# REPORT PART 1

The clean_kaggle_data.csv file is used to study the survey result to find out the nature of women's representation in Data Science and Machine Learning and the effects of education on income level. The coding part is in the ipynb file, and this report provides detailed analysis along with the plots and tables presented in the appendix.

First, we start with performing exploratory data analysis to summarize main characteristics. The dataset contains question answers in 369 columns answered by the participants of the survey and with table 1 in appendix, we could see a basic description of the dataset with mean, standard deviation, min, max etc. Since in the analysis we are going to perform, we will focus on Age, Gender, Country, Education, Professional Experience and Salary, we build a new data frame which contains these columns only and this data frame along with its descriptions are shown in table 2 and 3 in appendix. To use plots to represent different trends in the data, line plots have been chosen because the trend can be visualized clearly in line plots. By considering the purpose of our study, the exploratory analysis is based on the average salary among different age groups, education levels and professional experiences' lengths. The plot of average salary trend among different age groups' is shown as plot 1 in appendix, and it shows that except for the decrease in 50-54 and 60-69-years old group, overall, average salary increases when age increases. The plot of average salary trend among different education levels is shown as plot 2 in appendix, and it shows that except for people who have no formal education past high school and people who studied some part of university/college but not earn a degree has higher average salary then people who own bachelor's degree, overall, average salary increases with education level. The plot of average salary trend among different professional experience's lengths is shown as plot 3 in appendix, and it shows that except for the first few groups where professional experience less then 1 year have less average salary then people who have never written code before, and people who have never written code before have similar average salary as people who have 1-3 years' experience, the overall trend of average salary increases when the length of professional experiences increases.

Next, we perform the estimation of the difference between average salary of men versus women. The descriptive statistics are shown as table 4 and 5 in appendix. Table 4 shows that among 12642 men in the dataset, the maximum salary is $1000000 and the minimum is $1000, while the mean is $51193.60 and the median is $20000. Table 5 shows that among 2482 women in the dataset, the maximum salary is $1000000 and the minimum is $1000, while the mean is $34816.88 and the median is $7500. From these results, the maximum and minimum are the same, which indicates that men and women both could gain the maximum salary if they are qualified. It also shows that the average salary of women is $16376.72 less than that of men, and the median is $12500 less than that of men, which indicates in general, women receive lower in this industry then men which might be an implication of current nature of women in this industry. Since we need to study whether the average salary between men and women are statistically significantly different, a two-sample t-test is performed with a 0.05 threshold. Before the t-test is performed, the variances of the two groups have been printed to test for equal variance and the result seems to be relatively equal. Since t-test is valid for large sample (sample size larger than 50) from non-normal distribution, we perform a two-sample t-test with equal variance. Since the p-value of t-test is

smaller than 0.05, we reject the null hypothesis that average salary of men is equal to average salary of women. Then we proceed to perform bootstrap on the data with 1000 replications to make it approximately normally distributed because of Central Limit Theorem (CLT) to avoid large type I error. Plots of two bootstrapped distribution and the distribution of the difference in means are shown as plot 4 and 5 in appendix. Then two-sample t-test is conducted using bootstrapped data, by the printed result of variances, it still has similar variance and after bootstrapping the plot is bell-shaped which implies it is approximately normal, then we conduct a t-test with equal variance and the p-value of t-test is smaller than 0.05, so we reject the null hypothesis that average salary of men is equal to average salary of women on bootstrapped data. The overall analysis in this part shows that the mean salary of men and women are not equal, and the mean salary of women is less than men which indicates that the nature of women in this industry is relatively hard especially in salary achieved as compared to men.

