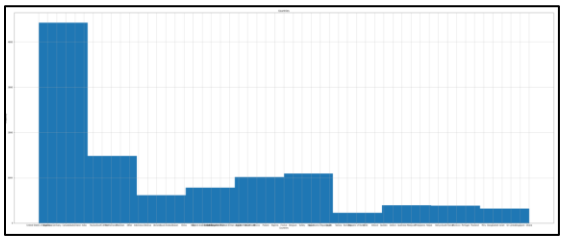
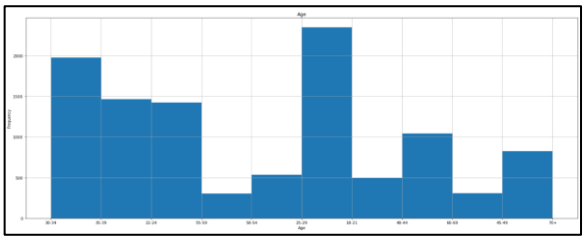
Question 1: Performing exploratory data

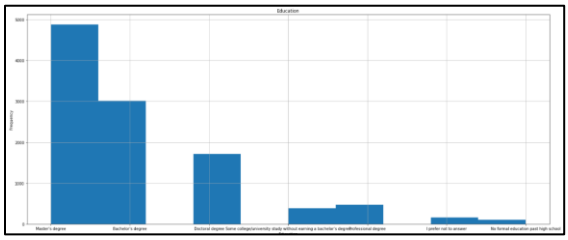
We can observe some characteristics about the subset of the data science community represented in this survey. Some characteristics analyzed in the dataset about the participants are country, age, education, professional experience, and salary.



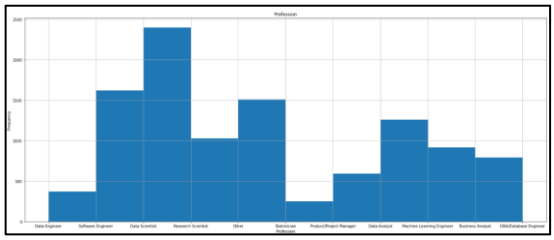
a) Country frequency



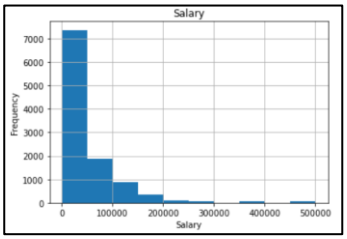
b) Age range frequency



c) Education level frequency



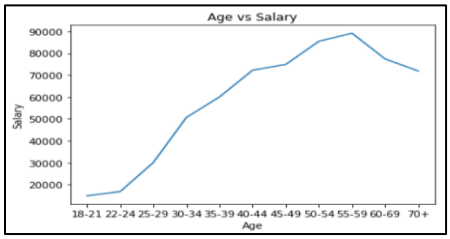
d) Profession frequency



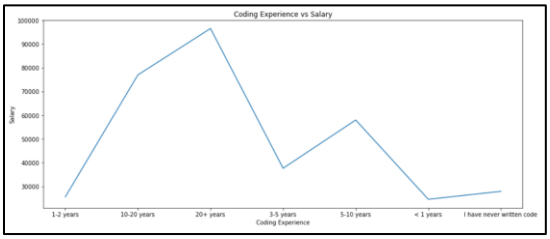
e) Salary frequency

Observations:

- There are many participants from across the world, with most of the participants coming from India and United States.
- 25–40-year-olds are most likely to participate in the survey.
- There is a wide variety of professions who answered the survey, mostly data scientists
- Most of the participants who answered the survey has obtained a master's degree
- Most common salaries are typically very low, perhaps they did not want to disclose their salary or does machine learning on the side



