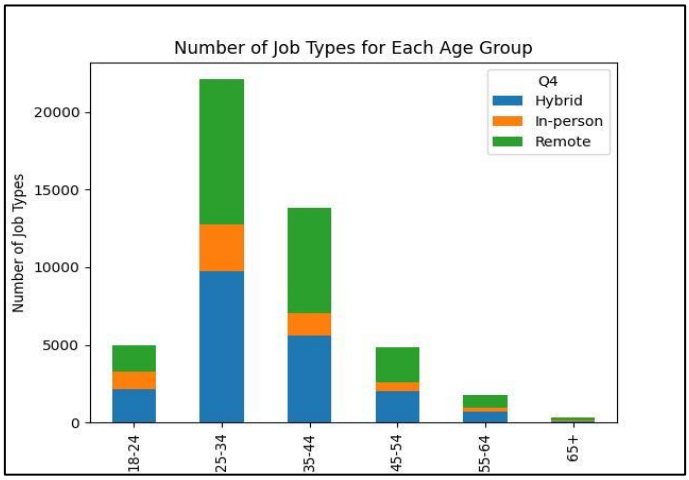
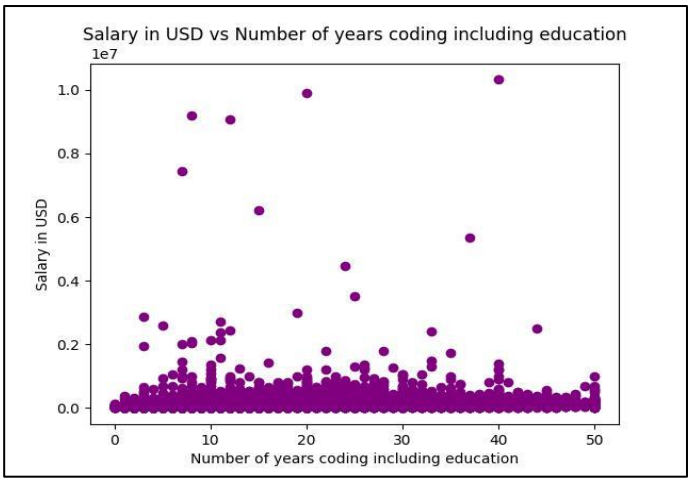
Bootstrapping and Central Limit Theorem for statistical understanding of education, jobs and salaries

Exploratory Data Analysis

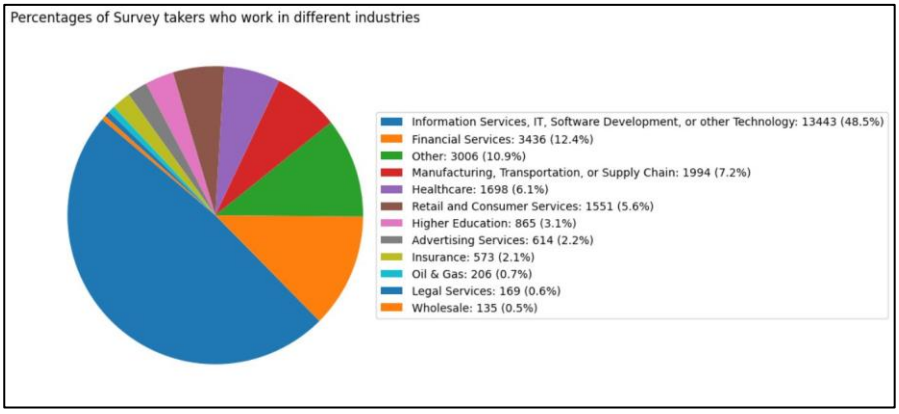
The descriptive statistics for the data are the means, standard deviations, data points, the quartiles and the minimum and maximum value for each numerical column of the data. The graphs discuss the **1.)** A stacked bar plot of frequency of different types of jobs for different age groups, **2.)** a scatter plot examining the relationship between Salary earned (USD) vs number of years of coding experience and **3.)** A pie chart dissecting percentages of industries where survey takers work.



Graph 1: Number of Job Types vs. Age groups



Graph 2: Salary (USD) vs No. of Coding Experience



Graph 3: Percentage of Survey takers working in different industries

	Q10	Q19	Q64	Q82
count	47840.00	47840.00	32567.00	47840.00
mean	15.66	13698363.37	11.44	95703.18
std	9.85	513739491.24	8.82	139607.97
min	0.00	1.00	0.00	1.00
25%	8.00	64000.00	5.00	44000.00
50%	13.00	115000.00	9.00	74963.00
75%	20.00	234000.00	16.00	121634.00
max	50.00	100000000000.00	50.00	10319366.00

