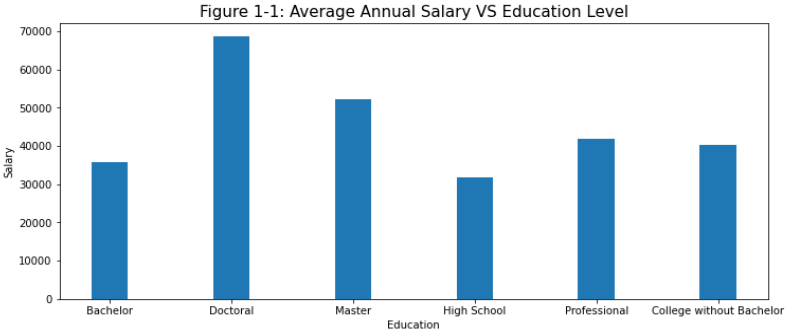
**Sicheng Zhu 1002194957**

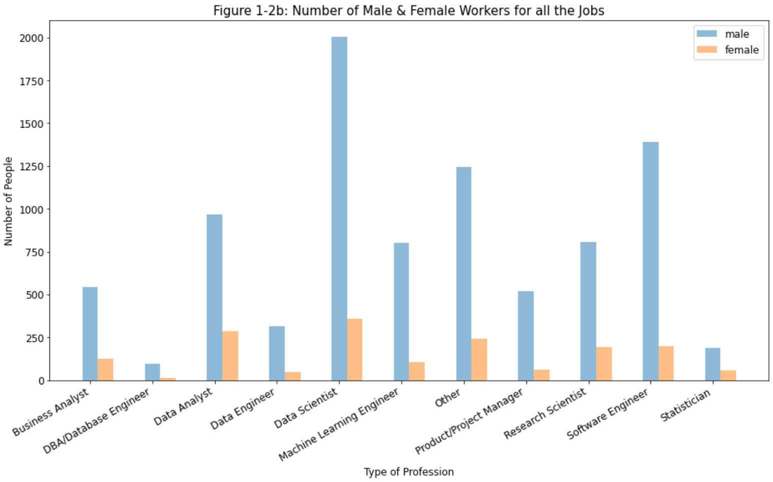
# MIE1624 Assignment 1

**Problem 1:** *Plot 1 - Average annual salary versus the education level.*



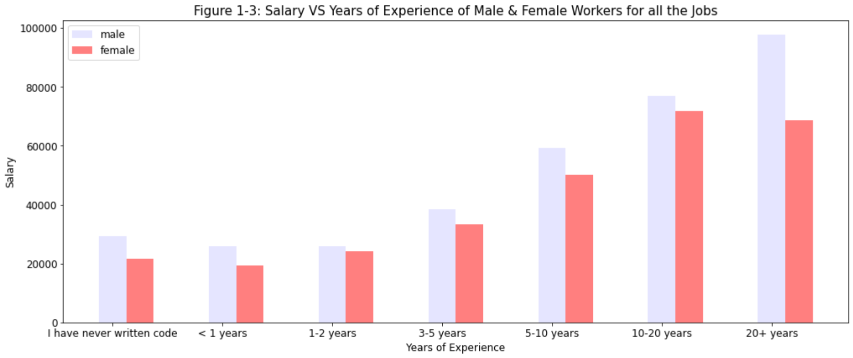
As shown in the plot, people with doctoral degree have the highest average annual salary ($68719) and people who does not go to college have the lowest average annual salary ($31854). In general, the average salary has a positive correlation with the education level. Another interesting finding is that people with professional degree has a little bit higher annual salary ($41892) than people with bachelor's degree ($35733).

**Problem 1:** *Plot 2 - Distribution of Men & Women of all Positions*



According to the pie chart (Figure: A-1) in appendix, the has the dominate portion (22.4%) among all positions. The machine learning engineer only takes 8.6%. From the Figure 1-2b, the number of men in all positions are much larger than women. The difference between the number of men and women are very large. The number of female workers in data analyst and data engineer are similar.

