Project By: Pallavi Saitu
Course : TECH 07 Stanford Continuing Studies
Submission Date: 5/31/2025

IBM Employee Performance & Attrition

Introduction

The focus of this project was to explore the IBM employee dataset, understand the patterns, and check the drivers behind employee attrition. I ran exploratory data analysis(EDA) and gained some insights that could potentially help enhance employee retention strategies.

Analysis Tools & Techniques

- Python Notebook (GoogleColab)
    - Pandas, Numpy for data handling
    - Seaborn, Matplotlib for data visualization

Dataset Methodology

This project uses the IBM HR Employee Attrition dataset, which contains structured information about employees, capturing personal details, job roles, compensation, satisfaction levels, and whether or not they left the organization. The dataset includes 1,470 rows (each representing an individual employee) and 35 columns of features that span across various attributes such as job involvement, work-life balance, income, experience, and performance. This dataset includes both categorical (e.g., JobRole, Department, MaritalStatus) and numerical (e.g., MonthlyIncome, TotalWorkingYears, YearsWithCurrManager) data.Some of the key data fields used in the analysis are:

- Attrition is the target variable indicating whether an employee left (Yes/No)
- MonthlyIncome, TotalWorkingYears, OverTime, YearsWithCurrManager, JobSatisfaction, and PerformanceRating are some other variables in the dataset

Research Question

What are the key factors that influence employee attrition within an organization, and how can these insights be used to support retention strategies?

The analysis is aimed to understand the roles of compensation, managerial relationships, work-life balance, and performance in predicting whether an employee is likely to leave. Furthermore, by identifying strong patterns and correlations between employee attributes and attrition outcomes, the findings can inform HR decisions such as offering competitive salaries, managing workload, and fostering stable management-employee relationships. These insights could ultimately support data-driven retention strategies, reduce turnover costs, and improve overall employee satisfaction.

Exploratory Data Analysis (EDA)

EDA was the central component of this project. I explored the dataset's structure, cleaned and preprocessed the data, and performed univariate and bivariate analyses.

Key Steps and Observations:

Data Inspection: Identified 35 columns including both numerical and categorical variables. No missing values were detected.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

See Figure 1:Summary Statistics Table and Data Structure Overview.

Attrition Distribution: Found that approximately 16% of employees had left the company, highlighting a class imbalance.

```
# Attrition count
sns.countplot(data=df, x='Attrition')
plt.title('Attrition Count')
plt.show()
```

See Figure 2 :Attrition Distribution Bar Chart.

Categorical Analysis: Bar plots were used to compare attrition across job roles and departments. This revealed that certain roles, like Sales Representative, had notably higher attrition rates.

```
# Attrition by Job Role
plt.figure(figsize=(10,6))
sns.countplot(data=df, y='JobRole', hue='Attrition')
plt.title('Attrition by Job Role')
plt.show()
```

See Figure 3 : Attrition by Job Role Chart.

Numerical Trends: Boxplots were used to examine how numerical variables like MonthlyIncome differ between employees who left and those who stayed. This revealed that employees with lower incomes were more likely to leave the organization.

```
sns.boxplot(data=df, x='Attrition', y='MonthlyIncome')
plt.title('Monthly Income vs Attrition')
plt.show()
```

See Figure 4 : Boxplot of Monthly Income by Attrition.

Performance vs Attrition Analysis:

To investigate whether performance rating correlates with attrition.This view helps assess if employees with higher or lower performance ratings are more or less likely to leave the company.

```
sns.countplot(data=df, x='PerformanceRating', hue='Attrition')
plt.title('Attrition by Performance Rating')
plt.show()
```

See Figure 5: Attrition by Performance Rating.

