MIE 1624 Introduction to Data Science and Analytics

Assignment 1: Salary Analysis Report

Yunjie Xu 1008601951

Department of Mechanical and Industrial Engineering, University of Toronto

October 15, 2022

Introduction

This assignment aims to explore the nature of women's representation in data science field. Q25 "What is your current yearly compensation?" is the target column in this assignment.

Question 1

I integrated those whose gender is not "Man" or "Woman" into "Gender Minority". I found that woman only accounted for 16.13% of people who working in data science field, while man constituted 82.14%. In general, Country(Figure 1), Age(Figure 2), Education(Figure 3), and Professional Experience(Figure 4) are all very important factors in influencing salary.In Gender, Men usually have higher salary than women with the same education level(Figure 5, Figure 6) or experience(Figure 7, Figure 8) or age(Figure 9, Figure 10). In education, people with higher degree are more easy to get higher salary generally. In experience and age, people with long experience usually can get higher rewards as well as the olders usually can earn more.

Question 2

a

The second part focuses on estimating the difference between men's average and women's(Figure 11). I observe that the mean salary for men is higher than women as well as median,and other quartiles, while men's salary fluctuate more than women's salary. I also plotted box plot of salary separately(Figure 12, Figure 13), I observed outliers.

b

According to the p-value from the T-test result is much smaller than 0.05 threshold from the code, thus reject the null hypothesis should be rejected. So, I may conclude that the difference of mean salaries of men and women is statistically significant. However,since two data samples are not normally distributed based on that both two histograms do not show like a bell shape (Figure 14). So the normality assumptions are violated. Thus, this conclusion may not be reliable and T-test is not suitable here.

c

I used random.sample function of Python to bootstrap the samples with sample sizes of men and women being relative to their sizes. After replicated 1000 times, it is found that the bootstrapped means are very close to the actual mean salaries of men and women. The two distributions, men and women's bootstrapped mean salary, have clear different means with a small overlap. It further proved our findings of higher salary for men previously(Figure 15). And Figure 16 shows the difference in means for men and women.

d

Both bootstrapped distribution distribute normally, the normality assumption can be satisfied in this time. The T-score is 98.14 and p-value of our t-test is statistically significant with $\alpha = 0.05$. I thus reject the null hypothesis that there is no difference between mean salary of men and women in the data science field, based on bootstrapped data. Thus I can conclude that the difference of mean salaries of men and women is statistically significant .

e

I found that the bootstrap is a method of doing inference in a way that does not require assuming a parametric form for the population distribution. It means the method can fix the dataset which fails the normality test. In addition, there is a high statistically significant difference between mean salary of men and women in the data science field, while the mean difference based on bootstrapping is 16467.47.

Question 3

a

I found that masters accounted the largest proportion of people in data science field, while the doctor accounted least. And I observe that the mean salary for doctor is higher than master and bachelor as well as median,and other quartiles, while doctor's salary fluctuate more(Figure 17). I also made box plot here(Figure 18, Figure 19, Figure 20).

b

ANOVA is not suitable here. There are three primary assumptions in ANOVA. Firstly, the responses for each factor level have a normal population distribution. Secondly, these distributions have the same variance. Thirdly, the data are independent. I plotted histograms like q2.b above (Figure 21), I found normality assumption failed. Then, I also calculated variances, I found equal variance assumption also failed. Though, the f-score is 109.76, thus the null hypothesis should be rejected. I conclude there is at least one group mean is different than others.

c

Similarly, using random.sample to bootstrap the samples with sample sizes of three groups being relative to their sizes. It is found that the bootstrapped means are very close to the actual means and among three groups, doctoral displays a higher variance(Figure 22). Figure 23, 24, 25 demonstrate the bootstrapped differences among the pairs within three groups. On average, mean salary for doctorates is $18268.67 higher than masters , while the mean salary for masters is $17123.06 higher than bachelors.

d

ANOVA is still not suitable here.Because all bootstrapped distribution distribute normally, the normality assumption can be satisfied after bootstrapping. But from doctoral displays a higher variance(Figure 22), thus equal variance assumption is still violated. Though, the f-score is 10938.95, larger than q3.b. So, the null hypothesis still be rejected. I conclude there is at least one group mean is different than others.

e

In future, I can use bootstrapping method do inference without assuming a parametric form for the population distribution. It is very useful when normality assumption is need but violate. However, bootstrapping method can only fix distributions to normal distribution with different variance, thus this method cannot help with fixing equal variance assumption.

