

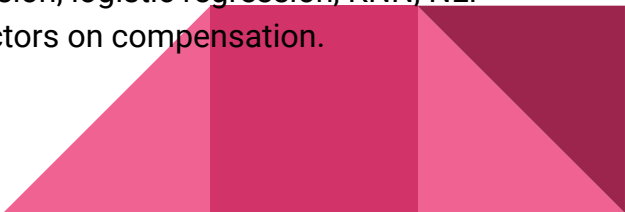
EMPLOYEE SALARY PREDICTION

Project by
Chetan Bhangare

Introduction

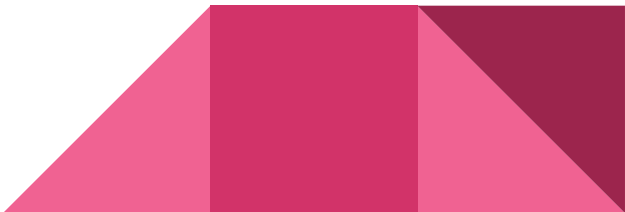
Objective: This project aims to leverage data analysis techniques to predict employee salaries based on various factors, including demographics, job roles, education, and performance metrics. The goal is to provide organizations with data-driven insights to optimize compensation strategies.

Overview:

- **Data Insights:** Using statistical and machine learning methods, the project explores key trends and correlations in employee data to identify the primary drivers behind salary decisions.
 - **Data Quality & Preprocessing:** The project addresses common data challenges, including missing values and outliers, ensuring that the dataset is clean and ready for analysis.
 - **Exploratory Data Analysis (EDA):** Visualizations and summary statistics are used to uncover hidden patterns and relationships in employee data, providing guidance for HR policy formulation.
 - **Predictive Modeling:** Various machine learning models, including linear regression, logistic regression, KNN, NLP are applied to predict employee salaries and assess the impact of different factors on compensation.
- 

Dataset Overview

The dataset contains multiple columns with information about employees. Here is an overview of the key columns:

- Employee_ID: A unique identifier for each employee.
 - First_Name: The first name of the employee.
 - Last_Name: The last name of the employee.
 - Department: The department in which the employee works (e.g., Finance, HR, IT).
 - Position: The job role or position the employee holds (e.g., Manager, Engineer, Coordinator).
 - Age: The age of the employee, providing demographic information.
 - Salary: The salary of the employee (target variable for prediction).
- 

Dataset Overview

- **Joining_Date:** The date when the employee joined the company.
- **City:** The city where the employee works.
- **Country:** The country of the employee's location (important for analyzing salary trends across regions).
- **Performance_Score:** A numeric score reflecting the employee's performance.
- **Experience_Years:** The total number of years the employee has worked in their current role or field.
- **Education_Level:** The highest level of education attained by the employee (e.g., High School, Bachelor's, PhD).
- **Gender:** The gender of the employee, used for diversity and equity analysis.
- **Marital_Status:** The marital status of the employee (Single, Married, Divorced, etc.).



Data Integrity

Data Integrity Issues:

- **Duplicate Data:**
 - **Problem:** Duplicate records from merging datasets can cause redundancy.
 - **Impact:** Inflates results, leading to biased analysis.
 - **Solution:** Implement deduplication checks to remove redundant records.
- **Inaccurate Labels:**
 - **Problem:** Incorrect or inconsistent labels (e.g., job titles, department names) can distort analysis.
 - **Impact:** Leads to misclassifications and incorrect salary predictions.
 - **Solution:** Standardize and validate labels to ensure consistency and accuracy.



Data Preprocessing

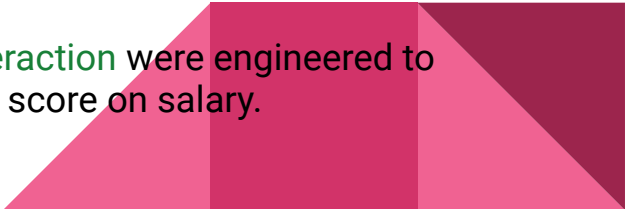
No Missing Values:

- **Data Integrity:** The dataset has been thoroughly checked, and there are no missing values. Each column has complete information, ensuring no gaps that would affect model accuracy.

No Outliers:

- **Outlier Detection:** Through statistical analysis, no significant outliers were found in any key numerical columns (e.g., salary, performance score). This ensures that the data distribution is normal and avoids skewing model predictions.

