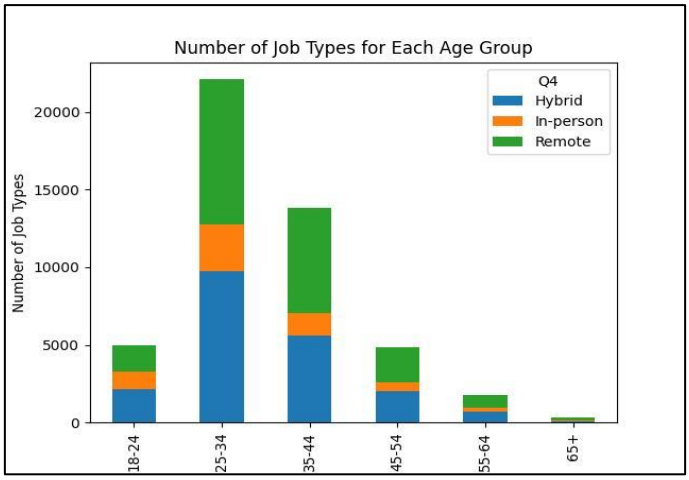
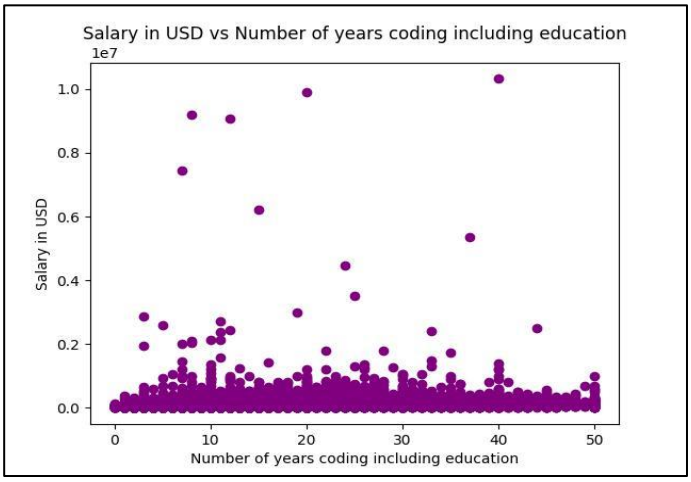
Name: Neil Sanjiv Punjani **Student Number: 1010615381**

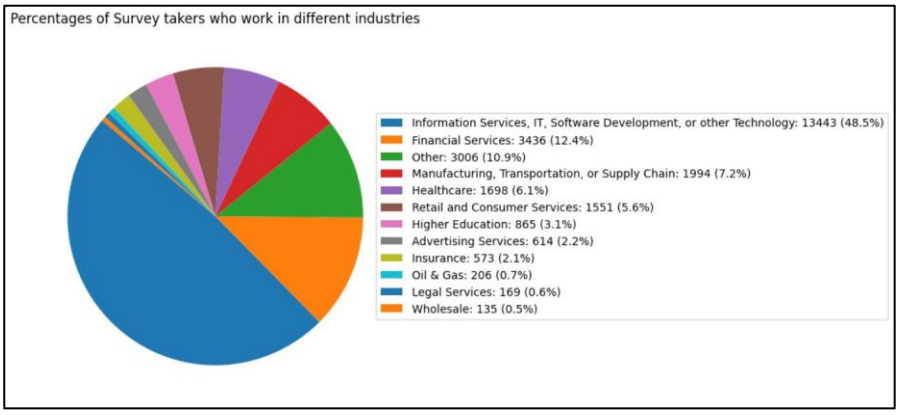
Q.1 The descriptive statistics for the data are the means, standard deviations, data points, the quartiles and the minimum and maximum value for each numerical column of the data. The graphs discuss the **1.)** A stacked bar plot of frequency of different types of jobs for different age groups, **2.)** 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

