Employee Salaries Analysis and Regression Model

Project Overview

- Analysed employee salary dataset.
- Cleaned data, visualized key insights, and built regression models.
- Predicted employee salaries using different machine learning algorithms.

Dataset

- Source: Kaggle `Employee_Salaries.csv`
- Contains: Employee salary, department, and other related attributes.

Libraries Used

- pandas
- numpy
- seaborn
- matplotlib
- sklearn

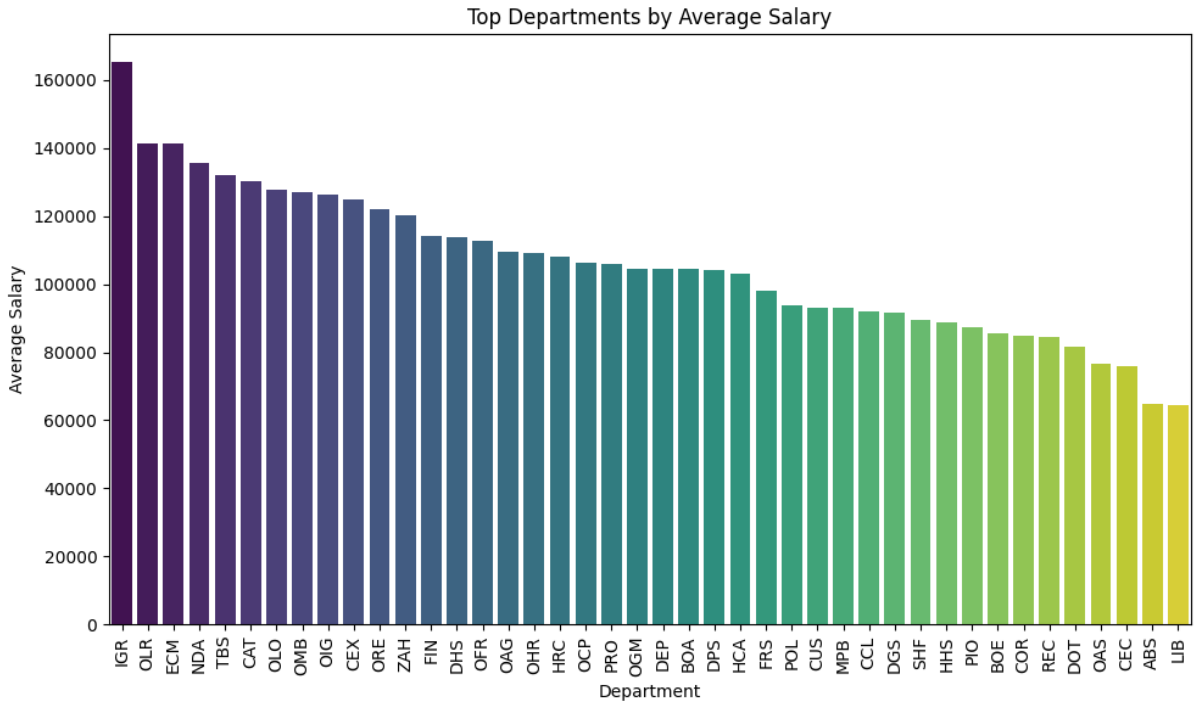
Data Cleaning

- **Null Values:**
  - Since there is only less than 1% of null values only there in the dataset. Checked and removed rows with null values.
- **Duplicates:**
  - Identified and removed duplicate rows (more than 5%).

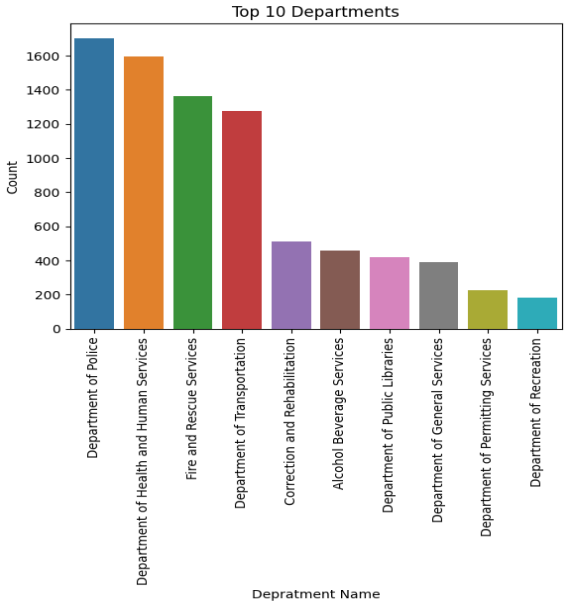
Data Visualization

- **Distribution and Outliers:**
  - Box plots and histograms for numeric columns.
  - Insight: Higher salaries could be justified by the roles; no need to treat outliers.