Feature Transformation:

- **Log Transformation:** The `log_salary` feature was created to normalize salary distribution, improving model performance and reducing skewness.
 - **Interaction Features:** New features like `experience_performance_interaction` were engineered to capture the combined effect of years of experience and performance score on salary.
- 

Feature Engineering

Key Features Added for Salary Prediction Model

- **Log Salary Transformation**
 - `log_salary = np.log(Salary + 1)`
 - Normalizes salary distribution, reduces skewness.
- **Experience and Performance Interaction**
 - `experience_performance_interaction = Experience_Years * Performance_Score`
 - Captures combined effect of experience and performance on salary.
- **Polynomial Features**
 - `experience_squared = Experience_Years ** 2`
 - Models non-linear relationships between experience and salary.
 - `performance_squared = Performance_Score ** 2`
 - Models non-linear impact of performance on salary.
- **Experience + Performance Score**
 - `experience_plus_performance = Experience_Years + Performance_Score`
 - Captures additive effects of experience and performance on salary.

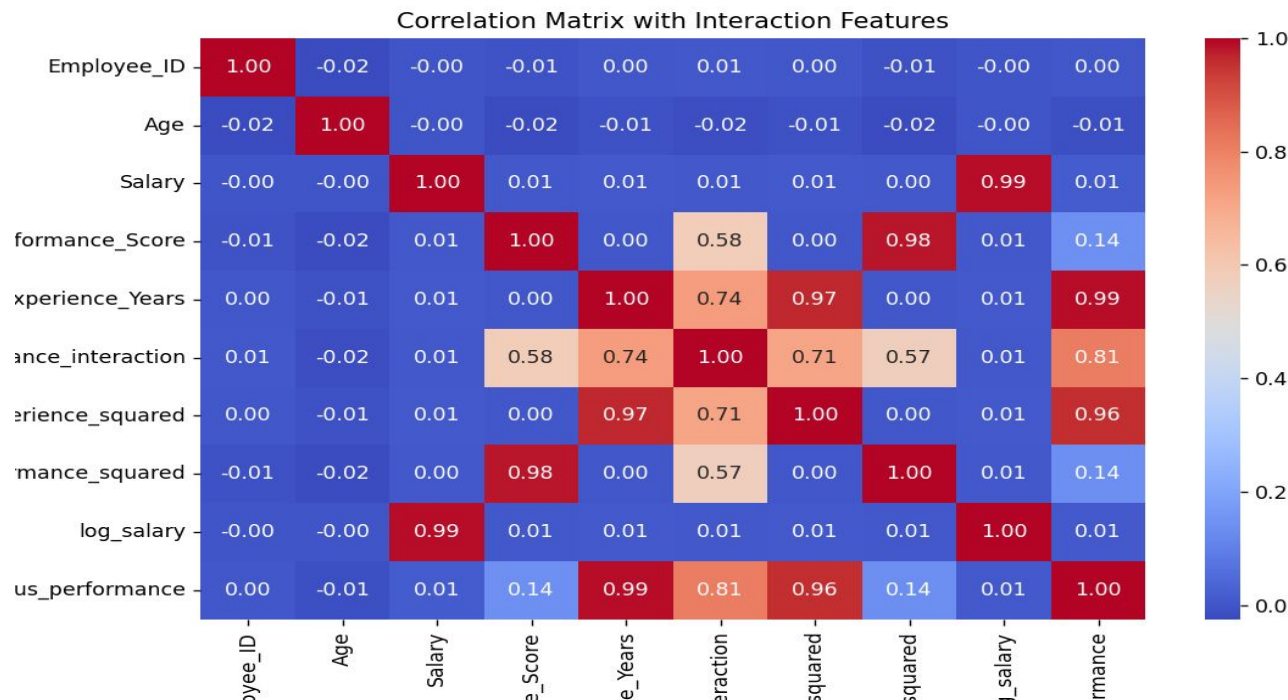
Data Mining

Before feature
engineering



Data Mining

After feature
engineering

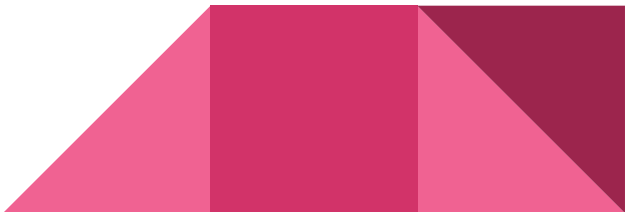


Data Mining

Clustering Process

- Used **K-Means Clustering** on these features:
 - **Employee_ID, Age, Salary, Performance_Score, Experience_Years.**
- Dataset details:
 - **10,000 rows with 5 numeric columns.**

Results

- **Optimal Clusters:**
 - Best cluster count: **3 groups** (Silhouette Coefficient = **0.55**).
 - Performance declines with more clusters (e.g., 9 clusters = **0.38**).
 - **Cluster Centroids** (average values for each feature):
 - Salary ranges: **\$48,000 to \$111,000.**
 - Performance Score: ~3 across clusters.
 - Experience Years: ~16-17 across clusters.
- 

