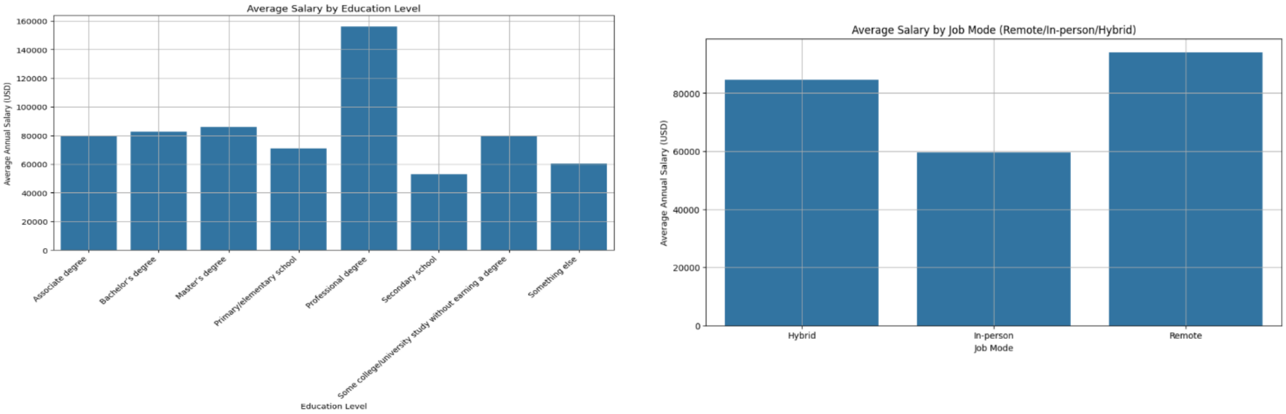
Omar Al-Hilawani
1008735978

Assignment 1 - MIE1624

Question 1



Figures 1 & 2: Education Level & Job Mode vs. Average Annual Salary (USD)

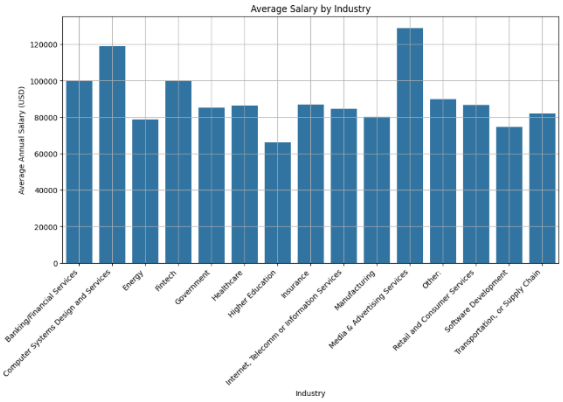


Figure 3: Industry vs. Average Annual Salary (USD)

Question 2

a. Below are the computed descriptive statistics for the two groups of in-person and remote job modes. Missing data was removed to avoid distorting the results such as the mean, median, and standard deviation, and to avoid inconsistencies which would not help us better understand the data. Outliers were removed to avoid distorting the results such as the mean, median, and standard deviation. The IQR method with a multiplier of 1.5 is widely used as it effectively identifies outliers while preserving the overall structure of the data. 1.5 has become the standard in statistical analysis.

Table 1: Descriptive Statistics for Remote & In-person Job Modes

Count	Mean	Std	Min	25%	50%	75%	Max
4852	82883.76	60326.92	104	35241.25	72000	120000	264264
1845	43218.50	37394.95	123	11963	32222	64544	160000

b. A welch t-test was utilized to compare the average salaries between in-person and remote job modes. Welch was used because the variances of the data were different.

Table 2: Manual Calculation of two-sample t-test

Mean (Remote)	Mean (In-person)	Variance (Remote)	Variance (In-person)	Pooled Standard Deviation	T-statistic (manual)	Degree of Freedom
82883.77	43218.51	3639337962.16	1398382896.35	54973.75	32.30	5319.37

When using welch's t-test via Python's built-in function, a t-statistic of 32.30 and a p-value of 0.00 were calculated. The assumptions that were utilized are normality, homogeneity of variance, and independence. The independence assumption is satisfied based on the results we have previously performed. The other two assumptions have been violated which means that the null hypothesis has been violated.

c. Below are the graphs of the bootstrapped results.

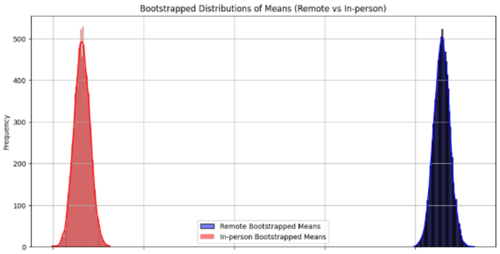


Figure 4: Bootstrapped Distributions of Means (Remote vs In-person)

As shown in figure 4, the distribution shows distinct separation between the bootstrapped means of in-person and remote workers. The remote workers' salaries are much higher than the in-person workers.

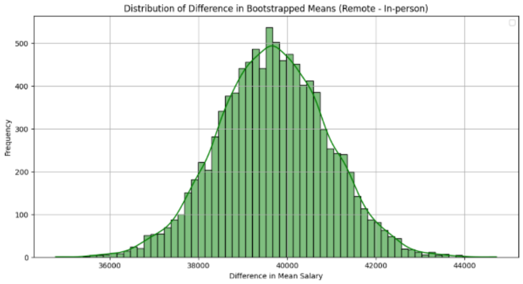


Figure 5: Distribution of Difference in Bootstrapped Means (Remote - In-person)

This graph illustrates the bootstrapped distribution of the difference between the mean salaries of remote and in-person workers. The mean difference is around $39,662.

