San Francisco Employee Salary Analysis

Project Overview

- Objective: Analyze salary data of SF public employees and predict TotalPayBenefits.
- Dataset: 312,882 records from 2011 to 2018
- Tools: Python (pandas, seaborn, sklearn), Jupyter Notebook
- Tasks: Data Cleaning, EDA, Visualization, ML Modeling

Data Cleaning Steps

- Converted pay columns to numeric values
- Removed rows with missing or negative pay values
- Filtered invalid or zero TotalPay entries

Exploratory Data Analysis (EDA)

- Top job titles by average pay include executive roles
- BasePay and Benefits are highly correlated with TotalPay
- Average salary has increased from 2011 to 2018

Machine Learning Model

- Model: Linear Regression
- Target Variable: TotalPayBenefits
- Features: BasePay, OvertimePay, JobTitle (encoded), etc.
- Train/Test Split: 70/30

Model Evaluation

- Mean Absolute Error (MAE): approx. 12,000-15,000 (depending on dataset)
- R-squared Score: approx. 0.85-0.90
- Good prediction accuracy for regression task

Conclusion

- Salaries vary widely by job title and year

- BasePay and Benefits are major factors in TotalPay

- Model effectively predicts overall compensation

- Project demonstrates end-to-end ML pipeline with real-world data