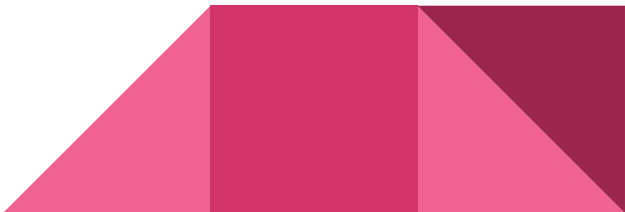
Data Mining

Linear Regression

Evaluation Metrics

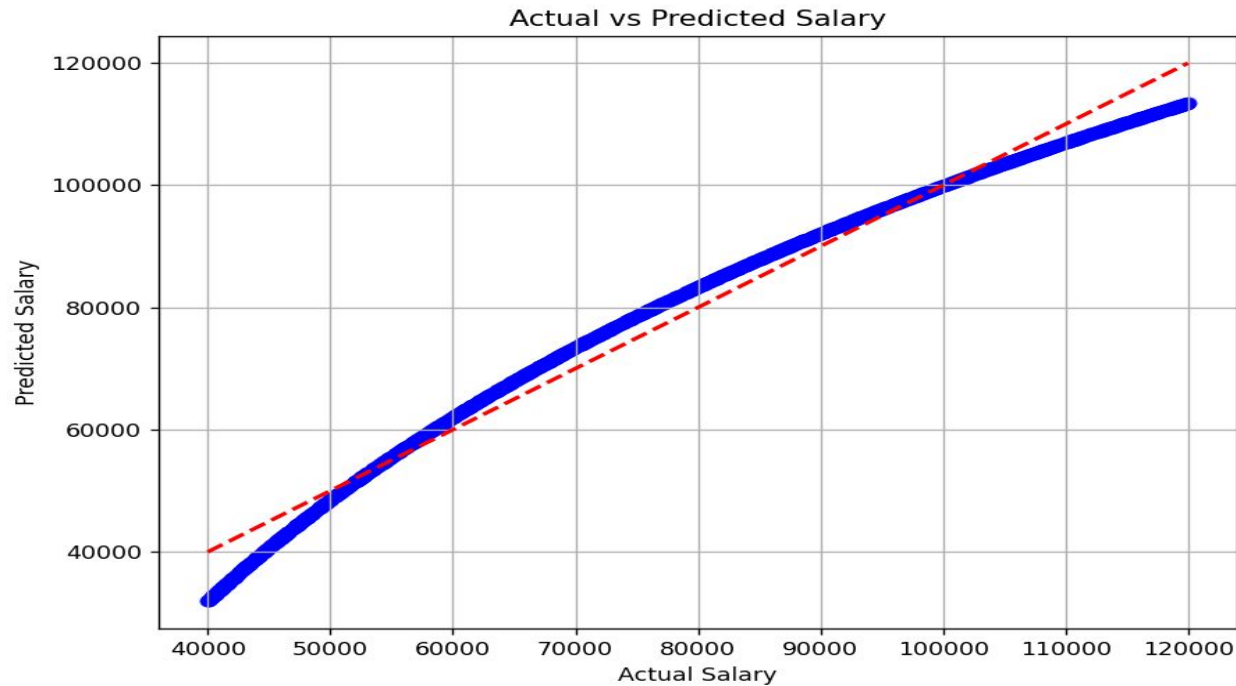
- **Adjusted R^2 : 0.9805 .**
- **Mean Absolute Error (MAE): 2808.22 .**
- **Mean Squared Error (MSE): 10,648,925.36 .**
- **Root Mean Squared Error (RMSE): 3,263.27 .**

Key Takeaways

- **Best-performing model** for salary prediction with minimal error.
 - High **Adjusted R^2** demonstrates strong explanatory power.
 - Suitable for salary forecasting and HR decision-making.
- 

Data Mining

Linear regression
graph



Data Mining

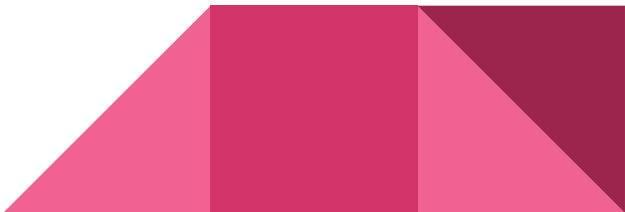
Logistic Regression

Objective: Classify employee salaries into "Low" ($<80,000$) and "High" ($\geq 80,000$).