a) Age vs Salary



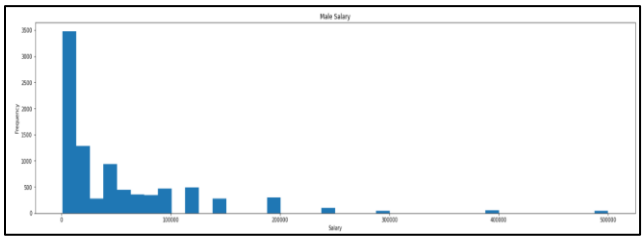
b) Coding Experience vs Salary

c) Profession vs salary

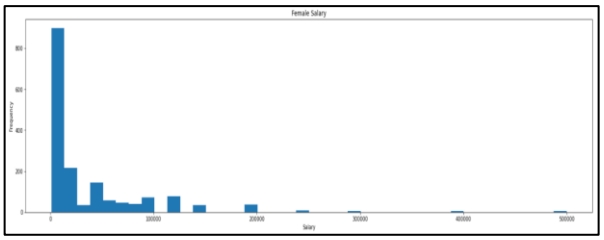
Trends in data: We can observe that those who are older have a higher salary, this is related to the fact that they will more experience in coding. With more experience, they can negotiate and seek careers that provide a higher salary. Another trend we see is that those in manager positions have higher salaries compared to other professions.

Question 2: Estimating the difference between average salary between men and women.

a) We reported the salary mean and variance between different genders. We can see a large difference in salary means between men and women but has similar variance.



ai) Men salary frequency



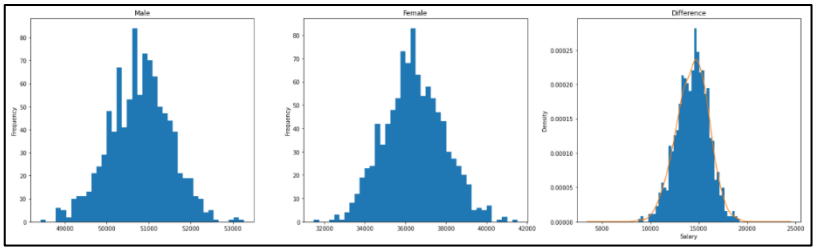
aii) Women salary frequency

b) We can visually see that the data doesn't fit normal distribution.

2-sample t-test has the assumptions required:

- Assumptions: 2 groups are independent, normally distributed, and similar variance
- Our data has similar variance and can assume independent; however, data is not normally distributed, therefore cannot run 2 sample t-test.

c) We will bootstrap the data for the data to converge to central limit theorem and fit a normal distribution. Procedure: Randomly select a sample from the dataset, repeating that multiple times to create a set which is used to calculate the mean. Bootstrapping was done 1,000 times, so 1,000 sets were made to calculate 1,000 means. The distribution of the means is plotted.

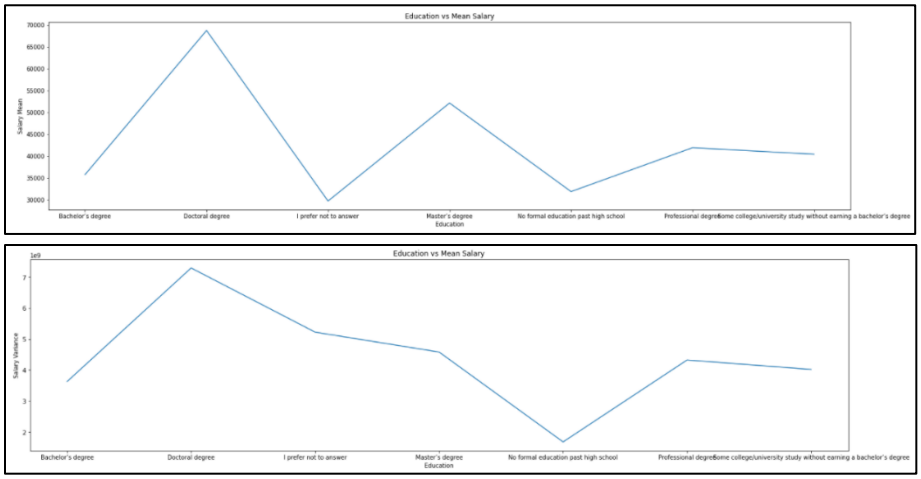


ci) Male salary vs frequency – bootstrapped data, cii) Women's salary vs frequency – bootstrapped data, ciii) Difference between mean and women's salary

d) Now that the data is normally distributed using bootstrapping, we can perform the 2-sample t-test. Python provides us the following results for t-test: t = -278.533, p = 0.

e) **Summarize Findings:** After applying bootstrapping to the male and female salaries, the mean salary data was normally distributed. This allowed us to perform a 2-sample t-test. Using the 2-sample t-test, we got a p value equal to 0 < 0.05, that shows the probability of the men salary mean and the women salary mean being equivalent. Equivalent to the null hypothesis stating that the difference in group means is zero. Thus, we reject the null hypothesis, and can conclude there is a relationship between gender and salary. Women are being underpaid compared to their men counterpart. if we solely look at gender and salaries.

