1.Exploratory data analysis (3 plots to explore potential trends)

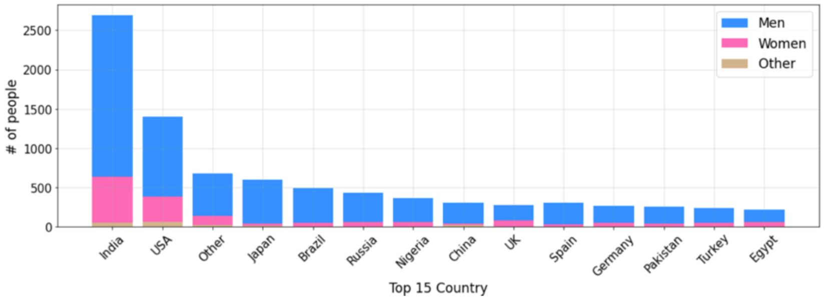
Figure 1.1:Number of people working in data science In Top 15 Countries By Gender

From **Figure 1.1**, we see that India have most people working in data science community, followed by United States. In each Top 15 country, the number of men in data science are much more than the number of women in that aspect.

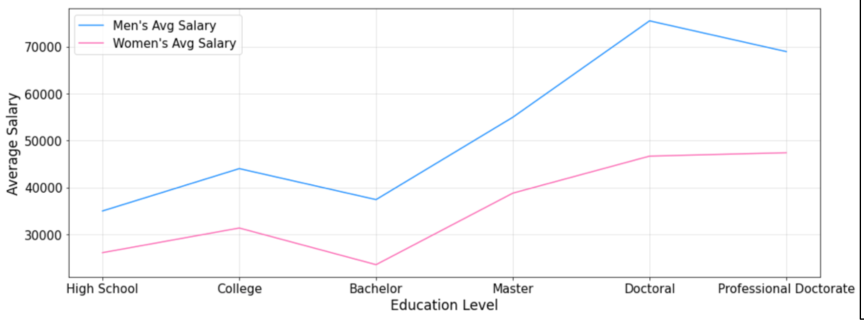
Figure 1.2: Average Salary of Data Scientists In Different Education Levels By Gender

From **Figure 1.2**, for people in different education levels, doctorals' mean salary is higher than masters' mean salary, masters' mean salary is higher than bachelors' mean salary (Assumption: Focusing on above three groups since above three groups occupy 90% data).

Furthermore, within the same education level, men's mean salary is higher than women's mean salary.

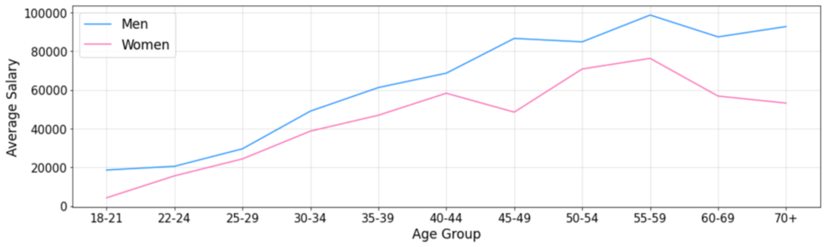
Figure 1.3: Average Salary Of Data Scientists In Age Groups By Gender

From **Figure 1.3**, in the data science community, within the same age group, men's mean salary is higher than women's mean salary. In the age group of 45-49, the wage gap between men and women is largest. Both of women and men's mean salary shows an upward trend overall with the age increase until 55 - 59.

2. Estimate the difference between average salary of men vs women

2.a: <u>Descriptive Statistics:</u> In original sample, mean salary ($) : μ_{men} = 51193.6, μ_{women} = 34816.88. Standard deviation: sd_{men} = 99979.3, sd_{women} = 72017.3. More statistics refer to the table in ipynb file. We found that among survey responses, the number of men is much more than the number of women in data science community (12642 men and 2482 women).