Class Distribution:

- Low: 5044 records
- High: 4956 records

Confusion Matrix:

- True Positives : 1250
 - True Negatives : 1243
 - False Positives : 0
 - False Negatives : 7
- 

Data Mining

KNN based of Logistic Regression

- **R-squared (Score):** 0.993

KNN Classification

- **Accuracy:** 98.4%
- **Confusion Matrix:**
 - True Positives: 853
 - True Negatives: 1115
 - False Positives: 17
 - False Negatives: 15

Key Observations

- Excellent performance for both regression and classification tasks.
- Model shows signs of **overfitting**, indicated by close alignment of training and test scores.
- Future steps: Apply hyperparameter tuning or cross-validation to mitigate overfitting risks.

Data Mining

Principal Component Analysis(PCA)

Objective: Applied Principal Component Analysis (PCA) for dimensionality reduction and compared with non-PCA Linear Regression.

PCA Transformation: Reduced features to 3 principal components.

Dataset Details:

- Total samples: 10,000.
- Training set: 8,000.
- Testing set: 2,000.

PCA Linear Regression Performance: $R^2 = 0.9805$.

Non-PCA Linear Regression Performance: $R^2 = 0.9805$.

Observation: PCA did not significantly impact performance but reduced dimensionality effectively.



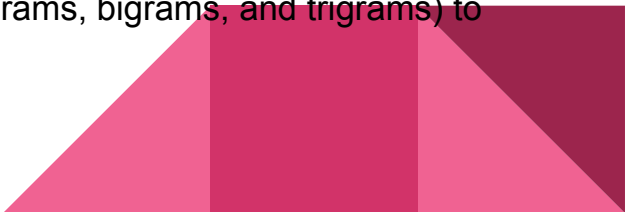
Data Mining

Natural Language Processing(NLP)

Text Feature Engineering:

- In this project, synthetic text data was created by combining key categorical attributes (Country, Department, Position) into a single text column `Employee_Info`. This text was used to extract meaningful features for model training.
- Text data is essential for Natural Language Processing (NLP) models, as it enables the system to understand patterns or structures from categorical data in a textual format.

Count Vectorization:

- A technique used to convert a collection of text documents into a matrix of token counts. Each word or phrase (n-gram) is represented as a feature, and its frequency within the document is counted.
 - The `CountVectorizer` captures features at different granularities (unigrams, bigrams, and trigrams) to capture more context and patterns in the data.
- 

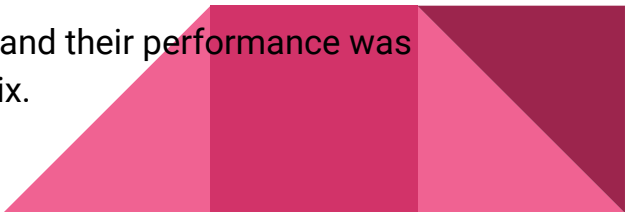
Data Mining

Natural Language Processing

TF-IDF Vectorization:

- The Term Frequency-Inverse Document Frequency (TF-IDF) model is used to convert text data into numerical features.
- Unlike simple counting, TF-IDF weighs each word based on its frequency in the document and the overall corpus, helping to prioritize more relevant words.

Modeling:

- **Logistic Regression** was applied to classify the salary class (**low**, **high**) based on the transformed text features.
 - Two vectorization techniques, **CountVectorizer** and **TF-IDF**, were applied, and their performance was compared in terms of accuracy, classification report, and confusion matrix.
- 

Data Mining

Natural Language Processing

Model Performance:

- **Count Vectorizer Model Accuracy: 54.2%**
 - Precision for 'high' salary class: 0.57
 - Recall for 'high' salary class: 0.81
 - F1-score for 'high' salary class: 0.67
 - Precision for 'low' salary class: 0.44
 - Recall for 'low' salary class: 0.19
 - F1-score for 'low' salary class: 0.26
- **TF-IDF Model Accuracy: 54.8%**
 - Precision for 'high' salary class: 0.57
 - Recall for 'high' salary class: 0.86
 - F1-score for 'high' salary class: 0.68
 - Precision for 'low' salary class: 0.44
 - Recall for 'low' salary class: 0.15
 - F1-score for 'low' salary class: 0.22