Appendix

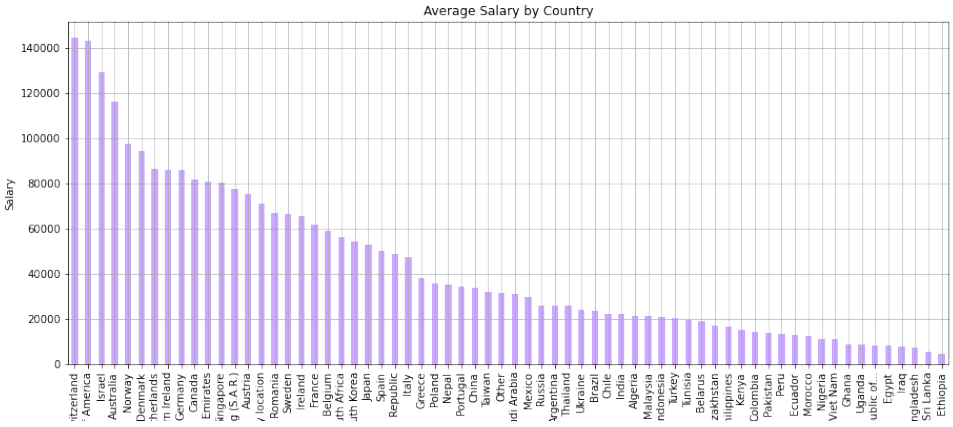

Figure 1: Average Salary by Country

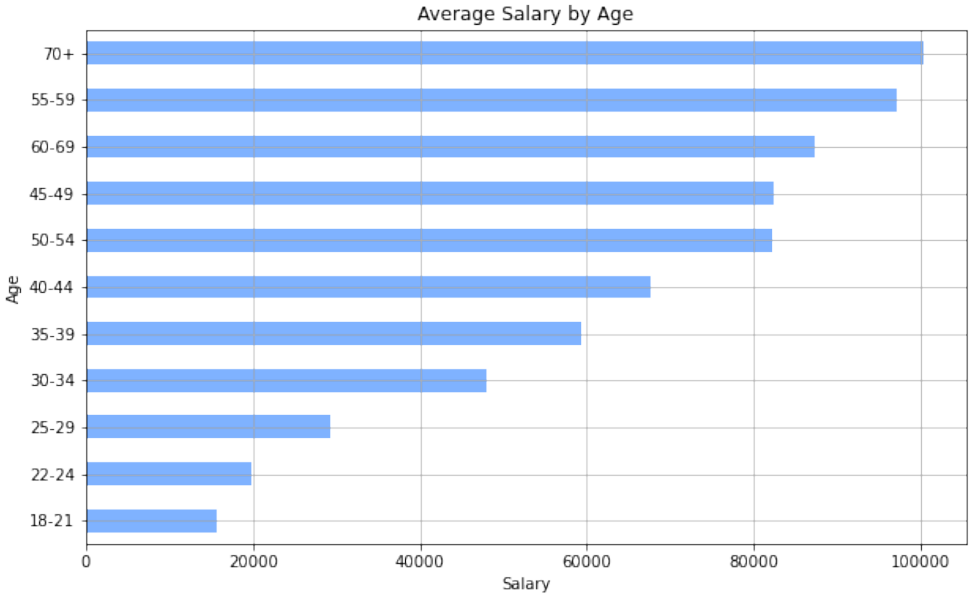

Figure 2: Average Salary by Age

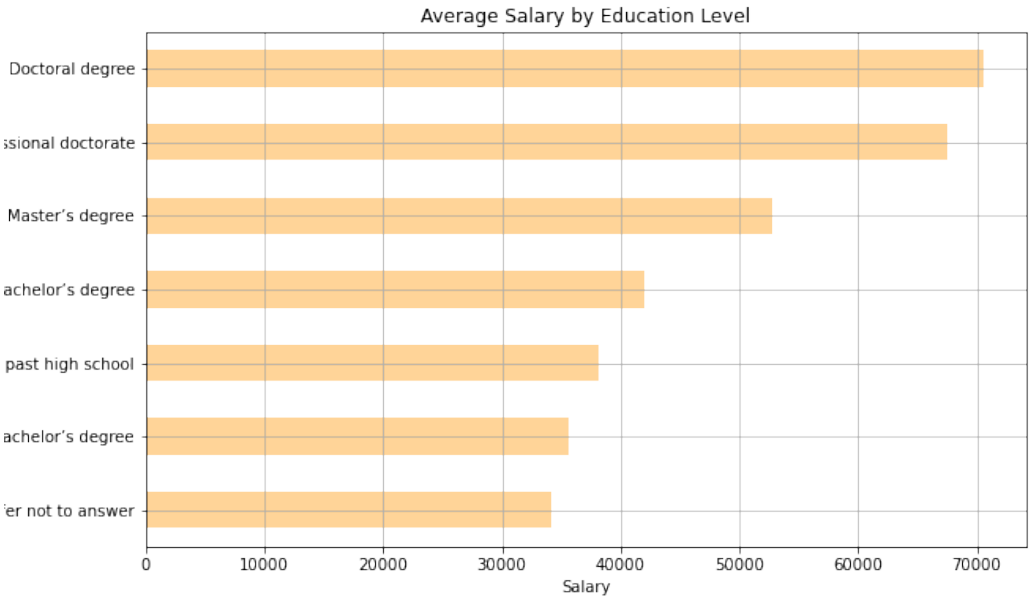

Figure 3: Average Salary by Education

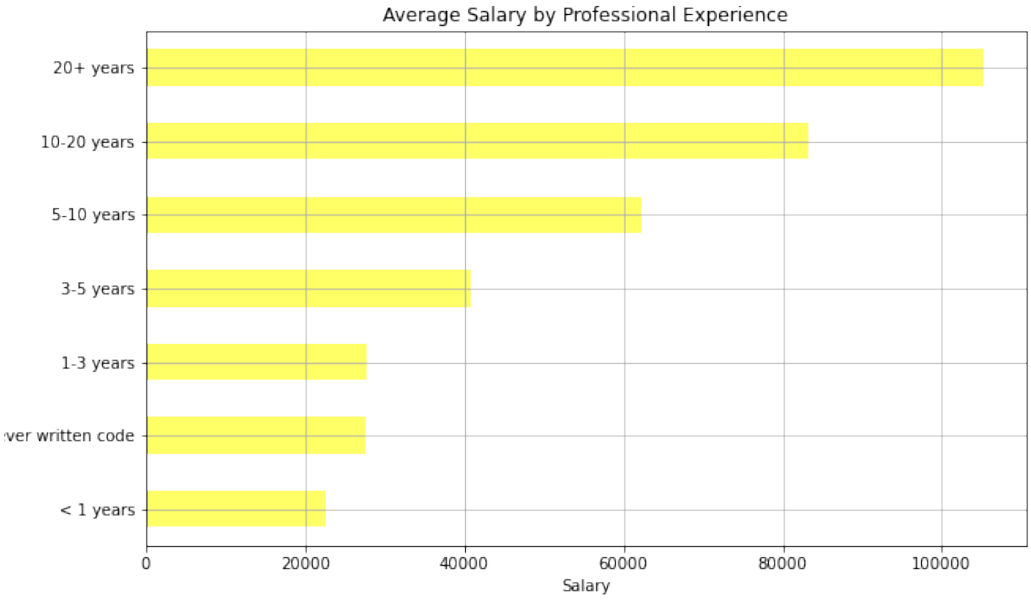

Figure 4: Average Salary by Experience

Education	Bachelor's degree	Doctoral degree	I prefer not to answer	Master's degree	No formal education past high school	Professional doctorate	Some college/university study without earning a bachelor's degree
Gender							
Gender minority	55381.36	91588.89	119847.83	82198.28	214400.00	216333.33	22230.77
