REPORT ON 2020 KAGGLE MACHINE LEARNING AND DATA SCIENCE SURVEY

This project explores the `2020 Kaggle Machine Learning (ML) & Data Science (DS) Survey` dataset to understand the nature of women's representation in Machine Learning (ML) and Data Science (DS) industry. It also gives an insight into how different levels of formal education affects annual income.

Exploratory Data Analysis (EDA)

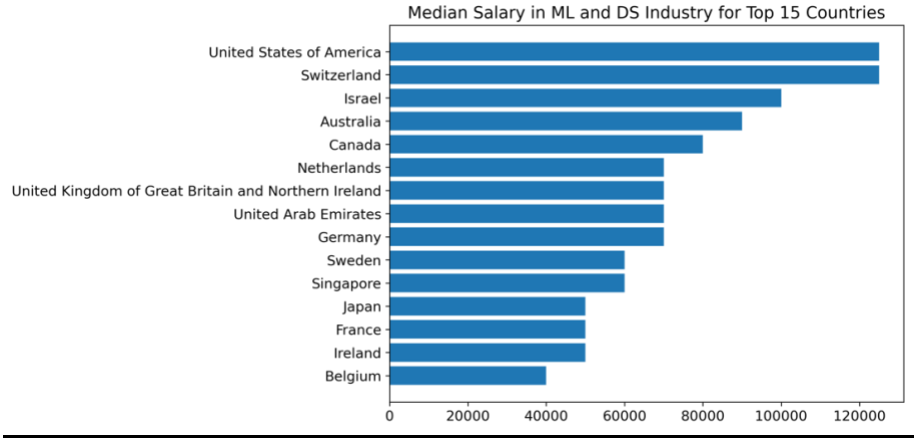

Figure 1. Median Salary in ML and DS Industry for Top 15 Countries

The bar chart in Figure 1 shows the median salary in ML & DS fields for the top 15 countries. The US and Switzerland are the top-paying countries in the industry with practitioners earning approximately $125,000 on average.

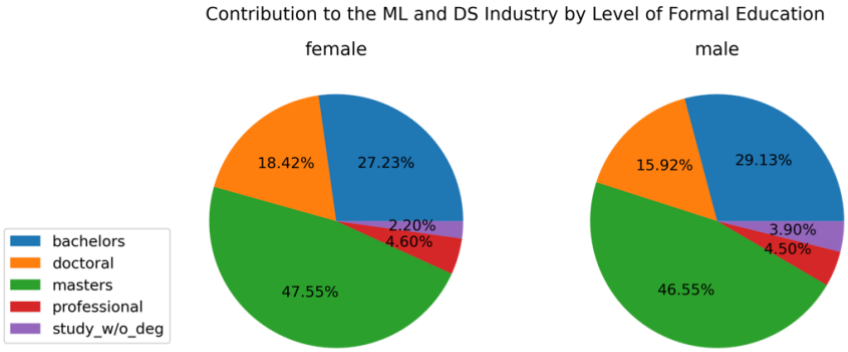

Figure 2. Contribution to the ML and DS Industry by Level of Formal Education

As shown in Figure 2, most of the practitioners in ML & DS professions are master's degree holder. They make up approximately 50% of the total distribution for both genders.

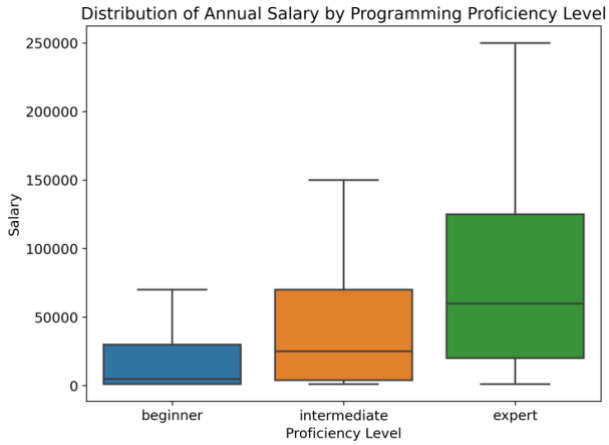

Figure 3. Distribution of Annual Salary by Programming Proficiency Level

The boxplot in Figure 3 shows the distribution of annual salary income among different levels of programming proficiency. It is interesting to note that some experts earn as low as what beginners earn, but they can also earn as high as $250,000 annually.

