MIE 1624: A1 Report

Date: 2023-02-12
Student: William Hazen
Student ID: 1009231225
Institution: University of Toronto
Faculty: Department of Mechanical & Industrial Engineering

Objective

This assignment aims to explore the survey data to understand (1) the nature of women's representation in Data Science and Machine Learning and (2) the effects of education on income level. Column Q25 "What is your current yearly compensation (approximate $USD)?" is the targeted column.

1 Data Exploration

The initial exploration of the clean Kaggle survey dataset was conducted through different graphical figures that brought insight into some of the characteristics of the survey. The categories chosen were Age (Q1), Country (Q3), and Experience (Q5) vs the yearly compensation (Q25) in $USD. Figures 1 and 2 represent salary based on age. Figure 1 is a boxplot which provides clear insights into the median and spread of each category and Figure 2 is a pandas DataFrame with general descriptive statistics. I decided to discard the outliers in the plot due to scaling, as shown in Figure 2; each category has a maximum value of 1 million, which may be inaccurate. Therefore, the information in the box and whiskers provide more insight. Using both figures, there appears to be a positive linear trend with yearly compensation and age – being on average, salary increases with age. Figures 3 and 4 show the descending average salary for each country. Looking at Figure 3, Switzerland and the USA have the highest average salaries ($144548, $143463) with a difference of $\approx$ $14000 to the third place, Israel ($129322). The country with the lowest average salary was Ethiopia ($4633), which is substantially different from the highest. The last figure (Figure 4) shows the average salary based on experience or job title in the form of a boxplot. Figures 3 and 4 show that product managers have the highest salary, with the mean and median being $90877 and $50000 respectively, in addition to the highest maximum whisker being $250000. Interestingly, the dataset shows that Data Scientists and Data Engineers make more than Software Engineers on average, and Data analysts have the lowest salary on average.

2 The Disparity in Average Salary between Men and Women

2a) Initially, I explored the gender category using a boxplot (Figure 7) and a pandas DataFrame with descriptive statistics using pd.describe() (Figure 8). The results showed large differences in salary. Figure 7 shows that the median salary for men is $20,000 and $7,500 for women. The mean salary for men was $51,193, and for women, the average salary was $34,816. Additionally, the salary spread and maximum whisker for men were larger. It is important to note that these statistics are using unbiased estimators, as their expected values are equal to the true population parameter being estimated, which I believe is more applicable to this dataset.

2b) I was unable to perform a two-sample t-test as several assumptions required for this test failed. To check if the men and women data were normally distributed, I used quantile-quantile plots (QQ) and the normality function from the pingouin library. The QQ plot of the normalized data should be approximately straight, however, Figure 9 shows non-normalized behaviour and the pingouin normality function indicated that both the men and women data were not normalized. Moreover, the t-test requires homoscedasticity in the variance. To check if the variances were similar, I conducted a Bartlett test. The p-value was $<< 0.05$, indicating a rejection of the null hypothesis and thus concluding that the variances were significantly different. Lastly, the sample size for each dataset was largely different, with 12642 men and 2482 women. Therefore, the t-test could not be conducted at this time.

2c) To account for non-normally distributed data, bootstrapping was an appropriate approach to obtain a more accurate estimate of the statistics for salary differences between men and women. After 1000 replications using

a sample size of 40% for each respective dataset, the result is shown in Figure 10. Both men and women have a bell-shaped curve with a mean (red vertical line) of $51263 and $34782 respectively. Notably, the bootstrapped averages are close to the original means of $51193 and $34816, suggesting that the sample mean is a good representation of the population mean. Figure 11 shows the difference in salary, which also appears to be normally distributed with a mean value of $16480. This difference implies that, on average, there is a $16480 gap in salary between men and women.

2d) Repeating the steps used in 2b) to check for normality, the bootstrapped data QQ plots showed normal behaviour (Figure 12) as well as bell-shaped histograms in Figure 10.The Bartlett test was used to check for variance, and the p-value was $<< 0.05$, indicating that the variance in each dataset was different. Therefore, a t-test could not be used, as the homoscedasticity assumption in variance was false. However, since the bootstrapped data was normally distributed and had equal sample size, the Welch's test was appropriate, as it accounts for differences in variance. The p-value obtained was 0, indicating that there is a notable difference between the datasets and it does not occur by chance. Thus, the null hypothesis can be rejected.

2e) After analyzing the salary data for men and women in the Kaggle survey, I found that, on average, there is a $16,480 difference in salary. This indicates that there is a disparity between salaries. To gain further insight, further analysis should be conducted to account for other gender identification options in the survey.

3 The Difference in Average Salary based on Formal Education

3a) The boxplot in Figure 13 illustrates the median and spread of each formal education degree (Bachelor, Master, and Doctoral). Figure 14 contains the unbiased statistics for the datasets, showing that the mean salaries for the respective degrees are $35578, $52706, and $70641. Notably, the Doctoral degree has the highest average salary. However, it also has the highest standard deviation of $117160, which is quite substantial. There is a linear trend of higher levels of education correlating with higher average salaries. It is worth noting that there was a high population of users with Master's degrees compared to Bachelor's and Doctoral.

