MIE1624 Assignment 1 Report

Kaiyan Jiang 1003848189

February 16, 2022

The assignment aims to explore the nature of women's representation in Data Science and Machine Learning and the effects of education on income level. Q25 "What is your current yearly compensation (approximate $USD)?" is the target column in this assignment.

Question 1:

An exploratory data analysis has been presented in the first part to analyze the dataset. So, here I selected three factors: age(Q3), the years of programming experience(Q6) and which industry is the participants currently employed(Q20) to explore the relationships between these characteristics with the yearly compensation (Q25).

Firstly, from the summary table salaries of different age groups, the mean salary increases with age. The standard deviation is very high, so, the situations are different between cases. A boxplot between salary and age groups is plotted, which is very right-skewed. Most people have an income below 200 thousand, while there are a few elites who can achieve higher income up to one million a year.

Then, I explored the effect of the years of programming experience on salary. By plotting the mean salary of different experience levels in a barplot, the salary increases with participants who have a longer programming experience in general. However, with one exception that people with no programming experience seem to have a higher average income than people with programming experience of less than one year.

I also plotted a barplot between the average salary and the participant's current industry. The plot and the mean data show that insurance/risk assessment and military/security/defense are the top two high salary industries, over 70 thousand USD. And academics/education is the industry that offers the lowest average salary of only around 27003 USD per year.

Question 2:

The second part focuses on estimating the difference between men's average and women's. As there are five categories of gender from the survey, only the categories men and women will be selected. Descriptive statistics are presented; male participants are far more than female participants. And from the statistics, the average salary of men participants is 10,000 more than the average salary of women participants.

A two-sample t-test will be chosen for the goal. Before carrying out the test, a fitness of assumption will be validified, mainly normality. Histograms and QQ Plots of two groups are plotted. Taking a log scale when plotting histograms provides wider bins, which is clearly to observe. From the histograms, the distribution of the mean income both for the men group and women group is not approximately normal distributed with a clear right-skewed. And in the QQ Plots, the points are far from the reference line for both groups. So, the assumptions of normality are not held. A two-sample t-test cannot be performed for now.

In the following steps, I first bootstrapped the data. I resampled the two data sets for women's and men's income 1000 times relative to their size and determined the mean of each time. Then two distribution plots of bootstrapped men's salaries and bootstrapped women's salaries have been plotted. The distribution of

bootstrapped data is approximately normally distributed. Also, a histogram of distribution for the difference data between men's average salaries and women's average salaries is drawn, which is also close to a bell shaped.

A two-sample t-test is performed on bootstrapped data. Although the histogram showed a bell shape, QQ plots are still used to test normality. From the QQ plots, the points are quite fit the reference line. Thus, the bootstrapped data is normally distributed. However, the variances between the two groups have a significant difference; the variance of women's salary is 2040119.8; the variance of men's salary is 762571.5. A Welch's two-sample t-test was performed as the variances are not homogeneous. The null hypothesis is that the women's average salary is the same as the men's average salary; the alternative hypothesis is that the women's average salary is not equal to the men's average salary. After carrying out the test, p-value equals 0, which is smaller than 0.05, the null hypothesis is rejected.

The histogram of two bootstrapped data also showed this. The women's average salary is centered at 35,000 while the men's average salary is 50,000, which is quite a big difference. And the t-test result, the women's average salary is not equal to the men's average salary, supports that finding.

Question 3:

The third part focuses on the effects of education on income level, three groups (bachelor's degree, master's degree, and doctoral degree) are selected for analysis. First, descriptive statistics are presented for each group. Participants with higher education background tend to have higher average income. Since there are three groups now, ANOVA is used instead of two-sample t-test. Before testing, assumptions must be checked. Thus, QQ plots of three groups are plotted to verify normality. The quantiles do not lie along with the reference line; thus, the normality is not held. ANOVA test cannot be performed.