Question 3: Estimating the difference between average salary between education level.

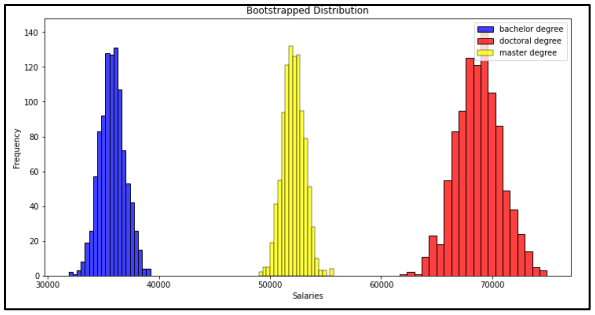


a) We reported the salary mean and variance between different education levels. We can see a large difference in salary means between bachelor's, master's, and doctoral degree. We can see they have similar variance.

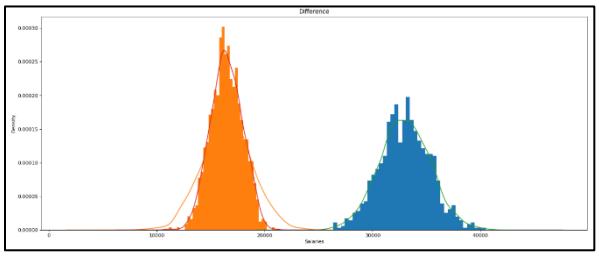
b) To perform ANOVA testing we need to check if it's suitable to use by checking the distribution first. Scipy.stats.shapiro( ) can test the null hypothesis that the data was drawn from a normal distribution. It returns a tuple of test statistic and p-value. If p-value is less (<) than the alpha (0.05), we reject the null hypothesis, which means the data is not normally distributed. The p-value for all three groups were 0, thus we conclude the data is not normally distributed and cannot apply ANOVA test.

c) We bootstrap the data for the data to converge to central limit theorem and fitting a normal distribution. Procedure: Randomly select a sample, repeating that multiple times to calculate the mean, and then plotting the distribution of those means.

- Check if the data is normally distributed after bootstrapping by using Scipy.stats.shapiro( ), since P-values are larger than 5%, we cannot reject null hypothesis that the data is normally distributed. We also graph the bootstrapped data to visually observe the distribution.



ci) Plotting the bootstrapped distribution of different education levels. We can see that the data now fits a normal distribution.



cii) Calculates the difference in salaries. Orange distribution line is the difference in salary between doctoral and master’s degree. Green distribution line is the difference in salary between doctoral and bachelor's degree. Red distribution line is the difference in salary between master and bachelor's degree.

d) We can now perform 1-way ANOVA after bootstrapping since the data fits a normal distribution. Our data has similar variance and can assume independent.

- Python provides us the following results for f-test: t = 128587, p = 0.
- Since p = 0 < 0.05, it is statistically significant, we reject the null hypothesis

e) **Summarize Findings:** Since the data was not normally distributed, we could not apply ANOVA test on it. After applying bootstrapping to the doctoral, master's, and bachelor's salaries, the mean salary data was normally distributed. This allowed us to perform a 1-way ANOVA test. Using the 1-way ANOVA test, we got a p value equal to 0, that shows the probability of the doctoral salary mean, the master's salary means, and the bachelor's salary mean being equivalent. We reject the null hypothesis that the means of all three groups are the same. Thus, we can conclude there is a relationship between education level and salary. Those who obtained their doctoral degree are being paid higher compared to if they have received a master's or bachelor's degree. This is if we solely look at education and salaries, not accounting for country, job position, etc.