**Sugumar Prabhakaran** (**Id #:** 994126815) **Due Date:** 16 Feb 2021

# ASSIGNMENT 1 - KAGGLE ML & DS SURVEY CHALLENGE

1. **Exploratory Data Analysis (EDA)**. For EDA, the following three graphs were generated:

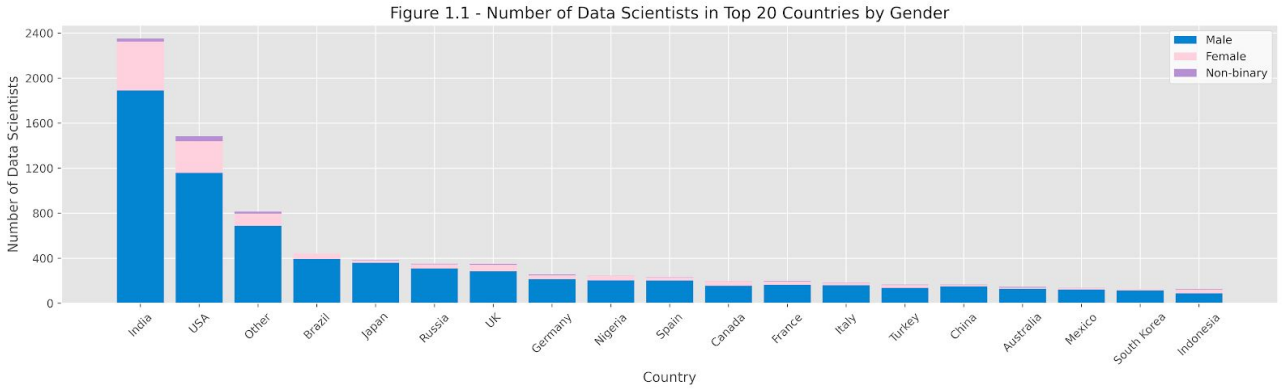



Figure 1.1 displays a breakdown of data scientists by gender for the top 20 countries. We observe that India and the United States have the most data scientists and they are predominantly male. All non-binary genders were grouped together.

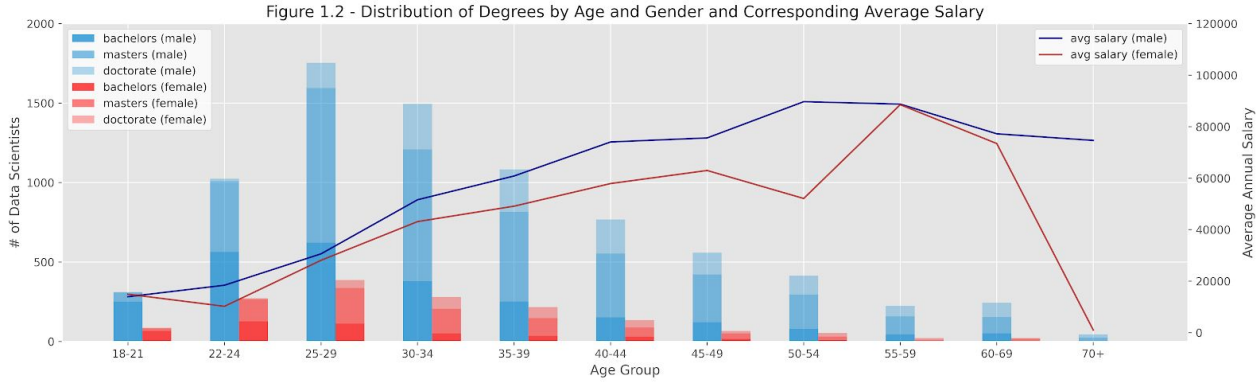



Figure 1.2 shows the average salaries of men and women grouped by age in the dark blue and red lines (right scale). There is a wage gap in the age range of 30 - 54 between men and women. The bars breakdown the number of data scientists (left scale) with bachelor's, master's and doctoral degrees by gender and age group, which could explain some of the gap.