Furthermore, three groups from 'highest level of formal education' are selected, which include Bachelor's, Master's, and Doctoral degree, to perform comparison of their average salary. The descriptive statistics tables are shown as table 6, 7 and 8 in appendix. From these tables, they show that the minimum and maximum salaries of three groups are the same, while the median salary is ranging from $7500 for Bachelor, $25000 for Master to $40000 for Doctoral, and the mean salary is ranging from $35578.29 for Bachelor, $52706.87 for Master to $70641.18 for Doctoral. From the descriptive statistics, it shows that people from all education levels could achieve the maximum amount of salary, however, overall, the higher education level people achieve, the higher the average salary. To perform a test to compare the means of salary among these three groups, one-way ANOVA F-test is used to test for the null hypothesis that the means of salary among these three groups are equal. Since assumptions need to be met before ANOVA F-test, Shapiro Wilk Test is used on three groups respectively to check for normality, and since all p-values are smaller than 0.05, we reject the null hypothesis of normality. Then the variances of these three groups are printed and the results are relatively homogeneous. However, since the normality assumption are not met, it is not suitable to perform ANOVA test at this stage. To let the data approximately normal, bootstrap is performed. Data has been bootstrapped with 1000 replications and the plots of three bootstrapped distributions (for Bachelor's, Master's and Doctoral degree) and the distribution of the difference in means are provided in plot 6, 7, 8, and 9 in appendix. From these plots, the data for each group are shown to be approximately normal and the average salary of Doctoral is the highest followed by Master and Bachelor. The difference in mean is approximately normally distributed as well. Since after bootstrap the data is approximately normal because of CLT, we could then proceed to the ANOVA F-test. Variance of each group after bootstrap has been printed and they are relatively homogeneous. Since they are also independent, ANOVA F-test is then performed, and since the p-value of ANOVA F-test is smaller than 0.05, we reject the null hypothesis that average salary for Bachelor, Master and Doctoral degree are the same on the bootstrapped data. From the plots, the result of the ANOVA F-test and the mean and median in the descriptive statistics, they all indicate that in this industry, people with higher education level could receive relatively higher income.

Thus, the nature of women in this industry is relatively harder especially in average salary achieved as compared to men and people with higher education level could receive relatively higher income.

# Appendix

## Tables:

| | Unnamed: 0 | Time from Start to Finish (seconds) | Q25 | Q30_B_Part_1 | Q30_B_Part_2 | Q30_B_Part_3 | Q30_B_Part_4 | Q30_B_Part_5 | Q30_B_Part_6 | Q30_B_Part_7 | Q30_B_OTHER |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 15391.000000 | 1.539100e+04 | 15391.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| mean | 12955.828926 | 9.260347e+03 | 49116.009356 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| std | 7493.072541 | 8.849740e+04 | 98090.207788 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| min | 1.000000 | 1.210000e+02 | 1000.000000 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 25% | 6511.000000 | 5.450000e+02 | 2000.000000 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 50% | 12916.000000 | 7.500000e+02 | 15000.000000 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| 75% | 19440.500000 | 1.140000e+03 | 60000.000000 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| max | 25973.000000 | 2.488653e+06 | 1000000.000000 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

Table 1

| | Age | Gender | Country | Education | Professional Experience | Salary |
|---|---|---|---|---|---|---|
| 0 | 50–54 | Man | India | Bachelor’s degree | 5–10 years | 30000 |
| 1 | 50–54 | Man | Indonesia | Master’s degree | 20+ years | 70000 |
| 2 | 22–24 | Man | Pakistan | Master’s degree | 1–3 years | 1000 |
| 3 | 45–49 | Man | Mexico | Doctoral degree | 20+ years | 40000 |
| 4 | 45–49 | Man | India | Doctoral degree | < 1 years | 40000 |
| ... | ... | ... | ... | ... | ... | ... |
| 15386 | 30–34 | Man | India | Bachelor’s degree | 1–3 years | 4000 |
| 15387 | 35–39 | Man | South Korea | Bachelor’s degree | 5–10 years | 90000 |
| 15388 | 30–34 | Man | Egypt | Bachelor’s degree | 1–3 years | 20000 |
| 15389 | 50–54 | Man | Sweden | Doctoral degree | I have never written code | 1000 |
| 15390 | 18–21 | Man | India | Bachelor’s degree | I have never written code | 1000 |

Table 2

| | Salary |
|---|---|
| count | 15391.000000 |
| mean | 49116.009356 |
| std | 98090.207788 |
| min | 1000.000000 |
| 25% | 2000.000000 |
| 50% | 15000.000000 |
| 75% | 60000.000000 |
| max | 1000000.000000 |

Table 3

| | Salary |
|---|---|
| count | 12642.000000 |
| mean | 51193.600696 |
| std | 99979.274378 |
| min | 1000.000000 |
| 25% | 2000.000000 |
| 50% | 20000.000000 |
| 75% | 60000.000000 |
| max | 1000000.000000 |

Table 4

| | Salary |
|---|---|
| count | 2482.000000 |
| mean | 34816.881547 |
| std | 72017.347888 |
| min | 1000.000000 |
| 25% | 1000.000000 |
| 50% | 7500.000000 |
| 75% | 50000.000000 |
| max | 1000000.000000 |

Table 5

| | Salary |
|---|---|
| count | 4777.000000 |
| mean | 35578.291815 |
| std | 89382.060777 |
| min | 1000.000000 |
| 25% | 1000.000000 |
| 50% | 7500.000000 |
| 75% | 40000.000000 |
| max | 1000000.000000 |