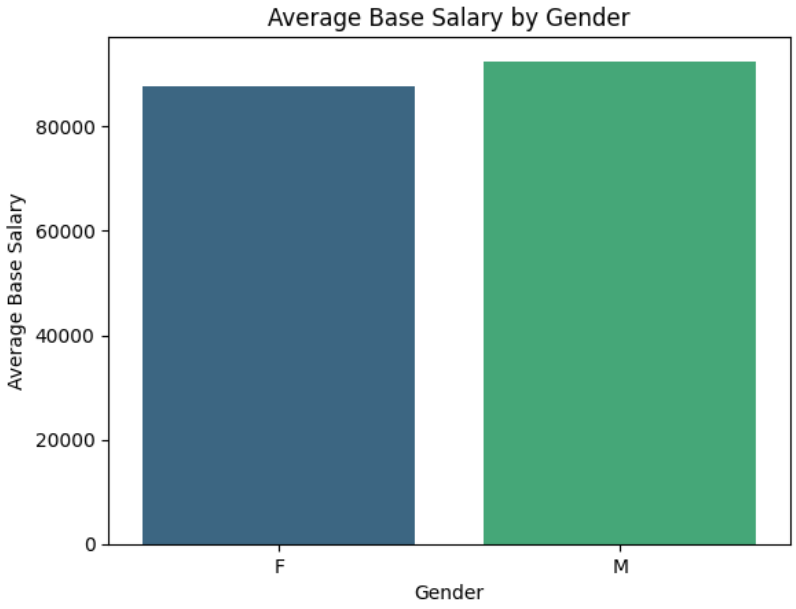
- **Department Analysis:**
  - Average salary per department.
  - Insight: Significant variation in average salary across departments.
  - Employees in the IGR,OLR,ECM departments get higher salary than other department employees.
  - Employees in the OEC,ABS,LIB department get lower salary.



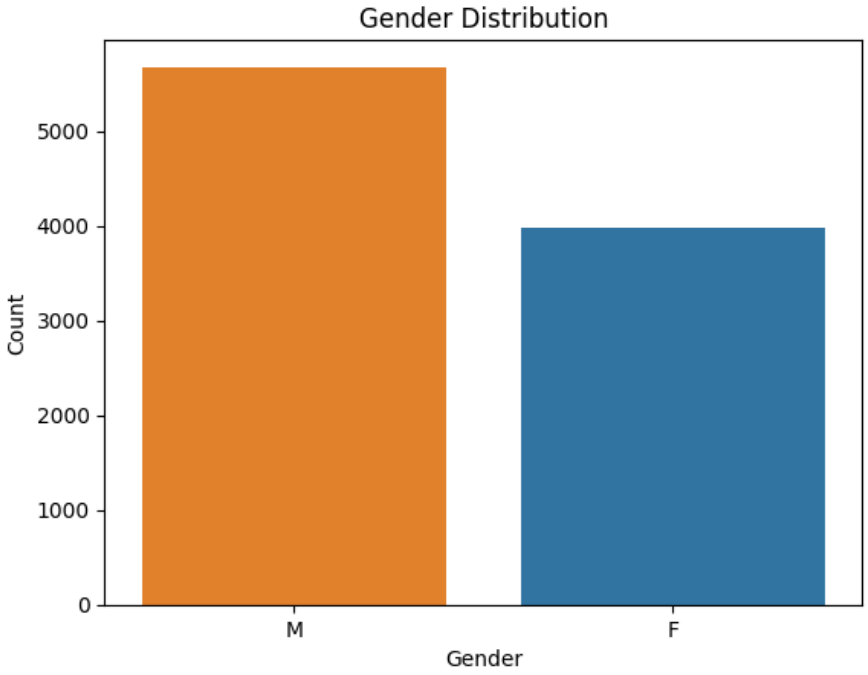
- **Department Occurrences:**
  - Count of top 10 departments.
  - Insight: Police, Health and Human Services, Fire and Rescue Service, and Transportation Department have the highest number of employees.



