Comprehensive Salary Analysis Report
- Gursimran

Introduction

The purpose of this analysis is to delve into a dataset that encompasses salary information for various positions within the fields of data science and engineering. The data is sourced from the years 2020 to 2023, inclusive of job categories, experience levels, and employment types. The study aims to understand the distributions, identify outliers, and gauge the relationship between years of experience and salary levels.

Dataset Summary

The dataset contains 9,355 entries with 12 attributes. Upon preliminary inspection, the data set is complete with no missing values or duplicate entries, providing a solid foundation for analysis.

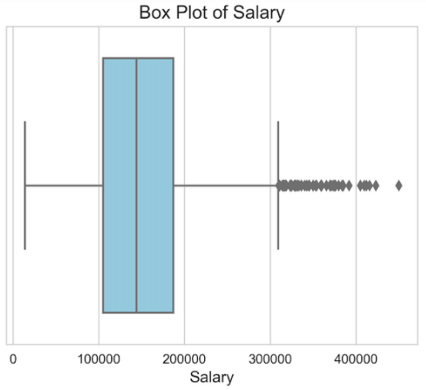
Descriptive Statistics

- **Years of Work**: From 2020 to 2023.
- **Salary**: Ranges from a minimum of 14,000 to a maximum of 450,000 with a mean of approximately 149,927.98 and a median of 143,860.
- **Salary in USD**: Mirrors the salary range with a mean of approximately 150,299.49 and a median of 143,000.

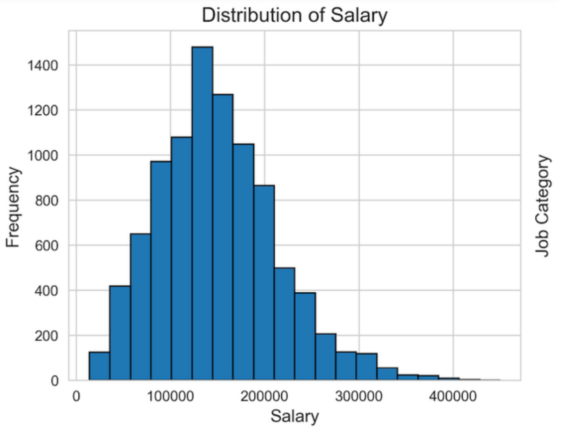
Visual Analysis

The following visual representations were derived from the data:

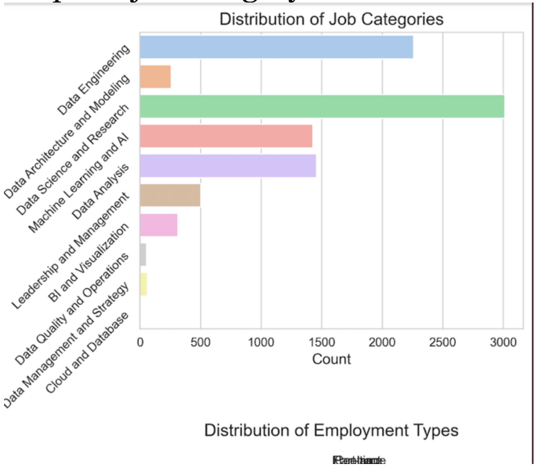
1. **Box Plot of Salary**: A spread of salaries with multiple outliers beyond the upper quartile was depicted.



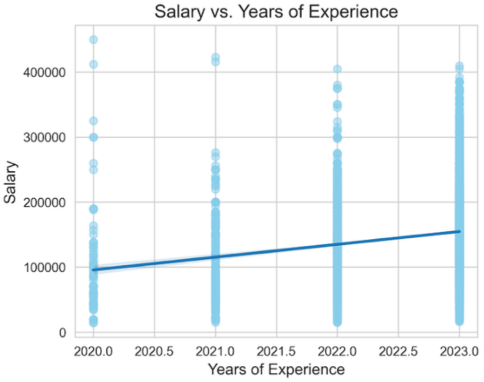
2. **Distribution of Salary**: Showcased a right-skewed histogram, implying that most data-related professionals earn less than the mean salary.