Estimating the difference between the average salary of men and women with two-sample t-test (**α**=0.05)

The analysis on the dataset shows that the average salary of men is 1.4 times larger than that of women. Also, it is apparent that women are under-represented in Machine Learning and Data Science fields as 16% of them participated in the survey.

As shown in Figure 4, the distributions of salary for both genders are positively skewed. Since these distributions fail to meet the assumption of Normality for a Two-Sample T-test, the test cannot be computed. A bootstrap of 1000 replications was performed maintaining the relative size for each group in the original sample.

According to Central Limit Theorem (CLT), the distribution of a bootstrapped sample means is normally distributed (as shown in Figure 5); therefore, no need to test for normality of distribution for the bootstrap samples.

Since the distributions are normal, a homoscedasticity test was performed with Bartlett's test which rejected the null hypothesis(H_0) of equal variances with a p-value of 1.20e-93.

With different variances, a Welch's Two-sample t-test was computed with a p-value of 0. This shows convincing evidence that there is a statistically significant difference between the average annual salary for male and female in ML & DS Industry.

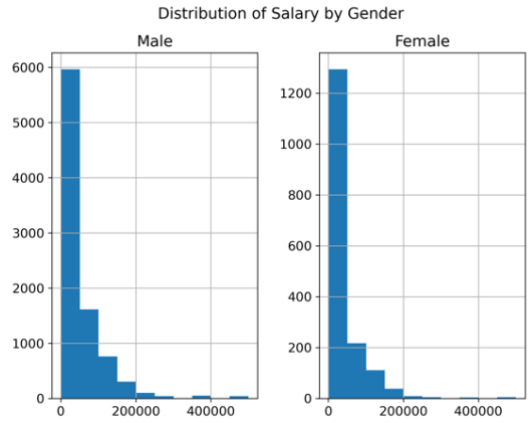

Figure 4. Distribution of Salary by Gender

Figure 5. Means distribution of Salary by Gender

Using analysis of variance (ANOVA) to compare the means of salary for levels of academic degree (**α**=0.05)

Conforming to the analysis carried out, the level of formal education is a determinant factor for annual salary. The average salary for doctoral degree approximately doubles that of the bachelor's degree with average salary for master's degree in-between. According to the histograms in Figure 6, none of the distribution looks normal. Hence, the ANOVA test cannot be performed. A bootstrap of 1000 replications was performed maintaining the relative size for each degree level in the original sample. According to CLT and the histogram shown in Figure 7, the bootstrapped samples are normally distributed.

A Bartlett's test for homogeneity of variance was computed with a p-value of 4.33e-150 thereby rejecting the null hypothesis (H_0) of equal variance among the groups. Welch's one-way ANOVA test is suitable for this test since it involves a single independent factor (three levels of academic degree), a dependent variable (means salary) and different variances. The result from test has a p-value of 0 which signifies that the difference in average annual salary among the three levels of academic degree is statistically significant.

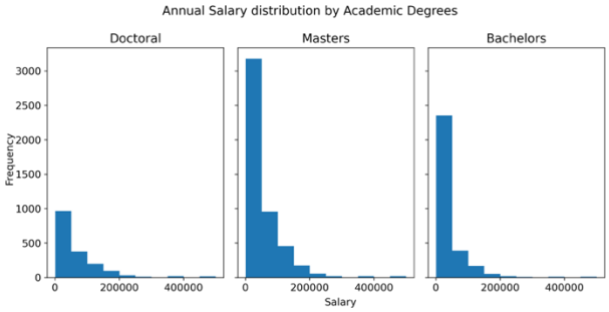

Figure 7. Annual Salary distribution by Academic Degrees

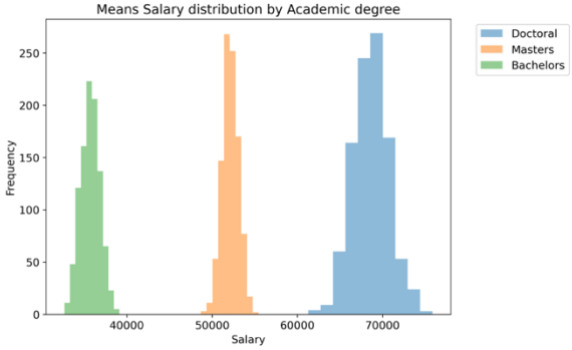

Figure 6. Means Salary distribution by Academic degree