Man	37394.30	75505.36	32386.83	54950.23	34975.61	68952.91	43999.22
Woman	23524.93	46664.16	11669.12	38769.77	26083.33	47385.25	31338.54

Figure 5: Table of Salary Based on Different Gender and Education Level

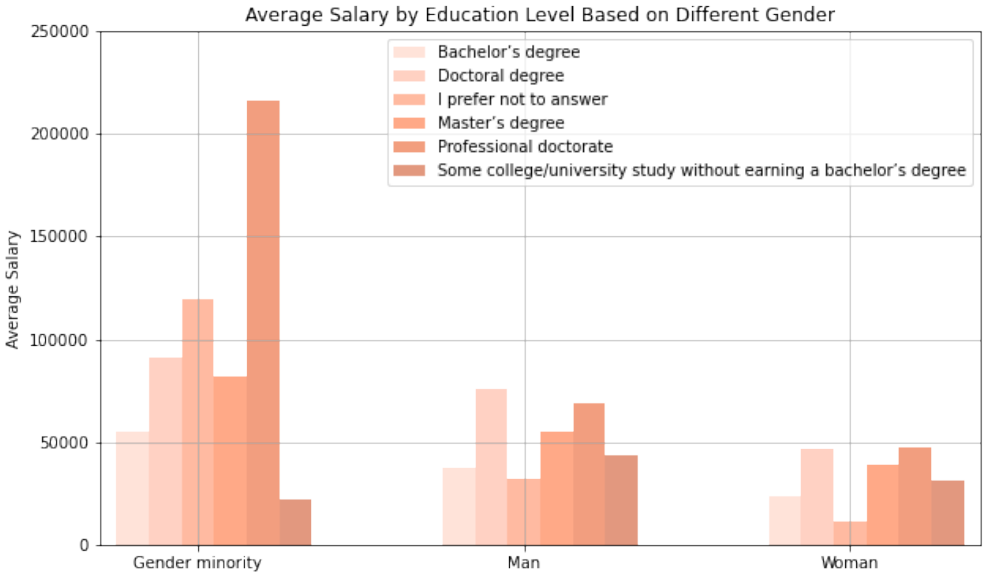

Figure 6: Plot of Salary Based on Different Gender and Education Level

Experience Gender	1-3 years	10-20 years	20+ years	3-5 years	5-10 years	< 1 years	I have never written code
Gender minority	41905.66	73166.67	133829.27	78841.46	125894.23	23267.86	97937.50
Man	29037.55	85957.47	107050.00	41105.76	62538.85	22540.01	26448.91
Woman	20479.33	66486.55	75792.68	35363.05	50103.86	22803.06	26494.23