**Problem 1:** *Plot 3 - Average annual salary versus year of experience of men and women*



From the graph, in general, the average annual salary is higher with longer working experience. However, the average salary is approximate the same for 0 to 2 years experience. Women always have less salary than men and women with 10-20 years experience have higher average salary than women who worked more than 20 years.

**Sicheng Zhu 1002194957**

**Problem 2a:** For the men dataset, the mean is 50750.62. The median is 25000 and the standard deviation is 70347.97. For the women dataset, the mean is 36417.11. The median is 7500.0 and the standard deviation is 59442.71. The full details are shown in the python notebook.

**Problem 2b:** The assumptions of two-sample t-test are: 1. Two datasets have equal variance 2. Two datasets are independent. 3. Two datasets are randomly sampled 4. Two datasets should have normal distribution. [1] From the distribution plot (Figure A-2), two datasets are positively skewed, and we cannot use t-test.

**Problem 2c**: The distribution of bootstrapped data of men and women are shown below. The distribution of difference of means is also shown below. In general, the mean distributions of men and women are normal distribution. This corresponds to the central limit theorem. The distribution of mean difference is also normal distribution.

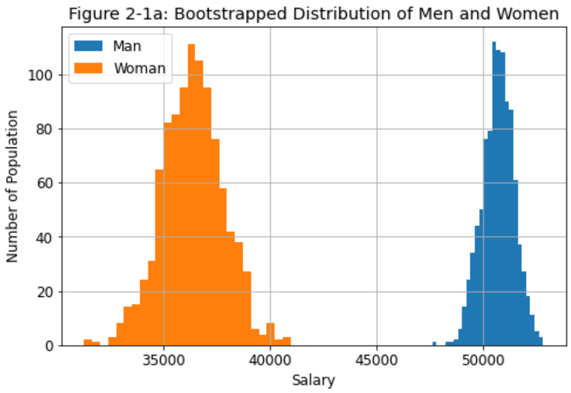


Figure 2-1a: Bootstrapped Distribution of Men and Women

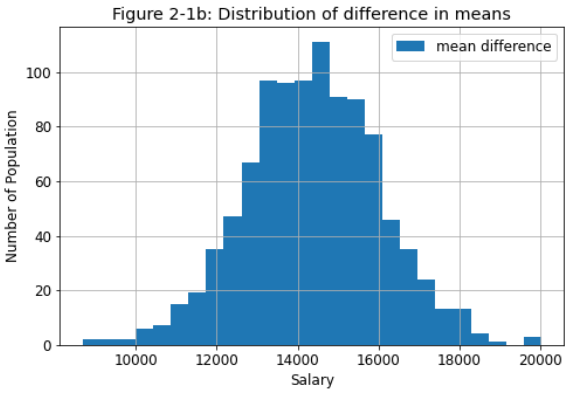


Figure 2-1b: Distribution of difference in means

**Problem 2d:** Since the mean distributions are normal distribution, we can apply two-sample t-test. We assume they have equal variance. The t-value is 278. The p-value is 0.0. The p-value is much smaller than the threshold (0.05).

**Problem 2e:** Thus, the null hypothesis (equal means) is rejected, and difference in means of two datasets are statically significant. Bootstrap is a method is to replicate the random sampling and to estimate the distribution of whole population. [2] The salary of men is always higher than women.

**Problem 3a:** For Bachelor dataset, the mean is 35732. The median is 10000 and the standard deviation is 60247. For Master dataset, the mean is 52120. The median is 25000 and the standard deviation is 67681. For Doctoral dataset, the mean is 68719. The median is 40000 and the standard deviation is 85403. The full details are shown in the python notebook.

**Problem 3b:** The ANOVA test has the same assumptions as t-test. 1. All datasets have equal variance 2. The datasets are independent. 3. The datasets are randomly sampled 4. Datasets should have normal distribution. [3] From the distribution plot (Figure A-3), all datasets are positively skewed, and we cannot use ANOVA test.

