DATA ANALYSIS EXPLORATION

SHAI BOOTCAMP TASK REPORT

MARYAM JAMAL

INTRODUCTION

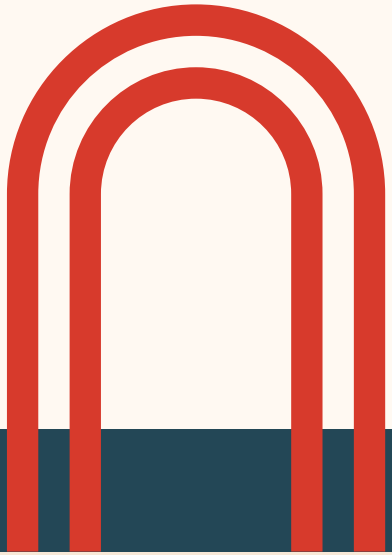
In the world of artificial intelligence (AI) and machine learning (ML), where data is crucial for making informed decisions, this report dives into the details of employee information within an organizational setting.

Throughout the ShAI Bootcamp internship, this report outlines the steps taken to carefully examine and interpret the provided employee dataset. Using a mix of advanced statistical methods and machine learning techniques, the aim is to uncover patterns, reveal insights, and derive actionable intelligence.

TASKS EXPLANATIONS AND RESULTS

In the initial phase of **data exploration**, the dataset was loaded, and a thorough examination of its key features commenced using the 'info' method. The insights obtained revealed that the dataset comprises 148,654 rows and 13 columns, encompassing a mix of integer, floating-point, and object data types. This comprehensive overview also facilitated the identification of non-null values in each column, enabling the assessment of missing values. Notably, two specific columns, namely 'Notes' and 'Status,' were observed to lack any recorded values.

Within the **Descriptive Statistics** task, an analysis focused on the 'Salary' (TotalPay) column unfolded. Employing the 'describe' method, fundamental statistical measures were derived. The mean was computed at 74,768.32, with a median value of 71,426.61. The range, determined by the disparity between the maximum and minimum values, stood at 568,213.56. Notably, the minimum salary was -618.13, while the maximum reached 567,595.43. The standard deviation was calculated at 50,517.01, offering insights into the dispersion of the data. Additionally, employing the 'mode' method, the mode of the 'Salary' column was identified as 0.



Within the **data cleaning** task, the focus turned to addressing missing values in specific columns. 'OvertimePay' and 'OtherPay,' both containing only four instances of missing data, were handled by dropping the corresponding rows for data integrity.

For the 'BasePay' column with 604 missing values, a non-normal, positively skewed distribution guided the decision to impute using the mode, the most frequently occurring value. Simultaneously, the 'Benefit' column, displaying a normally distributed histogram, saw missing values imputed with the median, chosen for its resilience to extreme outliers.

Moreover, recognizing that both 'Notes' and 'Status' columns contained no values, they were promptly removed to streamline the dataset and focus on pertinent information.

In the **data visualization** task, a comprehensive exploration of employee-related insights was pursued. Initially, a histogram depicting the salary distribution was crafted to provide a visual representation of the diverse salary ranges within the dataset.
Furthering the exploration, a pie chart was employed to delineate the proportion of employees across different departments. Recognizing that department names were embedded within the 'JobTitle' column, a meticulous process ensued. Utilizing the 'extract' method, department names were extracted from job titles enclosed in parentheses. Subsequently, a custom function was employed to address inconsistencies arising from variations in letter capitalization.
Upon successfully extracting and standardizing department names, the resulting proportions were effectively visualized through the pie chart. Given that a considerable portion of records lacked departmental information, a strategic decision was made to streamline the dataset by dropping the 'JobTitle' column upon task completion.

In the grouped analysis task, a systematic exploration of salary variations across different years was undertaken. Employing the 'groupBy' method, the dataset was segmented based on the 'year' attribute. Subsequently, aggregation statistics were derived using the 'agg' method to capture the average salary for each year.

The ensuing step involved visualizing the mean salary differences over the years. This was achieved by leveraging appropriate plotting techniques to provide a clear representation of how average salaries evolved annually. This grouped analysis offers valuable insights into the dynamic trends of salary changes throughout the dataset's temporal scop

In the conclusive task of correlation analysis, the focus was on examining the relationship between 'Salary' and 'BasePay.' Through correlation calculations, a robust quantitative measure of their association was obtained. This correlation was then visually represented using a scatter plot.

The correlation findings indicate a highly positive linear relationship between an employee's base salary ('BasePay') and their total compensation ('TotalPay'). Specifically, employees with higher base salaries consistently exhibit higher total compensation. This relationship is nearly perfectly linear, underscoring the direct and strong connection between these two key variables. The scatter plot serves as a visual testament to this compelling correlation, providing a comprehensive understanding of the interplay between base salaries and total compensation.