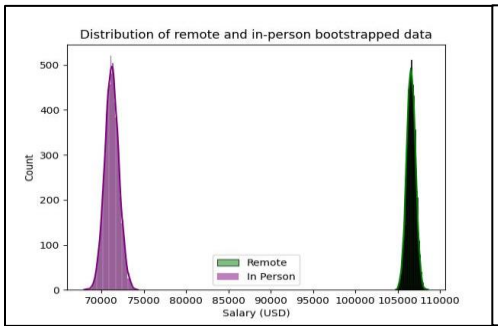
Q2. a) No missing data was observed for either column of remote salaries or in person data so there was no removal or replacement required. There were outliers for both remote and in person data. For remote salaries and in person salaries a minimum of $1 and $1 and a maximum of $10.3 million and $9.9 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

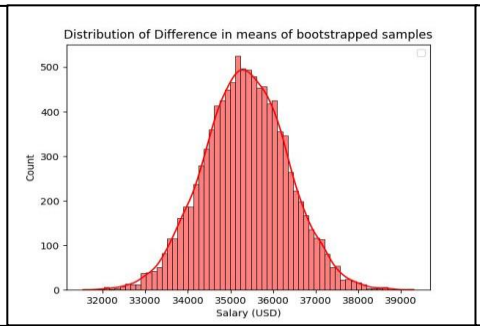
Table 1: Descriptive Statistics for Remote Salaries and In-Person Salaries after removal of outliers

Q2.b) Yes, 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 which is what the t-test is based on. The results from t-test are discussed in section **Q2.e)**

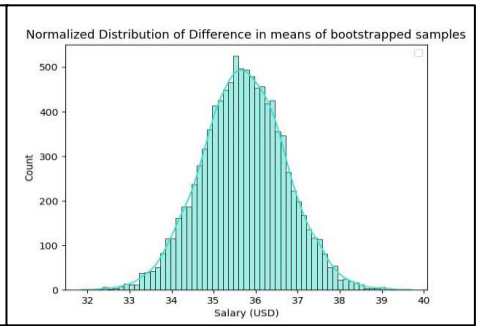
Q2.c)



Graph 4: Distribution of Bootstrapped Data



Graph 5: Distribution of difference in means



Graph 6: Normalized Distribution diff. in means

Q2.d) It is not suitable to perform a t-test as we did in **Q2.b)** again on the obtained bootstrapped data. The t-test is based on the assumption that the sampling distribution of means of our original sample will be normal. However, now we already have a sampling means of distribution and hence by performing a t-test again you would be doing so on the assumption that the sampling distribution of the sampling distribution of the means will be normal, which is redundant. Also, for t-test, samples need to be independent. However, by bootstrapping with replacement, we violate this assumption because our samples our dependent on the original data. This test is performed in the python notebook but is not suitable [results are discussed in **Q2.e)].** To conduct a hypothesis test on bootstrapped data, the p-value of the overlap using normalized difference in means can be computed which is discussed in **Q2.e).**

Q2. e) Firstly, performing a t-test **Q2.b)** on our original sample gave us a **t statistic of 36.17** and a **p value of** $\mathbf{7.37*10^{-271}}$ 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. When performing the t-test on bootstrapped data we receive a **t-statistic** of **3616.47** and a **p-value** of **0** but this t-test is not suitable as discussed previously. Furthermore, the **mean** of the **normalized difference of distribution in means** was observed as **36.16.** [this will vary because bootstrapping is random.] 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 in **Q2.b).**

To perform the hypothesis test using bootstrapped data the number of data points whose **normalized difference in means was 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]. These results confirm the results we obtained from the t-test performed in **Q2.b).**

