VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

EDA on Datasets of IBM Watson Employees

(SDG – Goal 8)

FINAL REVIEW

SUBMITTED BY

TEAM – G

HEMA B – 23MIA1101

PRIYANKA S – 23MIA1032

KAVYA N – 23MIA1125

A FINAL REVIEW REPORT SUBMITTED TO

Prof. Dr. ASNATH VICTY PHAMILA Y – SCOPE

IN PARTIAL FULLFILMENT OF THE REQUIREMENTS FOR THE COURSE OF

CSE3040 – EXPLORATORY DATAANALYSIS

IN MIA (M.Tech Integrated CSE with Specialization in Business Analytics)

S.No	Section Title
1	Introduction
2	Objectives
3	Dataset Overview
4	Data Preprocessing
5	Exploratory Data Analysis
6	Outlier Detection and Clustering
7	Feature Selection and Dimensionality Reduction
8	Regression Modeling
9	Frequent Pattern Mining
10	Conclusion and Future Scope

1. Introduction

The primary focus of this project is to apply comprehensive Exploratory Data Analysis (EDA) on an organizational dataset from IBM Watson, with the goal of identifying patterns and insights that align with the objectives of **Sustainable Development Goal (SDG) 8 – Decent Work and Economic Growth**. In modern organizational settings, data-driven approaches are increasingly vital for making informed decisions, particularly in the area of human resource (HR) management. Employee attrition, satisfaction, and career progression are crucial indicators of workplace stability and economic growth. Hence, this EDA project aims to explore how various employee attributes influence such outcomes.

The analysis emphasizes uncovering both visible and hidden relationships among employee characteristics, such as demographics, education, job roles, performance indicators, and compensation. The project not only employs traditional EDA methods like visualizations and descriptive statistics but also integrates more advanced techniques like feature selection, outlier detection, clustering, and association rule mining. This multi-layered approach offers an in-depth understanding of organizational dynamics and provides a foundation for building predictive models in the future. Additionally, this report offers structured documentation that mirrors industry standards, enabling replication and scalability in real-world HR analytics systems.

2. Objectives

The overarching goal of this study is to leverage EDA techniques to reveal actionable insights from structured HR data. Specifically, the project aims to dissect various employee attributes to determine what influences outcomes like attrition, job satisfaction, salary, and promotion trends. A thorough inspection of data integrity, feature behavior, and correlation is also conducted to ensure model readiness. Another objective is to demonstrate a comprehensive data pipeline—beginning from data acquisition and preprocessing, all the way to pattern discovery using unsupervised learning. Furthermore, the project aims to build a simple regression model to explore how demographic attributes such as age impact financial indicators like monthly income. This exercise serves to understand the limitations of linear predictors in compensation analysis. Additionally, frequent pattern mining is employed to detect latent categorical associations among employee roles and departments, offering unique insights into structural dependencies. Overall, the EDA process is designed to simulate the first phase of an industrial machine learning pipeline, providing both insight generation and modeling readiness.

