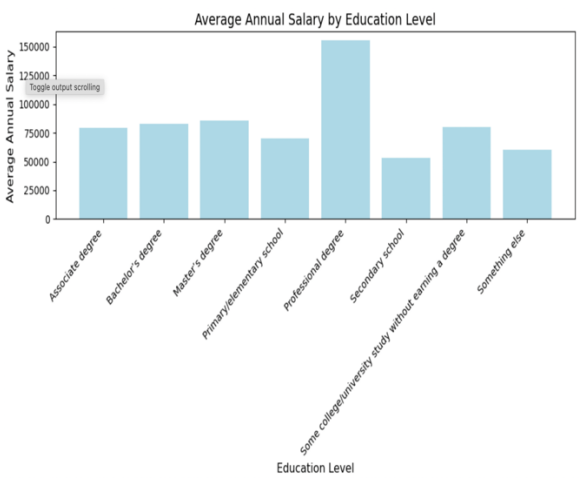
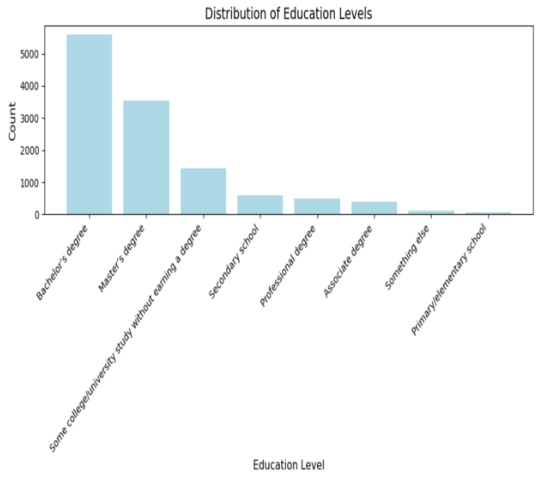
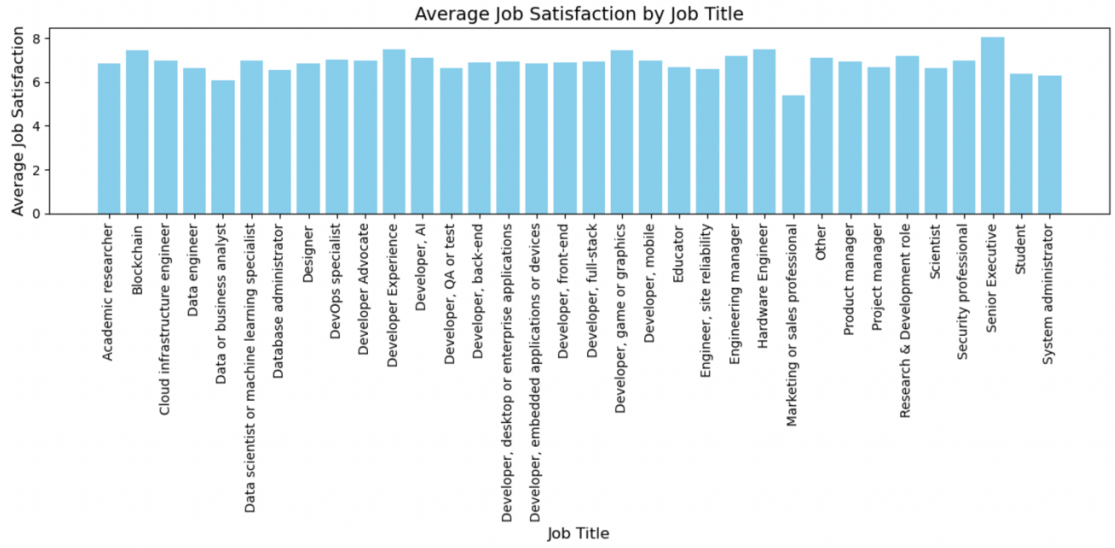
A1: Data Analytics Report

Shuman Zhao

Q1. Exploratory Data Analysis







Q2. Estimating the difference between average salaries of two job modes: in-person and remote

a. Descriptive statistics

	ConvertedCompYearly							
	count	mean	std	min	25%	50%	75%	max
RemoteWork								
Hybrid	5272.0	84515.642830	241457.587966	109.0	37081.75	64444.0	101910.00	13818022.0
In-person	1948.0	59382.189938	123254.766188	123.0	12889.00	36518.0	72000.25	3367716.0
Remote	4989.0	93850.616356	130228.134228	104.0	36000.00	74595.0	127388.00	6340564.0

