HR Data Cleaning and Analysis Report

# 1. Introduction

This report presents the steps taken to clean and analyze an unstructured HR dataset, which includes employee information, job roles, departments, salaries, and performance evaluations. The primary objective was to clean and transform the dataset to ensure accuracy, consistency, and readiness for insightful analysis. Visualizations were created using Power BI to present key findings.

# 2. Objective

The goal of this project was to:  
- Clean an unstructured HR dataset.  
- Standardize data formats and remove inconsistencies.  
- Eliminate duplicates and handle missing data.  
- Create insightful visualizations using Power BI to support data-driven decision-making in HR.

# 3. Dataset Overview

The original dataset consisted of the following fields:  
- Employee ID  
- Full Name  
- Gender  
- Date of Birth  
- Date of Hire  
- Job Title  
- Department  
- Annual Salary  
- Employment Status  
- Performance Rating  
- Years at Company  
- Location  
- Email  
- Phone Number  
- Last Promotion Date  
- Supervisor  
  
There were 200 rows in the dataset, with various issues such as missing data, duplicated entries, inconsistent formatting, and outliers in fields like salary and performance ratings.

# 4. Data Cleaning Process

The data cleaning process was carried out through several steps:

## 4.1 Removing Duplicates

Duplicate rows based on the "Employee ID" were identified and removed to ensure each employee had only one unique record.

## 4.2 Handling Missing Data

Missing values in key fields like "Employee ID" were removed. Fields such as "Gender" and "Performance Rating" were filled with default values like 'U' (Unknown) and 'Average', respectively.

## 4.3 Standardizing Formats

Dates in the fields "Date of Birth," "Date of Hire," and "Last Promotion Date" were converted to the standardized format (YYYY-MM-DD). Fields like "Gender," "Job Title," and "Department" were standardized to fix inconsistent entries.

## 4.4 Correcting Inconsistent Data

Inconsistent data for fields like "Gender," "Department," and "Job Title" were standardized to ensure uniformity. For example, "M" and "Male" were both converted to "M," and variations of department names such as "HR" and "hr" were standardized.

## 4.5 Removing Outliers

Outliers in fields like salary (below $10,000 or above $200,000) and negative values in "Years at Company" were identified and removed.

## 4.6 Fixing Email and Phone Number Formats

Invalid email addresses (e.g., missing '@' symbols) and phone numbers with fewer than 10 digits were corrected or removed.

## 4.7 Updating Calculated Fields

The "Years at Company" field was recalculated based on the difference between the "Date of Hire" and the current year.

# 5. Analysis Using Power BI

After cleaning the dataset, Power BI was used to perform Exploratory Data Analysis (EDA) and visualize key insights. The visualizations helped reveal trends and patterns in employee data, offering actionable insights for HR decision-makers.

## 5.1 Visualizations Created in Power BI

The following visualizations were created in Power BI to illustrate key insights:  
1. Salary Distribution by Department  
2. Employee Gender Breakdown  
3. Performance Rating Distribution  
4. Years at Company  
5. Department-Wise Headcount  
6. Promotion Trends Over Time

## 5.2 Summary Statistics

Power BI also provided summary statistics for key metrics:  
- Average Salary: $65,000  
- Average Tenure: 5 years  
- Gender Distribution: 55% male, 45% female employees

# 6. Key Findings

From the cleaned dataset and Power BI visualizations, several insights were uncovered:  
- Salary Gaps: Disparity in salaries across departments, with Engineering and IT earning more than HR and Sales.  
- Gender Diversity: Gender imbalances in departments like Engineering (male-dominated) and HR (female-dominated).  
- Performance Ratings: Most employees rated 'Good' or 'Average', indicating room for improvement.  
- Turnover Rates: Majority of employees with tenure below 5 years, suggesting possible turnover issues.

# 7. Conclusion

The data cleaning process successfully transformed the unstructured HR dataset into a reliable and analyzable resource. The Power BI visualizations provided clear, actionable insights into key HR metrics such as salary trends, employee performance, and tenure. These insights can help the organization optimize its HR strategies and make data-driven decisions to improve employee satisfaction and retention.

# 8. Recommendations

Based on the findings from the cleaned dataset and Power BI analysis, the following recommendations are made:  
- Salary Adjustments: Review salary structures for underpaid departments like HR and Sales.  
- Performance Improvement Programs: Implement training programs for employees rated 'Average' or 'Needs Improvement'.  
- Diversity Initiatives: Promote gender diversity in male-dominated departments like Engineering.  
- Retention Strategies: Investigate high turnover in departments like Sales and develop retention strategies.