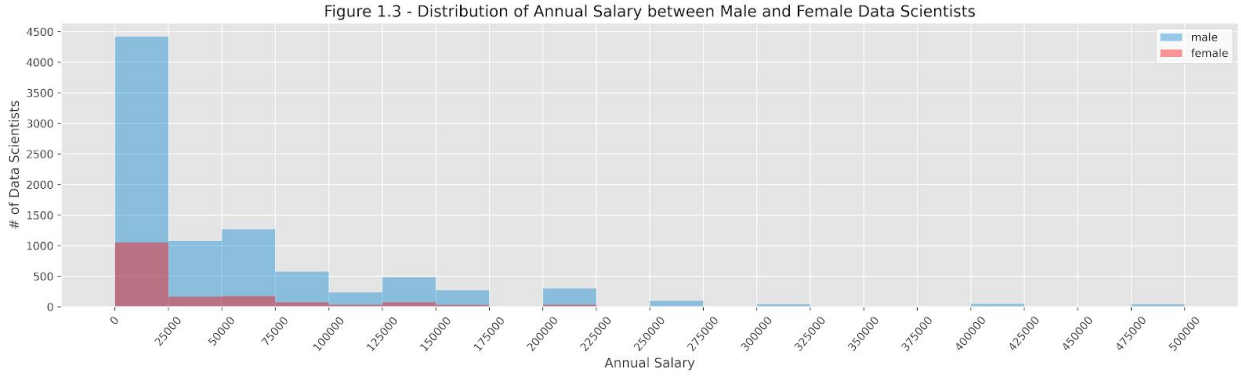



Figure 1.3 shows that the distribution of salaries for both men and women is skewed to the right. This is likely due to the large number of data scientists in India (see Figure 1.1), where wages are generally lower than western countries, which compose about 50% of the top 20 data science countries.

2. **Estimated Difference Between Male and Female Salaries**

(a) Descriptive Statistics. The sample mean salary for men ($\mu_{male}$) = **$50750.62** (standard deviation ($s_{male}$) = **$70347.97**) and women ($\mu_{female}$) = **$36417.11** ($s_{female}$= **$59442.72**). For the full table of descriptive statistics, refer to the .ipynb file.

(b) Two-Sample t-test. A two-sample t-test was performed because all underlying assumptions are met: (1) means are normally distributed as per the Central Limit Theorem (CLT), (2) variances are assumed equal and (3) male and female data is independent of each other. A **t-value = 7.84** and **p-value = 4.7e-15** were obtained. Using a threshold ($\alpha$) = 0.05, the null hypothesis ($\mu_{male} = \mu_{female}$) was rejected since $p < \alpha$. Therefore, the difference in means is statistically significant.

(c) Bootstrap Data to Compare Male and Female Salaries. Figures 1.4 and 1.5 show that all three: men's, women's and bootstrapped differences mean salaries are normally distributed.

Figure 1.4 - Bootstrap Distribution of Mean Annual Salary for Men and Women (1000 iterations)

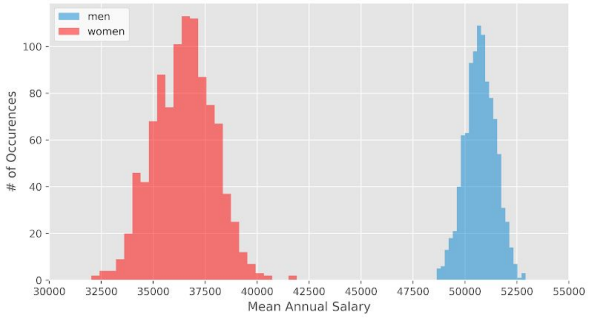
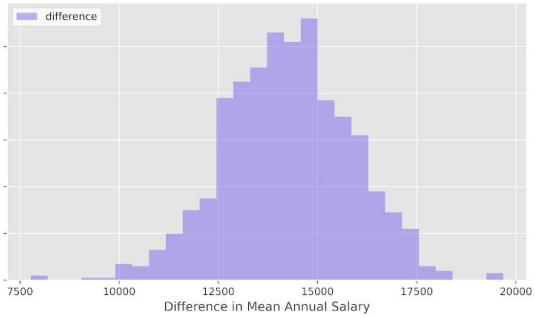



Figure 1.5 - Bootstrap Distribution of Difference in Mean Annual Salary between Men and Women (1000 iterations)




(d) Bootstrap Data t-test. A t-test was performed since same t-test conditions (para 2.b) are met and in particular, data is normally distributed as per bootstrap definition. The resulting **t-value = 278.73** and **p-value = 0.0**, and we still have $p < \alpha = 0.05$.