Table 1: Descriptive Statistics for our Data

Data Pre-Processing

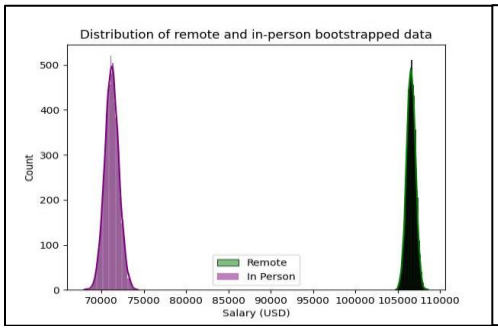
No missing data was observed for either column of remote salaries or in person data so there was no removal or replacement required. There were outliers for both remote and in person data. For remote salaries and in person salaries a minimum of $1 and $1 and a maximum of $10.3 million and $9.9 million were observed. A range of different extremely low and high values were also observed which suggested that this data was indeed a mistake by the survey takers, as it is very unrealistic for people to earn this low and this high salary for a job in the real world. Outliers were not removed to smoothen the distribution but rather because it suggested the data was faulty and would present incorrect results for our tests. Outliers were eliminated that were higher than $600,000 [Extremely unlikely to have a salary higher than this] and less than $5000 [Data is from survey takers across different countries but extremely unlikely for anyone to earn lower than this.] Subsequently descriptive statistics are reported for the data without outliers.

	count	mean	std	min	25%	50%	75%	max
Remote Stats	20197.00	106580.75	74024.80	5004.00	53545.00	90919.00	145000.00	580000.00
In Person Stats	5923.00	71252.94	63586.63	5020.00	30661.50	55687.00	91026.00	555582.00

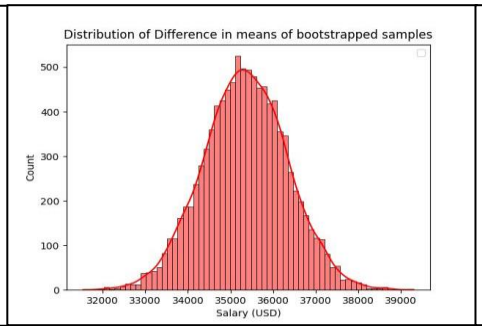
Table 1: Descriptive Statistics for Remote Salaries and In-Person Salaries after removal of outliers

In order to understand if there is a statistical difference in the means, it is suitable for us to perform a t-test because of the Central Limit theorem which states that a sampling distribution of means will approach close to a normal distribution if the sample size is high enough (usually > 30). Our sample size for remote and in-person stats are **20197** and **5923** respectively which are quite high and hence robust to the t-test violations as the sampling distribution of means is extremely likely to be a normal distribution.

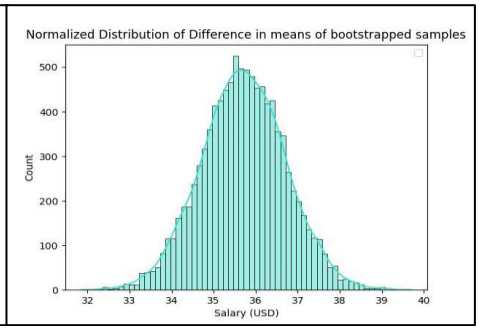
Bootstrapping to understand statistical significance for In-Person and Remote Jobs



Graph 4: Distribution of Bootstrapped Data



Graph 5: Distribution of difference in means



Graph 6: Normalized Distribution diff. in means

In order to estimate the exact distributions, and not get carried away by any assumptions, a bootstrapping approach was used to estimate the distribution of the sample means between in-person and remote jobs using **10000 resamples.** Subsequently, the difference in means of bootstrapped data was plotted and normalized to get a thorough understand of the data and its connection to the applicability of the t-test.

Performing a t-test on our original sample gave us a **t statistic of 36.17** and a **p value of $7.37*10^{-271}$** using the **SciPy stats t-test** with **unequal variances (Bartlett's statistical test)** provided evidence that our variances are unequal for the two groups)**.** As our p-value is less than our threshold of **0.05**, we reject the null hypothesis that the means of the two groups are equal. Furthermore, the **mean** of the **normalized difference of distribution in means** was observed as **36.16** of our bootstrapped data, which is very close to our obtained **t-statistic** of **36.17** which suggest our sampling distribution in means is indeed extremely normal and verifies the t-statistic obtained.

The p value can also be calculated using our bootstrapped distributions to perform hypothesis testing. The **normalized difference in means less than zero was divided by the total number of normalized difference samples** (which is **10000**). This gave approximately the area of overlap (**p value**) between each group of remote and in-person salaries which is **0** and lower than our threshold of **0.05**, providing enough evidence to reject the null hypothesis [means equal]. The benefit here is that by using bootstrapping, we ensure robustness within our hypothesis testing without any assumptions about our underlying data.

