AHMED MOATAZ 22-101092
ELAMIR ELSADY 22-101029
OMAR MAHMOUD 22-101059

DATA ANALYSIS REPORT

DOES EDUCATION & EXPERIENCE IMPACT DEVELOPER COMPENSATION

TABLE OF CONTENTS

INTRODUCTION

In today's rapidly evolving tech landscape, understanding the factors that influence developer compensation is crucial for both employers and professionals in the field. This project delves into the significant impact of education level and professional coding experience on developers' yearly compensation. Using data meticulously gathered in a CSV file, various parameters such as developers' experience, education level, employment status, industry type, age, and yearly salary are analyzed.

This report provides valuable insights into how different factors contribute to salary variations. For emerging developers planning their career paths, seasoned professionals seeking to maximize earning potential, and employers aiming to attract and retain top talent, these findings offer critical information for making informed decisions in a competitive market.

By elucidating the relationship between education, experience, and compensation, this analysis aids in understanding the key determinants of developer salaries. The goal of this report is to empower stakeholders with knowledge to achieve their objectives and maintain a competitive edge in the tech industry. The subsequent sections will detail the methodology, analysis, and findings, highlighting the significant trends and patterns observed in the data.

Research Question:
Does the education level and years of experience impact the income of developers?

Hypothesis:
The level of education and years of professional coding experience significantly impacts the yearly compensation of developers

METHODOLOGY & BIAS CONSIDERATIONS

Population of Interest

The population of interest in this study comprises developers from various industries, with differing levels of education and professional experience. The data includes developers working full-time, part-time, and self-employed, spanning a range of ages and industries. The dataset aims to cover the whole population of developers in general.

Sampling Method

The data used for this analysis is sourced from the 2023 Stack Overflow Developer Survey, conducted from May 8 to May 19, 2023. Respondents were primarily recruited through Stack Overflow channels, such as onsite messaging, blog posts, email lists, meta.stackoverflow posts, banner ads, and social media. This sampling method can be considered convenience sampling, as respondents were recruited through easily accessible channels, potentially leading to a sample that is skewed towards highly engaged users on Stack Overflow. While this sampling method provides a large and diverse dataset of developer responses, it is important to note the potential bias in the sample composition.

Bias Identification

This study considers potential biases, including self-reporting and sampling biases. To address these, anonymity was ensured, and honest responses were encouraged. Additionally, the data was segmented by different categories, such as employment type and industry, to enable meaningful comparisons. This way, salary comparisons between groups of inherently different earning potentials, like full-time and self-employed individuals, are avoided.

DATASET

The dataset used in this analysis is derived from the 2023 Stack Overflow Developer Survey and has been streamlined to include only the most pertinent columns for our study. Originally, the dataset consisted of 89,185 entries, but after cleaning to remove null values, it was reduced to 6,351 entries. The important columns retained in the dataset are as follows:

01 — Employment

This column indicates the employment status of the respondents, specifying whether they are employed full-time, part-time, or self-employed. This information is crucial for analyzing the impact of employment type on yearly compensation.

02 — EdLevel

This column records the highest level of education attained by the respondents, ranging from primary school to PhD. Understanding the educational background is essential to assess its influence on professional coding experience and compensation.

03 — YearsCodePro

This column captures the number of years the respondents have been coding professionally. It provides insight into how professional experience affects yearly compensation.

04 — Industry

This column specifies the type of industry in which the respondents are employed. Different industries often have varying compensation standards, making this an important factor in the analysis.