3b) I used QQ plots and the normality function to check if the data was normally distributed. The results showed non-normal behaviour. Additionally, the variance/std for each degree was not equal, and the sample sizes for each group were largely different. Therefore, the ANOVA test is not a reliable metric at this time.

3c) After bootstrapping the datasets, I combined the histograms for each degree into one plot (see Figure 16). There is a clear difference in salaries based on education level, with bachelor's degrees having considerably lower salaries than those with master's and doctoral degrees. The means for the bootstrapped data were $35527, $52773, and $70852, which were quite similar to the original means, indicating that the sample mean is a good estimate of the population means. Figure 17 shows the difference in salaries between each degree, and Figure 18 shows the overlay differences. Notably, the difference between bachelor's and master's degrees was similar to the difference between master's and doctoral degrees, being approximately $17500. However, the difference between bachelor's and doctoral degrees was $35325 (see Figure 18). Lastly, the mean difference between all the degrees was $70651, suggesting that higher levels of education generally imply higher salaries.

3d) Figure 19 shows that the bootstrapped data is normalized. The Bartlett test reveals that the sample sizes are the same, but the variance differs between the groups, with a p-value $<< 0.05$. This means that the homoscedasticity assumption fails. Therefore, the alternative approach is Welch's ANOVA test, provided by the pingouin library. The result is a p-value $= 0$, which indicates that the null hypothesis can be confidently rejected, indicating that there is a significant difference in the means of the 3 groups.

3e) From 3c) the average difference in salary between formal education degrees was $\approx$ $17453. Moreover, from the histogram plots and Welch ANOVA test, it is clear that there is a difference in salaries based on education levels. Further investigation into non-degree and boot camp candidates could have produced interesting results.

Appendix

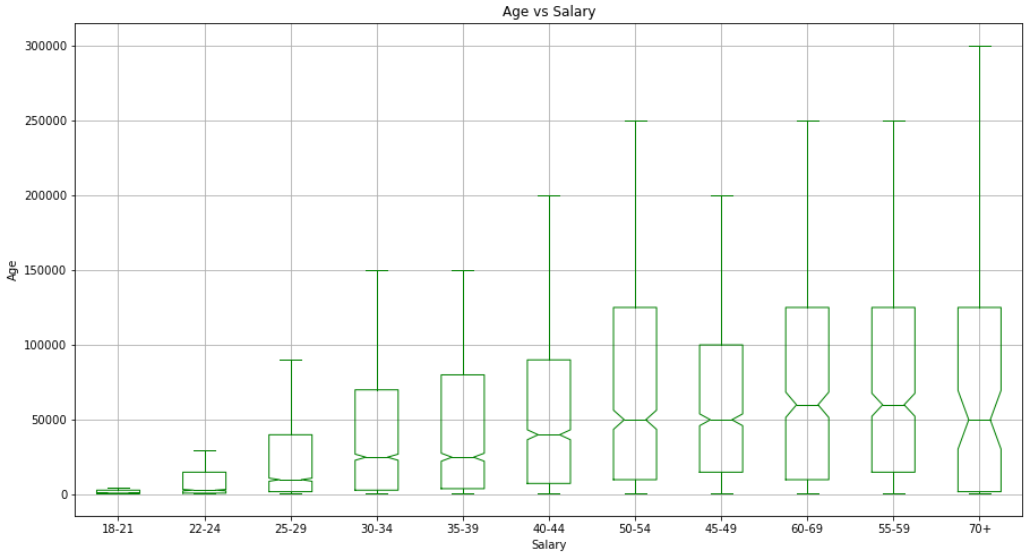

Figure 1: Q1 GF1

	Q25							
	count	mean	std	min	25%	50%	75%	max
Q1								
70+	97.0	100469.072165	161287.153576	1000.0	2000.0	50000.0	125000.0	1000000.0
55-59	504.0	97216.269841	140218.945713	1000.0	15000.0	60000.0	125000.0	1000000.0
60-69	454.0	87435.022026	103111.079789	1000.0	10000.0	60000.0	125000.0	1000000.0
45-49	1141.0	82403.593339	121998.820440	1000.0	15000.0	50000.0	100000.0	1000000.0
50-54	791.0	82304.677623	112676.455554	1000.0	10000.0	50000.0	125000.0	1000000.0
40-44	1528.0	67760.798429	109596.664315	1000.0	7500.0	40000.0	90000.0	1000000.0
35-39	1992.0	59316.265060	103367.663264	1000.0	4000.0	25000.0	80000.0	1000000.0
30-34	2626.0	47932.025895	84957.880022	1000.0	3000.0	25000.0	70000.0	1000000.0
25-29	3235.0	29213.910355	68629.908266	1000.0	2000.0	10000.0	40000.0	1000000.0
22-24	2092.0	19918.738050	81903.864589	1000.0	1000.0	3000.0	15000.0	1000000.0
18-21	931.0	15722.878625	86677.395961	1000.0	1000.0	1000.0	3000.0	1000000.0

