Report: Exploring the Impact of Education on Income Levels in Data Science

The objective of this report is to explore the relationship between education levels and income in the field of Data Science, Machine Learning, and Artificial Intelligence (AI). The analysis focuses on two geographical regions: North America and Eastern Europe. We investigate the effects of education on income, specifically comparing the mean salaries of individuals with Bachelor's, Master's, and Doctoral degrees.

We utilized the Kaggle ML & DS Survey Challenge dataset, which contains survey responses from 25,973 participants across 369 columns representing various survey questions. To conduct our analysis, we transformed the "current yearly compensation" column (Q25) into a numerical target variable. We excluded rows with missing data and handled outliers.

Data Preprocessing

Step 1: Data Cleaning

- We began by removing rows with missing data to ensure data quality.
- We also converted the salary buckets into numerical values for further analysis.

Step 2: Outlier Removal

- To maintain data relevance, we removed outliers in salary values that exceeded certain thresholds based on geographical regions.

Step 3: Education Level Grouping

- We combined categories "No formal education past high school" and "Some college/university study without earning a bachelor’s degree" into the "Bachelor’s degree" category.

Analysis for North America & Eastern Europe

The analysis methods for North America and Eastern Europe were no different, so I will present them only once. We'll talk about the results further.

Hypothesis Testing: Bachelor's vs. Master's Degree

- We conducted a two-sample t-test comparing the mean salaries of individuals with a Bachelor's degree and those with a Master's degree.

Hypothesis Testing: Master's vs. Doctoral Degree

- A second t-test comparing the mean salaries of individuals with a Master's degree and those with a Doctoral degree.

Bootstrapping

- We bootstrapped the data for comparing the mean of salary.

Comparing

- We compared t-test results to bootstrapping results.

Results

North America

Hypothesis Testing - Bachelor's vs. Master's in North America

	T-Test Results	Bootstrapping Results
t-statistic (SciPy)	-5.57706	-5.57706
p-value (SciPy)	2.8662e-08	0.0

The negative t-statistic (-5.57706) indicates that, on average, individuals with a bachelor's degree tend to earn less than those with a master's degree.

The very small p-value (2.8662e-08) indicates that this difference in average salaries is highly statistically significant. In practical terms, it means that the observed disparity in earnings between individuals with bachelor's and master's degrees is unlikely to be due to random chance.

Based on these results, we reject the null hypothesis and conclude that there is strong evidence to suggest a significant difference in average salaries between individuals with a bachelor's degree and those with a master's degree. The negative t-value indicates the direction of the difference, with bachelor's degree holders earning less on average, and the small p-value underscores the statistical significance of this difference.

Hypothesis Testing - Master's vs. Doctoral in North America

	T-Test Results	Bootstrapping Results
t-statistic (SciPy)	-4.25023	-4.25023
p-value (SciPy)	2.27947e-05	0.0

The second test showed similar results (t-statistic = -4.25023, p-value = 2.27947e-05), so we also reject the null hypothesis and conclude that there is a significant difference in average salary between people with a master's degree and a doctorate. A negative t value indicates the direction of the difference: master's degree holders earn less on average, and a small p value highlights the statistical significance of the difference.

Eastern Europe

Hypothesis Testing - Bachelor's vs. Master's in Eastern Europe

	T-Test Results	Bootstrapping Results
t-statistic (SciPy)	-4.44903	-4.44903
p-value (SciPy)	9.98104e-06	0.0

The first test for Eastern Europe showed the results that we saw when examining data from North America (t-statistic = -4.44903, p-value = 9.98104e-06), so we reject the null hypothesis as well and conclude that there is a significant difference in average salary between people with a master's degree and a doctorate. A negative t value indicates the direction of the difference: master's degree holders earn less on average, and a small p value highlights the statistical significance of the difference.

Hypothesis Testing - Master's vs. Doctoral in Eastern Europe

	T-Test Results	**Bootstrapping Results**
t-statistic (SciPy)	0.219216	0.219216
p-value (SciPy)	0.826556	0.3985

In this case, the t-statistic is close to zero, and the p-value is relatively high (greater than the typical significance level of 0.05), indicating that there is no statistically significant difference in the average salaries between individuals with a 'Master's degree' and those with a 'Doctoral degree' in Eastern Europe.

Conclusion

In North America, our analysis reveals a significant income disparity based on education levels. Individuals with Master's degrees earn significantly more than those with Bachelor's degrees, as indicated by both t-test and bootstrapping results. Additionally, there is a substantial income difference between Master's and Doctoral degree holders.

In Eastern Europe, a similar pattern emerges with Bachelor's and Master's degrees, showing a significant income gap. However, the difference in income between Master's and Doctoral degree holders is not statistically significant.

The results suggest that individuals in North America might benefit from pursuing higher degrees (Master's or Doctoral) in Data Science, ML, and AI, given the substantial income differences observed. In contrast, individuals in Eastern Europe may find that a Master's degree alone can lead to competitive salaries, as the income boost from obtaining a Doctoral degree is not statistically significant.

It's important to consider regional factors and the cost of education programs when making decisions about further education. Additionally, education level alone may not be the sole determinant of income. Other variables, such as years of experience and skills possessed, should also be taken into account in salary negotiations.

This analysis is based on the available dataset, and there may be limitations in the representativeness of the survey respondents. Future research could explore the role of additional variables in explaining income disparities.