Three groups are bootstrapped, by resampling the three data sets for bachelor's degree, master's degree, and doctoral degree income 1000 times relative to their size and determined the mean of each time. Also, the difference between each two groups is calculated. Three histograms are plotted for the bootstrapped data. The graphs show a bell shape, which can conclude that the data now is approximately normally distributed. The similar distribution also occurred on the bootstrapped difference data.

ANOVA test performed on the three bootstrapped education data. First, the normality and homogeneous variance assumptions must be tested. Three QQ plots are drawn for the datasets to see the normality. The quantiles points are along with the reference red line. The assumption of normality held. And from the descriptive statistics, the variances for bootstrapped average salary for bachelor's degree, master's degree, and doctoral degree are 1280.9, 1089.7 and 2467.6. As the variances are in the same order, the homogeneity of variances held. A one-way ANOVA test can be performed.

For this one-way ANOVA test, the null hypothesis is that the mean salaries for three groups are the same. The alternative hypothesis is that the mean salaries for different education background are different. From the ANOVA test result, the p-value is 0 which is smaller than 0.05. Thus, the null hypothesis is rejected. The average salaries of bachelor's degree, master's degree, and doctoral degree are not equal.

So, from the ANOVA test, the conclusion is that the average salaries of bachelor's degree, master's degree, and doctoral degree are not the same. When plotting the three histograms of bootstrapped data together, it is clearly showed that participants with higher education background will have a higher average salary when working is data science and machine learning area.

Appendix:

Table 1: Summary Table of Salary of Different Age Groups

	Salary							
	count	mean	std	min	25%	50%	75%	max
Age Groups								
18-21	931.0	15722.88	86677.40	1000.0	1000.0	1000.0	3000.0	1000000.0
22-24	2092.0	19918.74	81903.86	1000.0	1000.0	3000.0	15000.0	1000000.0
25-29	3235.0	29213.91	68629.91	1000.0	2000.0	10000.0	40000.0	1000000.0
30-34	2626.0	47932.03	84957.88	1000.0	3000.0	25000.0	70000.0	1000000.0
35-39	1992.0	59316.27	103367.66	1000.0	4000.0	25000.0	80000.0	1000000.0
40-44	1528.0	67760.80	109596.66	1000.0	7500.0	40000.0	90000.0	1000000.0
45-49	1141.0	82403.59	121998.82	1000.0	15000.0	50000.0	100000.0	1000000.0
50-54	791.0	82304.68	112676.46	1000.0	10000.0	50000.0	125000.0	1000000.0
55-59	504.0	97216.27	140218.95	1000.0	15000.0	60000.0	125000.0	1000000.0
60-69	454.0	87435.02	103111.08	1000.0	10000.0	60000.0	125000.0	1000000.0
70+	97.0	100469.07	161287.15	1000.0	2000.0	50000.0	125000.0	1000000.0