Figure 2: Q1 GF1 DF

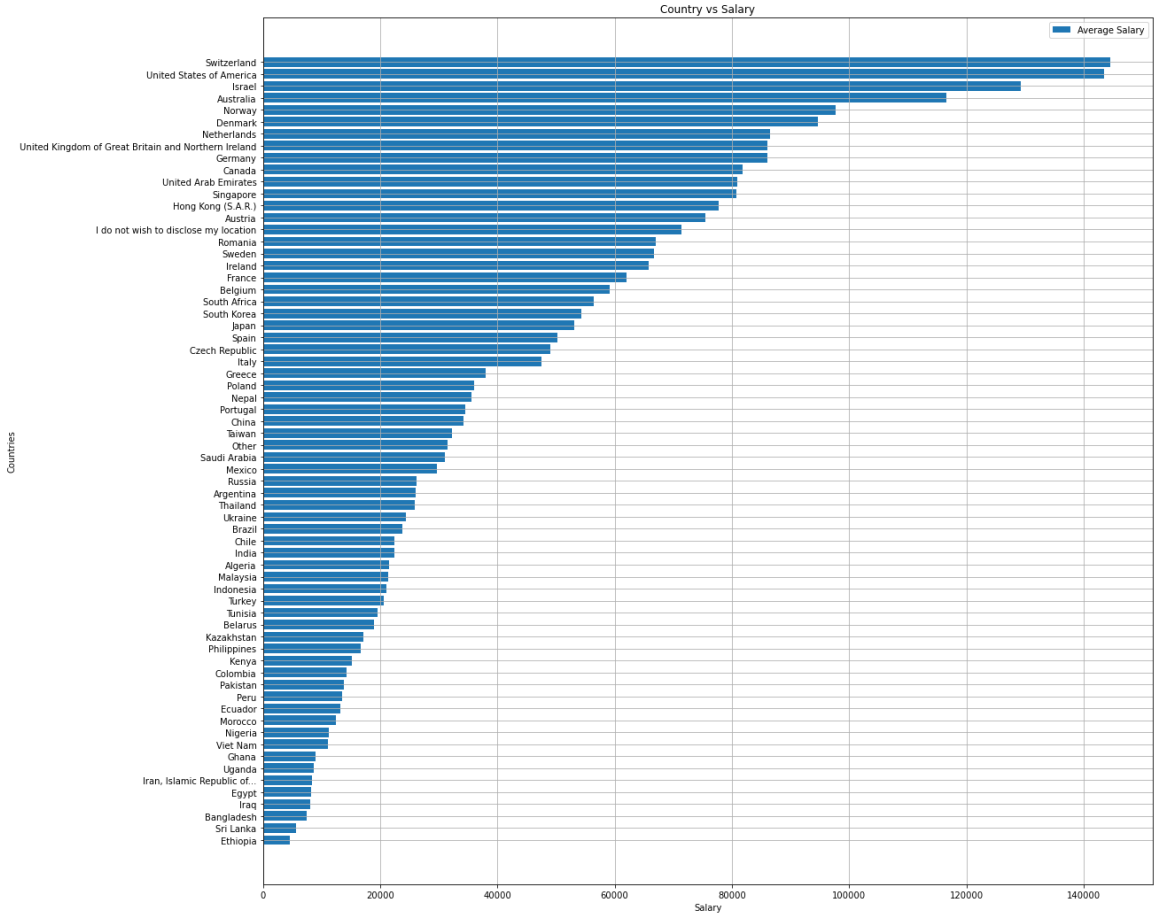

Figure 3: Q1 GF2

	Q25							
	count	mean	std	min	25%	50%	75%	max
Q3								
Switzerland	52.0	144548.076923	142394.345408	1000.0	87500.0	125000.0	200000.0	1000000.0
United States of America	1845.0	143463.685637	151989.614018	1000.0	70000.0	125000.0	200000.0	1000000.0
Israel	96.0	129322.916667	180981.766735	1000.0	40000.0	90000.0	150000.0	1000000.0
Australia	195.0	116520.512821	128057.642827	1000.0	70000.0	100000.0	125000.0	1000000.0
Norway	26.0	97653.846154	52985.992343	1000.0	70000.0	100000.0	125000.0	200000.0
...	...	...	...	...	...	...	...	...
Egypt	285.0	8175.438596	18796.981334	1000.0	1000.0	1000.0	7500.0	150000.0
Iraq	28.0	8053.571429	12177.954646	1000.0	1000.0	2000.0	10000.0	50000.0
Bangladesh	145.0	7479.310345	15226.858108	1000.0	1000.0	2000.0	7500.0	100000.0
Sri Lanka	55.0	5600.000000	6861.324783	1000.0	1000.0	2000.0	7500.0	25000.0
Ethiopia	30.0	4633.333333	5991.277184	1000.0	1000.0	2000.0	3750.0	20000.0