Education Level vs Salary – ANOVA vs Bootstrapping Approaches

	count	mean	std	min	25%	50%	75%	max
Master's Stats	12576.00	94771.70	68282.16	5029.00	52192.75	76511.00	117798.00	594811.00
Bachelor's Stats	21373.00	98892.84	73613.36	5004.00	48190.00	82458.00	132208.00	575000.00
Professional's Stats	2205.00	103944.04	79218.18	5072.00	53379.00	82459.00	130000.00	564086.00

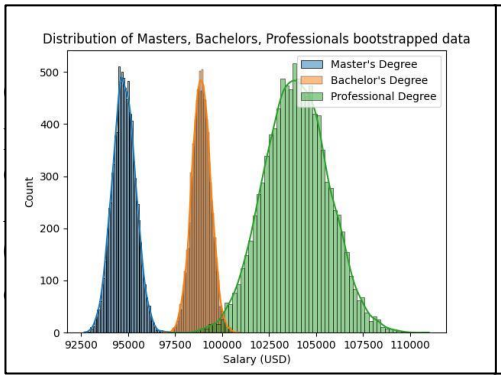
Table 2: Descriptive Statistics for Master's, Bachelor's and Professional's salaries after removal of outliers

We perform this study using data corresponding to individuals with Bachelor's, Master's and Professional Degrees. There were no empty values of salaries for either group of data so missing data was not required to be removed.

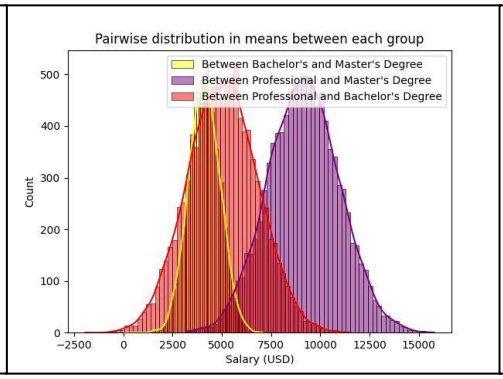
For masters, bachelor and professional salaries the minimum was $1, $1, and $3 while the maximum was $6.2 million, $10.3 million and $5.3 million respectively. There were several other unrealistic outliers and the justification for removing them is the same as described with the same bounds. The following descriptive statistics in **Table 2** were computed after removing the outliers:

It is suitable to perform an ANOVA test on the data-set. As we are comparing the means of more than two groups of data, a t-test is not a valid approach to confirm the hypothesis that all 3 means are equal to each other. ANOVA is also based on the assumptions that the sampling distribution of means of all 3 groups will be normal. Because of Central Limit Theorem, this is very likely to hold true for all 3 of our samples, because their sample size is quite large, as described in **table 2**, much more than the requirement (generally n >30). ANOVA test can be performed as the sampling distribution of means is likely to be close to normal. Also, each sample is assumed to be independent, which is another requirement. **Welch's unequal variances ANOVA test** was performed as the **Bartlett stat test** stated **variances are unequal** on the data set as variances were found to be unequal. This was achieved using pingouin library in Python.

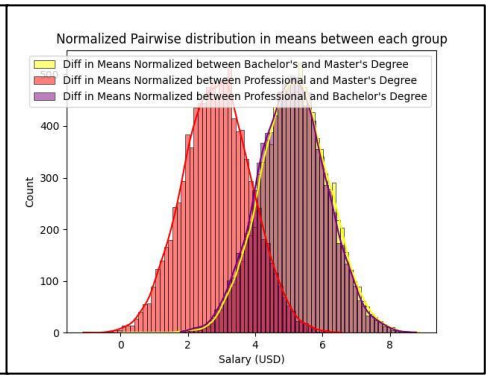
Bootstrapping to understand statistical significance between salary and education levels



Graph 7: 3 Group Bootstrap data distribution



Graph 8: Pairwise difference in means for bootstrap data



Graph 9: Normalized Difference in Means

Similar to the previous study, we will repeat the study by sampling the distribution of means for 10000 different samples, in order to perform our hypothesis testing and avoid any underlying assumptions of data. The graphs above plot the bootstrapped distributions for each salary group, the pairwise distribution in means and the normalized pairwise distribution in means.

In order to get more results of our hypothesis test, where the null hypothesis is that the means of the three groups corresponding to each education level are significantly different, the p value was calculated using the bootstrapped data. Values were computed using 2 bootstrapped distributions each time [Same approach where we computing the number of points less than 0 and divided by the total sample size to get overlapping area from bootstrap difference in means].

p values of **0, 0.0013** and **0** for each pairwise hypothesis test were calculated, stating that none of the groups have the same salary mean as all 3 null hypotheses are rejected. The results are similar to what was obtained in the **ANOVA** test, giving credence to the use of ANOVA when comparing means of three different groups.