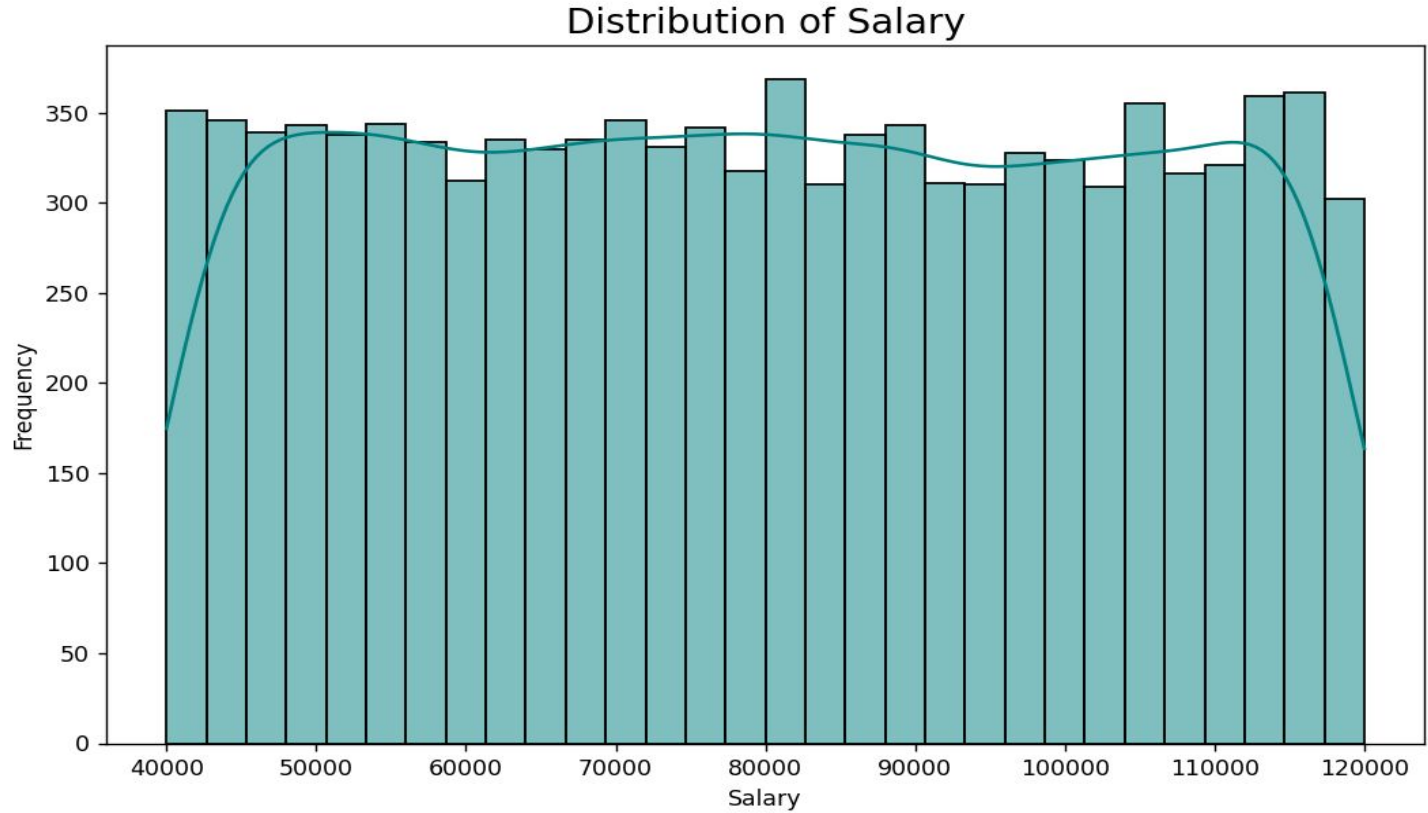
Data Mining

Data Mining Summary:

- **Cluster Analysis (K-Means):**
 - K-Means clustering identified patterns in employee demographics, salary, and performance.
 - Silhouette coefficient of 0.55 indicated reasonable clustering, but did not improve salary prediction accuracy.
- **Feature Engineering:**
 - Created new features like interaction terms, polynomial features, and log transformations to improve model performance.
 - Addressed weak correlations with salary.
- **Modeling:**
 - **Linear Regression:** Achieved high R^2 (0.98), highly effective in salary prediction.
 - **Logistic Regression & KNN Classification:** High accuracy (~99%) for salary classification, limited insights.
 - **PCA:** Showed similar results to non-PCA regression.
- **NLP Models:**
 - CountVectorizer and TF-IDF models achieved 54-55% accuracy, less effective for salary prediction.
- **Best Model:**
 - Linear regression, enhanced with feature engineering, provided the best predictive performance for salary prediction.