Figure 4: Q1 GF2 DF

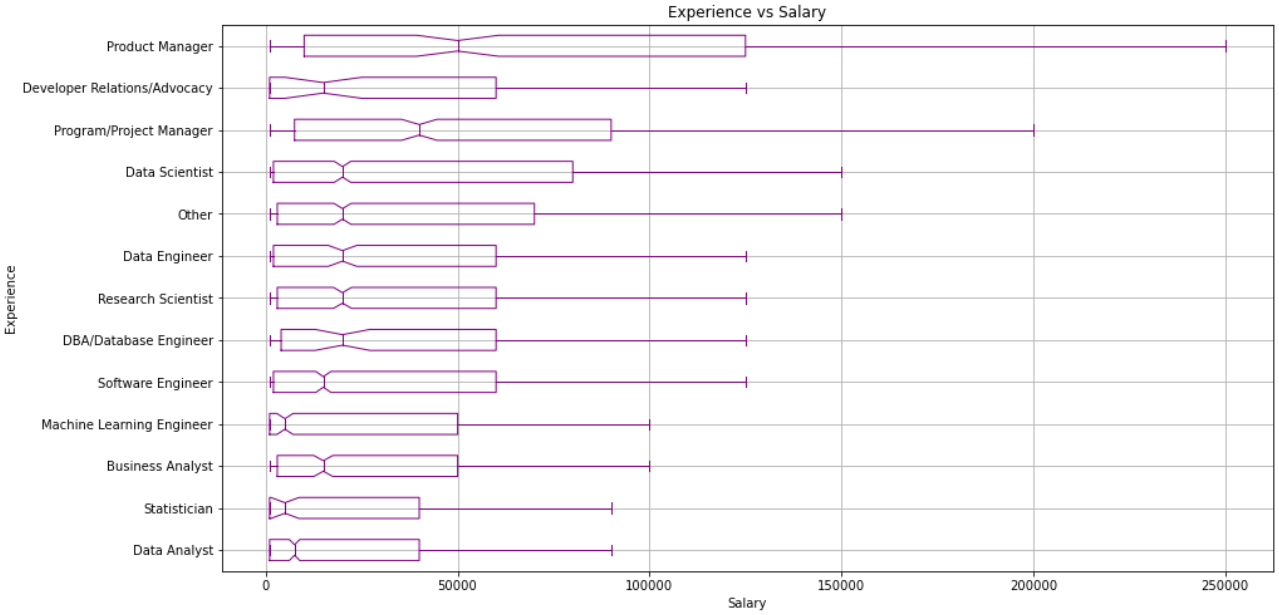

Figure 5: Q1 GF3

	Q25							
	count	mean	std	min	25%	50%	75%	max
Q5								
Product Manager	285.0	90877.192982	143369.772112	1000.0	10000.0	50000.0	125000.0	1000000.0
Developer Relations/Advocacy	86.0	72656.976744	171850.127276	1000.0	1000.0	15000.0	60000.0	1000000.0
Program/Project Manager	784.0	65728.954082	102745.212731	1000.0	7500.0	40000.0	90000.0	1000000.0
Data Scientist	3240.0	57771.296296	106170.740600	1000.0	2000.0	20000.0	80000.0	1000000.0
Other	2204.0	56941.923775	115403.732022	1000.0	3000.0	20000.0	70000.0	1000000.0
Data Engineer	597.0	49226.968174	87071.250580	1000.0	2000.0	20000.0	60000.0	1000000.0
Research Scientist	1404.0	47183.048433	80804.503293	1000.0	3000.0	20000.0	60000.0	1000000.0
DBA/Database Engineer	151.0	46023.178808	59467.227326	1000.0	4000.0	20000.0	60000.0	250000.0
Software Engineer	2110.0	45505.450237	101319.783478	1000.0	2000.0	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
Business Analyst	885.0	39983.050847	79124.960207	1000.0	3000.0	15000.0	50000.0	1000000.0
Statistician	279.0	35992.831541	78035.974126	1000.0	1000.0	5000.0	40000.0	1000000.0
Data Analyst	2039.0	28827.856793	56755.093315	1000.0	1000.0	7500.0	40000.0	1000000.0