**Problem 3c:** The distribution of mean values of bootstrapped data is shown below. They form normal distributions. The mean difference distribution is shown in the appendix (Figure A-4), and it is also normal distribution.

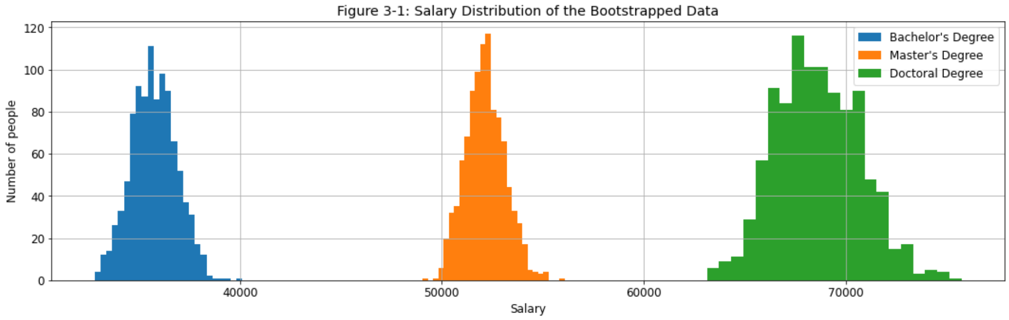


Figure 3-1: Salary Distribution of the Bootstrapped Data

**Problem 3d:** Since the mean distributions are normal distribution, we can apply one-way ANOVA test. We assume they have equal variance. The f-value is 133894. The p-value is 0.0. The p-value is much smaller than the threshold (0.05).

**Problem 3e:** The null hypothesis (equal means) is rejected and difference in means of three datasets are statically significant. People with higher education level will have higher average annual salary.

