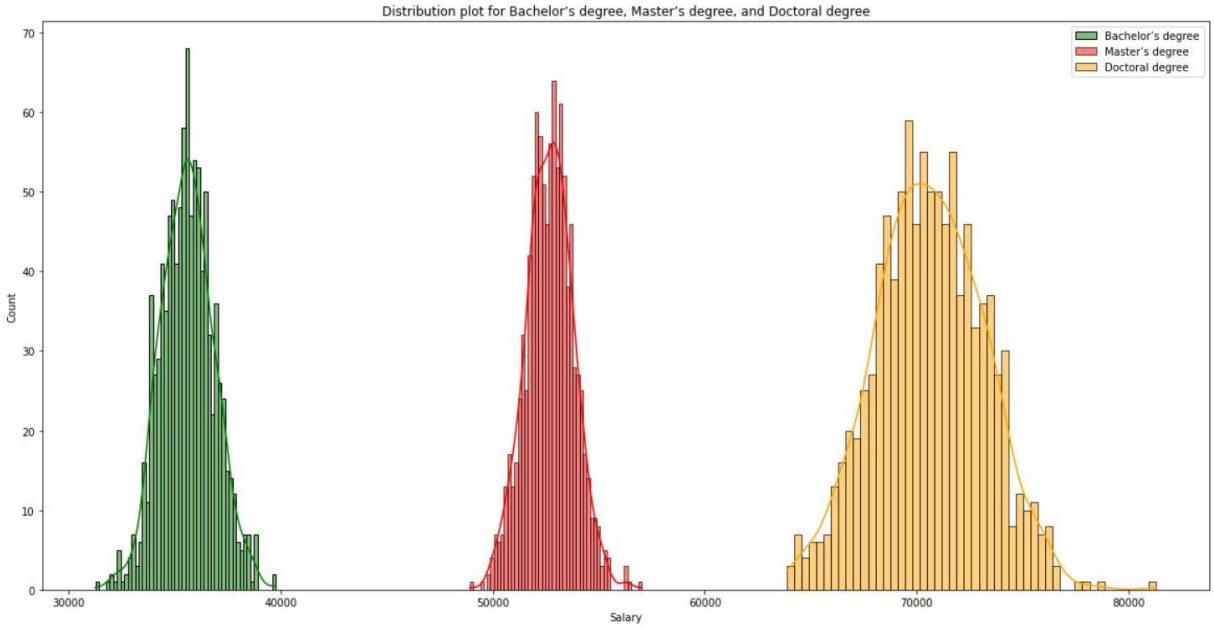

Figure 9: Distribution plot of difference in means of bachelor's, master's, and doctoral degree salaries