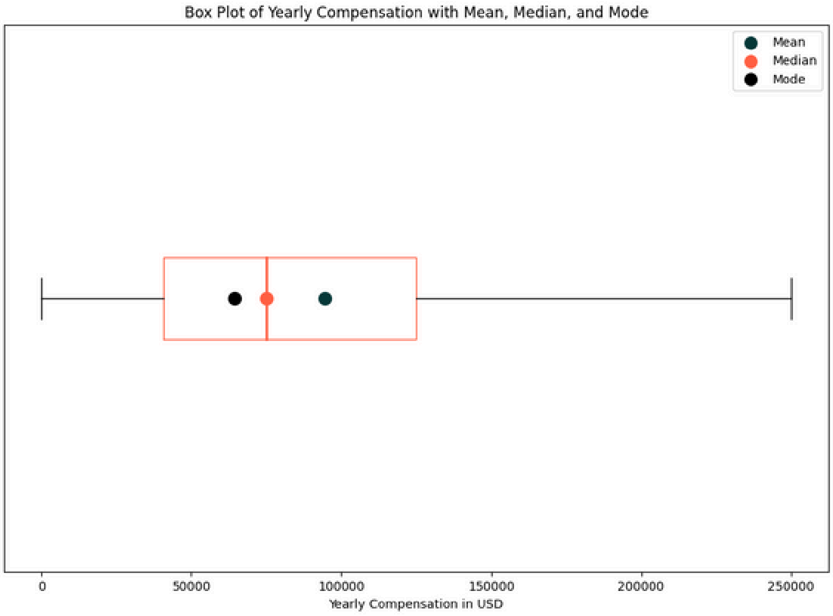
05 — ConvertedCompYearly

This column represents the yearly compensation of the respondents, converted to USD for consistency. It serves as the primary metric for analyzing the impact of education, professional experience, and employment type on earnings.

By focusing on these key features, the dataset allows for a comprehensive analysis of the factors influencing developer compensation. Each feature plays a vital role in understanding the interplay between education, experience, industry, and employment status in determining yearly earnings.

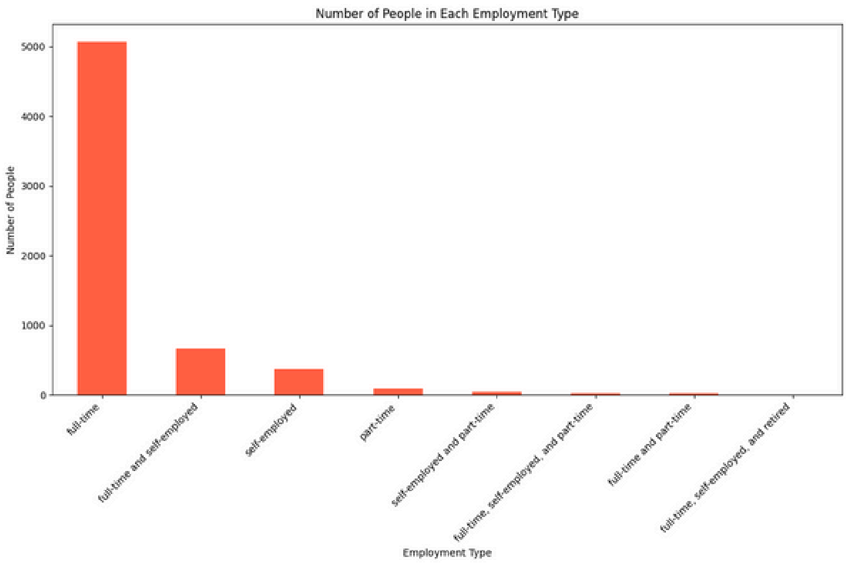
ANALYSIS

Before analyzing our dataset, we conducted thorough cleaning to ensure accuracy. We removed records with null values and those that didn't fit expected ranges, eliminating potential biases. We then mapped similar answers to key elements, streamlining the dataset. To ensure fairness and reduce skewness, we split certain categories, like employment and industry types, to avoid comparing individuals with inherently different earning possibilities. This preprocessing provided a cleaner, more consistent dataset for reliable analysis.



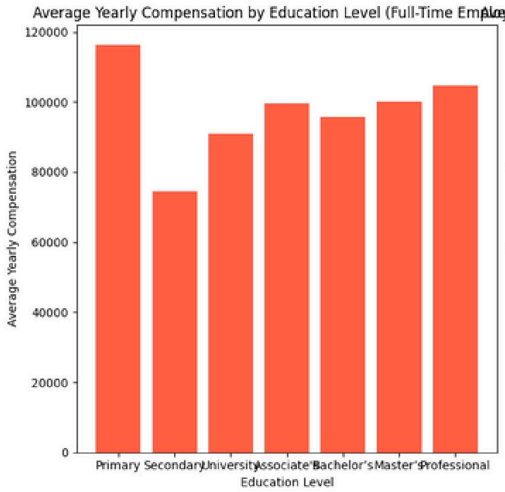
This boxplot shows the mean, median, and mode of the yearly compensation in the whole data set without outliers. This shows that the mean yearly compensation in the dataset is about 94592 USD.

