Project Report: HR Dataset Analysis

Introduction

This project focuses on analyzing the HR dataset (Version 14) to understand various aspects such as salary distribution, employee performance, reasons for termination, and more. The dataset contains information about employees, including their department, position, salary, employment status, satisfaction levels, and more. By conducting this analysis, we aim to uncover insights that can help improve HR decision-making and overall employee management.

Dataset Overview

The dataset comprises various features related to employees, including:

Employee Information:ID, Name, DOB, Sex, Department, Position, Salary, Hire Date, Manager Name, etc.

Employment Status:Active/Terminated, Termination Reason, Termination Date, etc.

Performance and Satisfaction: EmpSatisfaction, PerformanceScore, Special Projects Count, etc.

Recruitment Source: The source from which the employee was recruited.

Data Preprocessing

Handling Missing Values: The dataset was checked for null values, and appropriate measures were taken to handle them.

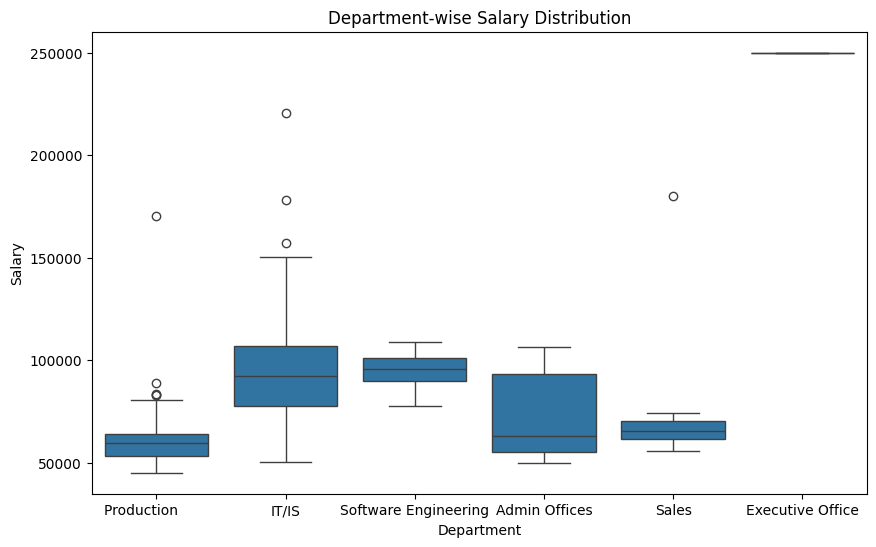
Handling Duplicates:Duplicate rows were identified and removed to ensure data integrity.

Feature Engineering:A new feature, 'Age', was created from the DOB column, and issues related to date formatting were resolved.

Salary Analysis

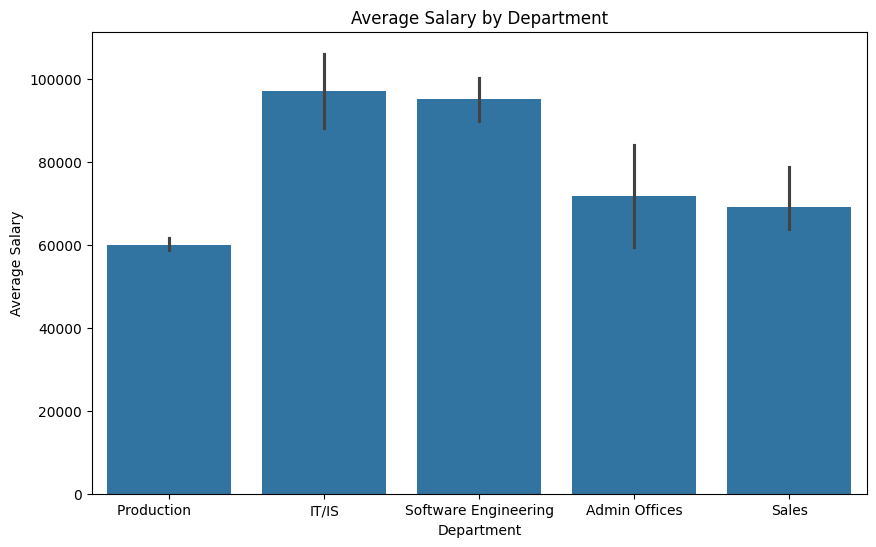
Salary Distribution by Department

- A boxplot was used to visualize the salary distribution across different departments. It was observed that the 'Executive Office' was an outlier, primarily due to the presence of the President & CEO, Janet King. Subsequent analyses excluded this department for a clearer view.