Figure 6: Q1 GF3 DF

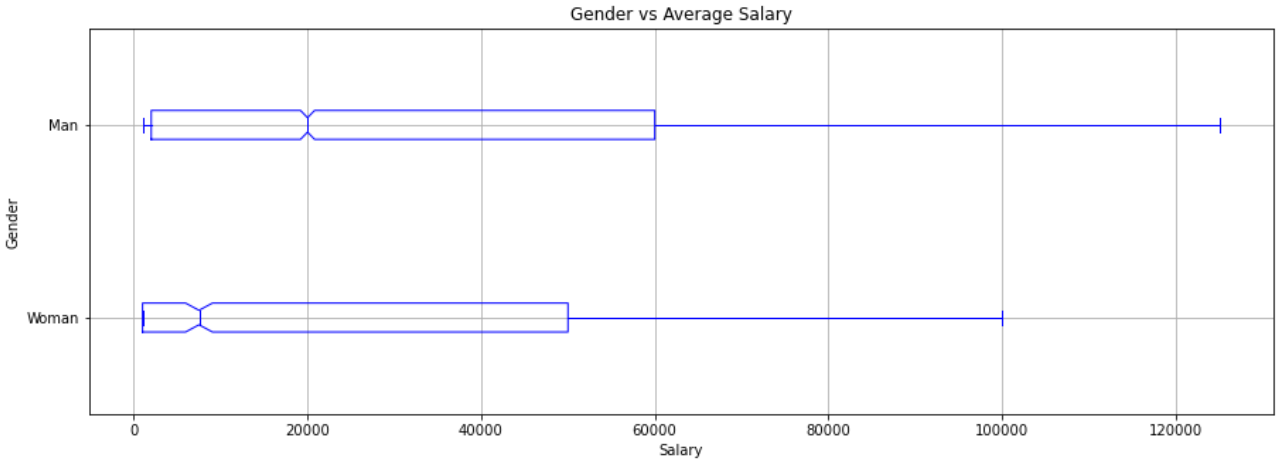

Figure 7: Q2a) Plot

	Man	Woman
Unbiased Estimator Statistics		
count	12642.000000	2482.000000
mean	51193.600696	34816.881547
std	99979.274378	72017.347888
min	1000.000000	1000.000000
25%	2000.000000	1000.000000
50%	20000.000000	7500.000000
75%	60000.000000	50000.000000
max	1000000.000000	1000000.000000

