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.

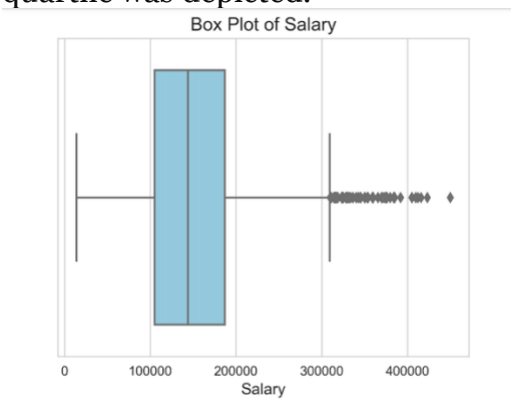
Descriptive Statistics

- **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

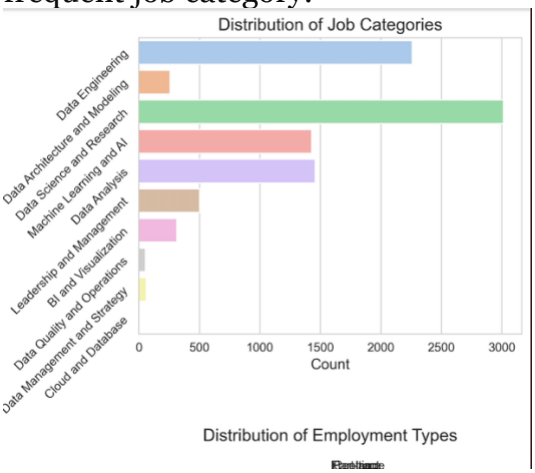
The following visual representations were derived from the data:

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

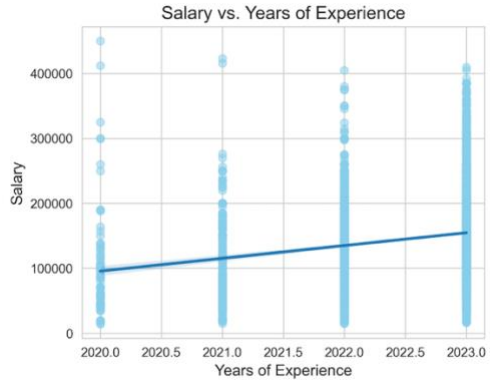


2. **Distribution of Salary**: Showcased a right-skewed histogram, implying that most data-related professionals earn less than the mean salary.

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

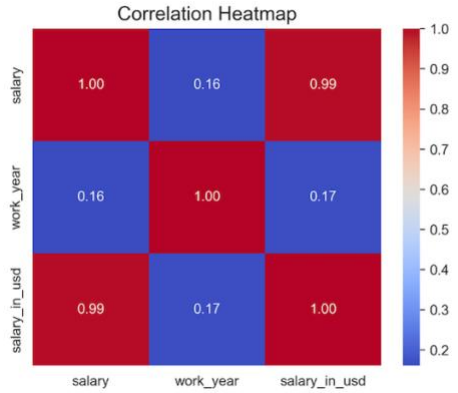


4. **Salary vs. Years of Experience**: A regression analysis indicated a modest positive correlation between years of experience and salary.

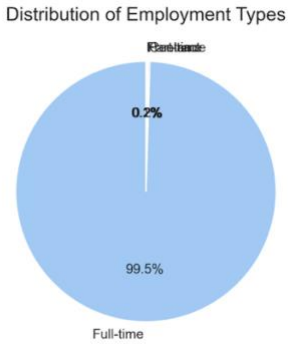


5. **Correlation Heatmap**: Exposed a weak correlation between work years and salary,

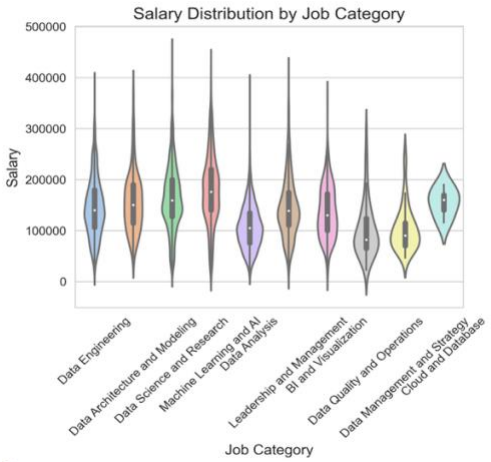
challenging conventional assumptions that more experience leads to higher pay.



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.

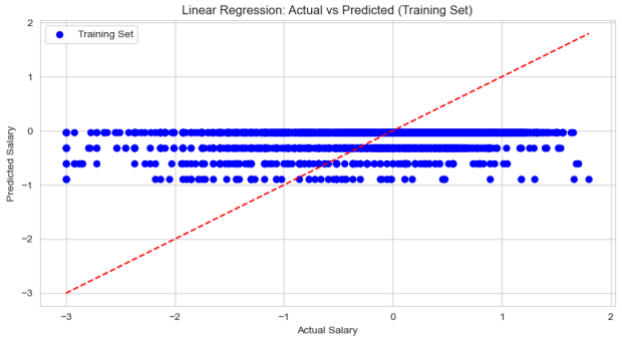
Analysis

The analysis involved building and evaluating three different regression models to predict salary based on years of experience: Linear Regression, Support Vector Regression (SVR), and XGBoost Regression. The models were assessed on their performance on both the training and testing datasets using Mean Squared Error (MSE) and R-squared metrics.