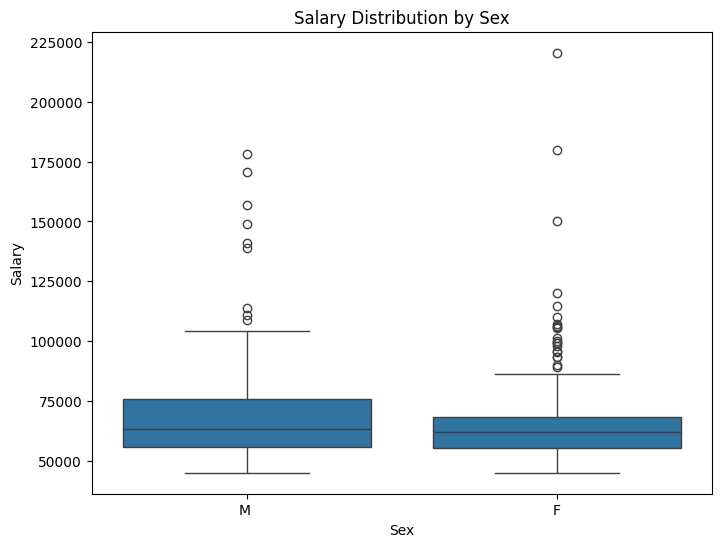
Average Salary by Department

- A bar plot showed the average salary across departments, excluding the Executive Office. This helped identify which departments had higher average salaries.



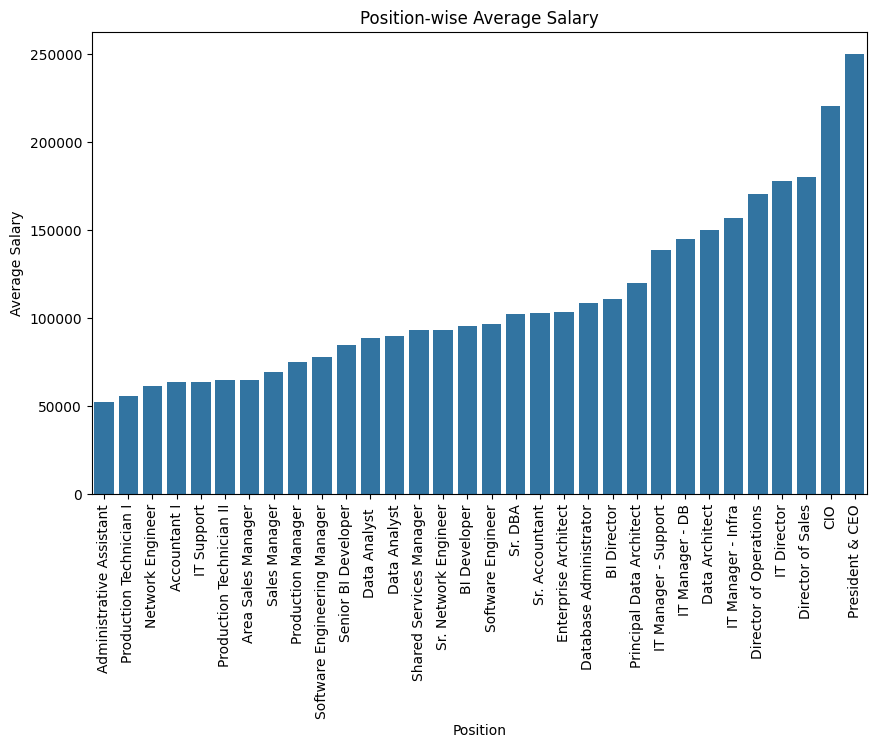
Salary Distribution by Sex

- The salary distribution by sex was analyzed, revealing any potential gender disparities in pay.



