Question 1:

Three graphical figures that represent different trends in the data.

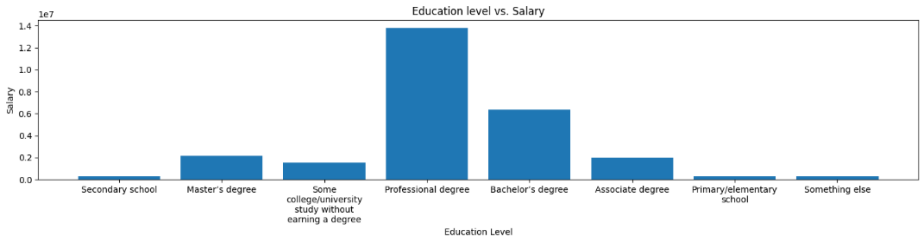

Figure 1: Education Level vs. Salary

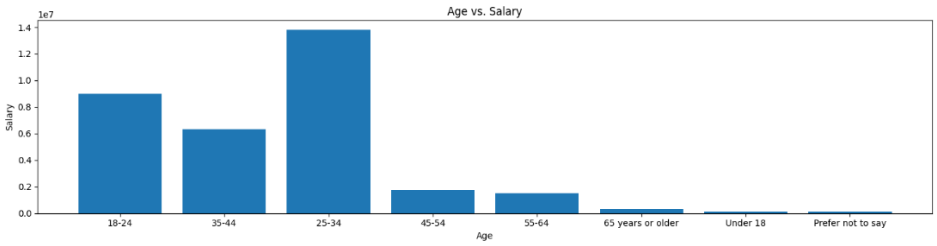

Figure 2: Age vs. Salary

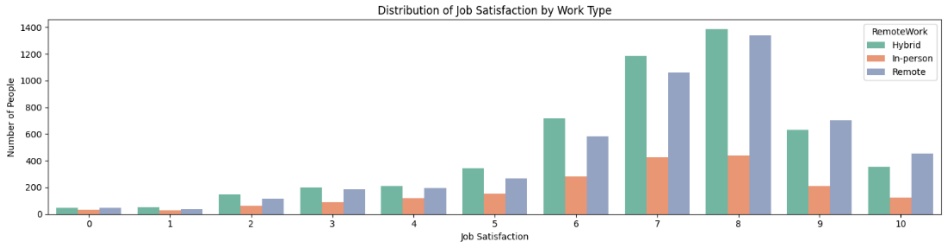

Figure 3: Distribution of Job Satisfaction by Work Type

Question 2:

a) Table 1 below shows the general statistics of remote jobs and in-person jobs after removing outliers from the data set. A straightforward method to remove outliers was applied: a box plot was used to visualize the salary distributions for both remote and in-person jobs, and all data points above the upper whisker of the box plot were removed as outliers.

Table 1: General Statistics of Remote Jobs and In-Person Jobs after removing outliers from the data

Work Type	Count	Mean ($)	STD	Min ($)	25% ($)	50% ($)	75% ($)	Max ($)
Remote Job	4808	81,328.48	58,358.26	104.00	34,679.25	71,381.00	120,000.00	244000.00
In-Person Job	1820	41,724.25	35,389.41	123.00	11,793.50	32,222.00	64,444.00	142,560.00

b) Table 2 below shows the manually calculated two-sample t-test with the means, variances, and pooled standard deviation for the two groups along with Python's built-in function of the two-sample t-test. The t-test value obtained manually and using Python's built-in function were the same therefore, the supplementary calculation used for the manual calculation such as the means and variances are correct.

Additionally, the assumptions required for the two-sample t-test are satisfied. The first assumption is that the data sets used are independent in which the two samples (remote and in-person salaries) are independent of each other. The second assumption is that the datasets are normally distributed. Although

they are not normally distributed, the Central Limit Theorem applies here due to the large sample size. Finally, the third assumption is homogeneity of variances; since the variances between the two work groups differ, Welch's t-test was used.