Results

Linear Regression Model

- **Training Set Performance:**
  - Mean Squared Error (MSE): 0.6226610849431234
  - R-squared: 0.0464807413172158
- **Testing Set Performance:**
  - Mean Squared Error (MSE): 0.6295876474759886
  - R-squared: 0.021344039568669926



Comprehensive Salary Analysis Report

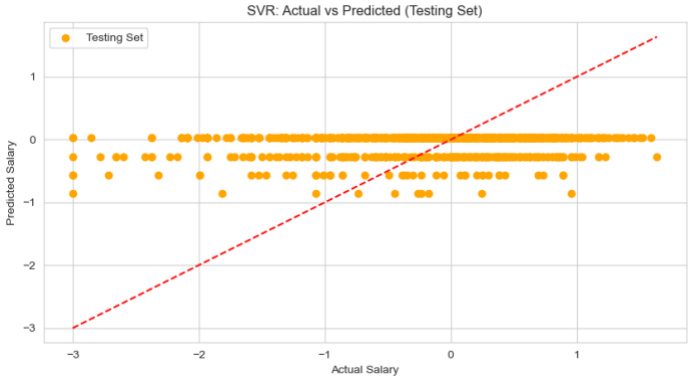
- **Gursimran**

The Linear Regression model shows low R-squared values for both the training and testing sets, indicating that the model explains only a small portion of the variance in the salary data based on years of experience. The MSE values suggest a moderate error in predictions.

Support Vector Regression (SVR) Model

- **Training Set Performance:**
  - Mean Squared Error (MSE): 0.6255912959218273
  - R-squared: 0.041993528822712234

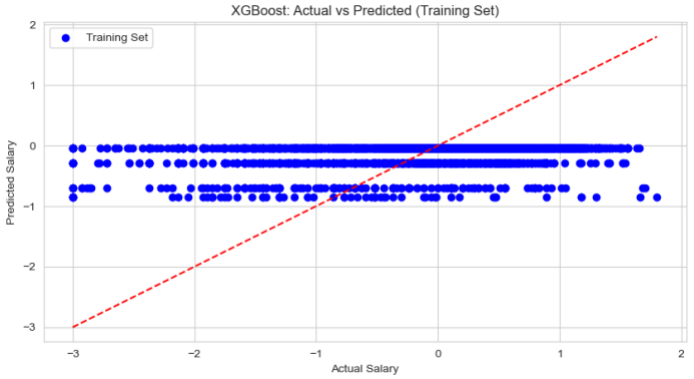
- **Testing Set Performance:**
  - Mean Squared Error (MSE): 0.634061304309229
  - R-squared: 0.014390010304710987



The SVR model performs similarly to the Linear Regression model, with slightly lower R-squared values, indicating that it also does not capture much of the variance in the salary data. The MSE values are close to those of the Linear Regression model.

XGBoost Model

- **Training Set Performance:**
  - Mean Squared Error (MSE): 0.6221726775763063
  - R-squared: 0.04722866959088523
- **Testing Set Performance:**
  - Mean Squared Error (MSE): 0.6281635273894057
  - R-squared: 0.023557748202712725

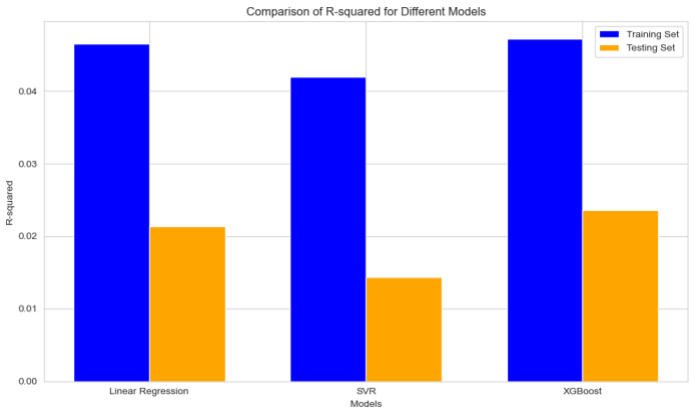


The XGBoost model shows the best performance among the three models in terms of R-squared values, although it is still relatively low. The MSE values are similar to those of the other models, indicating comparable prediction errors.

Conclusions

The predictive analysis of salary based on years of experience using Linear Regression, SVR, and XGBoost models revealed that:

- All three models have low R-squared values, suggesting that years of experience alone is not a strong predictor of salary in this dataset.
- The XGBoost model performed slightly better in terms of R-squared values but still did not capture a significant portion of the variance.
- The Mean Squared Error values are similar across all models, indicating that the prediction errors are comparable.



These results highlight the complexity of predicting salary and suggest that other factors beyond years of experience need to be considered to improve the predictive power of the models. Future analyses could incorporate additional

variables such as job category, education level, and company size to build more robust models.

Future Scope

The future scope of this project includes incorporating additional variables such as job category, education level, company size, location, and specific skills to enhance model accuracy. Advanced modeling techniques, including ensemble methods, neural networks, and extensive hyperparameter tuning, should be employed to improve prediction performance. Multivariate regression and time-series analyses will be conducted to understand the combined impacts of various factors and study salary trends over time. Geospatial analysis will be performed to explore regional salary differences, along with industry-specific analyses for tailored insights. Finally, developing interactive dashboards for dynamic data exploration and implementing automated updates will ensure that the models and insights remain current and relevant.