3. Dataset Overview

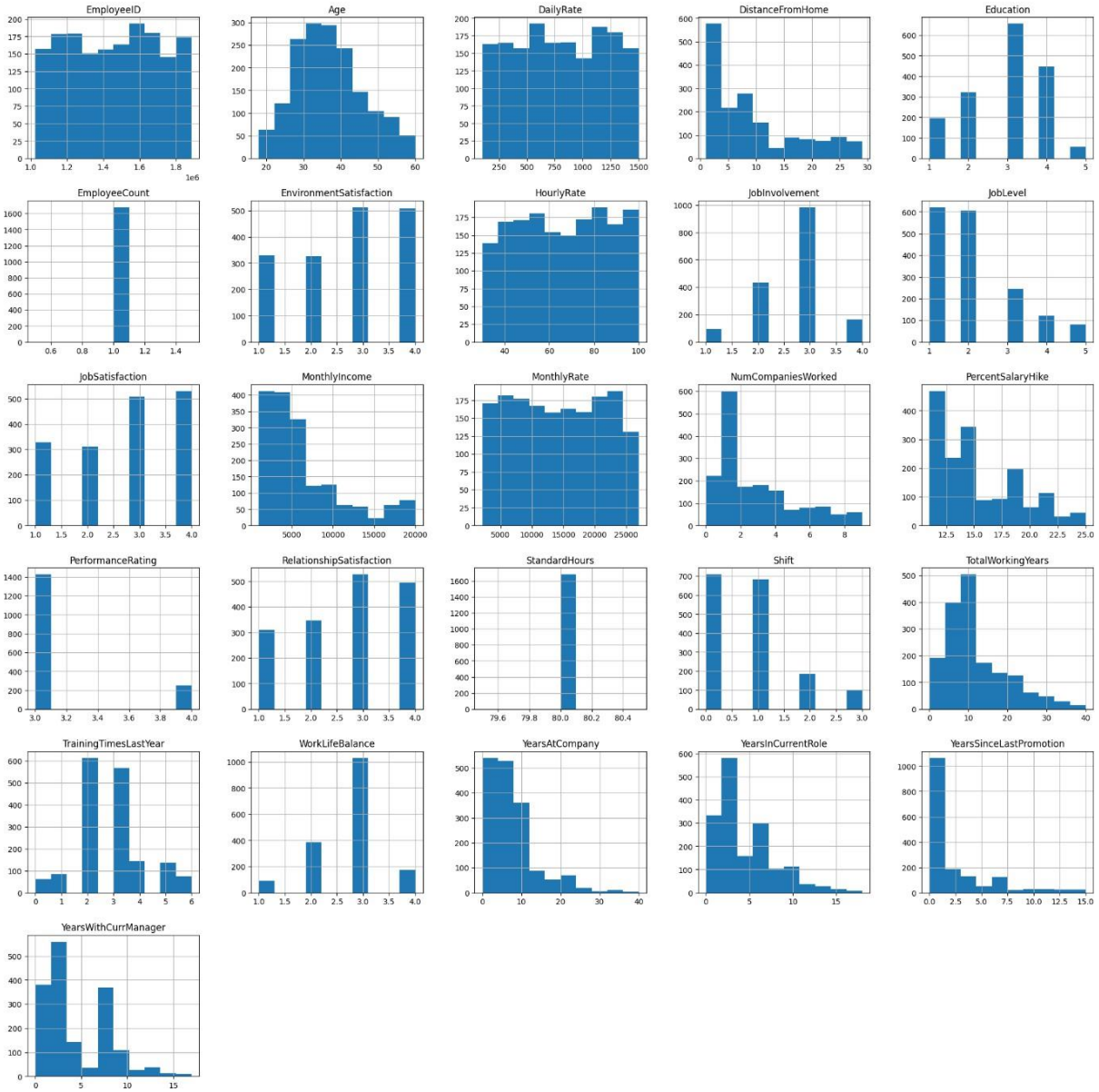
The dataset used in this project consists of **1676 rows and 35 columns**, each representing unique records of employees from IBM Watson. The features cover a wide range of variables categorized into demographic (e.g., Age, Gender, MaritalStatus), professional (e.g., JobRole, Department, YearsAtCompany), compensation-related (e.g., MonthlyIncome, HourlyRate, PercentSalaryHike), and organizational indicators (e.g., PerformanceRating, Attrition, WorkLifeBalance). There are no missing values or duplicate records in the dataset, which simplifies preprocessing and ensures a high level of data quality and consistency.

Additionally, the dataset contains a mix of categorical and numerical variables, offering a rich landscape for various EDA techniques. Some variables like StandardHours and EmployeeCount were observed to be constants, providing no informational value. Others like MonthlyIncome and YearsSinceLastPromotion displayed significant variance and were later found to be impactful in further analyses. The well-structured nature of this dataset allows for the application of both basic and advanced statistical methods, supporting a deep exploration of internal company dynamics.

4.Data Preprocessing

The preprocessing phase of this study involved validating the data quality, classifying features, and applying various imputation strategies to simulate a robust pipeline. Although no missing values were detected, **three imputation methods** were applied—mean/mode imputation, **MICE (Multiple Imputation by Chained Equations)**, and **K-Nearest Neighbor Imputation (KNN)**. These methods were used not out of necessity, but to prepare a generalizable framework that can handle datasets with real-world imperfections. Categorical and numerical columns were clearly separated to facilitate tailored preprocessing techniques. Categorical variables were transformed using label encoding, while numerical features were scaled using standardization methods to ensure uniform contribution during modeling and visualization. Descriptive statistics such as mean, median, mode, standard deviation, and percentiles were computed to understand the central tendencies and variability across attributes. The analysis also checked for unique values, confirming which columns were constant and thus irrelevant for further exploration. These preprocessing steps not only helped in cleaning and standardizing the dataset but also laid the groundwork for subsequent clustering and feature selection techniques.