Figure 1: Age Group vs Salary

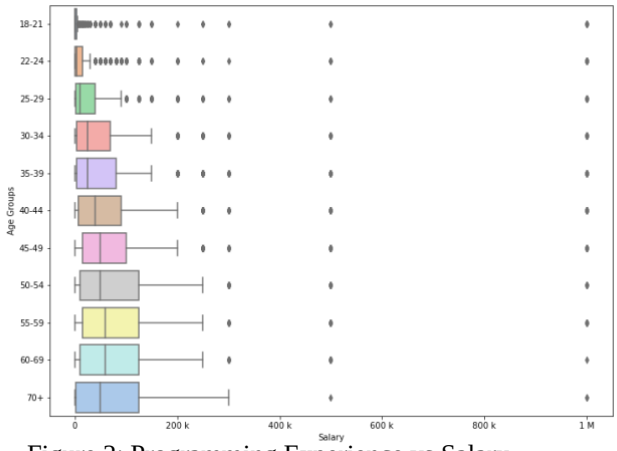

Figure 2: Programming Experience vs Salary

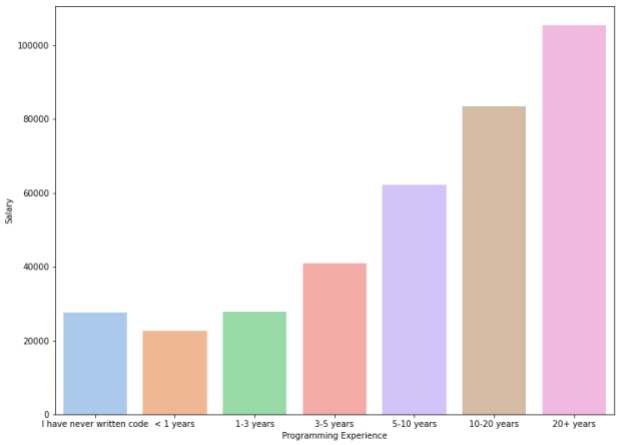

Figure 3: Industry of Current Employer vs Salary

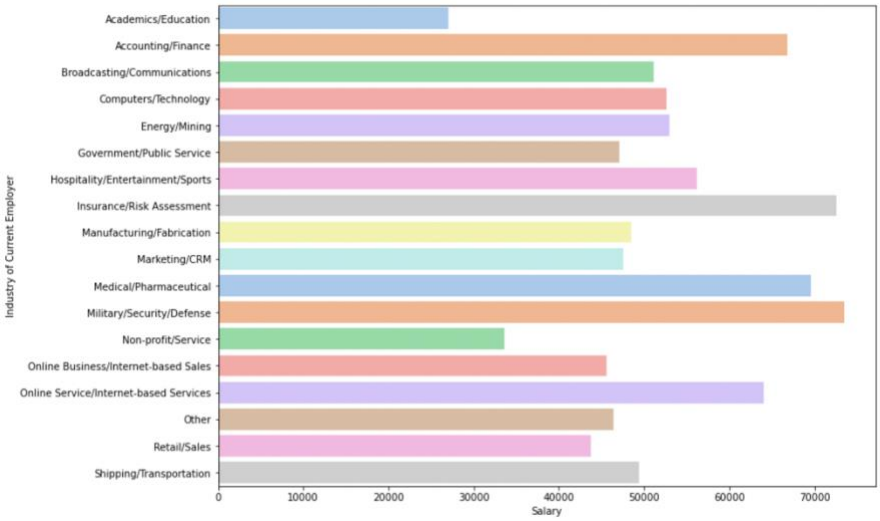

Table 2: Average yearly Salary of Industries of Current Employer

	Industry of Current Employer	Salary
0	Academics/Education	27003.51
1	Accounting/Finance	66753.43
2	Broadcasting/Communications	51079.94
3	Computers/Technology	52603.24
4	Energy/Mining	52928.71
5	Government/Public Service	47073.68
6	Hospitality/Entertainment/Sports	56149.39
7	Insurance/Risk Assessment	72529.34
8	Manufacturing/Fabrication	48511.93
9	Marketing/CRM	47544.03
10	Medical/Pharmaceutical	69558.11
11	Military/Security/Defense	73528.80
12	Non-profit/Service	33642.60
13	Online Business/Internet-based Sales	45589.15
14	Online Service/Internet-based Services	64057.25
15	Other	46449.81
16	Retail/Sales	43751.95
17	Shipping/Transportation	49363.33

Table 3: Descriptive statistics for Men's Salary and Women's Salary

	Salary							
	count	mean	std	min	25%	50%	75%	max
Gender								
Man	12642.0	51193.600696	99979.274378	1000.0	2000.0	20000.0	60000.0	1000000.0
Woman	2482.0	34816.881547	72017.347888	1000.0	1000.0	7500.0	50000.0	1000000.0

