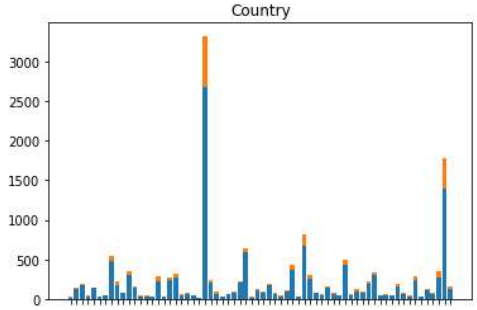
Introduction

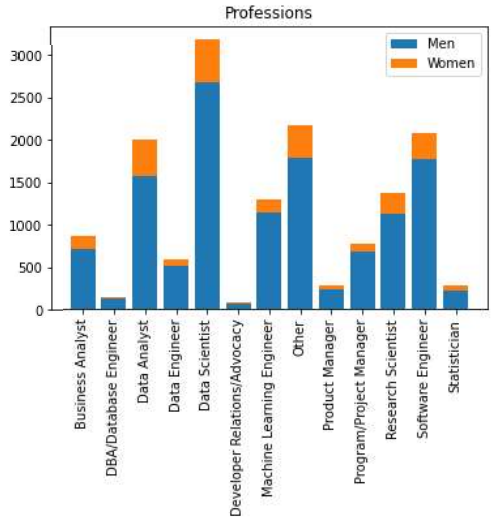
Kaggle, a very large online platform for hosting datasets and performing analysis, conducts an annual survey of its users. This paper seeks to, using this dataset, ascertain whether gender and education level are factors for income disparity according to the dataset.

Exploratory Data Analysis

The data used for this analysis is second party data from Kaggle who obtained the primary data by surveying their users in all countries and removing duplicate entries. The data is thus suitable for an analysis that is reflective of the data science and analysis community but suffers from self-selection bias. .

The cleaned data consists of 15391 entries from over fifty countries with various professions and age groups from 18 to 70+. There is a trend of increasing average salary by education, and a higher for men versus women. The data is unbalanced however as only 16% of respondents are women(2482), and the majority of respondents are from India or the United States.

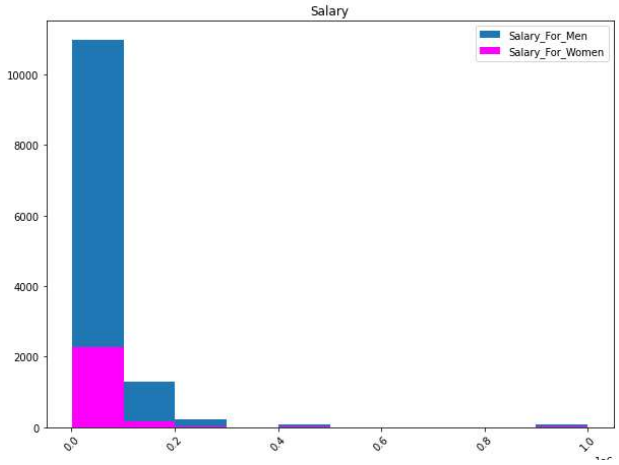




Analysis of Gender as a Factor (t- Test, confidence level 0.05)

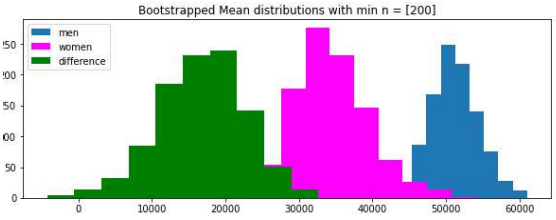
To compare the effect of two gender levels - men and women, on salary the t-test is to be performed. Student's t test requires independent sampling, similar variance and normally distributed samples.

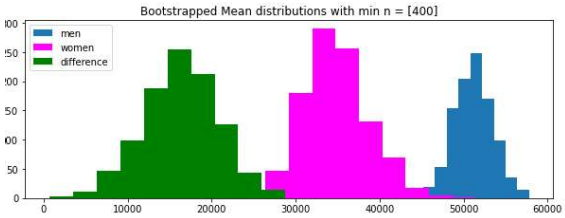
	Women	Men
Count	2482	12642
mean	34816.88	51193.60
Standard deviation	72017.35	99979.27
min	1000	1000
median	12630	20000
max	1000000	1000000



The distribution of salary by gender is non-normal and the application of Levene's variance test shows a p value under 0.05. With normality violated and heteroskedasticity present the t-test is unsuitable here.

In order to perform the t-test, the mean distribution was generated via bootstrapping similar percentages of the groups(men and women) to obtain a 1000 data point normal distribution. The variances remain statistically different, hence Welch's t-test for unequal variances is applied for a more reliable result



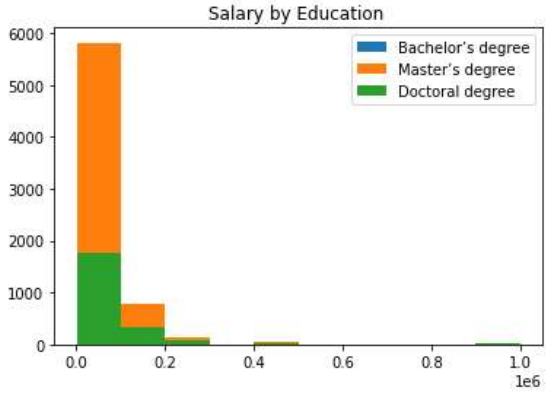


The t statistic is 116.65 with a p value close to 0, hence the null hypothesis that there is no difference in the mean salaries between men and women is not supported by the data.

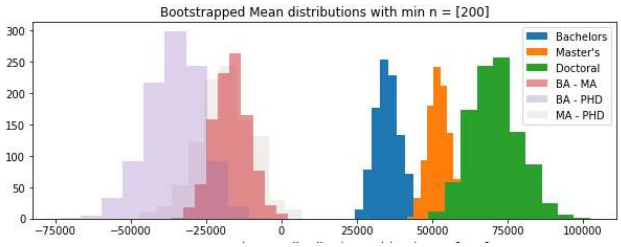
Analysis of Education as a Factor (ANOVA F test, confidence level 0.05)

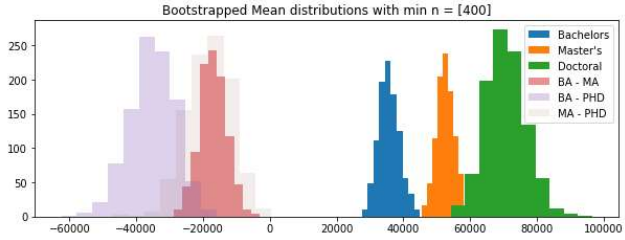
To compare the effect of education levels - Bachelors, Masters and PhD the same methodology is followed with analysis of variance rather than the t-test as there are more than two levels.

	Bachelors	Masters	PhD
Count	4777	6799	2217
mean	35578.29	52706.87	70641.18
Standard deviation	89382.06	90928.79	1117160.95
min	1000	1000	1000
median	7500	25000	40000
max	1000000	1000000	1000000



As with gender, the distribution is non-normal, and Levene's test shows statistical difference in the variance hence the F test is not applied to the raw data.





The mean distributions were generated as before to obtain normality. The differences in variances remain, but as the violation against normality has been removed the results from the ANOVA on this data would be more reliable than on the original data. The F statistic for the ANOVA is 109 with a p-value near 0 hence the null hypothesis that there is no difference in the mean salary between educational levels is not supported by this data. Due to the limitations in data this result may hold only for India and the U.S.