Average Salary by Position

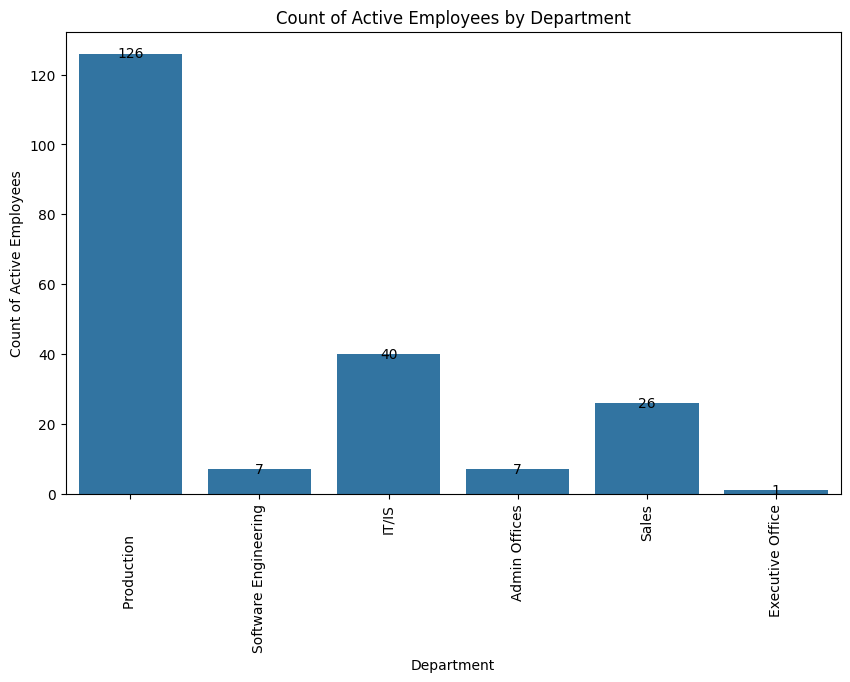
- A detailed analysis was conducted to understand the average salary for each position. This helped identify which roles were the highest or lowest paying within the organization.



Employee Count Analysis

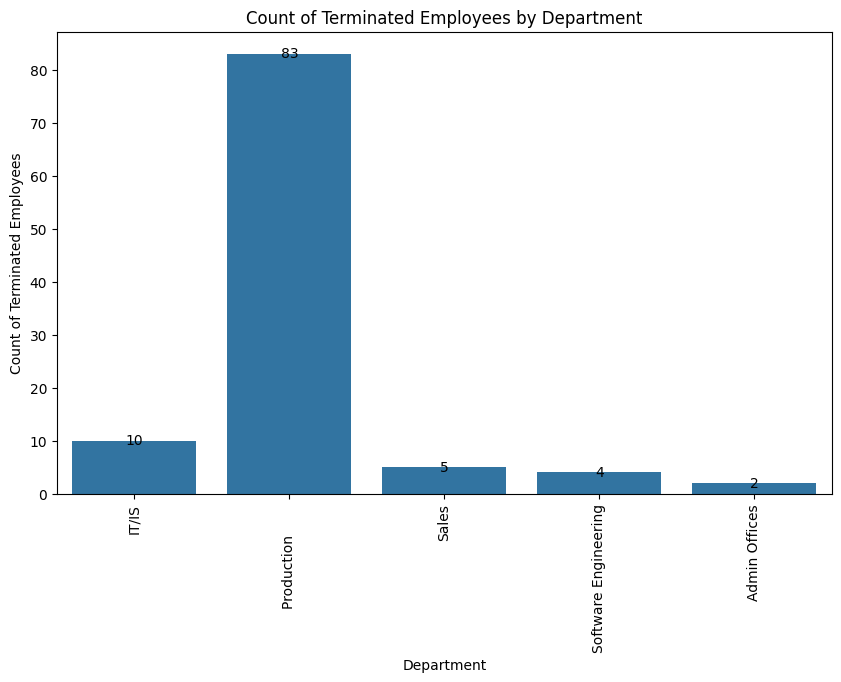
Active Employees by Department

- A count plot was generated to show the number of active employees in each department.