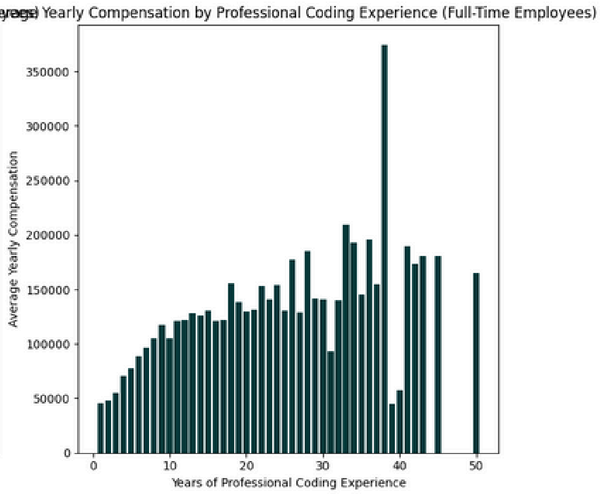
The bar graph on the right shows that most people who filled out the survey were reportedly working full-time. The next leading employment type was self-employed. Because the two might earn differently, we will analyze each one.



EDLEVEL & YEARSCODEPRO VS COMPENSATION

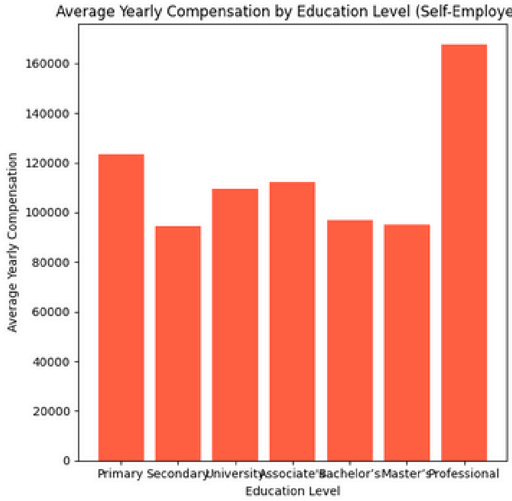
Full-Time

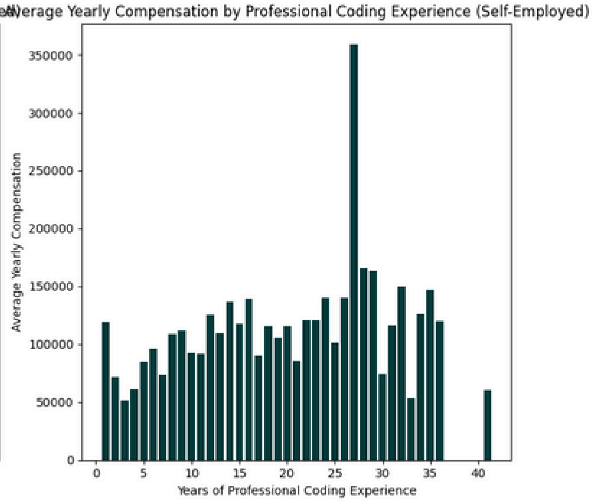




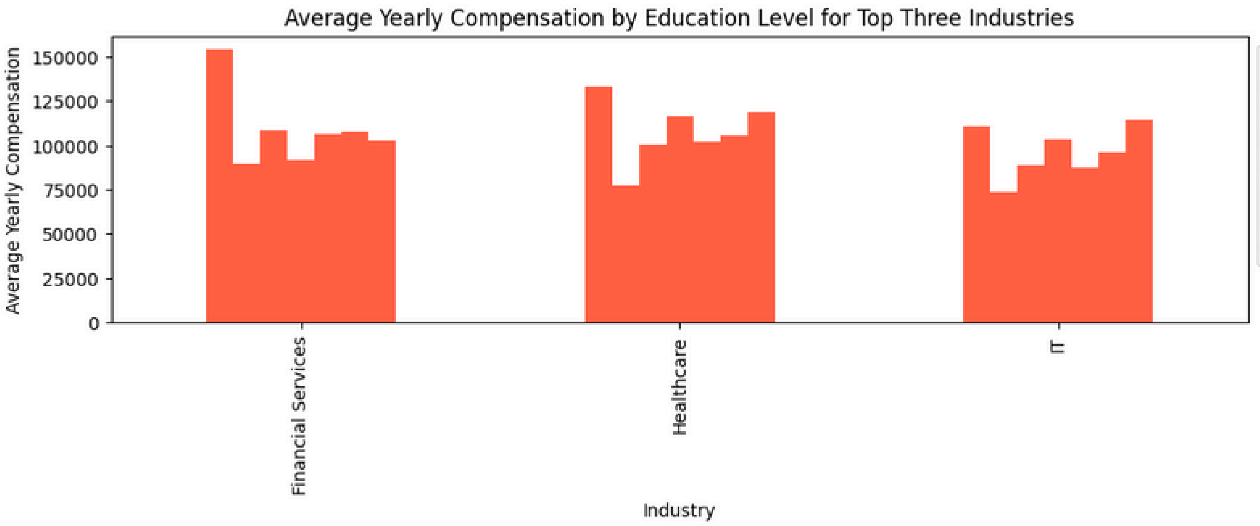
For full time employees, we plot two bar graphs for EdLevel and YearsCodingPro against Compensation. From the graphs, we cannot clearly see a correlation between education and compensation apart from the fact that