Table 2: Manually Calculated two-sample Welch's T-test and Python's built-in function of T-test.

Calc Type	Mean Remote ($)	Mean In-Person ($)	Var Remote	Var In-Person	Pooled STD	T-test	P-Value
Manual	81,328.48	41,724.25	3,405,687,136.09	1,252,410,499.57	53,052.42	33.51	
Python						33.51	1.08e-223

c) The three figures below are the bootstrapped data for comparing the mean of salary for the two groups of in-person and remote workers.

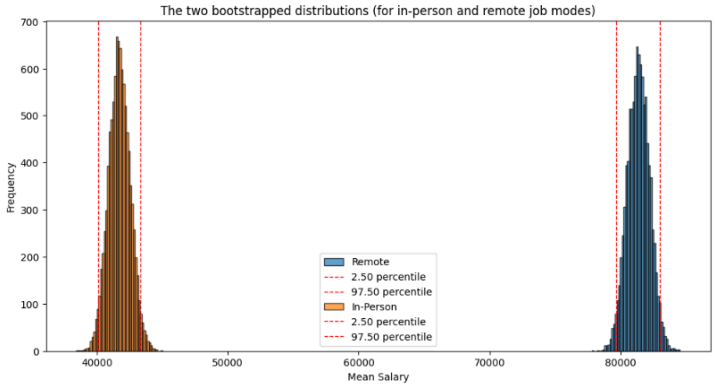

Figure 4: The two bootstrapped distributions (for in-person and remote job modes)

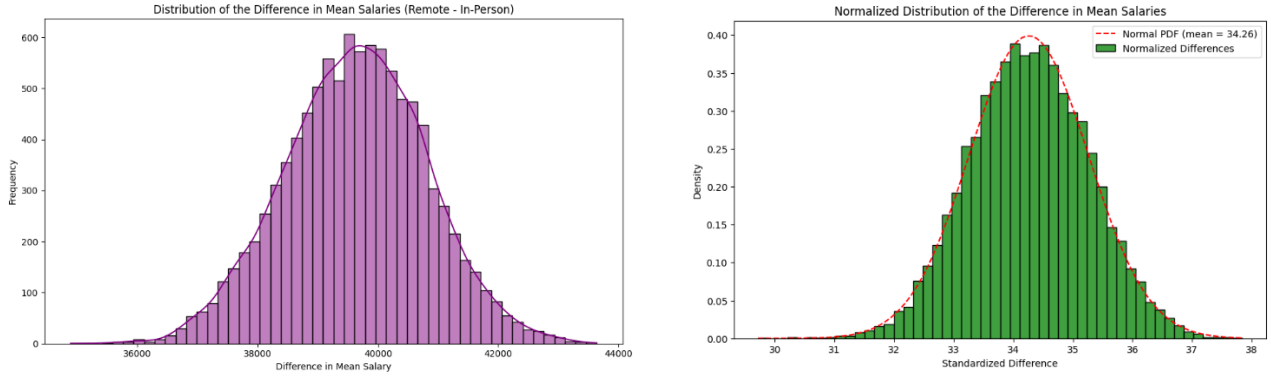

Figure 5&6: The Distribution and Normalized distribution of the difference in means

d) Calculating the t-static on the bootstrapped data, yields a t-value of 3367.64 when compared to the t-value of 33.51 of the original data. The reason for this difference is because the original data is heavily skewed to the right therefore, when calculating the standard error it is a much larger value (1181.72) when compared to the standard error of the bootstrapped data (11.75) while the means are roughly the same.

Question 3:

a) Table 3 below shows the general statistics of Bachelor's, Master's, and Professional degree salaries after removing outliers from the data set. The same method for removing outliers was used as in Q 2a).