Q3.a) 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

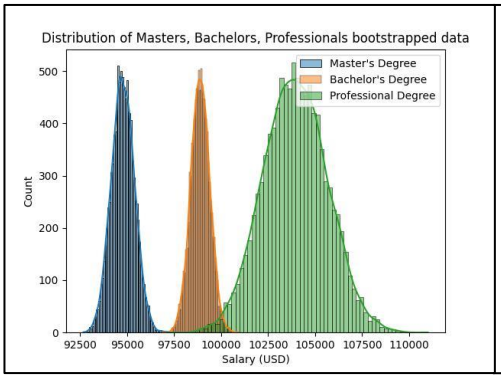
	count	mean	std	min	25%	50%	75%	max
Master's Stats	12576.00	94771.70	68282.16	5029.00	52192.75	76511.00	117798.00	594811.00
Bachelor's Stats	21373.00	98892.84	73613.36	5004.00	48190.00	82458.00	132208.00	575000.00
Professional's Stats	2205.00	103944.04	79218.18	5072.00	53379.00	82459.00	130000.00	564086.00

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

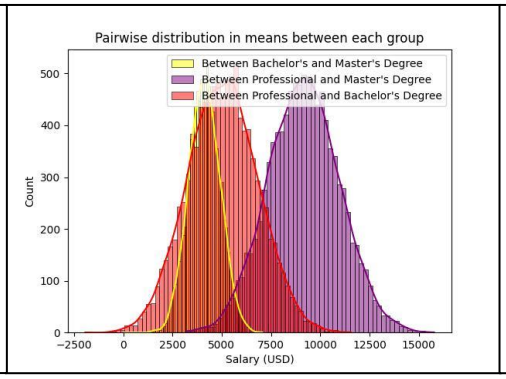
same as described in **Q2.a)** with the same bounds. The following descriptive statistics were computed after removing the outliers:

Q3. b) Yes, 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**. [Results are discussed in Q3.e)] on the data set as variances were found to be unequal. This was achieved using pingouin library in Python.

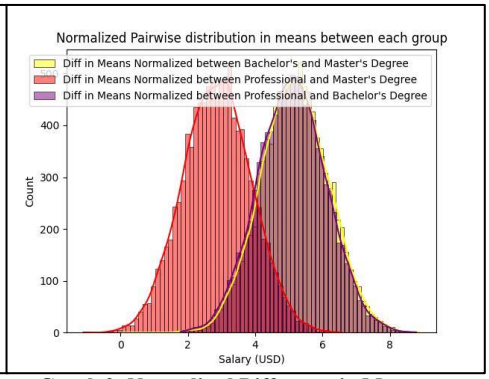
Q3. c)



Graph 7: 3 Group Bootstrap data distribution



Graph 8: Pairwise difference in means for bootstrap data



Graph 9: Normalized Difference in Means

Q3.d) No, it is not suitable to perform an ANOVA test as we did in **Q3.b)** test on the bootstrapped data. This is because the ANOVA test makes the assumption that the sampling distribution of means for all groups is normally distributed. Now that we have our bootstrapped data for all 3 groups, we already have our sampling distribution of means and it does not make sense to perform an ANOVA test again on this data as then the assumption will be that the sampling distribution of the sampling distribution of the means is normal which is redundant. Also, when we resample our data and get our bootstrapped samples, this bootstrapped samples are not independent samples because we sampled with replacement from our original sample and hence it is dependent on our original sample itself. One of the assumptions on ANOVA is that each group must be independently sampled and we are breaking this assumption now. [This test has been conducted in python notebook and discussed in **Q3. e),** but it is not a suitable test].

Q3.e) In question **3.b),** our unequal variances Welch's ANOVA test results state that the **F value** is **21.14** and the **p value** obtained is $\mathbf{7.06*10^{-10}}$. As our p value obtained is less than our **0.05 threshold**, we have sufficient evidence to reject the null hypothesis that the means of all three groups are equal. When performing the ANOVA test on the bootstrapped data as explained in **Q3.d)**, we get values of **209993.73** as our **F statistic** and **0** as our **p** value. This result confirms the hypothesis result we got in **Q3.b),** but it is not suitable due to reasons described before.

In order to confirm the results obtained in **Q3.b),** p values were computed using 2 bootstrapped different samples 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], where I got a **p value** of **0, 0.0013** and **0** for each pairwise hypothesis test, stating that none of the groups have the same salary mean as all 3 null hypothesis are rejected. This gives additional confirmation to the results obtained in **Q3.b).**