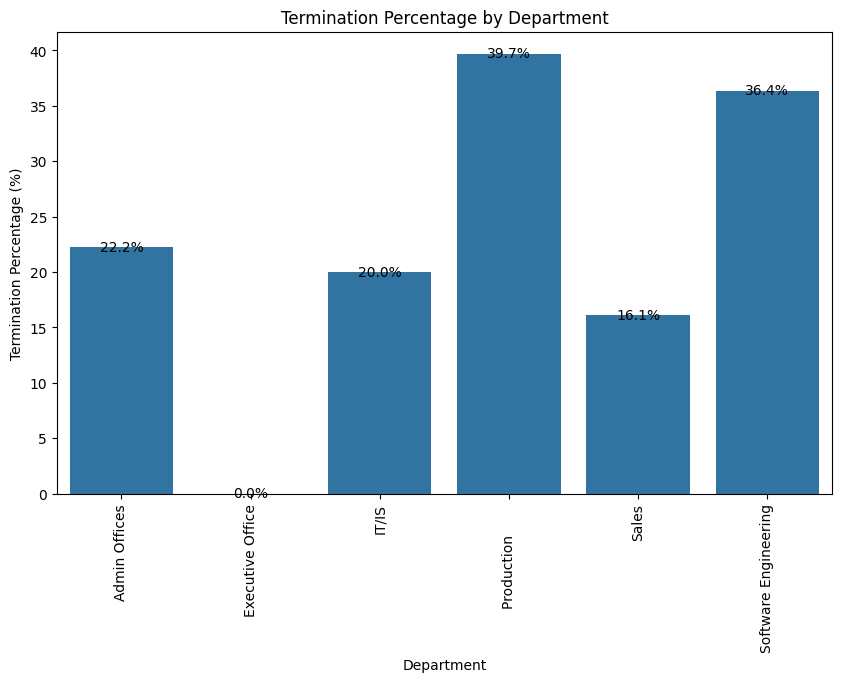
Terminated Employees by Department

- A similar plot was created to visualize the number of terminated employees per department.