Figure 7: Table of Salary Based on Different Gender and Experience

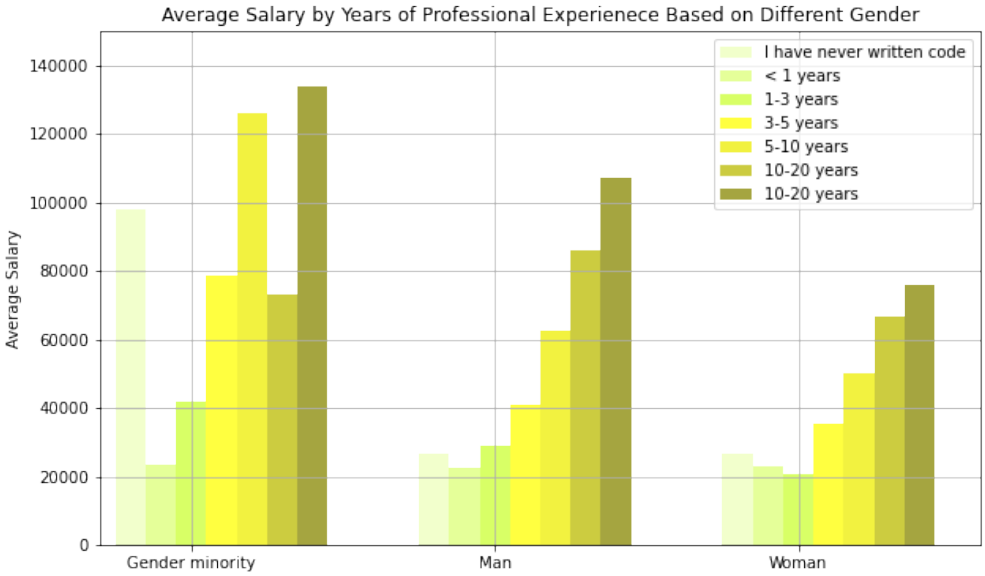

Figure 8: Plot of Salary Based on Different Gender and Experience

Age	18-21	22-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-69	70+
Gender											
Gender minority	20687.50	41882.35	69221.15	82434.21	83700.00	113782.61	109500.00	43833.33	152800.00	162818.18	250333.33
Man	18570.74	20501.80	29552.17	49040.06	61222.60	68662.74	86640.45	84874.81	98720.18	87429.09	92761.63
Woman	4211.23	15560.91	24333.88	38696.78	46903.96	58293.99	48517.86	70841.58	76327.59	56814.81	53200.00

Figure 9: Table of Salary Based on Different Gender and Age

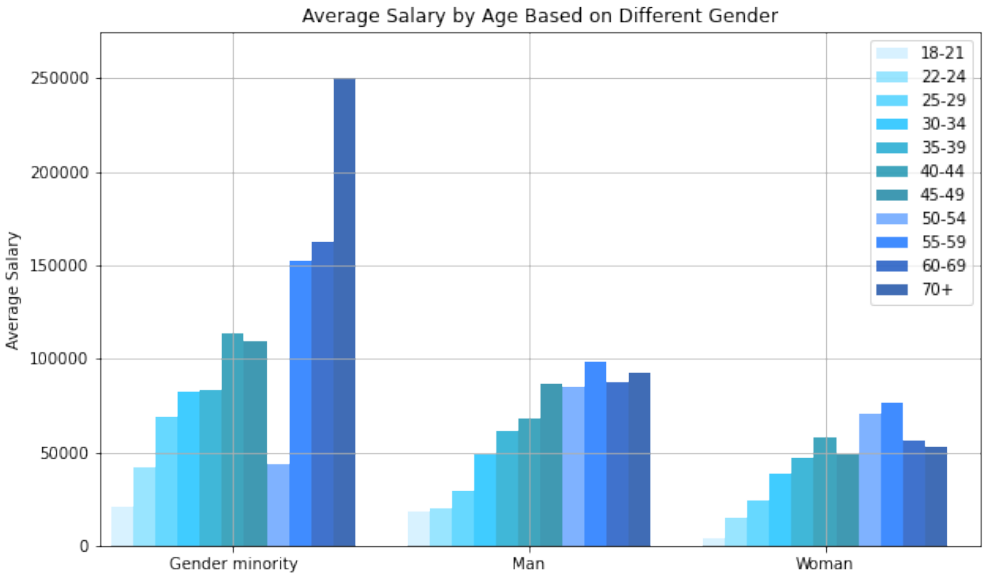

Figure 10: Plot of Salary Based on Different Gender and Age

Salary	Man	Woman
count	12642.00	2482.00
mean	51193.60	34816.88
std	99979.27	72017.35
min	1000.00	1000.00
25%	2000.00	1000.00
50%	20000.00	7500.00
75%	60000.00	50000.00
max	1000000.00	1000000.00