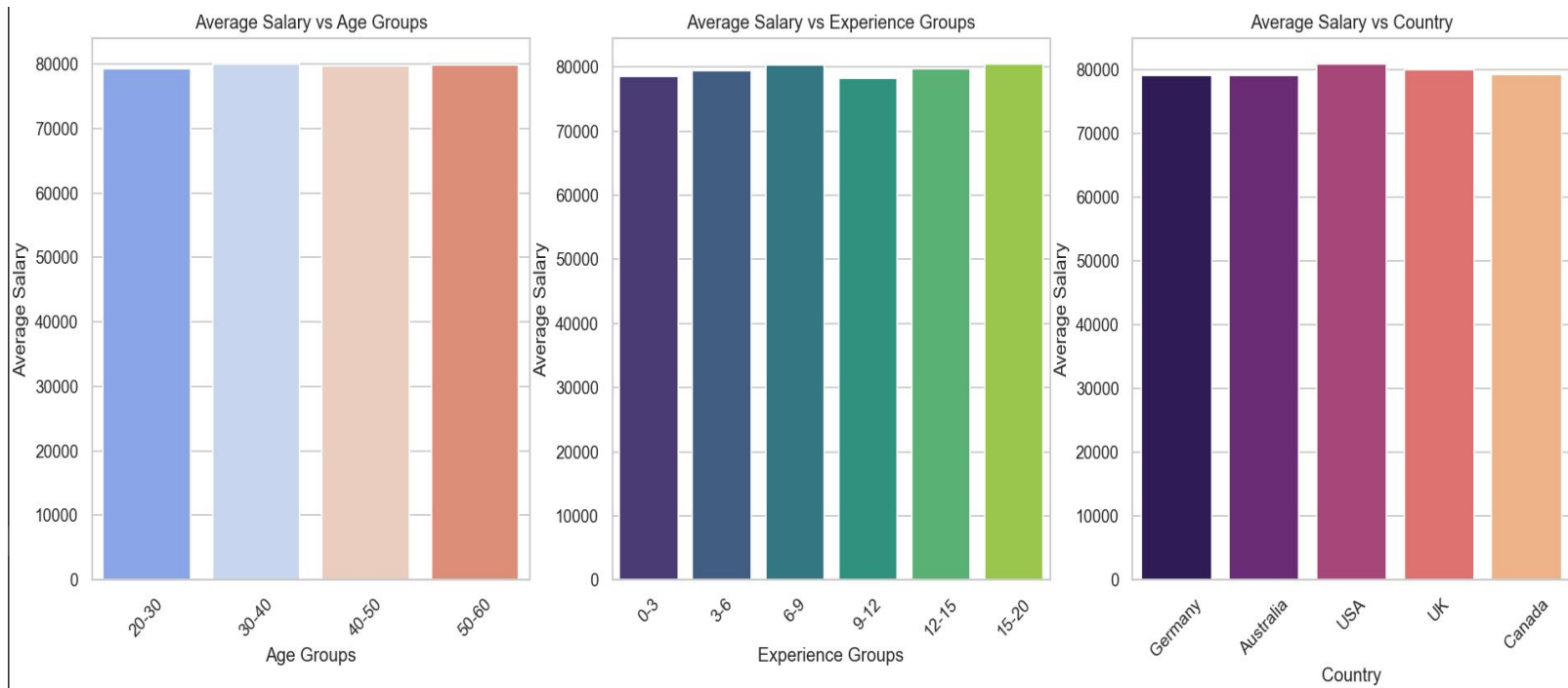
Data Visualization

Histogram



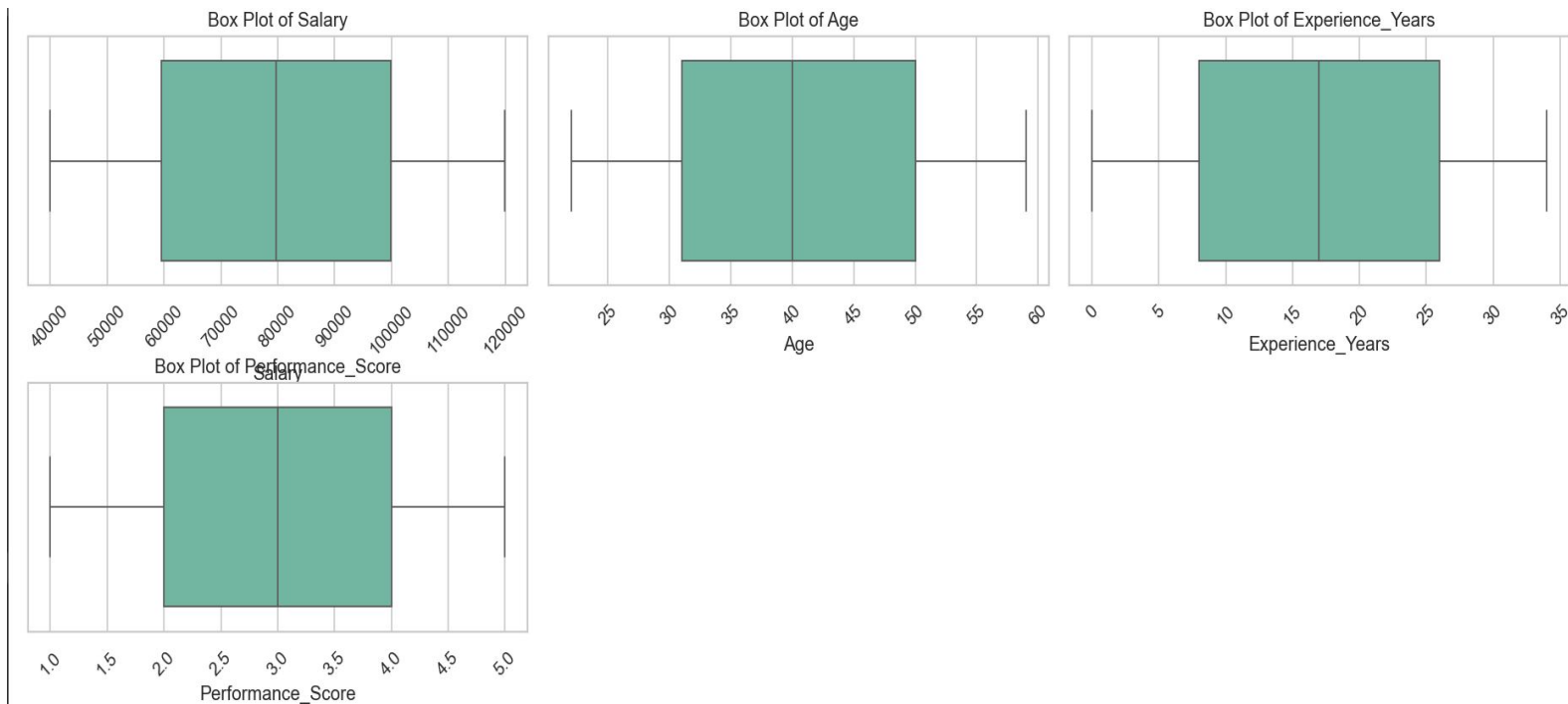
Data Visualization

Bar chart



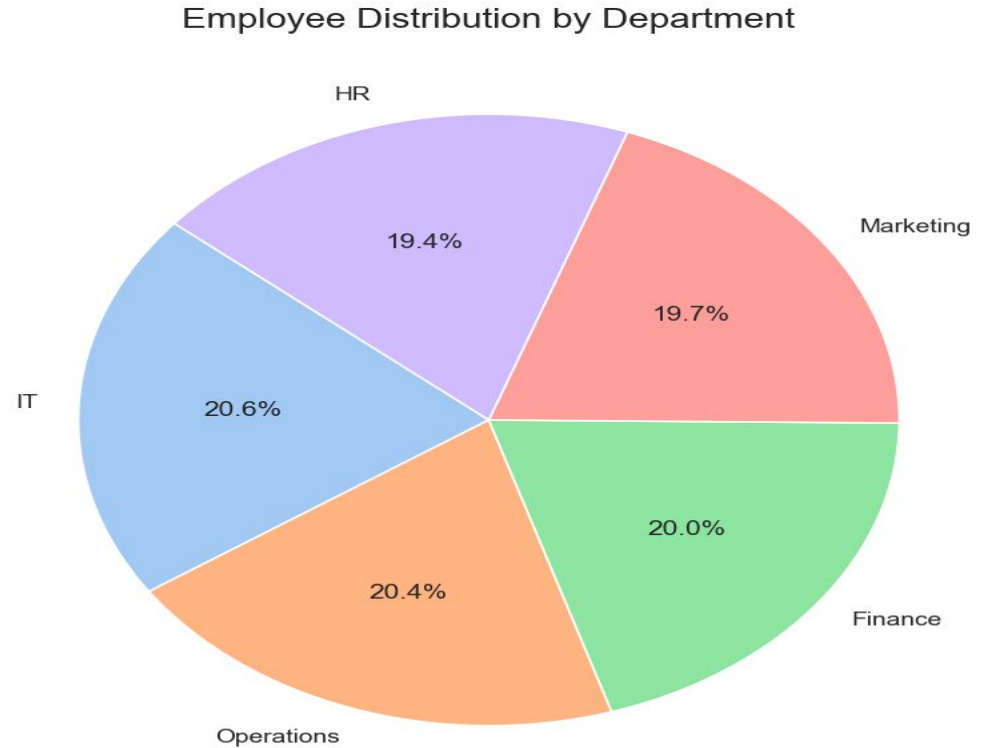
Data Visualization

Box plot



Data Visualization

Pie chart



Data Visualization

Violin plot



Data Visualization

Multi Line chart



Data Visualization

Data Visualization Summary

- **Histogram:**
 - **Distribution of Salaries:** Salary range from \$40,000 to \$120,000 with relatively uniform frequency across intervals.
 - A density plot overlay reveals a smooth distribution with no extreme concentrations.
- **Bar Charts:**
 - **Age Groups:** Slight differences in average salaries across various age groups.
 - **Experience:** Individuals with higher experience tend to earn slightly higher salaries.
 - **Countries:** The USA has the highest average salary among all countries compared.
- **Box Plot:**
 - No outliers are present in the data, suggesting a consistent salary distribution.
- **Pie Chart:**
 - The IT department has the highest number of employees, while the HR department has the lowest, indicating workforce distribution across departments.
- **Violin Plot:**
 - Visualizes salary distribution, capturing both density and spread for deeper insights.
- **Multiline Chart:**
 - Highlights that the HR department has the highest average salary compared to other departments, revealing a notable trend.