(e) Findings. We also reject the null hypothesis ($\mu_{male} = \mu_{female}$) using a t-test on bootstrap data and find that both methods determine that the difference in means is statistically significant.

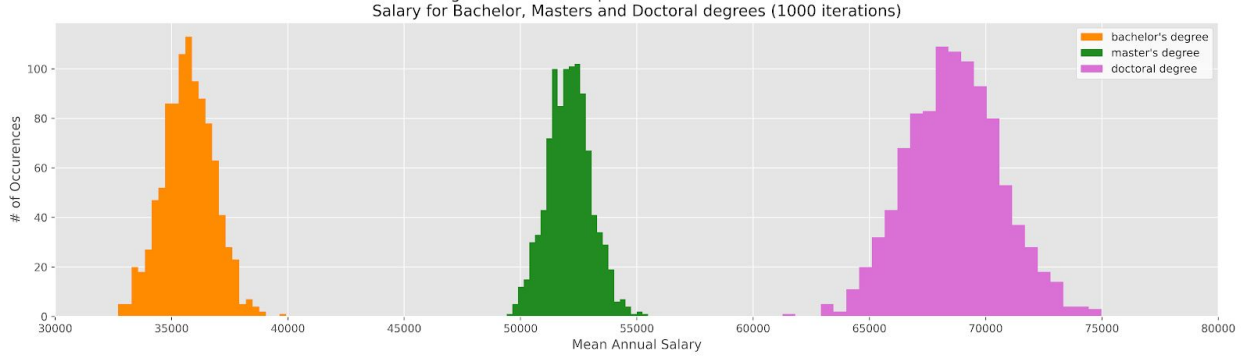
## 3. Estimated Differences in Salary Between Levels of Education

(a) Descriptive Statistics. The means for respective education levels are: $\mu_{Bach}$ = **$35732.82**, $\mu_{Mast}$ = **$52120.11**, $\mu_{PhD}$= **$68719.44**. The .ipynb file contains more statistics and Figure 1.6, which shows their histograms, which are all skewed to the right.

(b) Perform ANOVA. ANOVA was performed as all assumptions were met: (1) means are normally distributed as per CLT, (2) variances are assumed to be sufficiently similar, and (3) data for each degree are mutually independent. The null hypothesis ($\mu_{Bach} = \mu_{Mast} = \mu_{PhD}$) was rejected as the following results were obtained: **F-value = 129.76, p-value = 2.49e-56**, so $p < \alpha$. Therefore, at least one of the means ($\mu_{Bach}$, $\mu_{Mast}$, $\mu_{PhD}$) is statistically significantly different from the others.

(c) Bootstrap Data to Compare Education Levels. Figure 1.7 shows the bootstrap (normal) distributions for each degree. Figure 1.8 containing the differences between degrees, which are also normally distributed, can be found in the .ipynb file.

Figure 1.7 - Bootstrap Distribution of Mean Annual Salary for Bachelor, Masters and Doctoral degrees (1000 iterations)




(d) Bootstrapped Data ANOVA. Similar to 3b, all assumptions were met, and in particular, means are normally distributed for bootstrap as per figure 1.7. We obtain: **F-value = 130122.04, p-value = 0.0**, so $p < \alpha$ and we reject the null hypothesis.

(e) Findings. For both the original samples (3.b) and the bootstrapped means (3.d), we obtain the same result that at least one of the means ($\mu_{Bach}$, $\mu_{Mast}$, $\mu_{PhD}$) is statistically significantly different than the others.