2.b: <u>Two-sample t-test on original sample:</u> As seen in the Figure 2.1(a) and (b) in .ipynb file, both distributions of men's salary and women'salary are not normal, thus the normality assumption is violated. Not all assumptions of t-test are met, we can't perform two-sample t-test on original samples.

2.c: <u>Bootstrapped Distribution (men vs women):</u> As seen in Figure 2.2 (a) and Figure 2.2 (b), both bootstrapped distribution of mean salary for men and women follow normal distribution. Bootstrapped distribution of difference in mean salary is also normal. Thus, after bootstrapping data, the normality assumption is satisfied.

2.d: **t-test on bootstrapped data:** **1.** Since original samples for men and women don't follow normal distributions, it is hard to do appropriate tests to compare mean salary of two groups. Thus, we need to introduce bootstrapping this distribution-independent tool to work on it. **2.** To study the difference between average salary of men vs women, I choose the variable salary(Q25) and gender with two levels (man & woman) for analysis. **3.** As seen in 2.c., normality assumption is satisfied. Also two boostrapped samples are independent. All assumptions of t-test are satisfied, we can perform two-sample t-test on bootstrapped data. Relevant analysis on t-test are results are shown in 2.e..

2.e: **Comment on findings:** The null hypothesis for t-test done in 2.d. is that there is no difference in men's mean salary and women's mean salary. From 2.d, results show t-statistic = 307.67 and P value = 0.0 (note: without setting seed, as approved by TA). Since P value < 0.05, we have strong evidence to reject the null hypothesis. The difference in men's salary and women's salary is statistically significant.

3. Estimate the difference between average salary of bachelor vs master vs doctoral

3.a: **Descriptive Statistics:** Mean salary ($):$\mu_{bachelor}$ = 35578.3, μ_{master} = 52706.9, $\mu_{doctoral}$ = 70641.2 Standard deviation:$sd_{bachelor}$ = 89392, sd_{master}=90928.8, $sd_{doctoral}$ = 117160.9;More refer to ipynb file. We found that among survey responses, most number of people working in data science community have Master degree. (4777 bachelors, 6799 masters and 2217 doctorals).

3.b: **ANOVA on original sample:** As seen in Figure 3.1 in .ipynb file, all original samples of salary for bachelors, masters and doctorals don't follow normal distribution, so normality assumption is violated. Not all assumptions of ANOVA are met, we cannot perform ANOVA test on original samples.

3.c: **Bootstrapped Distribution (bachelor vs master vs doctoral):** As seen in Figure 3.2 in.ipynb file, all bootstrapped distributions of mean salary for bachelor, master and doctoral follow normal distributions. As seen in Figure 3.3, All three bootstrapped distribution of difference in mean salary also approximately follow normal distributions. Thus, the normality assumption is satisfied after bootstrapping data.

3.d. **ANOVA on bootstrapped data:** **1.** Since original samples for bachelors, masters and doctorals don't follow normal distribution, we need to bootsrap the data to facilitate latter ANOVA tests. **2.** To study difference in salary between bachelors, masters and doctorals, I choose variable salary and variable education. **3.** After bootstrapping the data, as analyzed in .ipynb file, all assumptions of ANOVA (normal distributions, homogenity of variances and independence) are met, we can perform ANOVA on three groups. **4.** Since here we have 'Education' this independent variable with three levels, I choose perform one-way ANOVA.

3.e **Comment on findings:** Null hypothesis for ANOVA: $\mu_{bachelor}$ = μ_{master} = $\mu_{doctoral}$ (μ refers to mean salary). In 3.d, results show that F-statistic = 99468.5855 and P value = 0.0. (Note: Without setting seed in the .ipynb file, as approved by TA). Since P value << 0.05, we have a strong evidence to reject the null hypothesis that there is no difference in mean salary among bachelors, masters and doctorals. We conclude that at least one education group from bachelors, masters and doctorals has statistically significant different average salary than the others.