Figure 8: Q2a) DF

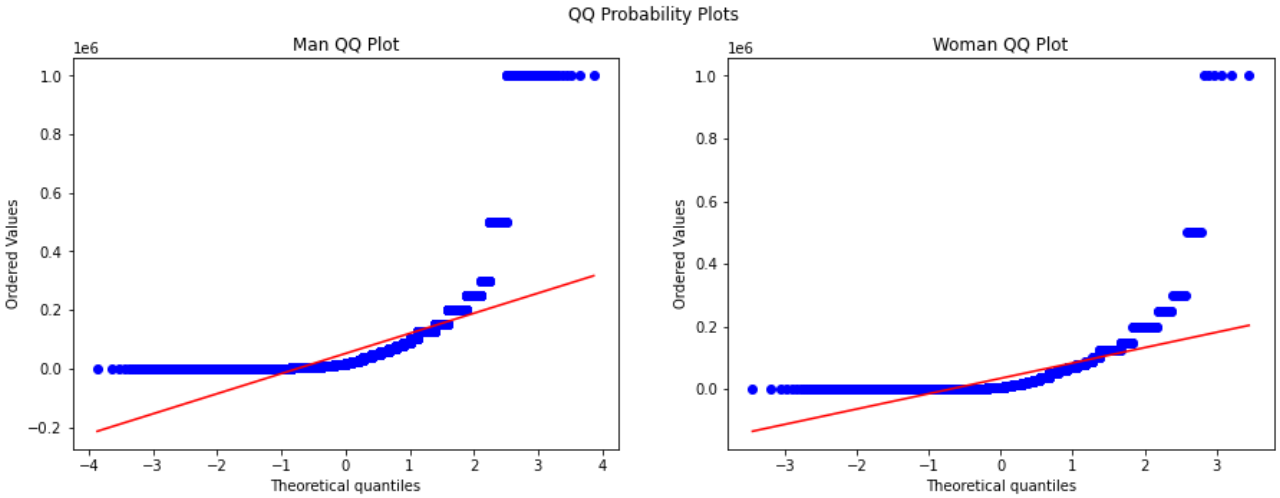

Figure 9: Q2b) QQ

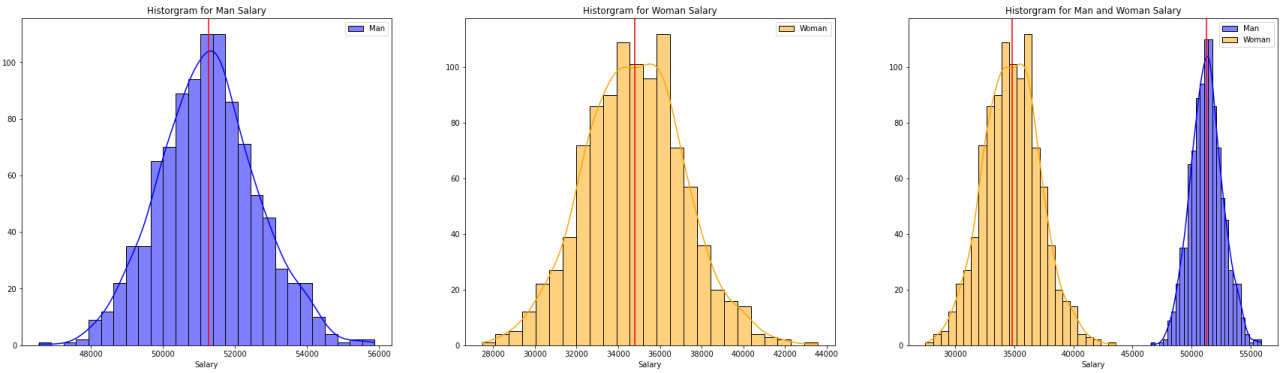

Figure 10: Q2c) Histograms

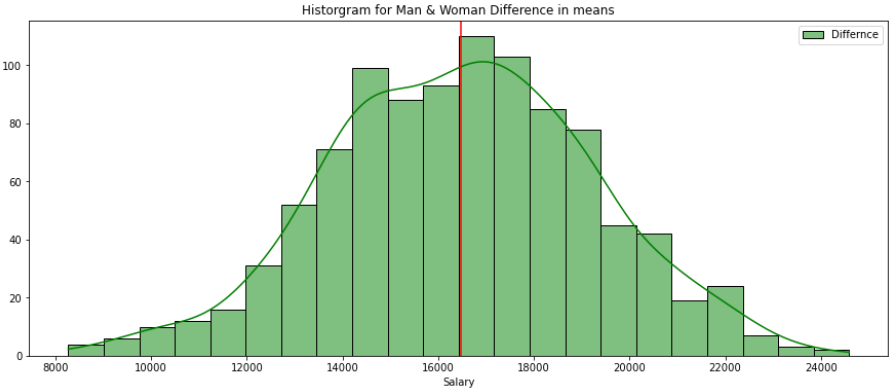

Figure 11: Q2c) Difference Histogram

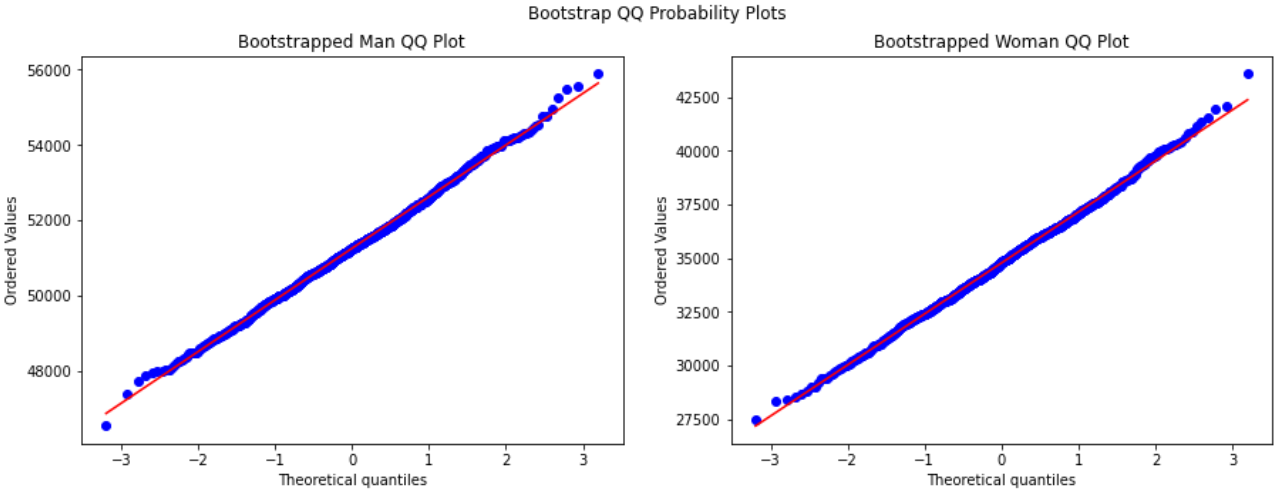

Figure 12: Q2d) Bootstrapped QQ

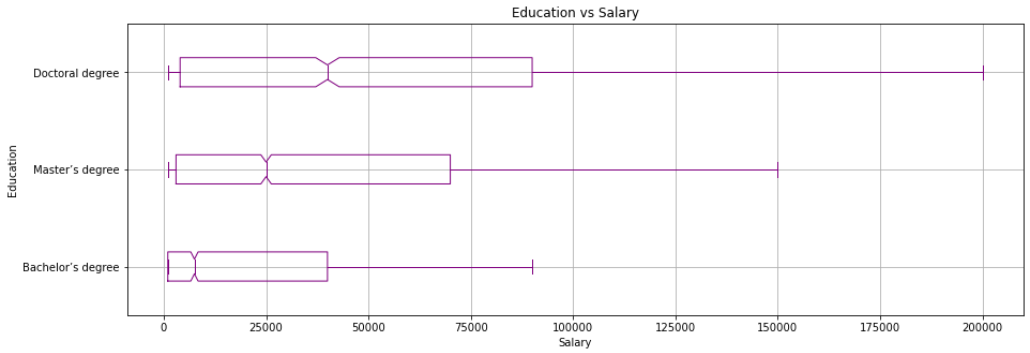

Figure 13: Q3a) Plot

	Bachelor	Master	Doctor
Unbiased Estimator Statistics			
count	4777.000000	6799.000000	2217.000000
mean	35578.291815	52706.868657	70641.181777
std	89382.060777	90928.786678	117160.947589
min	1000.000000	1000.000000	1000.000000
25%	1000.000000	3000.000000	4000.000000
50%	7500.000000	25000.000000	40000.000000
75%	40000.000000	70000.000000	90000.000000
max	1000000.000000	1000000.000000	1000000.000000