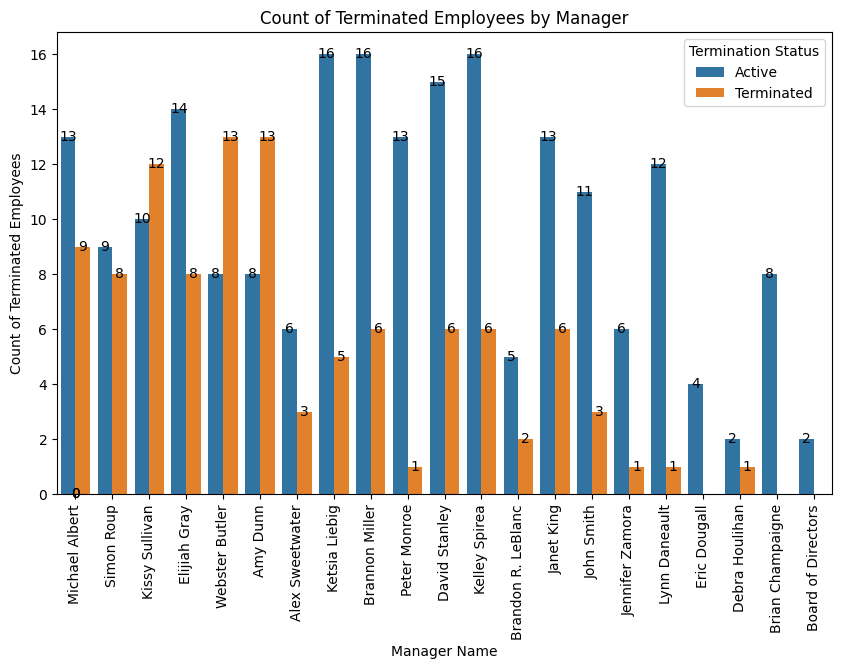
Termination Percentage by Department

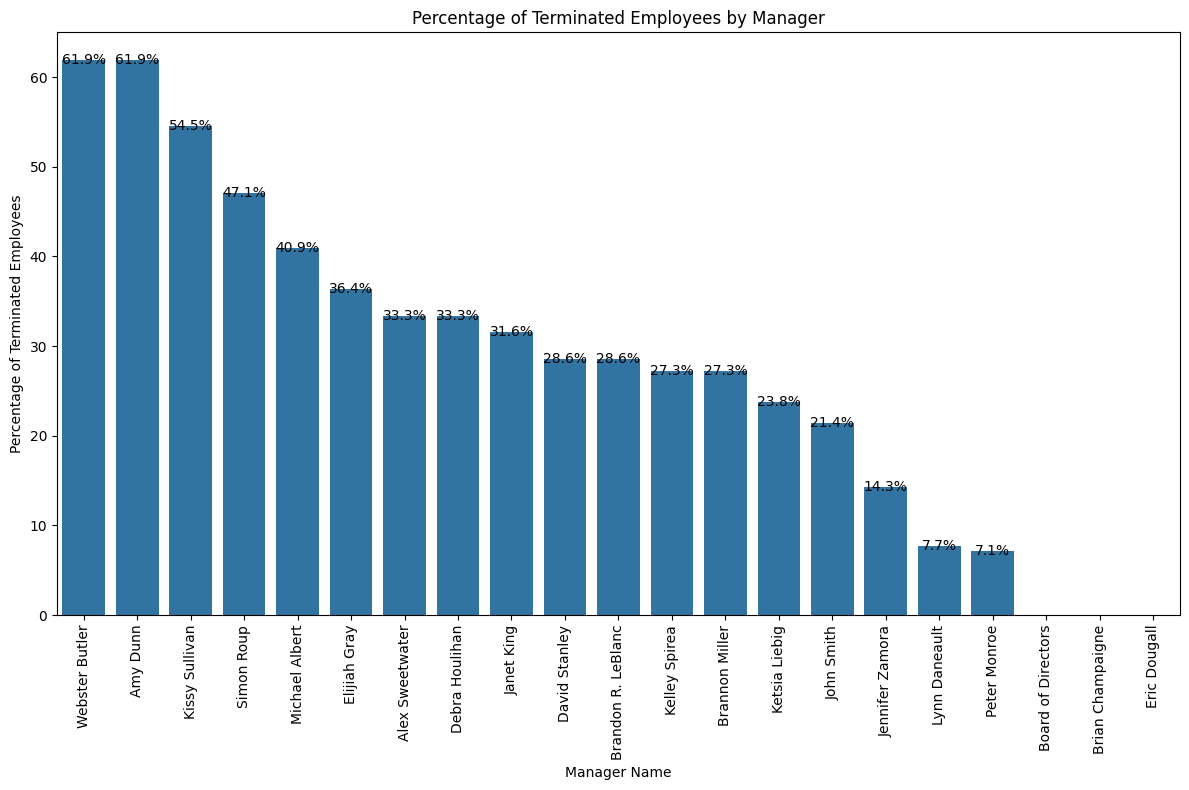
- A bar plot was used to calculate and display the termination percentage for each department, providing insights into departmental turnover rates.



Terminated Employees by Manager

- The analysis was extended to understand the count and percentage of terminated employees under each manager. This highlighted potential managerial issues contributing to employee turnover.