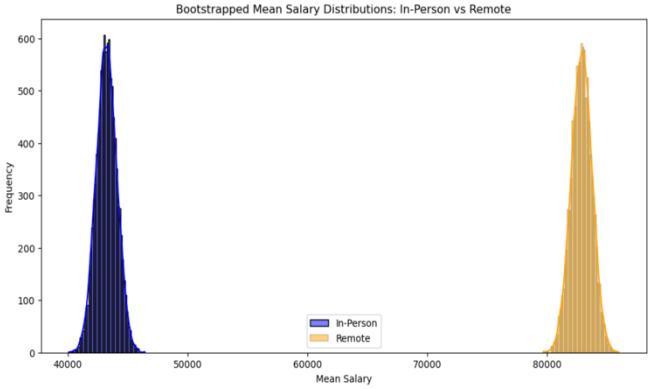
To remove outliers, I use IQR method, which focuses on the middle 50% of the data and robust to outliers. Removing outliers can minimize the impact of extreme values and improve the accuracy of analysis.

b. two-sample t-test to compare average salaries between in-person and remote job modes

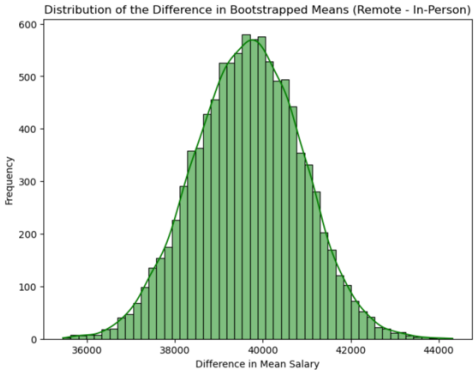
```
manual two-sample t-statistic: 26.37991194667647
Welch's t-statistic: 32.30051734614891
Shapiro-Wilk Test for In-Person: ShapiroResult(statistic=0.8996555926189317, pvalue=3.892618122022896e-33)
Shapiro-Wilk Test for Remote: ShapiroResult(statistic=0.9453917417113694, pvalue=4.190207885606065e-39)
Levene's Test for Equal Variance: LeveneResult(statistic=391.86751850465316, pvalue=8.334115672795999e-85)
```

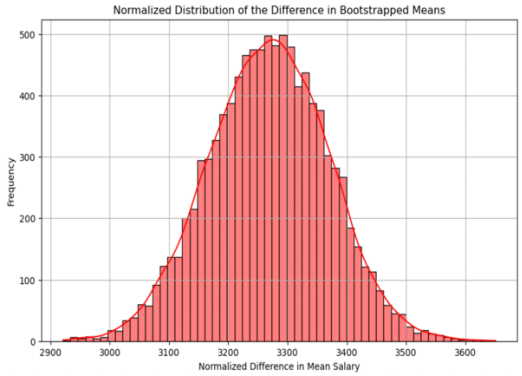
By using Shapiro-Wilk test to check Normality and Leven's test to check Equal-variance, I conclude that the two assumptions are both invalid (p-value smaller than 0.05). Then I use two-sample t-test for manual calculation (as required) and Welch's t-test for Python's method. Two result t-statistics are different. This is because Welch's t-test does not assume equal-variance. The p-value close to 0 from Welch's test shows that the average salaries between two job modes are significantly different.

c. Bootstrap



Pros of Bootstrap are no assumptions required for distribution and works well for small sample size. Cons are that it can be easily influenced by outliers and extreme small sample size. After bootstrapping with 10,000 replications, I find out that the mean salary for remote mode is 39663.84 on average larger than the in-person mode. The normalization graph shows the value of t-statistic on its mean.





d. Two-sample t-test after Bootstrapping

```
Two-sample t-test: t = 3235.6  p = 0
Welch's t-test: t = 3235.6  p = 0
```

The results are the same for two test methods because bootstrapping does not need to check assumptions of Normality and Equal-variance. P-value equal to 0 means that the null hypothesis is rejected and the average salaries between two job modes are significantly different. The conclusion is the same compared to 2b. But the t-statistics are different.