Figure 4: Histogram and QQ plot of Men and Women data

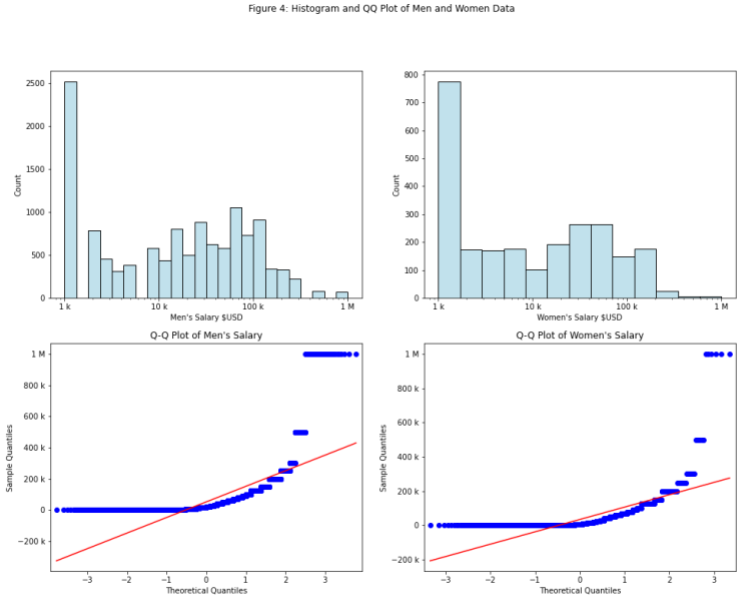

Figure 5 & 6: Histogram of Bootstrapped Data

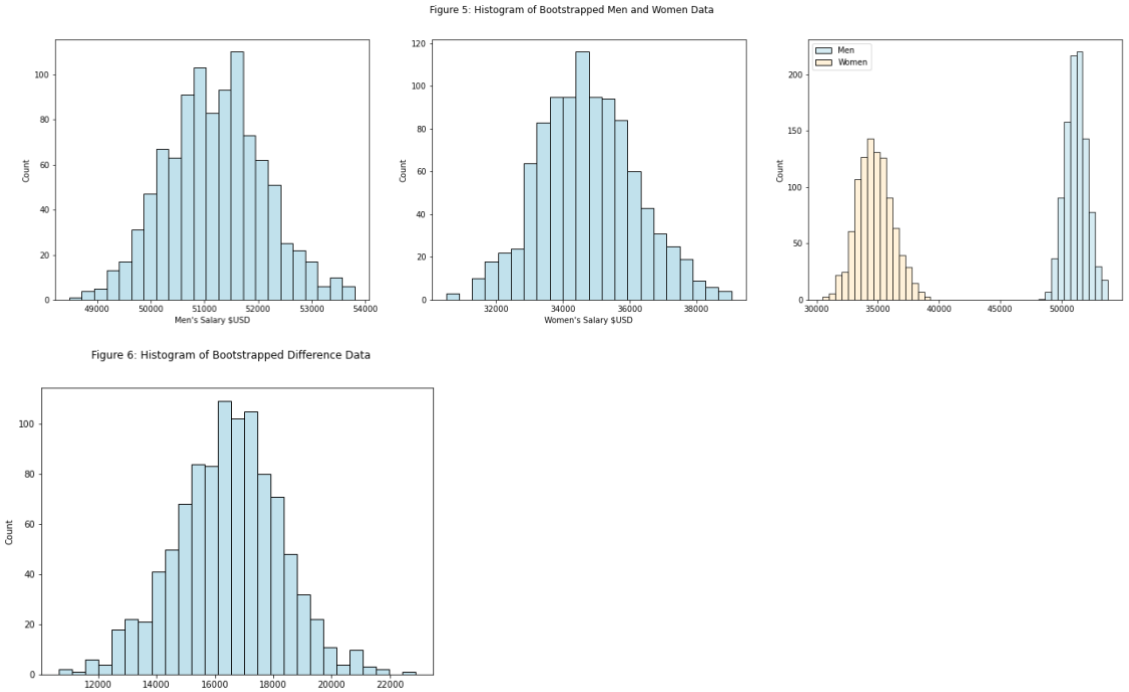

Figure 7: QQ plot of Bootstrapped Data

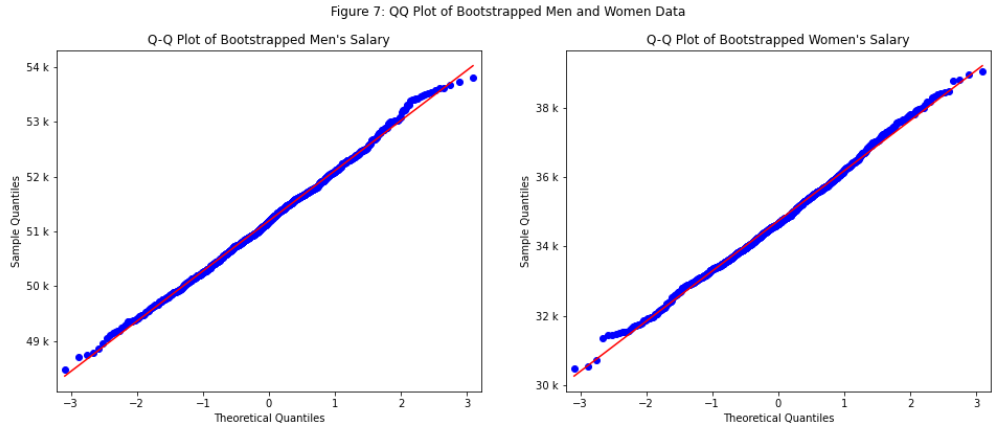

Table 4: Descriptive statistics for Salary of three degrees

	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
Master's degree	6799.0	52706.868657	90928.786678	1000.0	3000.0	25000.0	70000.0	1000000.0