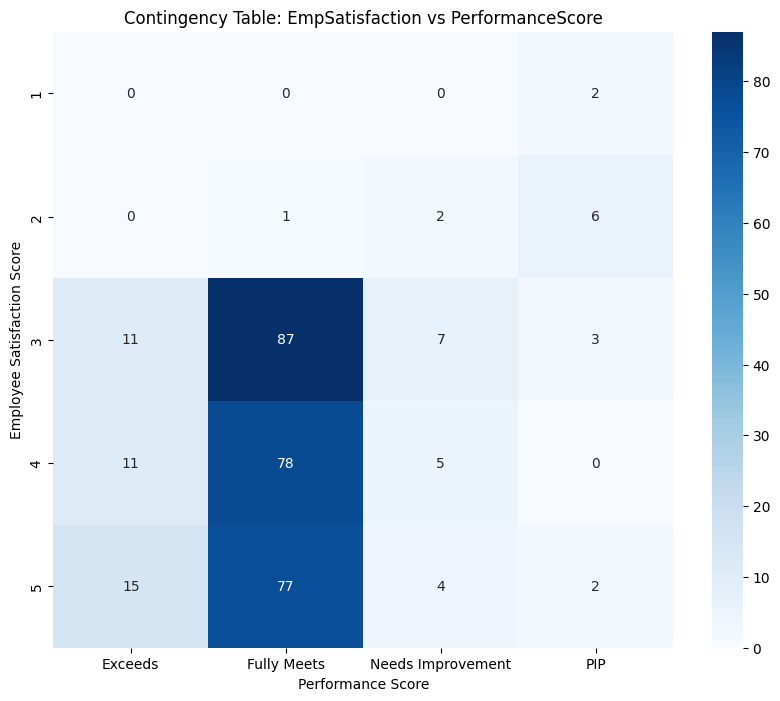




Employee Performance Analysis

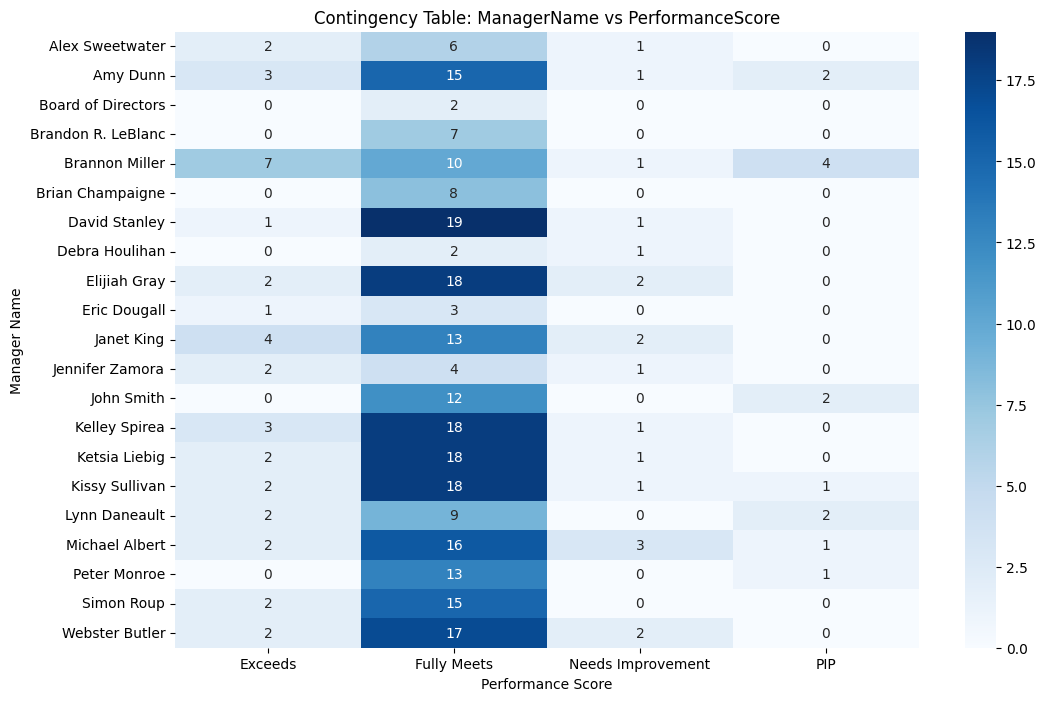
EmpSatisfaction vs Performance Score

- A contingency table and heatmap were used to examine the relationship between employee satisfaction and performance scores. A chi-squared test was conducted to determine if there was a statistically significant association between these variables.