starting from secondary school (EdLevel 2), as the education increases, the compensation gradually increases too. However, on the YearsCodePro graph, there is a clear increase of compensation as the years of proffessional coding increases.

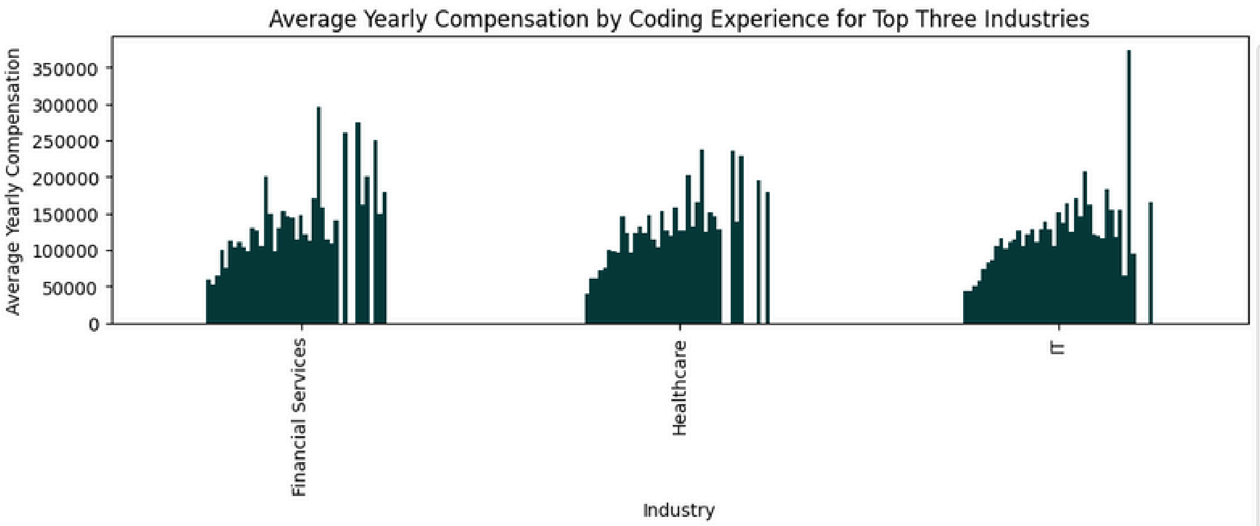
Self-Employed





For self employed individuals, we cant really see a correlation between education and compensation. As for the YearsCodePro graph, there also doesnt seem to be a clear correlation between years of professional coding and compensation. Only from about 2-15 years do we see a correaltion in that compensation increases as years of professional coding increases.