Correlation Matrix:

A heatmap was used to visualize correlations between numeric features to identify strong relationships and potential redundancies. The results showed a strong correlation between Age and TotalWorkingYears (0.66), and between YearsWithCurrManager and YearsInCurrentRole (0.76), highlighting how experience and managerial stability may relate to attrition.

```
numeric_df = df.select_dtypes(include='number')
plt.figure(figsize=(12, 10))
sns.heatmap(numeric_df.corr(), annot=True, fmt='.2f', cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

See Figure 6 :Correlation Heatmap of Numeric Features.

Visualizations:

- Count Plots for Categorical Features (e.g., JobRole, Department)
  Takeaway: Sales Representative and Laboratory Technician roles showed higher attrition rates (See Figure 3).

- Boxplots and Histograms for Numerical data (e.g., MonthlyIncome, TotalWorkingYears)
  Takeaway: Lower MonthlyIncome and fewer working years were consistently linked to employees leaving the company (See Figure 4).
- Bar Plot for PerformanceRating
  Takeaway: Attrition rates were consistent across different performance levels, indicating this variable had limited impact on predicting attrition (See Figure 5).
- Heatmap for Correlation Analysis
  Takeaway: Strong positive correlation was observed between Age and TotalWorkingYears (0.66), and YearsWithCurrManager and YearsInCurrentRole (0.76), emphasizing the effect of employee experience and managerial stability (See Figure 6).

Key Insights

Employees with lower monthly incomes were most likely to leave, highlighting the role of fair compensation in retention. Frequent overtime was another major factor, indicating that high workload may lead to burnout and attrition.

Short managerial relationships and low job satisfaction further contributed to employee turnover, signaling a need for leadership stability. Sales Representatives experienced the highest attrition among job roles, pointing to possible role-specific challenges.

Interestingly, performance rating had little effect, suggesting that even high performers are at risk of leaving if other key needs aren't addressed.

Conclusion

The analysis identified several key drivers of employee attrition that are highly relevant to HR strategy. Compensation emerged as a major factor, along with the effects of workload (e.g., overtime) and the importance of consistent managerial relationships. These findings provide a foundation for predictive modeling and can guide targeted, data-informed retention strategies.

****Appendix****

Figure 1: Summary Statistics Table and Data Structure Overview

```
# # Basic statistics
df.describe(include='all')
```

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	RelationshipSat
count	1470.000000	1470	1470	1470.000000	1470	1470.000000	1470.000000	1470	1470.0	1470.000000	...	1
unique	NaN	2	3	NaN	3	NaN	NaN	6	NaN	NaN	...	
top	NaN	No	Travel_Rarely	NaN	Research & Development	NaN	NaN	Life Sciences	NaN	NaN	...	
freq	NaN	1233	1043	NaN	961	NaN	NaN	606	NaN	NaN	...	
mean	36.923810	NaN	NaN	802.485714	NaN	9.192517	2.912925	NaN	1.0	1024.865306	...	
std	9.135373	NaN	NaN	403.509100	NaN	8.106864	1.024165	NaN	0.0	602.024335	...	
min	18.000000	NaN	NaN	102.000000	NaN	1.000000	1.000000	NaN	1.0	1.000000	...	
25%	30.000000	NaN	NaN	465.000000	NaN	2.000000	2.000000	NaN	1.0	491.250000	...	
50%	36.000000	NaN	NaN	802.000000	NaN	7.000000	3.000000	NaN	1.0	1020.500000	...	
75%	43.000000	NaN	NaN	1157.000000	NaN	14.000000	4.000000	NaN	1.0	1555.750000	...	
max	60.000000	NaN	NaN	1499.000000	NaN	29.000000	5.000000	NaN	1.0	2068.000000	...	

11 rows × 35 columns

Figure 2 : Attrition Distribution Bar Chart.