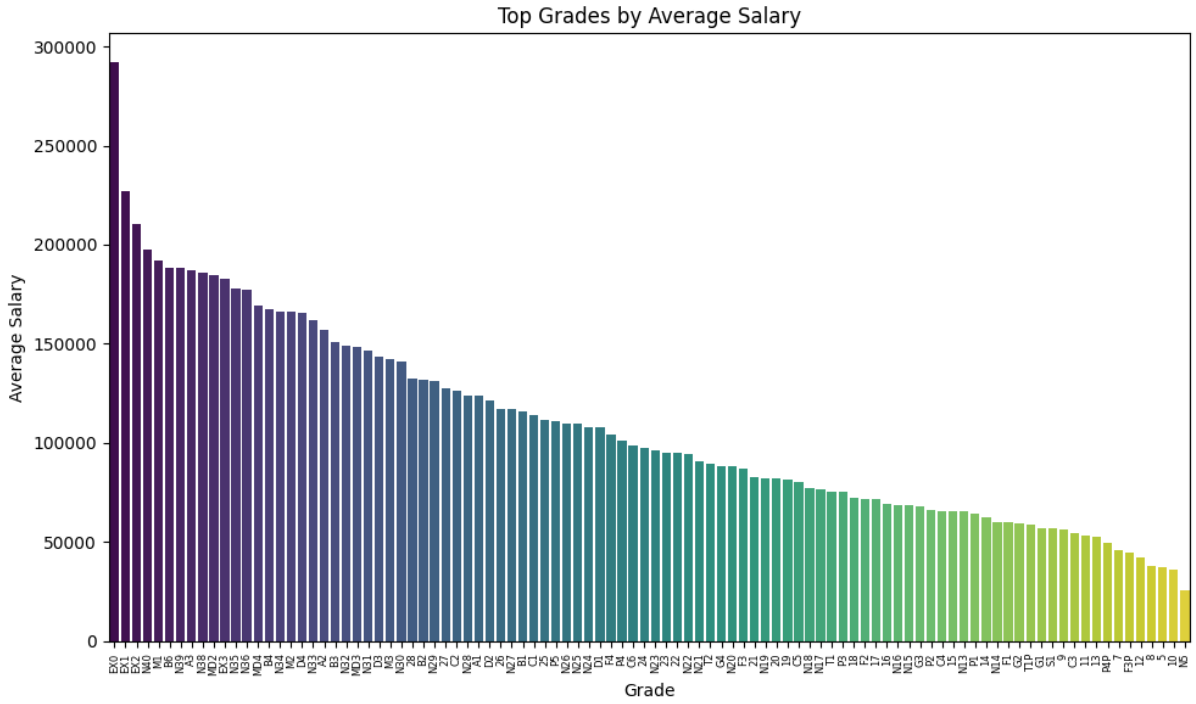
- **Gender Distribution and Pay Gap:**
  - Gender distribution and average base salary by gender.



  - Insight: Number of male employees is 17% higher than female employees, and male employees have slightly higher salaries.

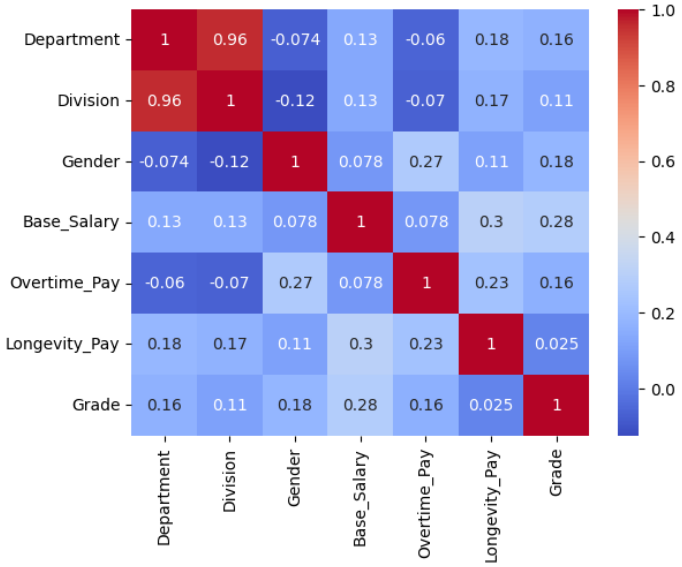


- **Grade Analysis:**
  - Average base salary by grade.
  - Insight: Salary varies significantly based on grade.