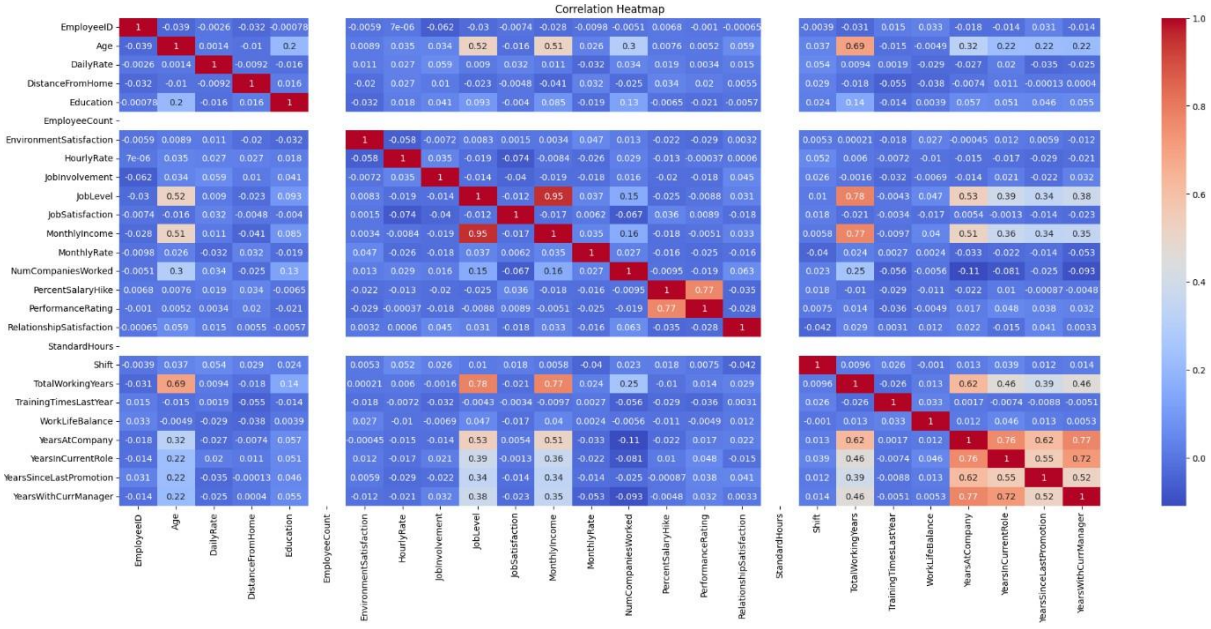
Histograms of Numerical Features



5.Exploratory Data Analysis

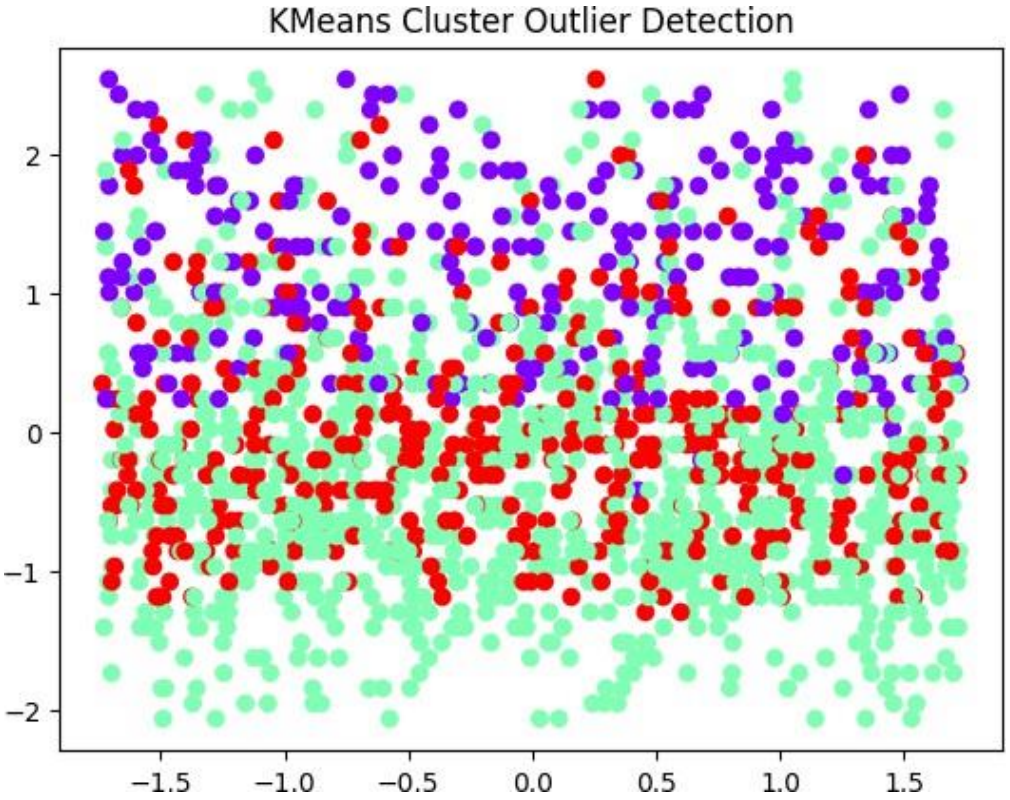
The EDA segment of the project combined descriptive statistics with visual analytics to derive insights into the dataset's structure and relationships. **Histograms** of numerical attributes revealed distributions that were mostly skewed to the right, indicating a concentration of values in the lower ranges. For example, MonthlyIncome and YearsAtCompany showed heavy right skewness, suggesting that most employees earn lower salaries and have relatively short tenures, which is typical in large organizations.