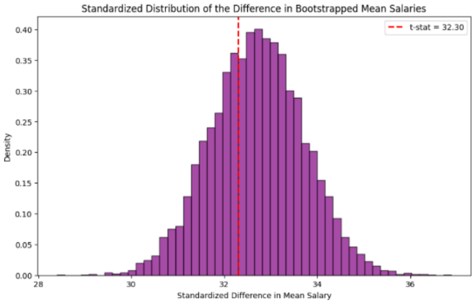


Figure 6: Standardized Distribution of the Difference in Bootstrapped Mean Salaries

For figure 6, the distribution shows the standardized difference in the means. As we know, after normalization occurs, the mean is equal to the t value which is proven in this figure. The manually calculated t-statistic was 32.30 which is close to the mean standardized difference of 32.23.

d. To remain consistent with what was done in 2b, welch's t-test was also used in 2d. The Bootstrapped Welch's T-test results were a t-statistic of 3250.0929 and a p-value of 0.00. The results for both the normal and bootstrapped data are statistically significant. Moreover, the bootstrapped t-statistic is significantly larger than 2b due to the standard error of the difference in means of the calculated bootstrapped data is much smaller than the normal data. This could be due to resampling which captured the variability much better and reduced the standard error.

Question 3

a. In this question, we will be performing the same data clean up that was done for question 2b.

Table 3: Descriptive Statistics for Bachelor's degree, Master's degree, & Professional degree

EdLevel	Count	Mean	Std	Min	25%	50%	75%	Max
Bachelor's degree	5427	71958.28	56435.21	115	24243	60147	108000	245000

Master's degree	3326	69625.83	43855.20	104	37861	64444	95592	197627
Professional degree	456	87121.32	54015.72	132	48789.5	76345.5	112777	244585

b. In this question, we will be assuming that same assumptions as question 2b. Also, as seen in table 8, the p value reinforces the null hypothesis being true.

Table 4: Calculation of welch's ANOVA via Python's built-in function

Source	Ddof1	Ddof2	F	P-Unc	Np2
EdLevel	2	1244.81	22.27	$3.89\text{x}10^{-33}$	0.00489

c. Below are the graphs of the bootstrapped results.

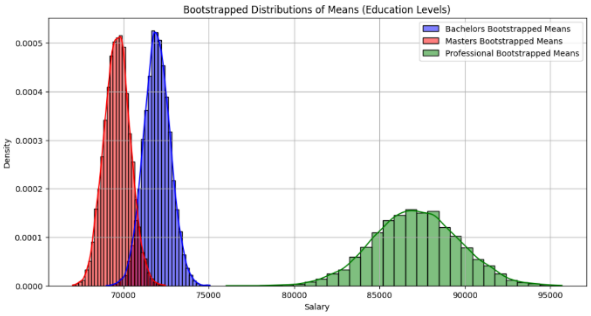


Figure 7: Bootstrapped Distributions of Means (Education Levels)

As seen in figure 7, the bachelor's and master's degree holders have close mean salaries. However, the professional degree holders salaries are much higher. Depending on the level of education someone has, it may be worth pursuing a higher degree.

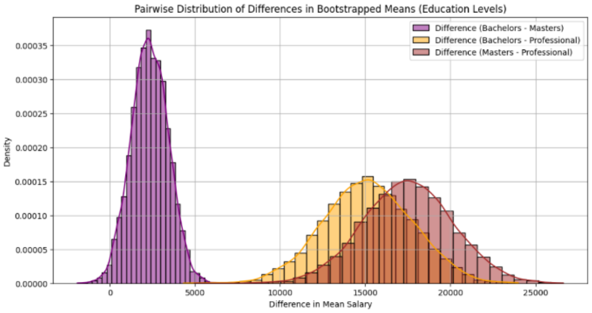


Figure 8: Pairwise Distribution of Differences in Bootstrapped Means (Education Levels)

The purple graph represents distribution of the difference between bachelor's and master's degree salaries which is not very much. However, the yellow and red graphs represent larger gaps between salaries. This indicates that pursuing a higher education is worth while.

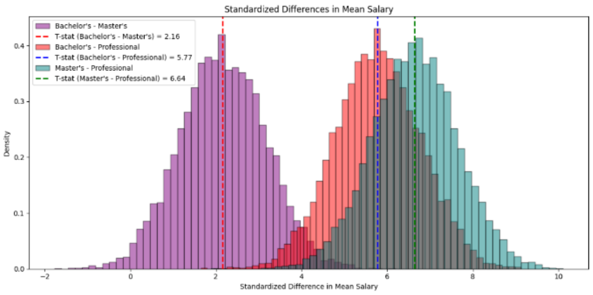


Figure 9: Standardized Differences in Mean Salary

As we know, after normalization occurs, the mean is equal to the t value which is proven in this figure.

d. As seen in the table below, the p value of the bootstrapped data is 0 which means that the null hypothesis is rejected. Also, the F value is magnitudes higher than 3b.

Table 5: Bootstrapped Welch ANOVA Results

	Source	Ddof1	Ddof2	F	P-Unc	Np2
Bootstrapped Welch's ANOVA	EdLevel	2	18269.18	224289.08	0.00	0.96