Figure 11: Table of Salary Based Gender

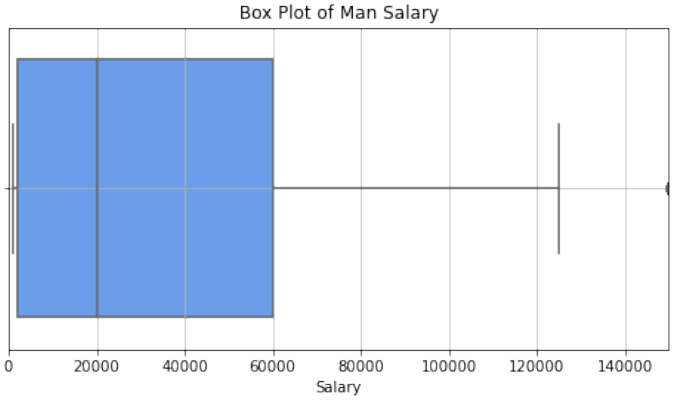

Figure 12: Box Plot of Men's Salary

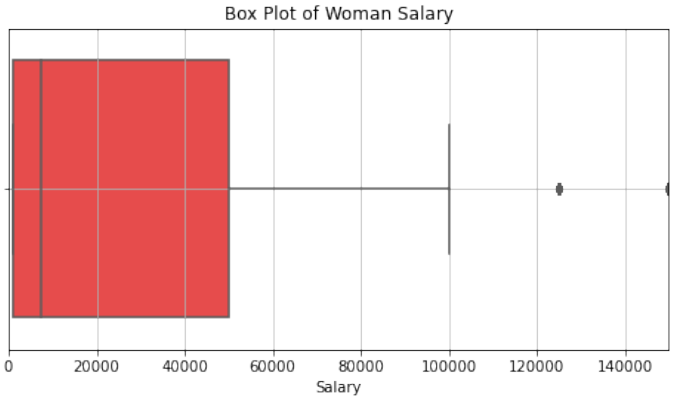

Figure 13: Box Plot of Women's Salary

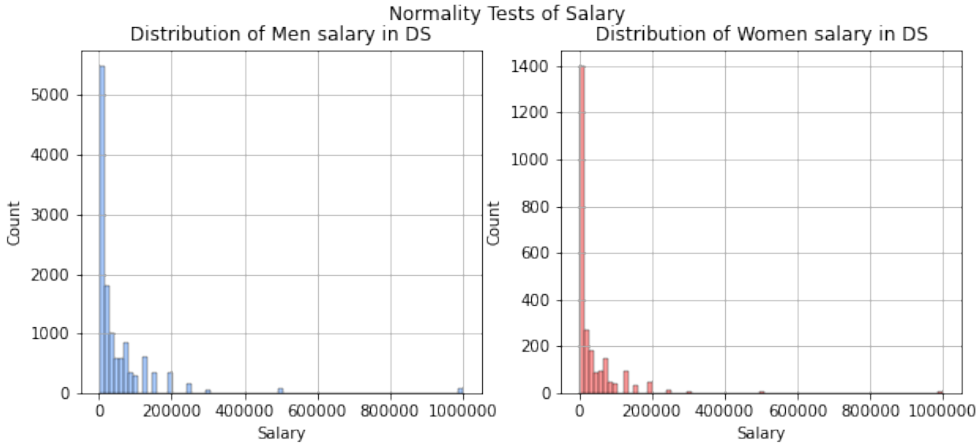

Figure 14: Normality Checking

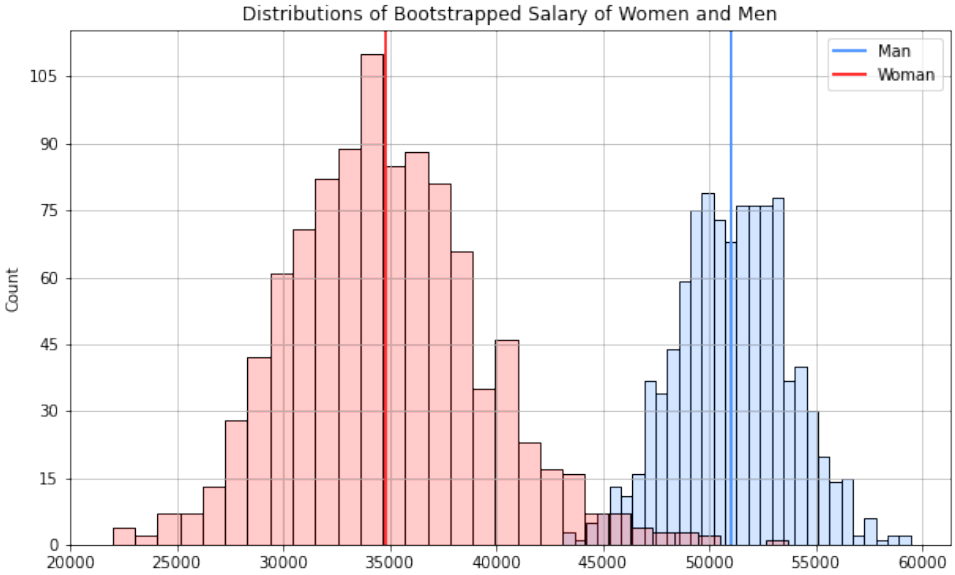

Figure 15: Distributions of Bootstrapped Salary of Women and Men