Data Visualization

EMPLOYEE ANALYTICS DASHBOARD



Associate Degree

Bachelor's

High School

Master's

PhD

Count of Employee

1009

Avg Age

41

Avg Salary

79K

Avg years

17

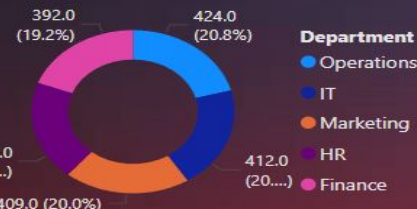
Male

664

Female

682

Count of Employee by Department



Count of Employee by Position



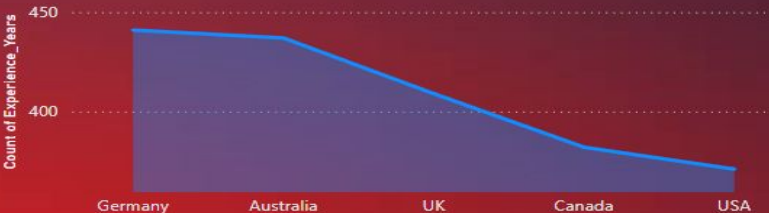
Sum of Salary by Performancescore

Position	Finance	HR	IT	Marketing	Operations	Total
Analyst	6724270	5596249	6415835	5811926	6863719	31411999
Coordinator	5625262	6296687	7377394	6633765	5964415	31897523
Director	5060987	8361437	5816084	5888776	7355863	32483147
Engineer	6564213	6647830	5996927	6772939	7121079	33102988
Manager	6957040	6009860	6381020	7206926	6516999	33071845
Total	30931772	32912063	31987260	32314332	33822075	161967502

Count of Employee by Age



Count of Experience_Years by Country



Data Visualization

Power BI Report Summary

- **Workforce Overview:**
 - 1,009 employees(Associate Degree) with an average age of 41 years, average salary of \$79K, and 17 years of experience.
 - Gender distribution: 664 males and 682 females.
- **Department & Position Breakdown:**
 - Employees evenly distributed across departments: Operations, IT, Marketing, HR, Finance (~20% each).
 - Positions: Director, Engineer, Manager, Analyst, Coordinator, each making up ~20% of the workforce.
- **Salary Insights:**
 - Total salaries highest in Operations (\$33.8M), followed by Marketing (\$32.1M), HR (\$32M), IT (\$31.9M), and Finance (\$31M).
- **Age & Experience Distribution:**
 - Employee ages range from 20 to 60, with a consistent spread.
 - Highest cumulative experience in Germany, followed by Australia, the UK, Canada, and the USA.
- **Interactivity:**
 - Dashboard allows filtering insights based on education levels (Associate's, Bachelor's, High School, Master's, PhD) for tailored analysis.

Conclusion

- **Goal:** Developed a model to predict employee salary based on demographic and performance features.
- **Data Preparation:** Extensive checks for null values, outliers, and feature engineering to enhance the dataset.
- **Exploratory Analysis:** K-Means clustering provided insights, but didn't significantly improve salary prediction accuracy.
- **Feature Engineering:** Interaction terms, polynomial transformations, and log transformations improved model performance by capturing non-linear relationships.
- **Model Evaluation:**
 - **Linear Regression:** Achieved R^2 of 0.98, the best-performing model for salary prediction.
 - **Logistic Regression & KNN:** High accuracy for salary classification but not for precise salary prediction.
 - **NLP:** Lower performance due to non-textual nature of the dataset.
- **Conclusion:** Feature engineering was critical to improving model accuracy. Linear regression, enhanced with feature engineering, was the most effective technique for predicting employee salaries.

Future Scope

Attrition Rate Analysis: Adding employee attrition data will enhance insights into turnover patterns, helping to build more comprehensive models for retention strategies.

Feature Expansion: Incorporating additional features like employee behavior, job satisfaction, and psychological factors could improve model robustness and accuracy.

Model Refinement: Feature engineering (interaction terms, polynomial features, and log transformations) has been key in improving linear regression model performance, and further refinements could enhance predictions.



THANK YOU