  - EX0 EX1 EX2 are the highest paid grades whereas 5,10 and N5 are the least paid grades.

- **Correlation Analysis:**
  - Heatmap for feature correlation.
  - Insight: No independent feature has a high correlation with the dependent feature; the Department attribute has a high correlation with Division.



Example Visualizations

Example Visualizations:

- Box plots and histograms of numeric columns.
- Bar plots for average salary per department.
- Count plots for top departments and divisions.
- Gender distribution and average salary by gender.
- Correlation heatmap.

Data Preprocessing

- **Encoding Categorical Variables:**
  - `OrdinalEncoder` for hierarchical features like `Department` and `Grade`.
  - `LabelEncoder` for other categorical features like `Gender` and `Division`.

Regression Models

1. Linear Regression
2. Decision Tree Regressor
3. Random Forest Regressor
4. Gradient Boosting Regressor

Model Evaluation Metrics

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- R-squared (R2) Score

Insights

- **Decision Tree Regressor:**
  - Higher R2 score for training data but lower for test data when compared with Random forest model.
- **Random Forest Regressor:**
  - Good R2 score for both training and test data, indicating better generalization.

Model Results

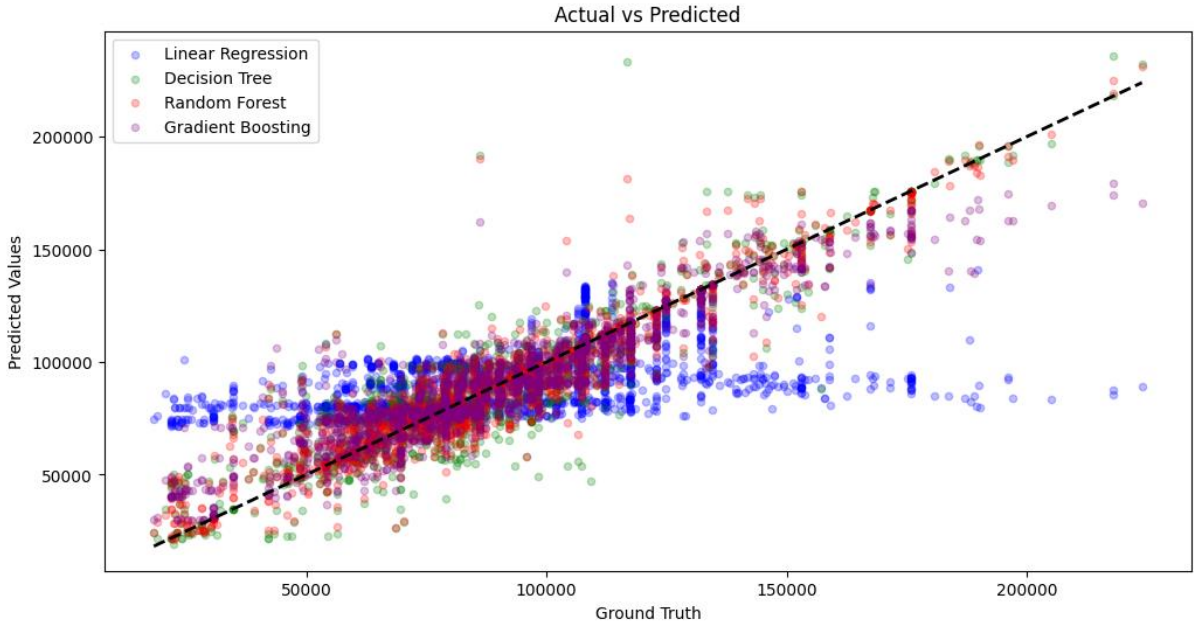
Random Forest Regressor Results:

- **Training Data Metrics:**

  R2 Score: 0.96

- **Test Data Metrics:**

  R2 Score: 0.89



- Values of test data predicted by the Random Forest model is closer to the actual values when comparing to the other models.

Conclusion

Conclusion:

- Analysis provides insights into salary factors.
- Demonstrates the effectiveness of ensemble methods like Random Forest.
- Future work: Explore additional features and refine models.

How to Run

How to Run:

1. Install required libraries.
2. Place `Employee_Salaries.csv` in the specified path.
3. Run the provided code for analysis and model building.

Thank You !