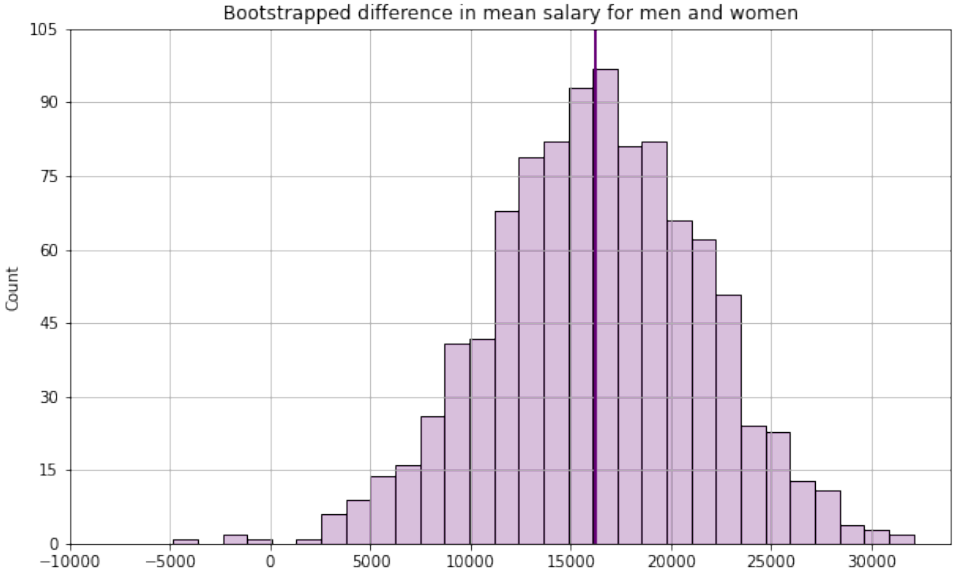

Figure 16: Bootstrapped difference in mean salary for men and women

Salary	Bachelor's degree	Master's degree	Doctoral degree
count	4777.00	6799.00	2217.00
mean	35578.29	52706.87	70641.18
std	89382.06	90928.79	117160.95
min	1000.00	1000.00	1000.00
25%	1000.00	3000.00	4000.00
50%	7500.00	25000.00	40000.00
75%	40000.00	70000.00	90000.00
max	1000000.00	1000000.00	1000000.00