Table 3: General Statistics of Bachelor's, Master's, and Professional degree after removing outliers

Degree Type	Count	Mean ($)	STD	Min ($)	25% ($)	50% ($)	75% ($)	Max ($)
Bachelor's	5409	71,399.92	55,691.04	115.00	24,033.00	60,000.00	107,406.00	231,000.00
Master's	3261.00	67294.90	41,022.13	104	3,7120	64,444.00	92,741.00	176,146.00
Professional	431	79493.70	44,809.87	132	4,8333	75,000.00	107,406.00	195,000.00

b) Table 4 below shows Welch's ANOVA calculation

Table 4: Welch's ANOVA results on original data.

Source	ddof1	ddof2	F	p-unc	np2
EdLevel	2	1206.44	18.38	1.36e-08	0.003122

c) The three figures below are the bootstrapped data for comparing the mean of salary for the two three groups of Bachelor's, Master's, and Professional degree.

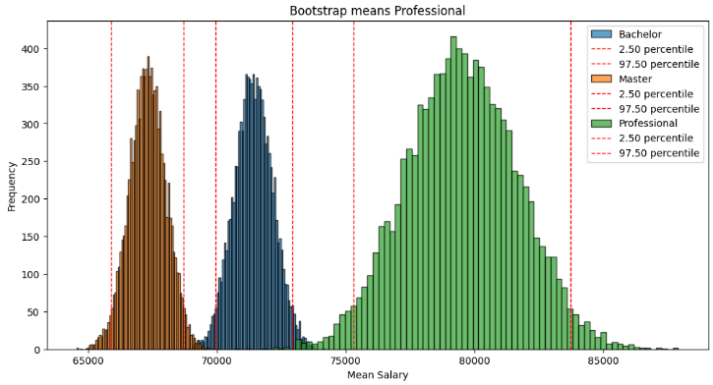

Figure 7: The three bootstrapped distributions (Bachelor's, Master's, and Professional degree)

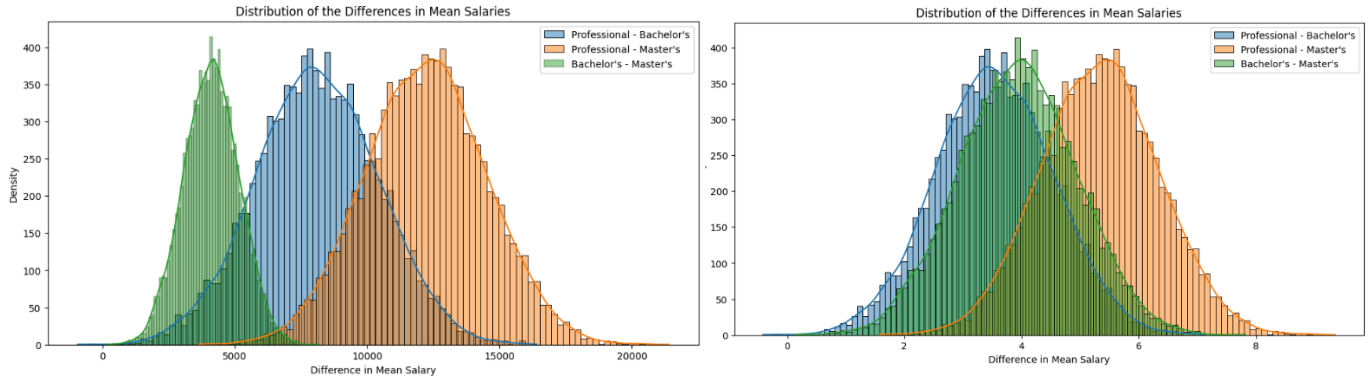

Figure 8&9: The Distribution and Normalized distribution of the difference in means

d) Table 5 below shows Welch's ANOVA calculation on the bootstrapped data. The F value is much larger than the F value on the original dataset because the bootstrapped data has reduced the variability within-group as seen in the normal distributions while the variability between groups are roughly the same.

Table 5: Welch's ANOVA results on the bootstrapped data.

Source	ddof1	ddof2	F	p-unc	np2
EdLevel	2	18373.57	183859.60	0	0.93