**Sicheng Zhu 1002194957**

## Appendix:

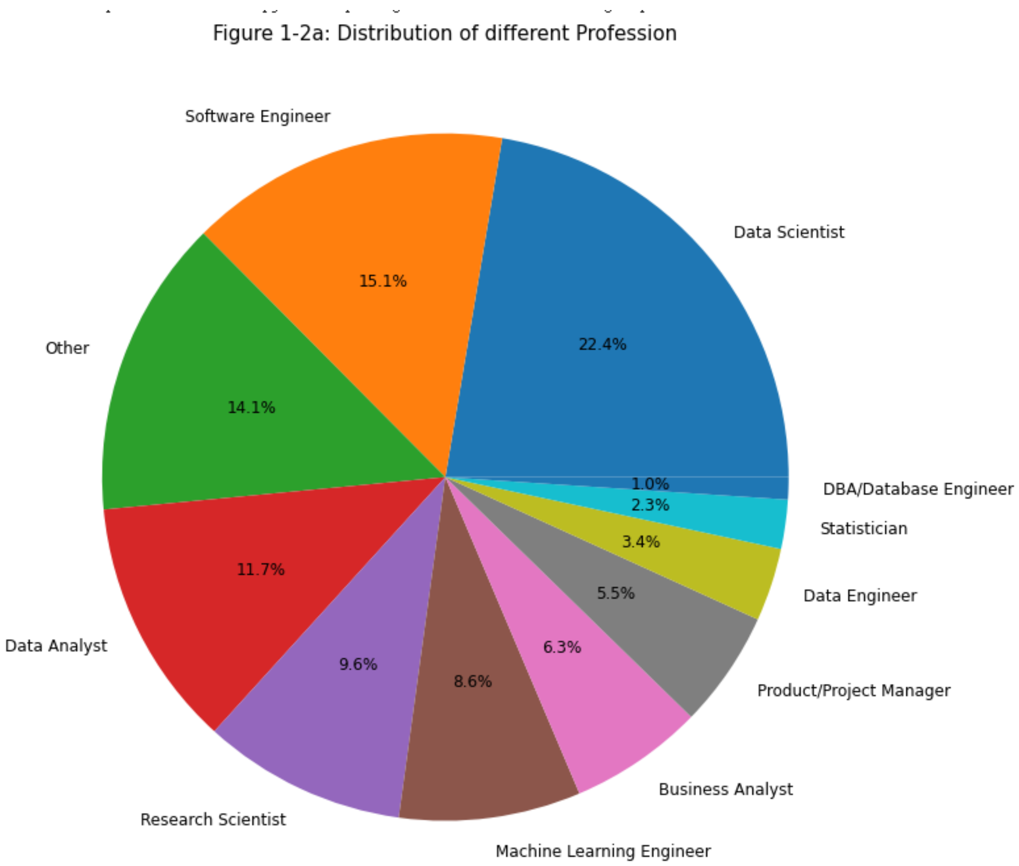


*Figure A-1: The Distribution of different Profession*

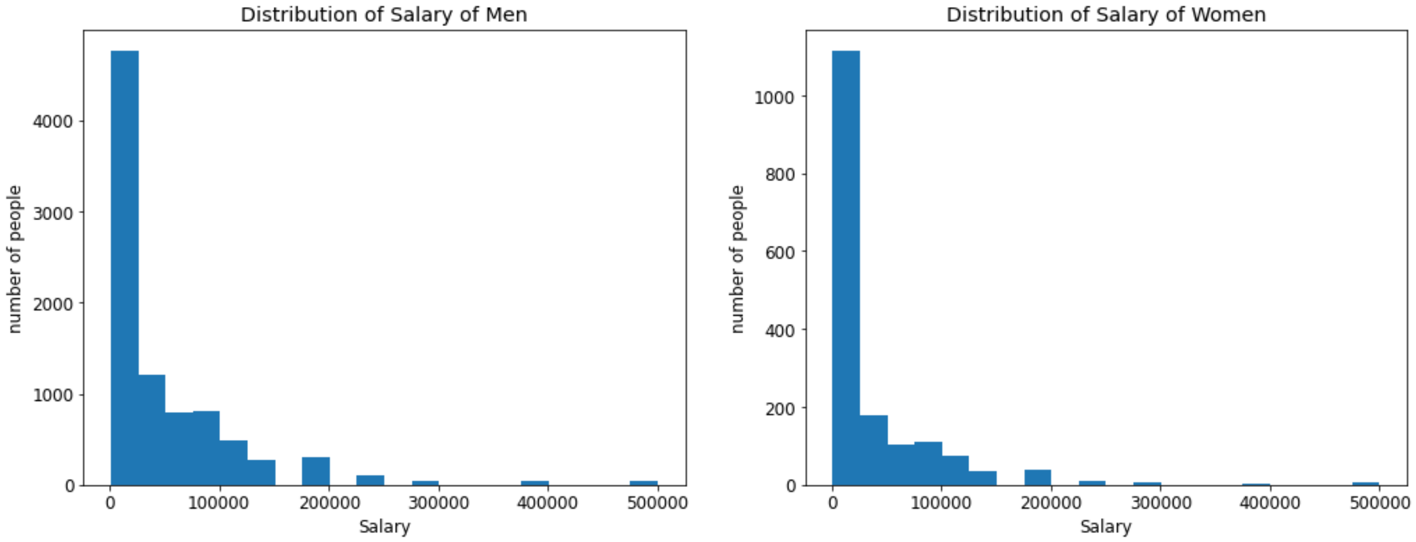


*Figure A-2: The Distribution of Annual Salary of Men & Women*

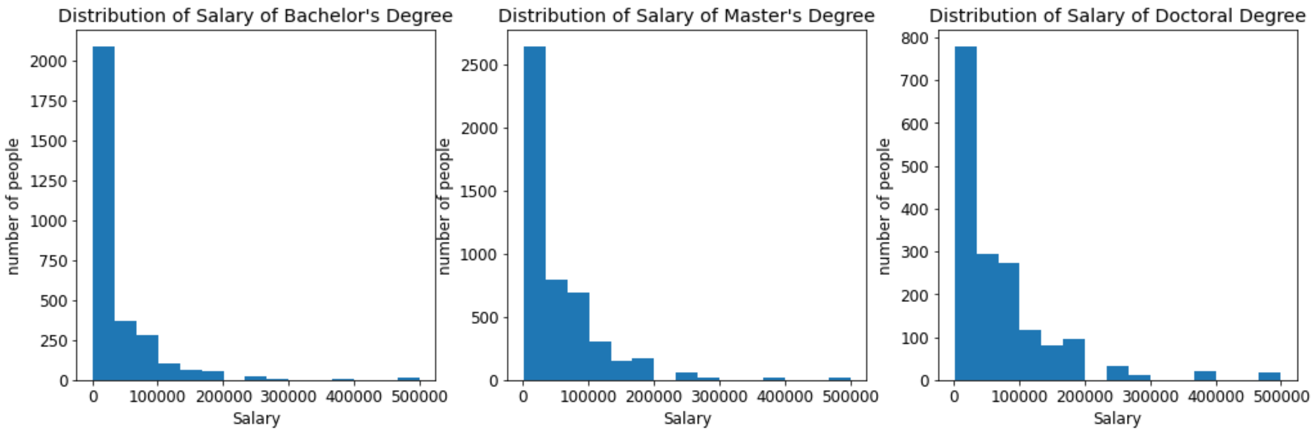


*Figure A-3: The Distribution of Annual Salary of Bachelor, Master and Doctoral*

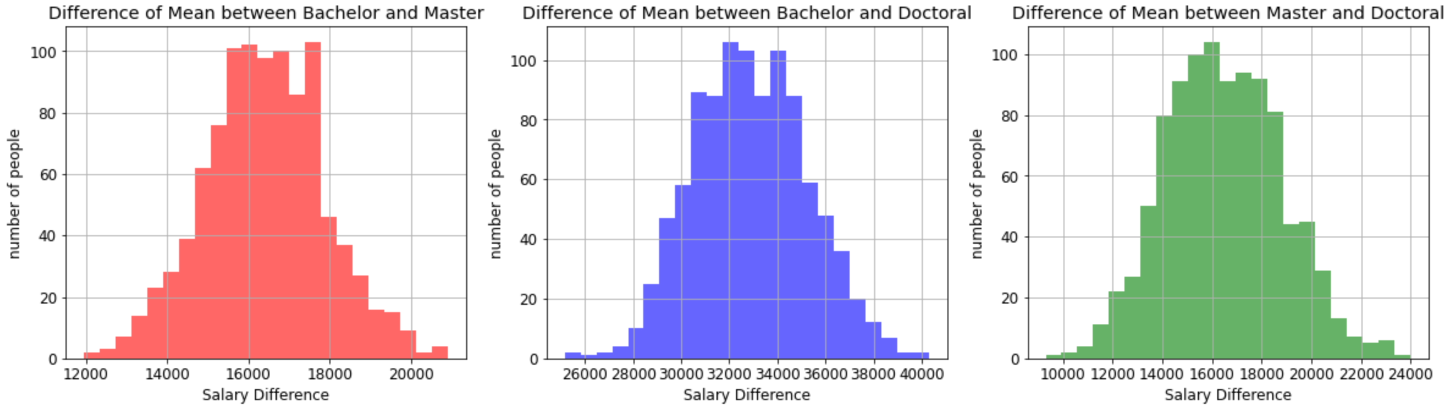


*Figure A-4: Distribution of Mean Difference of Different Education Levels*

## Reference:

[1] “Two-sample T-test,” JMP. [Online]. Available: https://www.jmp.com/en_ca/statistics-knowledge-portal/t-test/two-sample-t-test.html. [Accessed: 11-Oct-2021].

[2] V. Sharma, “Bootstrap sampling using Python's numpy,” Medium, 11-Jun-2020. [Online]. Available: https://medium.com/swlh/bootstrap-sampling-using-pythons-numpy-85822d868977. [Accessed: 11-Oct-2021].

[3] Renesh Bedre, “ANOVA using python (with examples),” Renesh Bedre, 26-Sep-2021. [Online]. Available: https://www.reneshbedre.com/blog/anova.html. [Accessed: 11-Oct-2021].