Figure 17: Table of Salary Based Degree

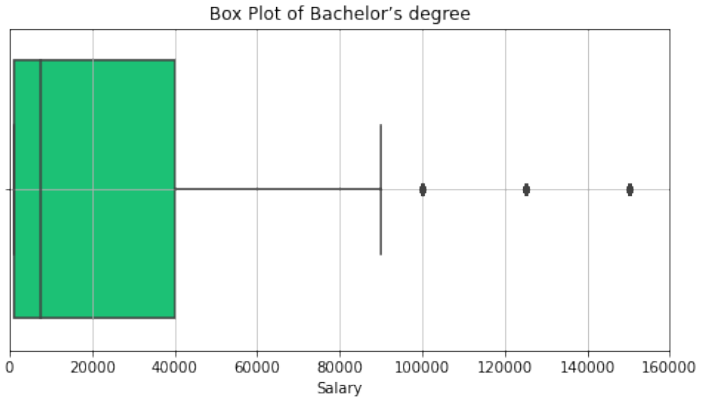

Figure 18: Box Plot of Bachelor's degree

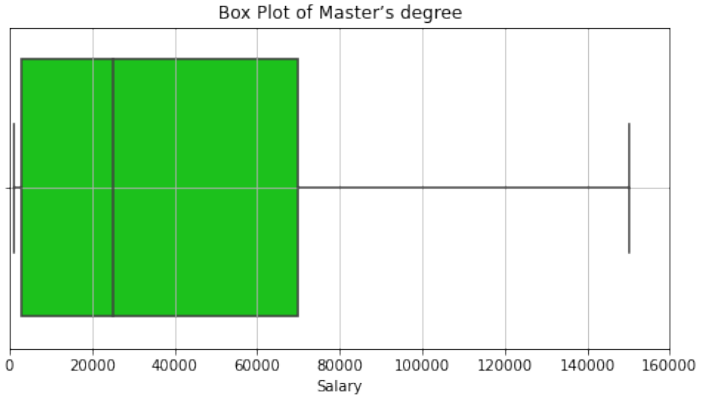

Figure 19: Box Plot of Master's degree

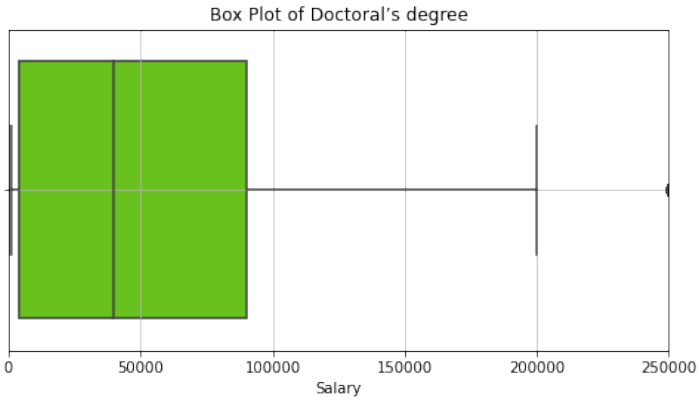

Figure 20: Box Plot of Doctor's degree

Figure 21: Assumption Checking for Q3

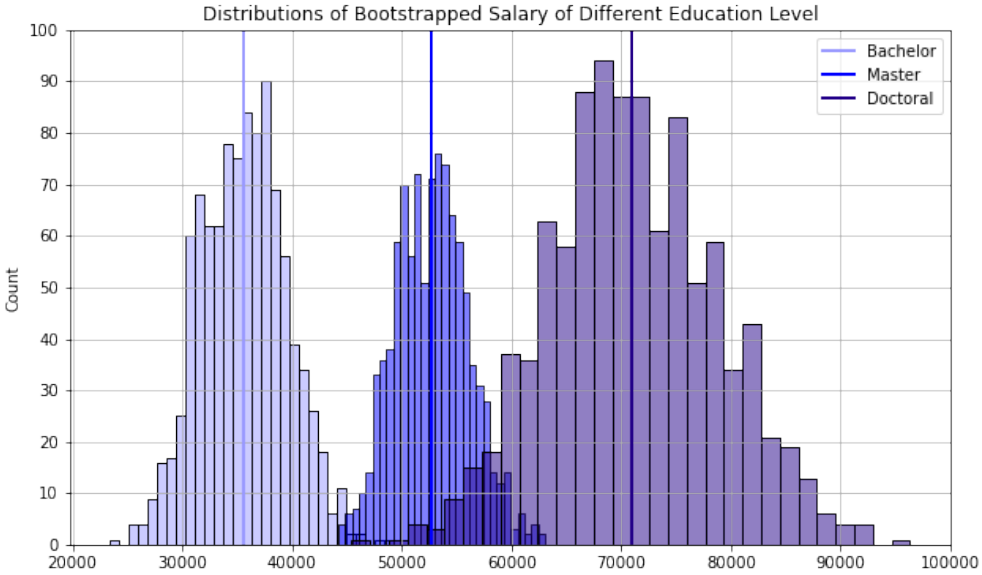

Figure 22: Distributions of Bootstrapped Salary of Different Education Level