Table 6

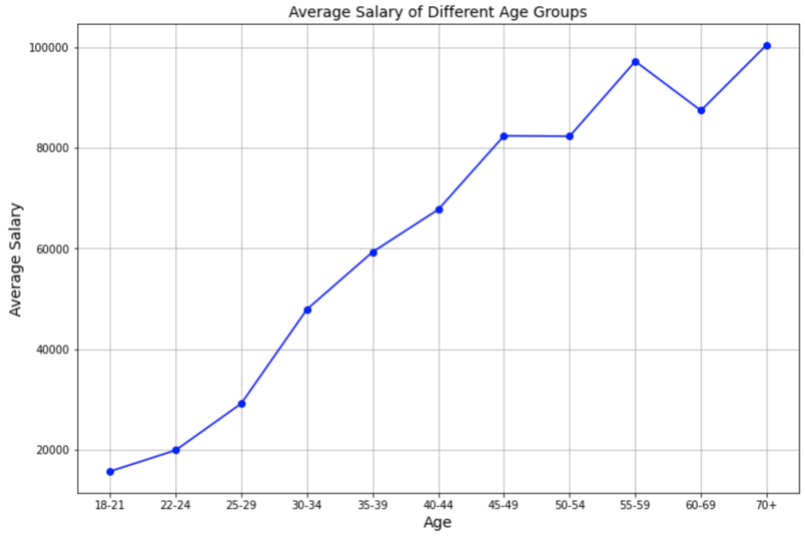
| | Salary |
|---|---|
| count | 6799.000000 |
| mean | 52706.868657 |
| std | 90928.786678 |
| min | 1000.000000 |
| 25% | 3000.000000 |
| 50% | 25000.000000 |
| 75% | 70000.000000 |
| max | 1000000.000000 |

Table 7

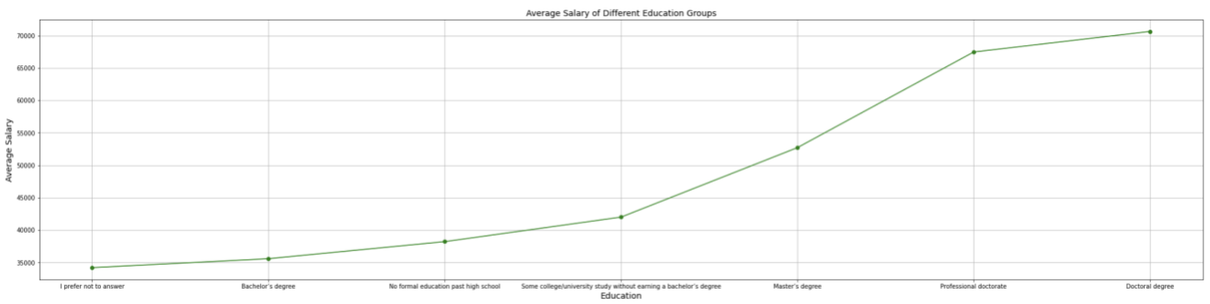
| | Salary |
|---|---|
| count | 2217.000000 |
| mean | 70641.181777 |
| std | 117160.947589 |
| min | 1000.000000 |
| 25% | 4000.000000 |
| 50% | 40000.000000 |
| 75% | 90000.000000 |
| max | 1000000.000000 |

Table 8

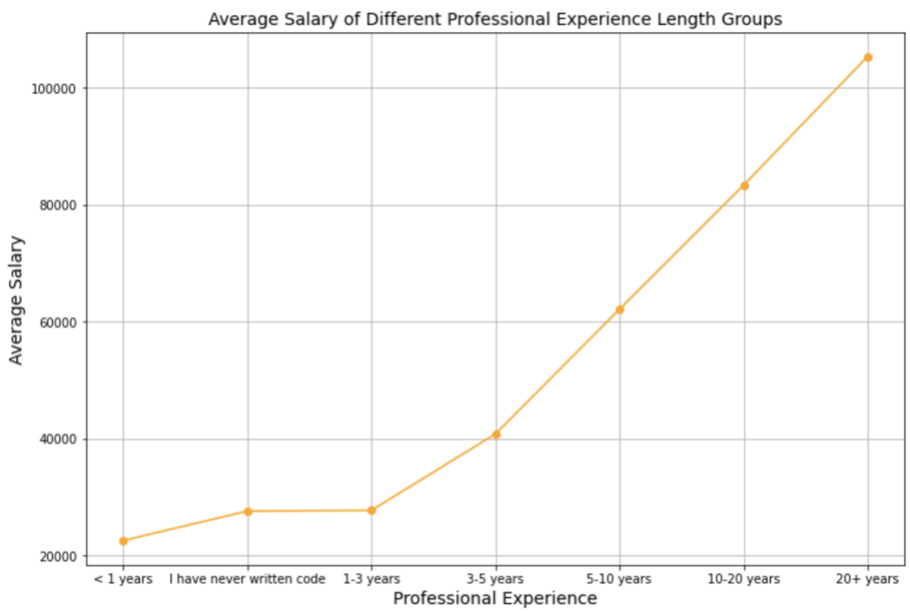
Plots:



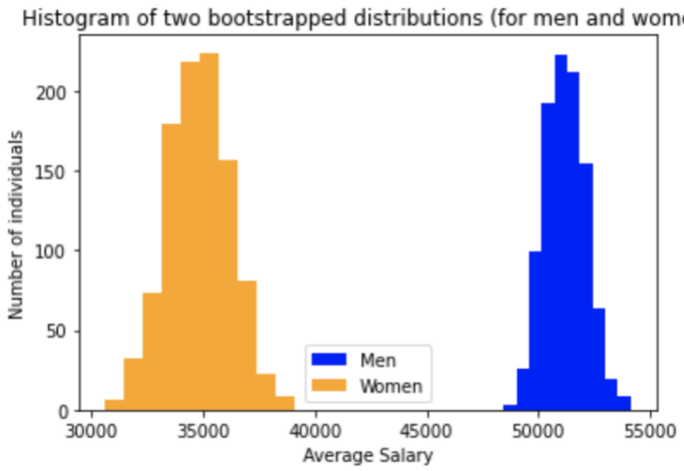

Plot 1



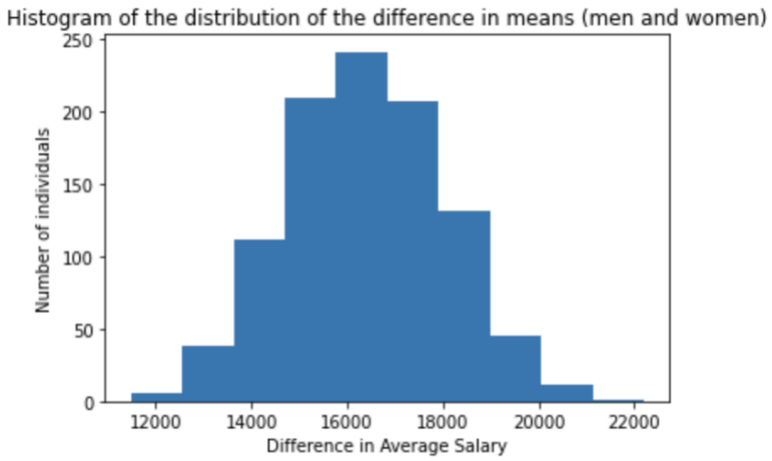

Plot 2



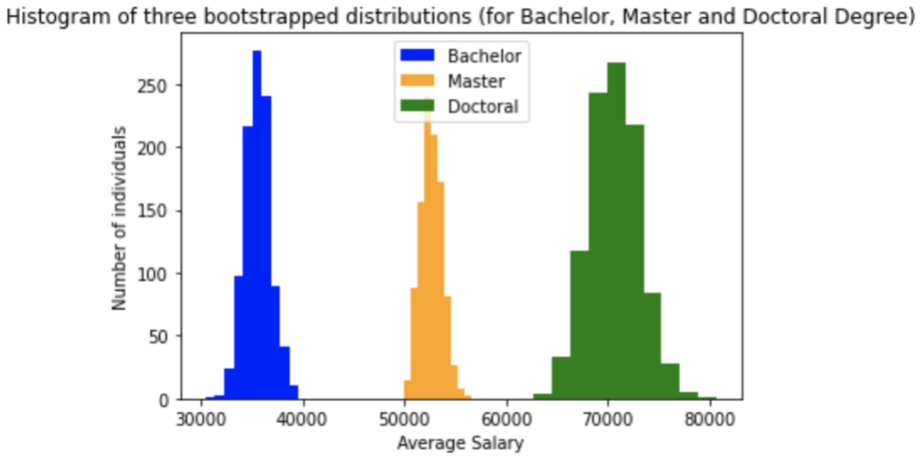

Plot 3



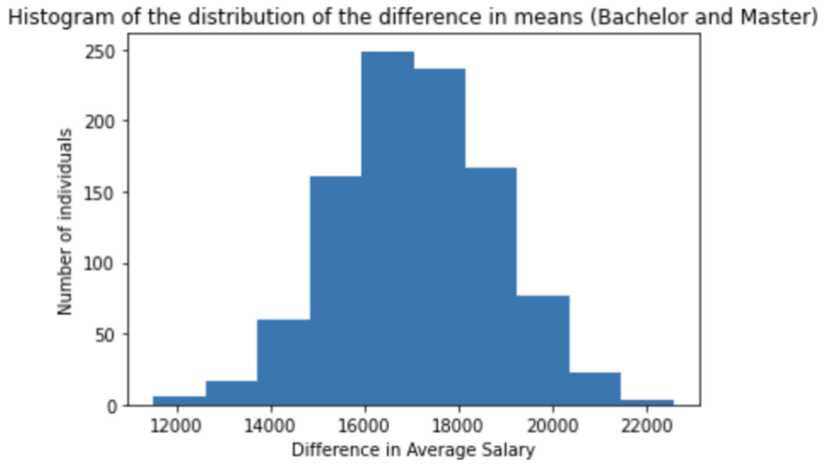

Plot 4



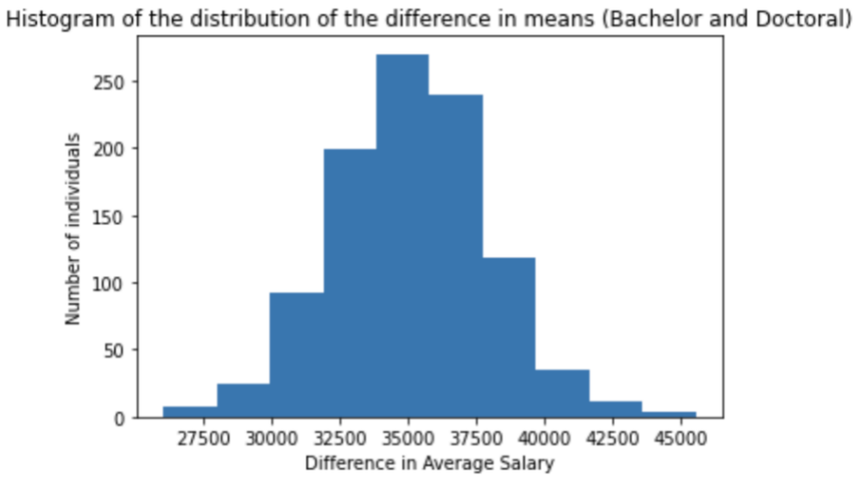

Plot 5



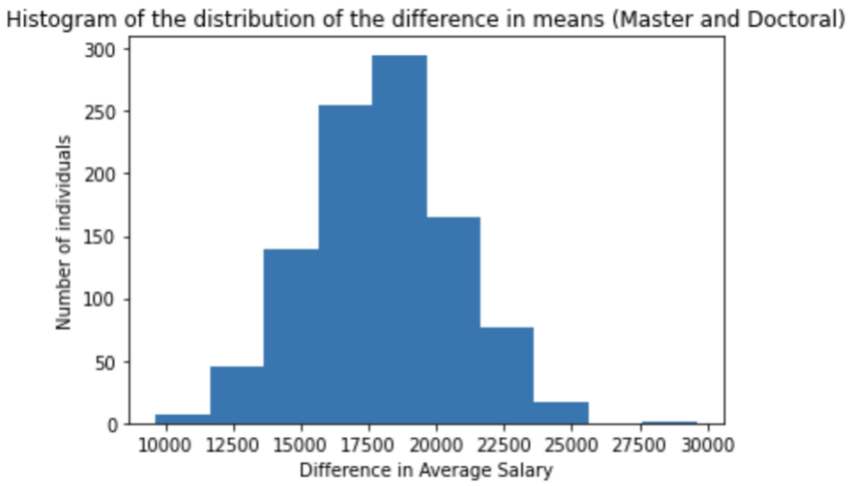

Plot 6




Plot 7


Histogram of the distribution of the difference in means (Bachelor and Doctoral)
Number of individuals
0
50
100
150
200
250
27500
30000
32500
35000
37500
40000
42500
45000
Difference in Average Salary

Plot 8


Histogram of the distribution of the difference in means (Master and Doctoral)
Number of individuals
0
50
100
150
200
250
300
10000
12500
15000
17500
20000
22500
25000
27500
30000
Difference in Average Salary

Plot 9