```
# Attrition count
sns.countplot(data=df, x='Attrition')
plt.title('Attrition Count')
plt.show()
```

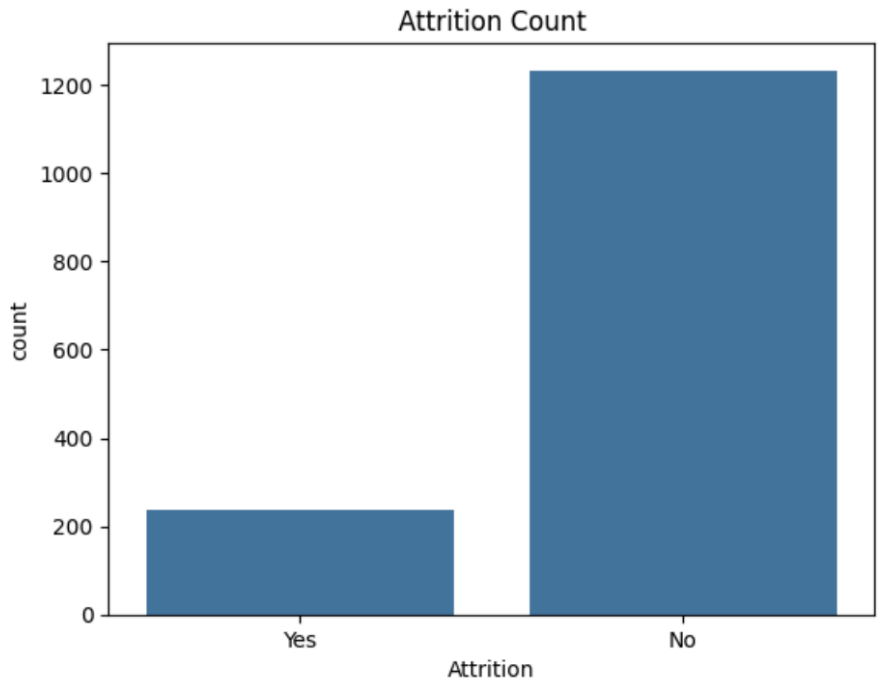

Figure 3 : Attrition by Job Role Chart.