Figure 14: Q3a) DF

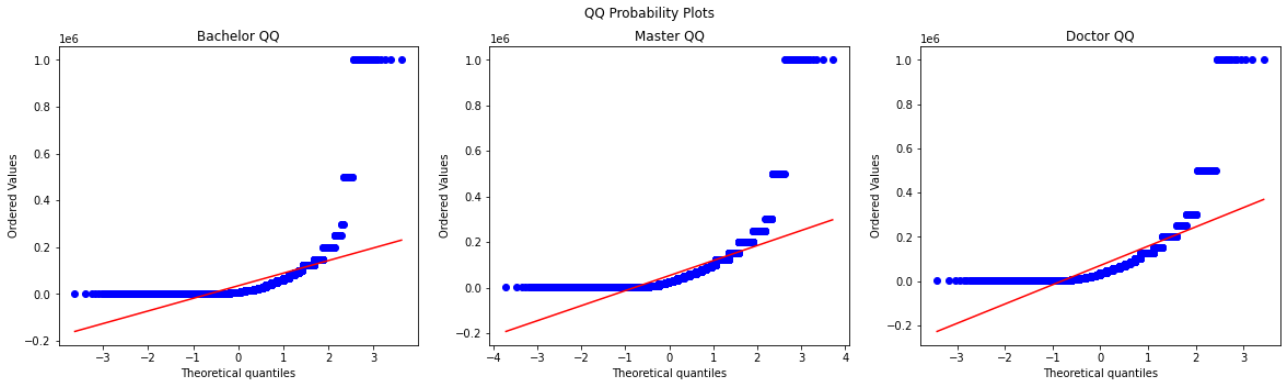

Figure 15: Q3b) QQ

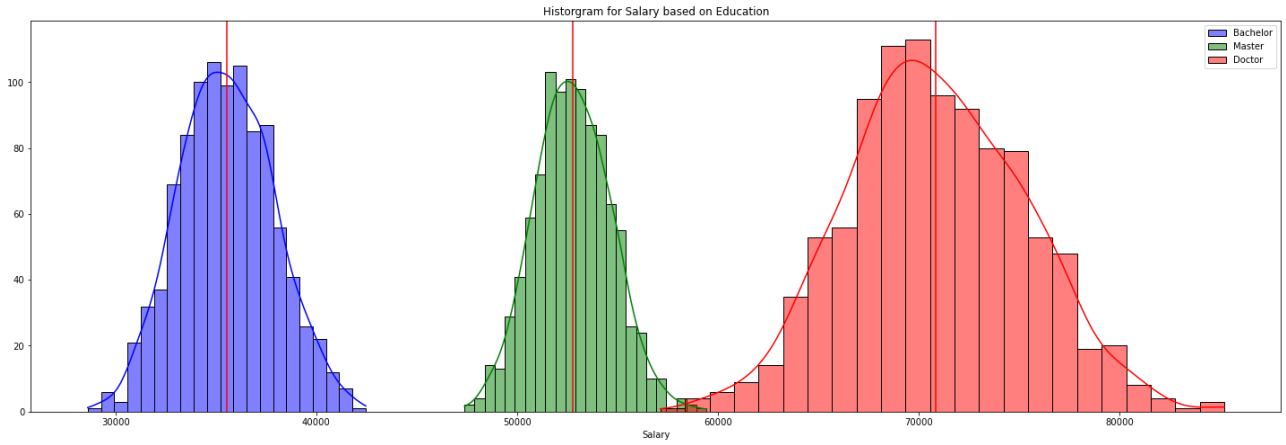

Figure 16: Q3c) Combined Histograms

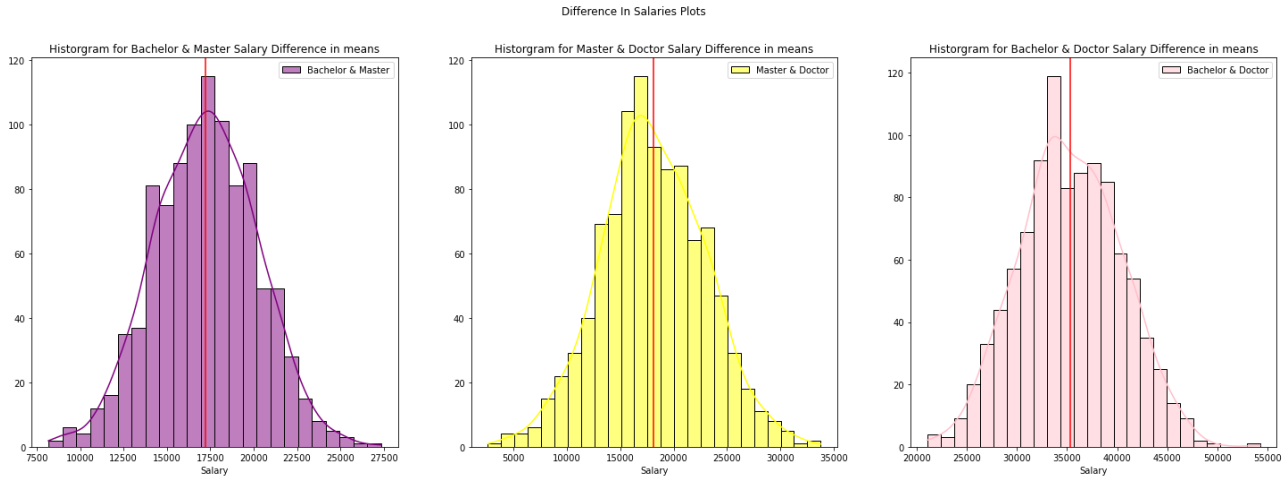