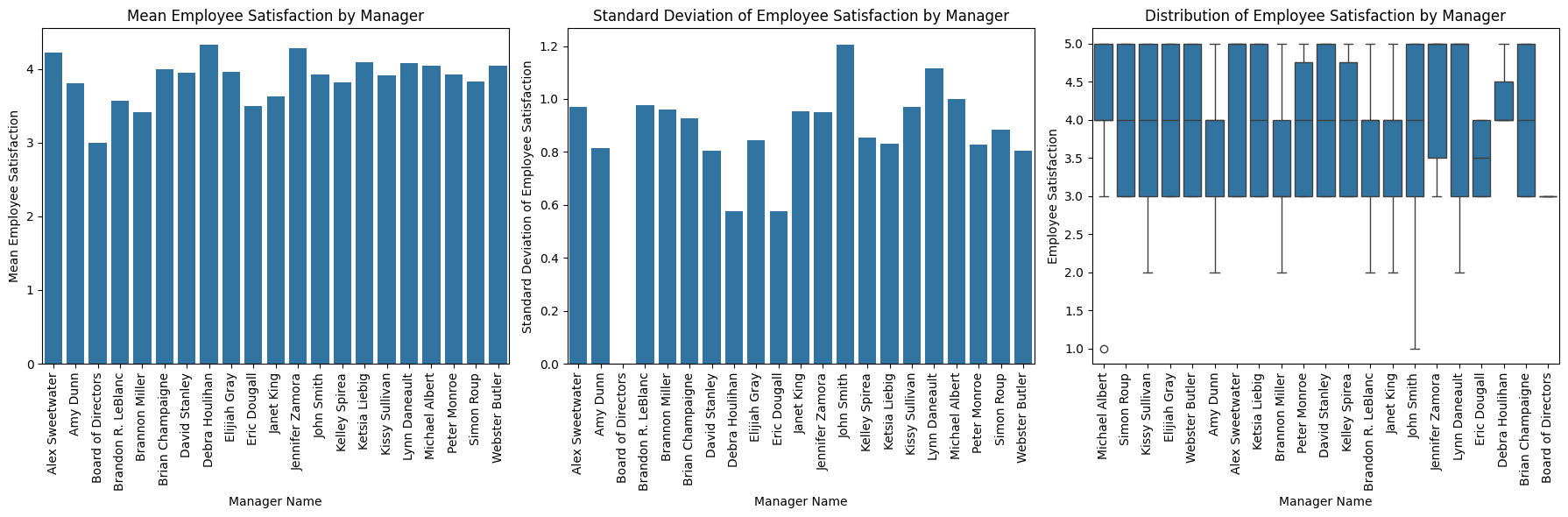
Manager Influence on Performance Score

- The relationship between the manager and employee performance was analyzed. A chi-squared test was also performed to check for a significant association.



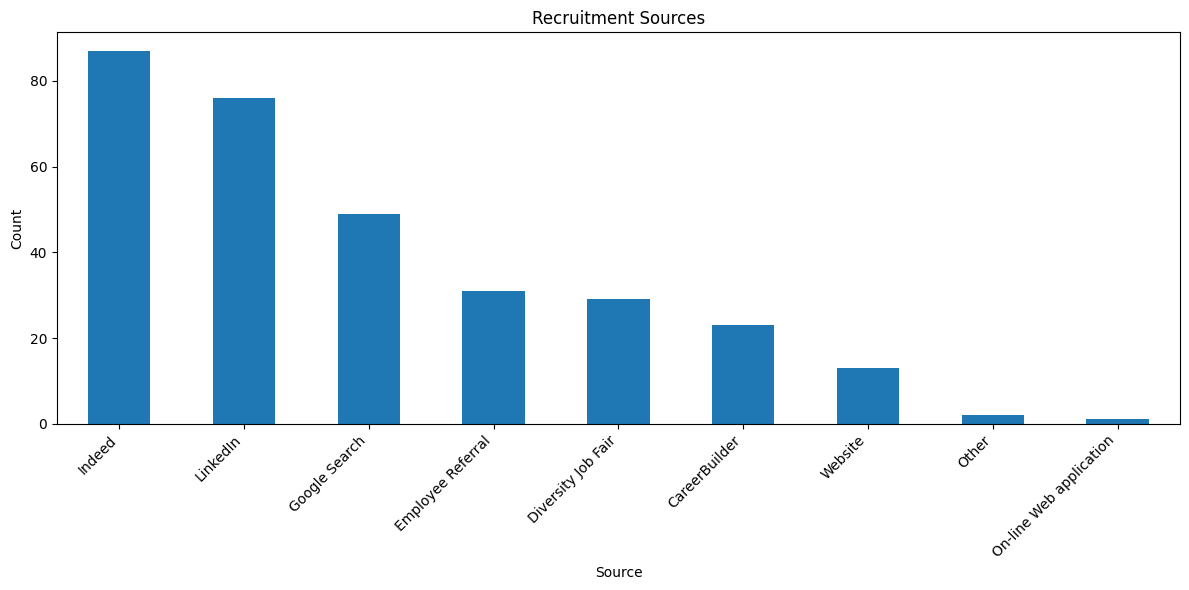
Employee Satisfaction Analysis

- Employee satisfaction scores were analyzed based on their managers. Bar plots and box plots were used to visualize the mean, standard deviation, and distribution of satisfaction scores for each manager.



