Exploratory Data Analysis (EDA) Report – Payroll Dataset

- **1.Dataset overview:** The dataset contains 1470 rows and 35 columns, The dataset contains employee demographic, job, compensation, experience and satisfaction-related features.
- **2.Initial inspection:** Used .info(), .head(), .describe() to see data types, non-null counts, summary statistics.
- **3. Feature Descriptions & Types:** The dataset consists of a mix of numerical and categorical features that describe employee demographics, job characteristics, compensation, experience, and satisfaction levels.
- Employee Demographic Features
- The demographic features include Age, Gender, MaritalStatus, Education, EducationField, DistanceFromHome, and Over18. Among these, Age and DistanceFromHome are numerical variables, while Gender, MaritalStatus, Education, EducationField, and Over18 are categorical. These features provide insights into workforce composition and help assess the impact of personal characteristics on employee attrition and satisfaction.
- job-related features describe the employee's role and position within the organization. This category includes Department, JobRole, JobLevel, EmployeeNumber, EmployeeCount, and StandardHours. JobLevel, StandardHours, and EmployeeCount are numerical, whereas Department and JobRole are categorical. These features help in understanding organizational hierarchy, departmental structure, and role distribution.
- Compensation features represent employee earnings and financial benefits. These include MonthlyIncome, MonthlyRate, DailyRate, HourlyRate, PercentSalaryHike, and StockOptionLevel. All these variables are numerical. They are essential for payroll analysis, salary prediction, and identifying anomalies in compensation patterns.
- Workload-related features capture working patterns and job demands. This category consists of OverTime, BusinessTravel, and DistanceFromHome. DistanceFromHome is numerical, while OverTime and BusinessTravel are categorical. These features are useful for evaluating work-life balance, employee stress levels, and workload-related attrition.
- Experience and career progression features describe an employee's professional history. These include TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, NumCompaniesWorked, and TrainingTimesLastYear. All variables in this category are numerical. These features are critical for analyzing employee stability, growth opportunities, and promotion trends.
- Satisfaction and performance features measure employee engagement and effectiveness at work. These include JobSatisfaction, EnvironmentSatisfaction, RelationshipSatisfaction, WorkLifeBalance, JobInvolvement, and PerformanceRating. All these variables are numerical and ordinal in nature. They are strong predictors of employee retention and performance.
- The Attrition variable is categorical and represents whether an employee has left the organization (Yes or No) and Monthly income is the salary of an employee per month. This feature serves as the target variable for attrition prediction andsalary prediction.
- **4. Missing Values & Data Quality and duplicates:** No missing values or duplicate records were found, indicating high data quality.
- **5. Univariate Analysis& Bivariate analysis:** For numerical features such as Age, MonthlyIncome, DailyRate, HourlyRate, MonthlyRate, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, PercentSalaryHike, TrainingTimesLastYear, and DistanceFromHome, summary statistics including mean, median, standard deviation, variance, skewness, and kurtosis were analyzed. Histograms and kernel density plots were

used to visualize the distributions of these variables, while boxplots helped identify outliers and assess data spread. Several compensation- and experience-related variables, particularly MonthlyIncome and TotalWorkingYears, exhibited strong positive skewness, indicating the presence of higher-end outliers.

- For categorical and boolean features such as Attrition, Gender, Department, JobRole, BusinessTravel, OverTime, frequency counts and bar plots were examined to understand category distributions. Boolean variables including OverTime and Attrition were analyzed using proportions to determine the percentage of employees working overtime and those who left the organization. For multi-category variables like JobRole and Department, relative frequency analysis highlighted dominant roles and departments as well as less represented categories. This univariate analysis provided essential insights into feature distributions and helped identify patterns relevant for further bivariate analysis and predictive modeling.
- MonthlyIncome shows a right-skewed distribution, with higher salaries concentrated in senior roles.
- Salary increases with job level, total working years, and years at company.
- Employees working overtime and having poor work-life balance show higher attrition.
- low job satisfaction are strong indicators of employee turnover.
- Stock options and salary hikes positively impact employee retention.
- **Outlier Detection:** Outliers found in columns MonthlyIncome, TotalWorkingYears, YearsAtCompany
- **6. Recommendations for Data Preparation & Modeling:** Based on the exploratory analysis, numerical features should be scaled to handle differences in magnitude, and categorical variables should be encoded using appropriate techniques such as one-hot encoding. Skewed variables, particularly compensation and experience-related features, may benefit from transformation or robust scaling to reduce the impact of outliers.remove outliers with various method like IQR. Feature selection should prioritize workload, compensation, and satisfaction variables, as they show strong relevance to attrition.and feature engineering can be done. Finally, pipeline-based modeling is recommended to ensure consistent preprocessing and reproducibility during model training and evaluation.