Figure 17: Q3c) Individual Difference Histogram

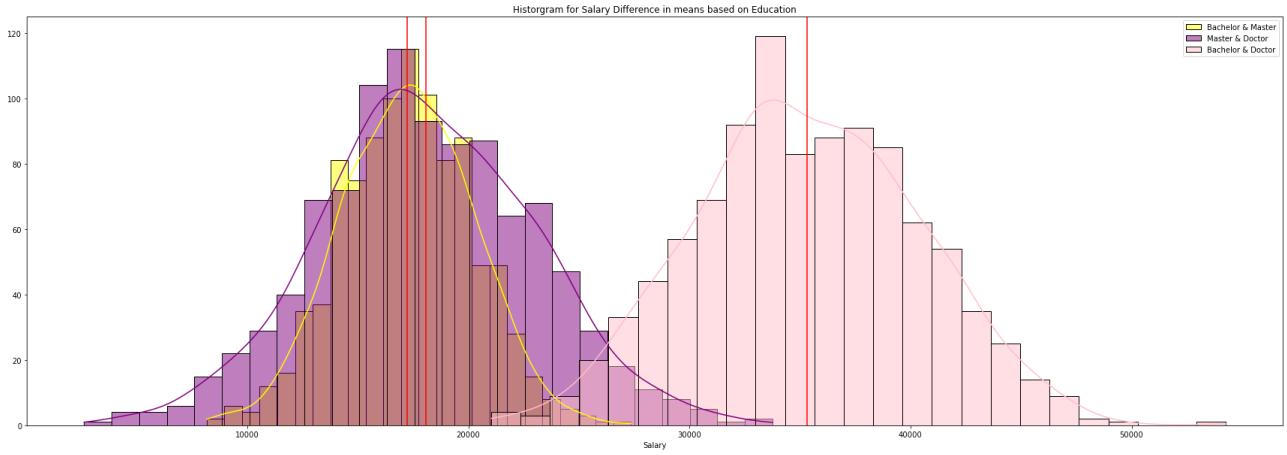

Figure 18: Q3c) Combined difference Histogram

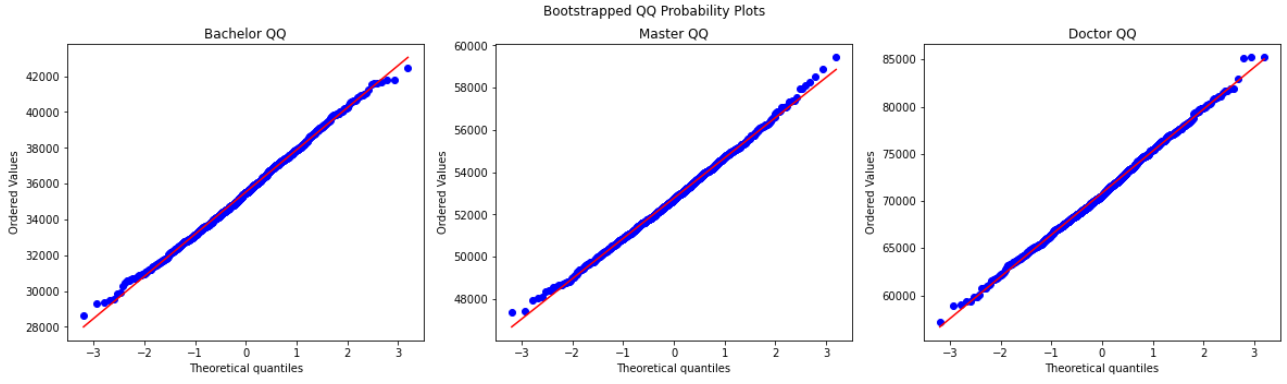

Figure 19: Q3d) QQ