Figure 8: QQ Plot of different Degrees' Salary

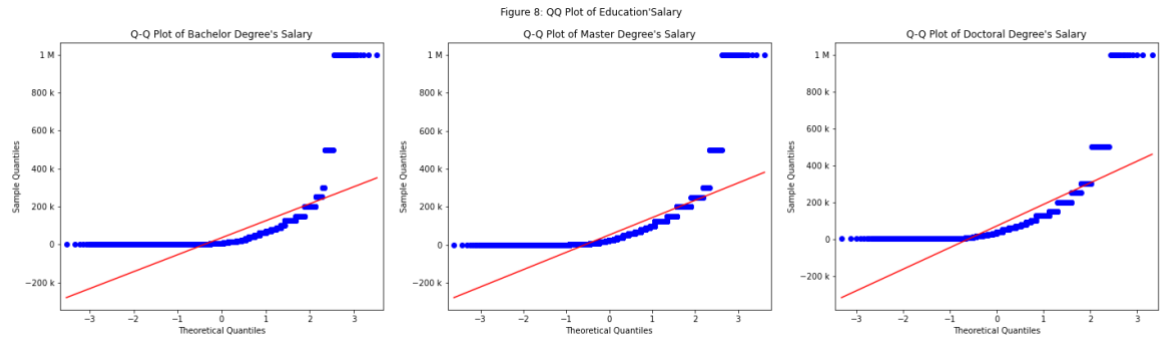

Figure 9 & 10: Distribution of Bootstrapped Data of Education

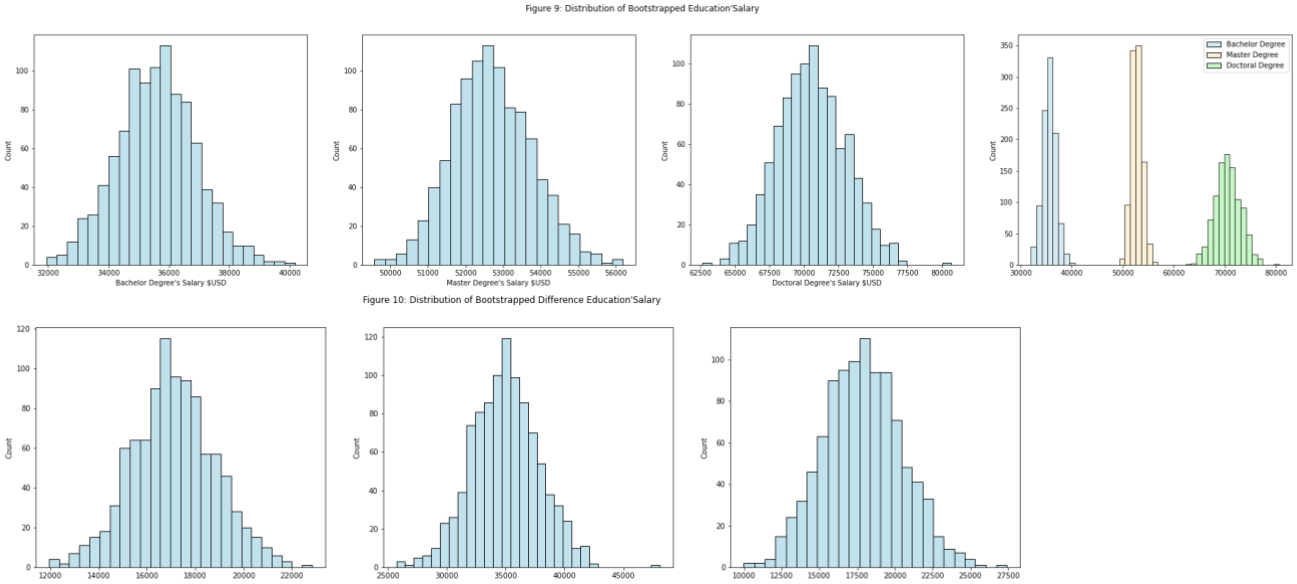

Figure 11: QQ Plot of Bootstrapped Education's Salary

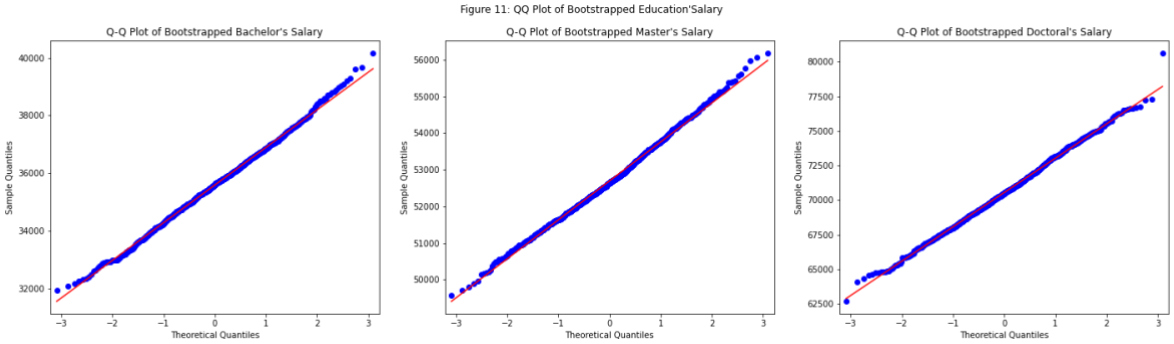

Table 5: Descriptive Table of Bootstrapped Education Data

	Bachelor Degree	Master Degree	Doctoral Degree
count	1000.000000	1000.000000	1000.000000
mean	35589.945049	52697.434549	70550.370095
std	1306.221223	1064.181240	2484.010615
min	31938.664434	49577.952640	62675.687866
25%	34730.741051	51951.665686	68888.080740
50%	35609.535273	52640.277982	70485.453315
75%	36455.882353	53427.544492	72204.442941
max	40175.528574	56191.425210	80656.292287