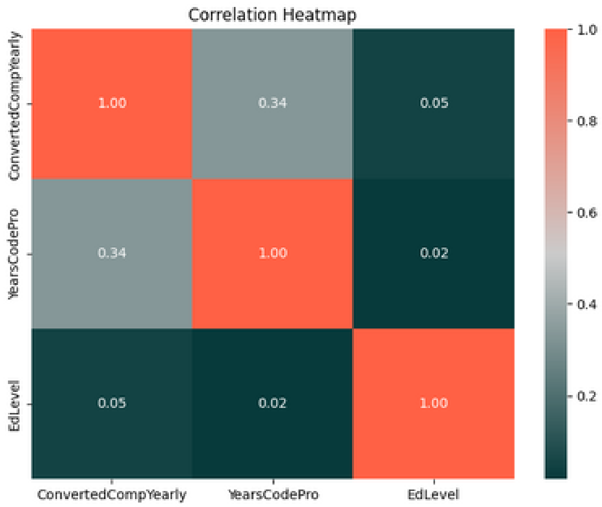


Additionally to splitting an analyzing the data set by full-time and self-employed individuals, different industries may also significantly different compensation, so we analyze the top three industries in the data set.

For the top three industries (IT, Healthcare, and Finance), we plot education level and years of professional coding against compensation. Throughout all three industries, we cannot see a clear correlation between education level and compensation when looking at the graphs. However, when looking at the graphs for years of professional programming vs compensation, it is clear that there is a slight correlation between the two for all industries, and as one gains more years of professional programming experience, their compensation tends to increase.

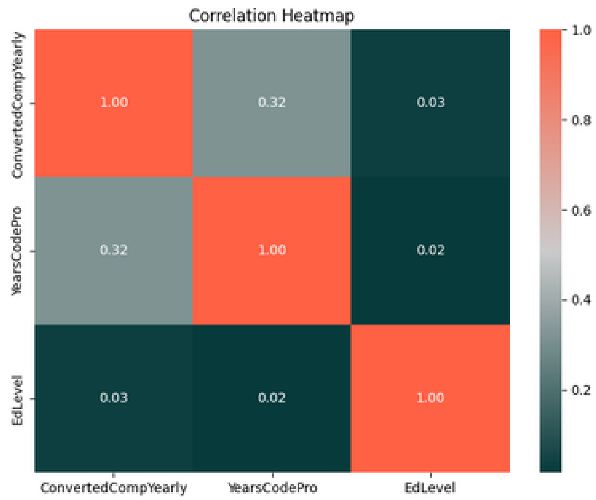
CORRELATION COEFICIANT

General



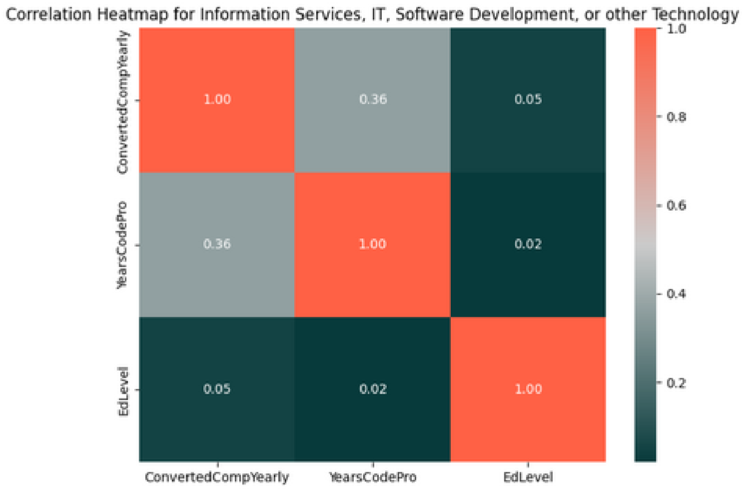
For the general data, we can see that the correlation between YearsCodePro and Compensation is 0.32, which is a weak correlation, and between EdLevel and Compensation is 0.03, which is an extremely weak correlation.