```
# Attrition by Job Role
plt.figure(figsize=(10,6))
sns.countplot(data=df, y='JobRole', hue='Attrition')
plt.title('Attrition by Job Role')
plt.show()
```

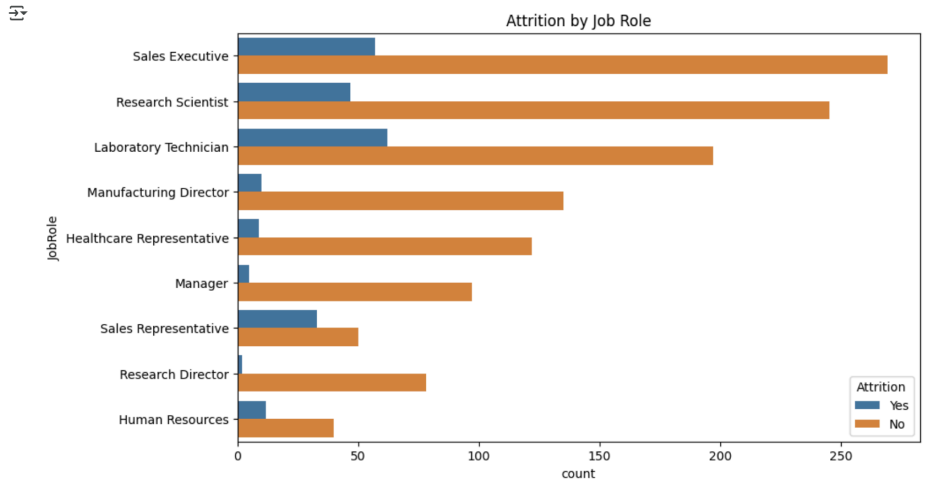

Figure 4 : Boxplot of Monthly Income by Attrition.

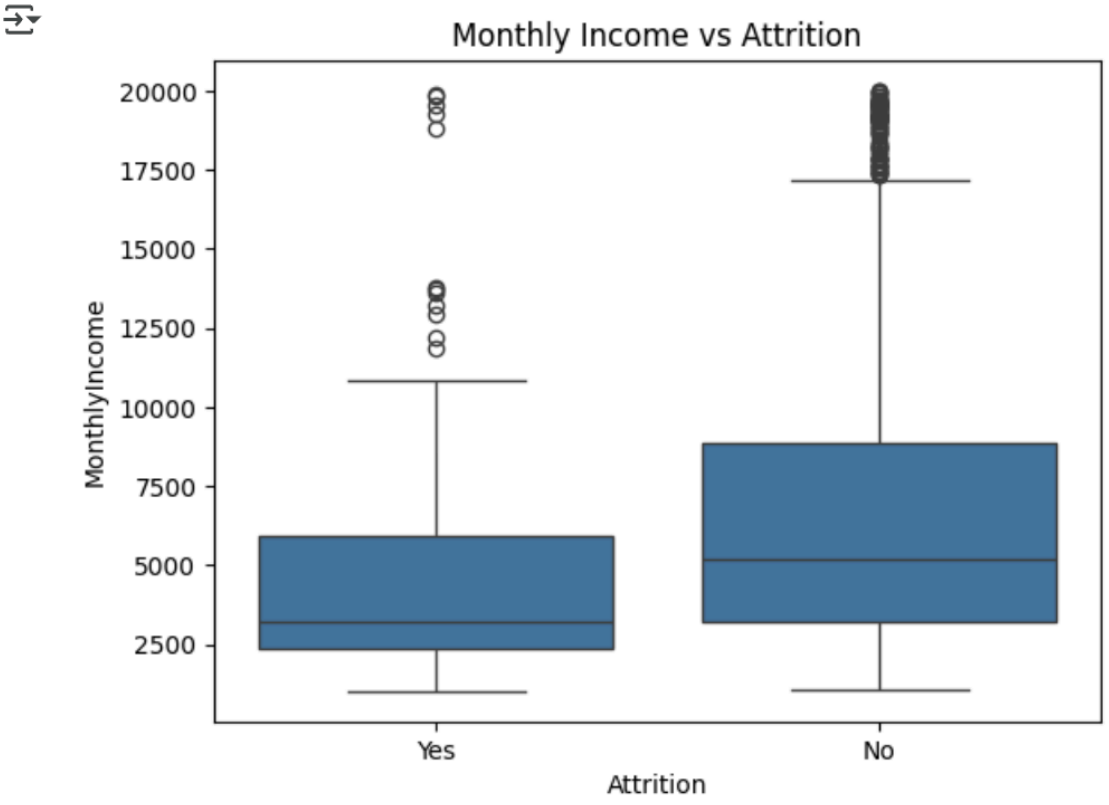

Figure 5 :Attrition by Performance Rating.

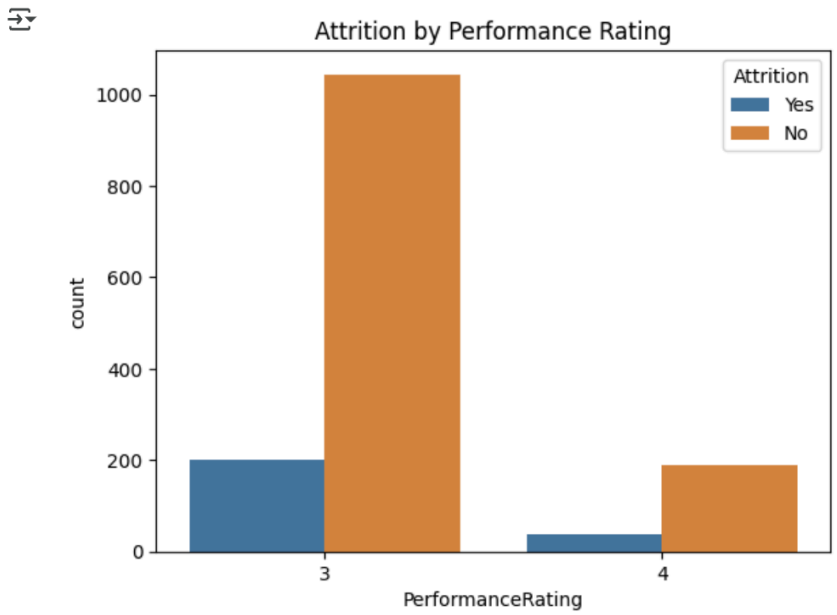

Figure 6 Correlation Heatmap.