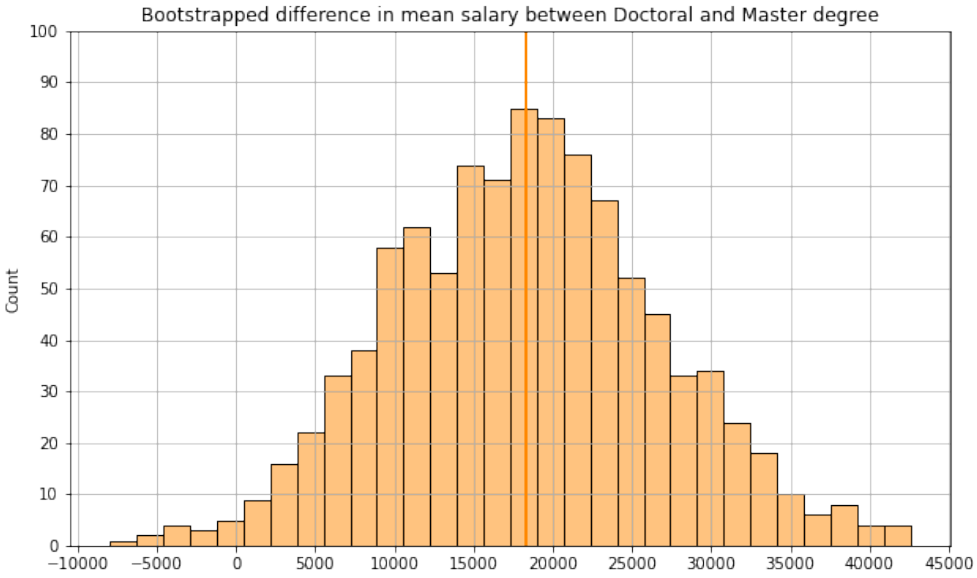

Figure 23: Bootstrapped difference in mean salary between Doctoral and Master degree

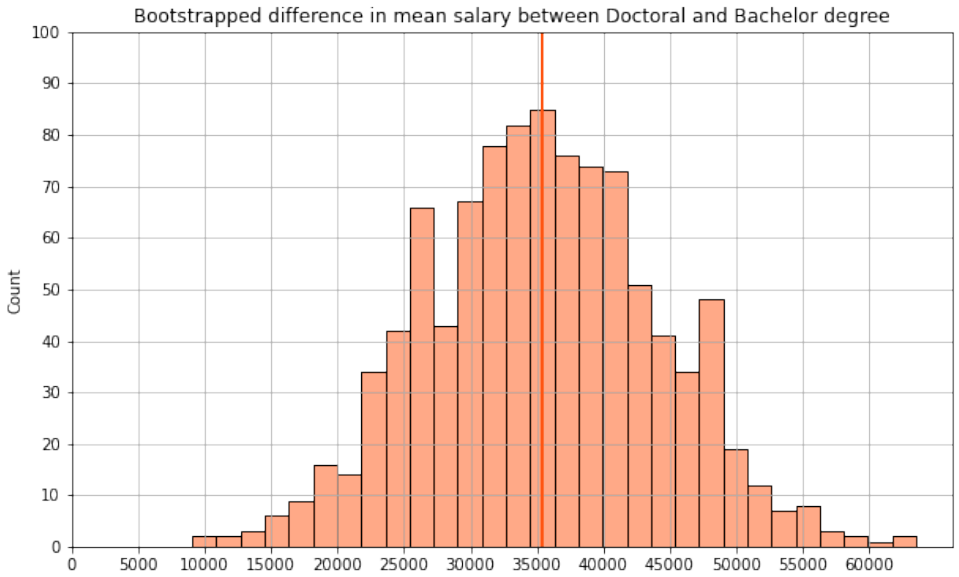

Figure 24: Bootstrapped difference in mean salary between Doctoral and Bachelor degree

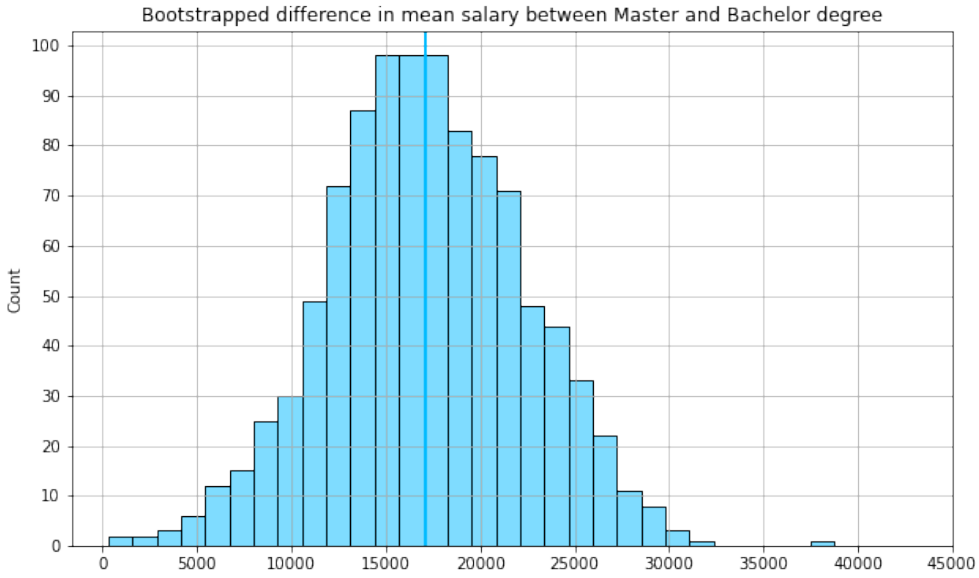

Figure 25: Bootstrapped difference in mean salary between Master and Bachelor degree