Recruitment Source Analysis

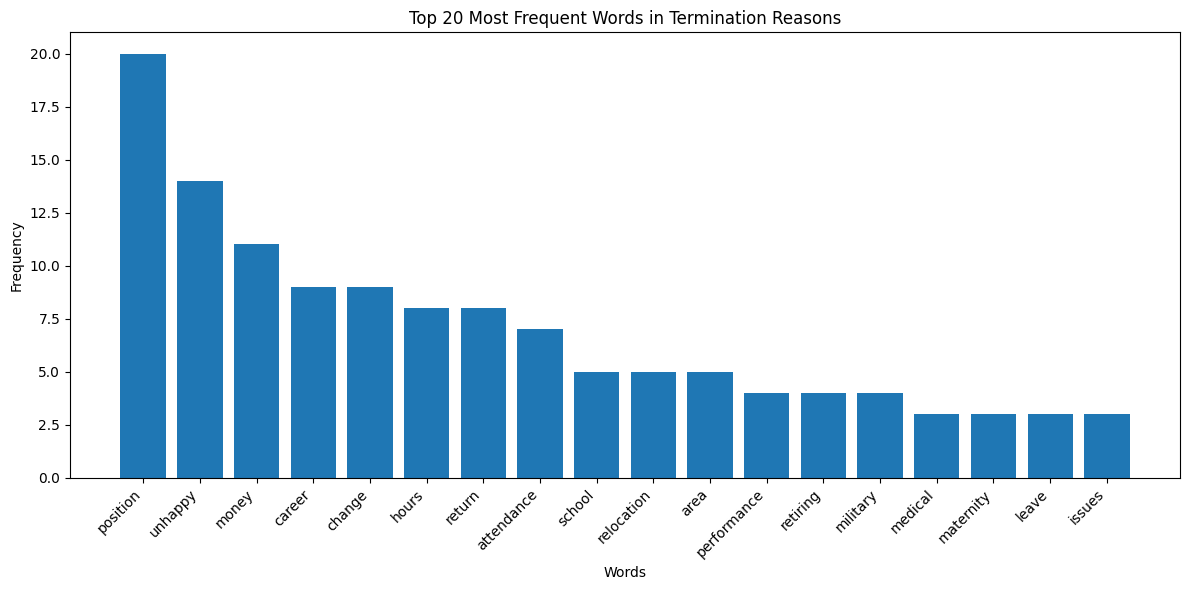
- The distribution of recruitment sources was analyzed to understand which sources were most effective in hiring employees. Both count and percentage distributions were displayed.



Termination Reasons Analysis

Text Processing:Text data related to termination reasons was processed to extract meaningful insights. This involved tokenization, stopword removal, and the creation of a document-term matrix.

Latent Dirichlet Allocation (LDA):LDA was applied to identify underlying topics in termination reasons. The top words for each topic were identified, and the most frequent words in termination reasons were visualized.



Feature Engineering

- A new feature, 'Age', was created from the 'DOB' column. The analysis uncovered issues with future dates, which were corrected by assuming they were meant to be 100 years earlier.

Conclusion

The analysis provided several key insights into the HR dataset:

Salary Discrepancies: Potential salary disparities were identified across departments, positions, and genders.

Employee Turnover: The analysis highlighted departments and managers with higher turnover rates.

Performance and Satisfaction: The relationship between employee satisfaction and performance was explored, with significant associations found in some cases.

Recruitment and Termination:The effectiveness of recruitment sources and common reasons for termination were analyzed, providing actionable insights for HR policies.

This analysis offers a comprehensive view of the HR dataset, helping to inform strategic decisions related to employee management and organizational improvement.