Full-Time



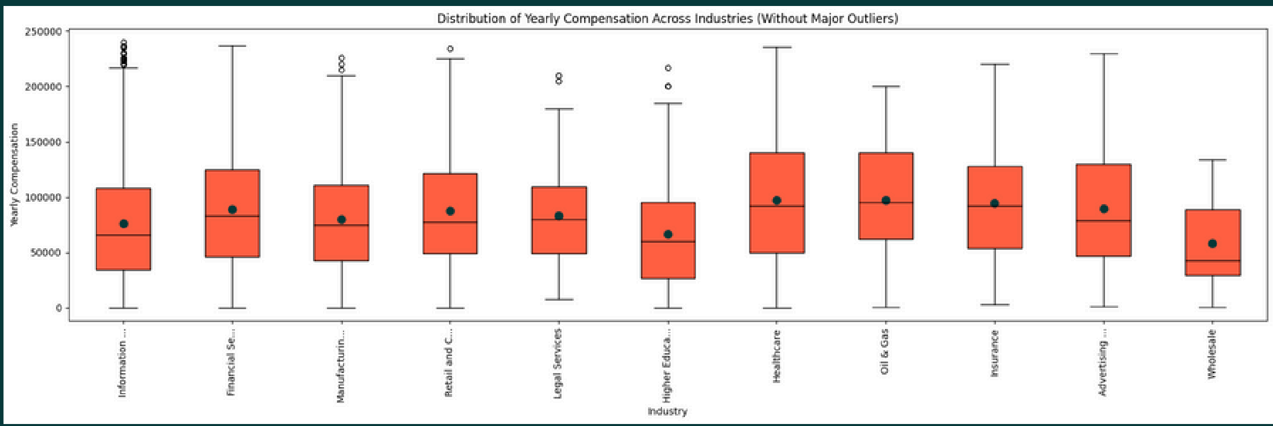
For the data reduced to people working full-time, we can see that the correlation between YearsCodePro and Compensation is 0.34, which is a weak correlation, and between EdLevel and Compensation is 0.05, which is an extremely weak correlation.

Leading Industry (IT)

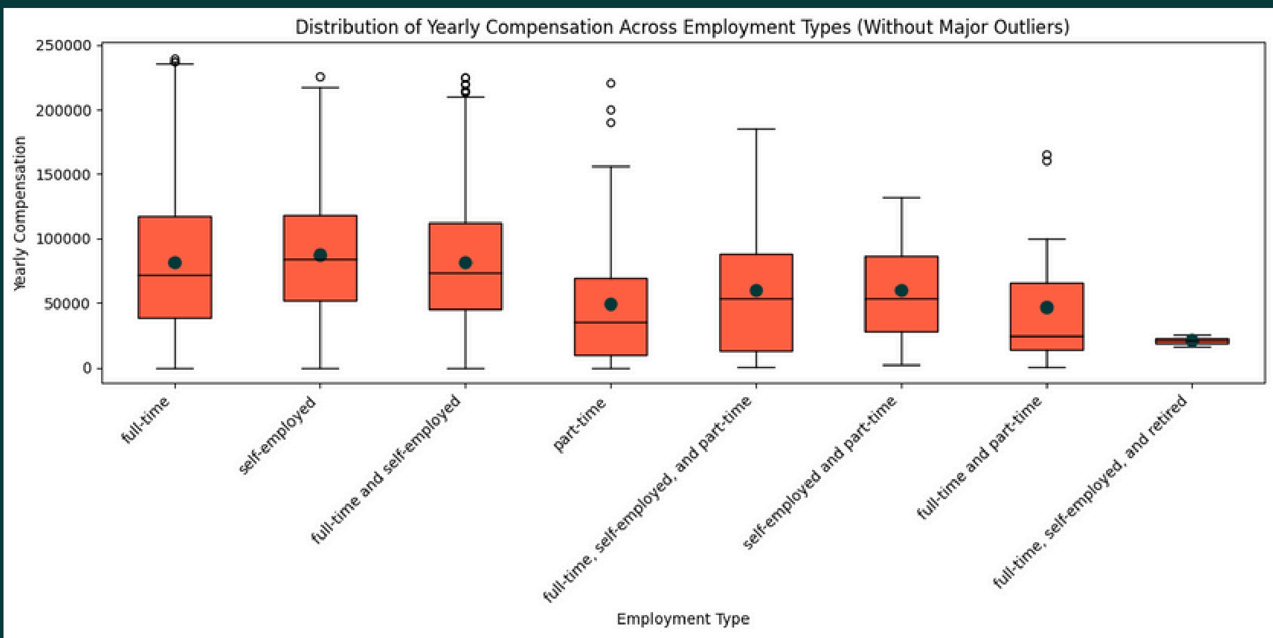


For the data reduced to people working in he IT industry, we can see that the correlation between YearsCodePro and Compensation is 0.36, which is a weak correlation, and between EdLevel and Compensation is 0.05, which is an extremely weak correlation.