Q3. “EdLevel” analysis for Bachelor, Master, and Professional degree

a. Descriptive statistics

ConvertedCompYearly	count	mean	std	min	25%	50%	75%	max
EdLevel								
Associate degree	403.0	79056.397022	114976.063034	371.0	29946.5	61840.0	100137.50	2014062.0
Bachelor's degree	5592.0	82685.918634	127595.706976	115.0	25525.5	63000.0	114018.50	6340564.0
Master's degree	3537.0	85714.912072	102478.997686	104.0	39740.0	68203.0	103000.00	2153432.0
Primary/elementary school	62.0	69933.322581	67621.184297	299.0	14583.5	50300.0	109680.75	305229.0
Professional degree	489.0	155296.527607	747206.128939	132.0	50885.0	82526.0	128887.00	13818022.0
Secondary school	603.0	53072.116086	43037.894298	123.0	19440.5	45784.0	73140.00	300000.0
Some college/university study without earning a degree	1422.0	79966.442335	93471.051062	132.0	28658.0	59180.5	102000.00	1562898.0
Something else	101.0	60000.356436	46724.259922	494.0	26852.0	53703.0	81629.00	306396.0

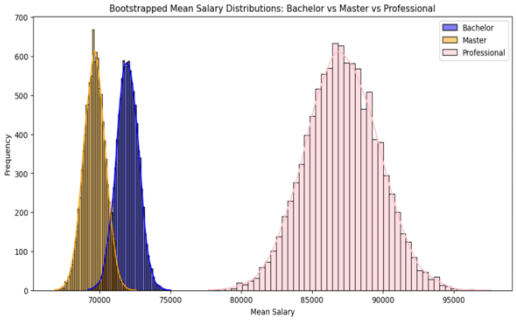
To remove outliers, I still use IQR method, which focuses on the middle 50% of the data and robust to outliers. There is no missing value for the three degrees.

b. ANOVA test to compare average salaries

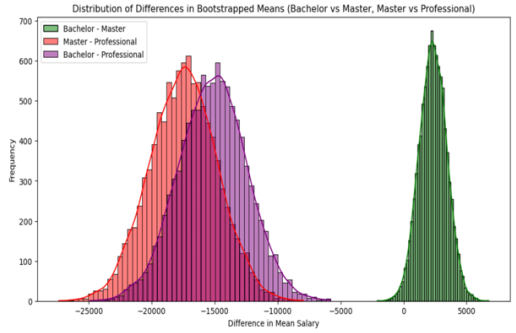
	sum_sq	df	F	PR(>F)
C(EdLevel)	1.228599e+11	2.0	22.617434	1.590109e-10
Residual	2.500390e+13	9206.0	NaN	NaN

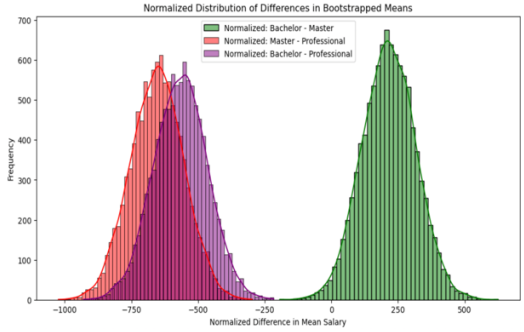
Pro of ANOVA is that it can handle multiple groups. Cons are it assumes Normality and Equal-Variance. The Shapiro-Wilk and Levene tests show that two assumptions are invalid, which means the ANOVA result may not be rational. P-value is smaller than 0.05, I conclude that there is a significant difference in the salary mean between at least two of the groups.

c. Bootstrap



After bootstrapping with 10,000 replications, I find out that the mean salary for Bachelor is 2316.70 on average larger than the Master, mean salary for Master is 17536.45 less than the Professional, and mean salary for Bachelor is 15219 less than the Professional. The normalization graph shows the value of t-statistic on its mean.





d. ANOVA after bootstrapping

	sum_sq	df	F	PR(>F)
C(EdLevel)	1.805460e+12	2.0	363112.725951	0.0
Residual	7.457516e+10	29997.0	NaN	NaN

P-value equal to 0. The conclusion is the same compared to 3b. But the F values are different.