3. **Job Category Distribution**: Data Science and Research emerged as the most frequent job category.

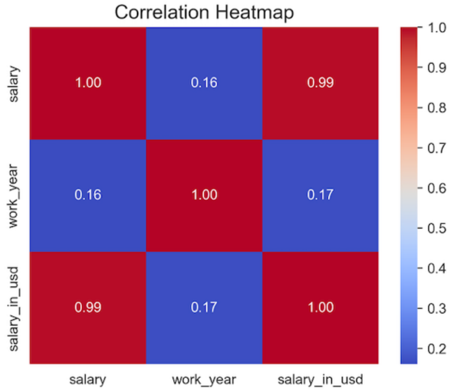


4. **Salary vs. Years of Experience**: A regression analysis indicated a modest positive correlation between years of experience and salary.

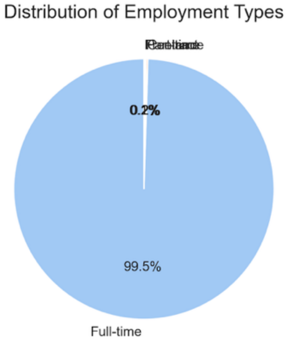


5. **Correlation Heatmap**: Exposed a weak correlation between work years and salary,

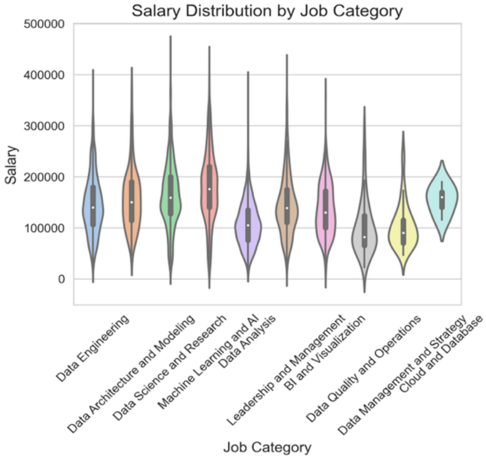
challenging conventional assumptions that more experience leads to higher pay.



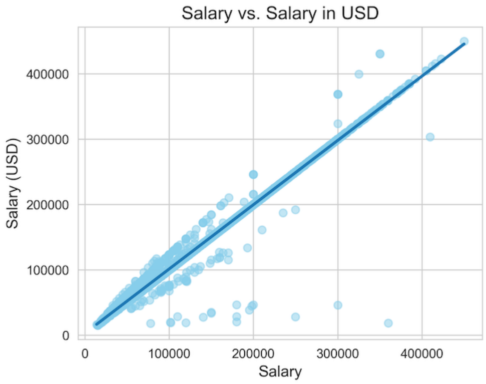
6. **Employment Type Distribution**: Highlighted a dominant prevalence of full-time employment at 99.5%.



7. **Salary by Job Category**: Variations in salary distributions across different job categories were visible through violin plots.



8. **Salary vs. Salary in USD**: Unsurprisingly, a direct correlation between salary and salary in USD was observed.



Outlier Analysis

Outliers, 144 in number, were identified and are considered to be salaries well above the norm, likely due to specialized roles, senior positions, or particularly high-demand skills.

Predictive Model Analysis

A linear regression model was applied to examine the relationship between years of experience and salary, resulting in:

- **Training Set Performance**: An MSE of 3,892,292,951.27 and an R-squared of 0.0277, suggesting a very low predictability from the model based on the training data.
- **Testing Set Performance**: An MSE of 4,138,383,131.07 and an R-squared of 0.0172, reaffirming the model's low predictability on unseen data.

Conclusions

The salary landscape within data-related professions is complex and influenced by a variety of factors. The analysis revealed significant variations in salary distribution & indicated that experience alone does not strongly predict salary.

Future Scope

Future analyses would incorporate additional variables such as job category, experience level, and company size into a multivariate regression model to potentially improve predictability. The insights gained from this study can serve as a reference for salary benchmarks in the data profession sector, aiding both employers and job seekers.