FURTHER FINDINGS



By plotting boxplots of compensation for every industry, we can determine the safest industry to enter if one is looking for a high paying job. As shown by the graph, the top three industries one should be looking to get into in terms of compensation are the healthcare, oil and gas, and infrastructure industries. The mean compensation for those three industries are higher than the rest



Similarly, plotting boxplots for the different employment types shows which one is more successful in terms of compensation. If one wants to earn higher compensation, according to the plots, they should look into getting hired full-time or working self-employed as the means for those box plots are higher than the rest.

HYPOTHESIS TESTING

01 — Derive Null and Alternative Hypothesis

- Null Hypothesis (H0): There is no significant relationship between yearly compensation and the predictors (education level and years of professional coding experience).
- Alternative Hypothesis (H1): There is a significant relationship between yearly compensation and at least one of the predictors (education level or years of professional coding experience).

02 — Choose Appropriate Test

Given that we want to determine if the level of education and years of professional coding experience significantly impact yearly compensation, we should consider a regression analysis and the p-values associated with the regression coefficients:

- Regression Analysis: This approach will allow us to understand the relationship between the dependent variable (yearly compensation) and the independent variables (education level and years of professional coding experience).

03 — Calculate P-Value

General:
- YearsCodePro 0.000 EdLevel 0.017

Full-time
- YearsCodePro 0.000 EdLevel 0.001

Self-Employed
- YearsCodePro 0.000 EdLevel 0.719

Leading Industry (IT)
- YearsCodePro 0.000 EdLevel 0.005

04 — Determine Statistical Significance

The commonly used significance level threshold is 0.05. Overall, these resulted p-values are all less than 0.05, which allows us o reject the null hypothesis and suggests that there is a significant relationship between yearly compensation and education level and years of professional coding experience, particularly in the general dataset and among full-time employees and in the IT industry. However, for self-employed individuals, only years of professional coding experience seems to have a significant impact on yearly compensation.

CONCLUSION

Based on the hypothesis testing conducted, it was found that education level and years of professional coding experience have a significant relationship with yearly compensation for developers in various contexts. In the general dataset, among full-time employees, and in the IT industry, both education level and experience were found to impact compensation. However, for self-employed individuals, education level was not significantly related to compensation, suggesting different factors may influence earnings in this group.

The analysis also revealed that while there is a relationship between years of professional coding experience and compensation, it is weak. Similarly, the correlation between education level and compensation was very weak, even when categories were split for a fair analysis. These findings indicate that while education and experience are factors in determining compensation, they are not the sole determinants, and other factors may play a role in how much a developer earns.

In conclusion, while education level and years of professional coding experience are important factors in determining yearly compensation for developers, their impact varies depending on the context. The findings suggest that while these factors are significant, they are not the only considerations, and other factors may also influence compensation levels.

IN SHORT

The analysis reveals a significant relationship between education level, years of professional coding experience, and yearly compensation for developers. However, this influence varies across different groups, with a weaker correlation found among self-employed individuals. Despite these factors playing a role, the study highlights that they are not the sole determinants of compensation, suggesting the presence of other influential variables.

POTENTIAL ISSUES

One potential issue with the project or dataset could be the presence of outliers in the data, especially in the 'ConvertedCompYearly' column. Outliers can significantly affect the results of statistical tests and analysis, leading to skewed conclusions. To address this, you can consider using more robust statistical methods that are less sensitive to outliers, such as using median instead of mean for central tendency measures or using non-parametric tests.

Another issue could be the limited number of variables considered in the analysis. While education level and years of professional coding experience are important factors, there may be other variables that could also influence yearly compensation, such as the size of the company, specific technologies used, or geographic location. Including these variables in the analysis could provide a more comprehensive understanding of the factors affecting compensation.

Additionally, the data cleaning process, while necessary, may have introduced bias or errors. It's important to carefully review the cleaning steps to ensure that they do not inadvertently remove or modify important data. Conducting sensitivity analyses or using multiple imputation techniques can help assess the impact of data cleaning on the results.

Lastly, the sample size, while large, may not be representative of the entire population of developers. Consideration should be given to the generalizability of the findings and whether they can be applied to a broader population. Increasing the diversity of the sample, such as including developers from different regions or industries, could improve the external validity of the study.

THANK YOU