Boxplots were used extensively to detect outliers and understand the spread of data. Notable outliers were found in YearsSinceLastPromotion, TrainingTimesLastYear, and MonthlyIncome. These variations indicate potential inequality in promotion opportunities and compensation, which could be valuable feedback for HR departments. A **heatmap of Pearson correlation coefficients** was constructed to observe relationships between numeric variables. Strong positive correlations were identified between JobLevel and MonthlyIncome, as well as between YearsWithCurrManager and YearsInCurrentRole, suggesting hierarchical alignment and employee-manager consistency. These correlations reinforce the idea that job level and experience are vital components of organizational success.



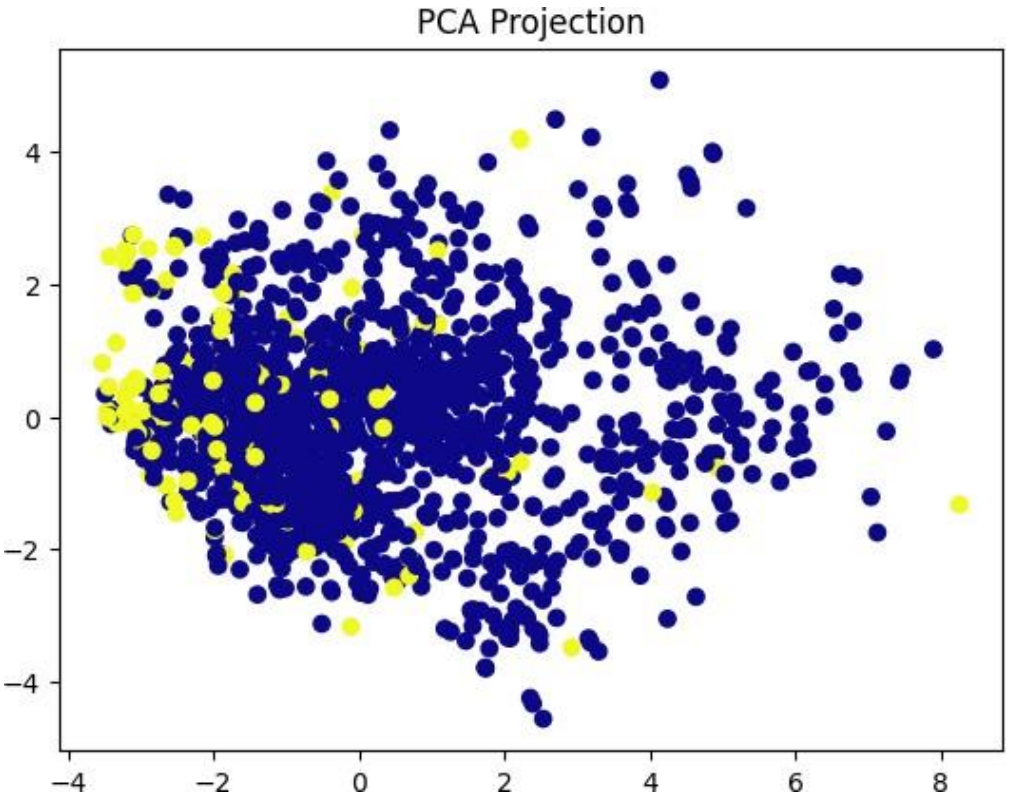
6.Outlier Detection and Clustering

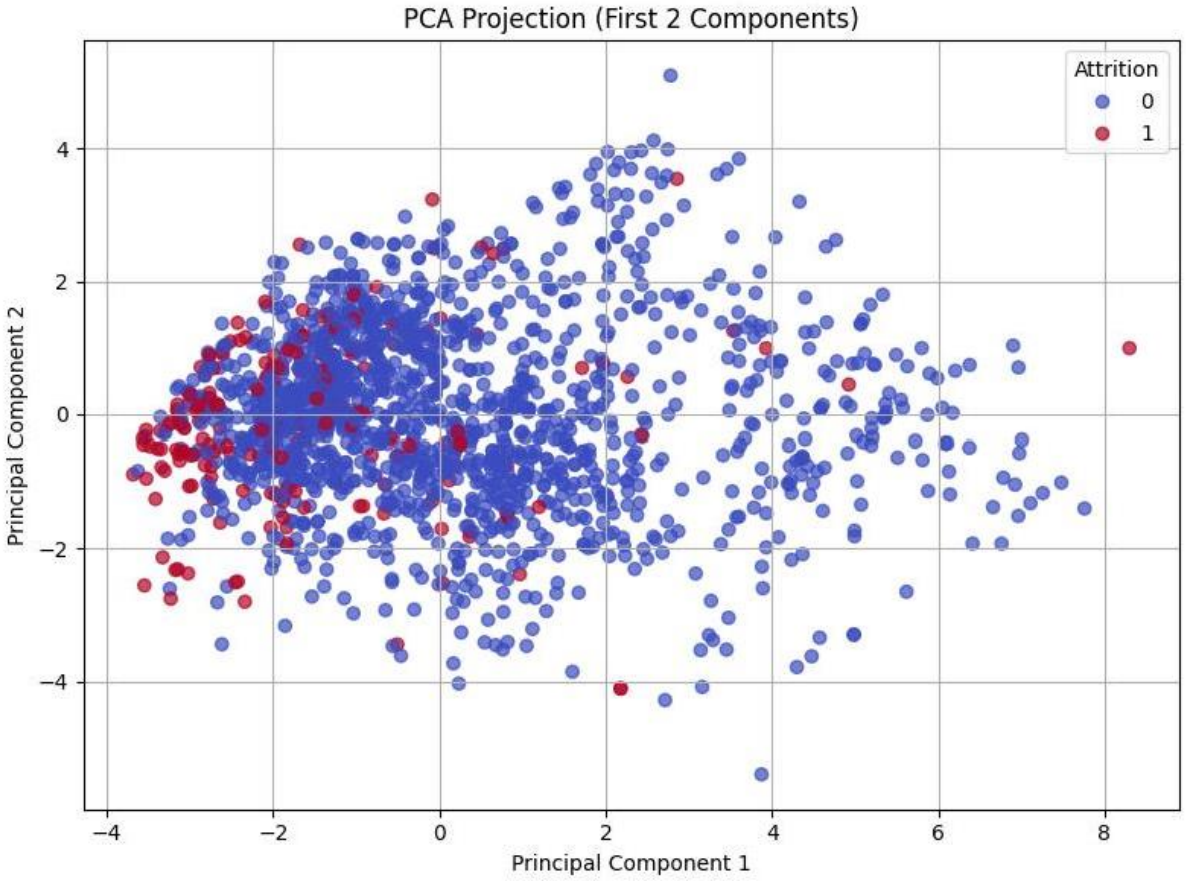
Outlier detection was conducted using both **Z-Score** and **Interquartile Range (IQR)** methods. The Z-score method identified few outliers—primarily in YearsAtCompany and TotalWorkingYears. In contrast, the IQR method revealed a larger number of outliers across multiple variables, including over **130 outliers in MonthlyIncome** and **over 250 in PerformanceRating**, suggesting irregularities in how performance is evaluated or rewarded.
To further identify structural anomalies and natural groupings, **KMeans clustering** was applied after standardizing the data. Three clusters were formed based on scaled numerical features. These clusters showed visible separations in employee experience and compensation levels. Visualization of the clusters indicated that some data points were distant from all cluster centroids, possibly representing organizational outliers or special employee categories such as executives or newly joined interns. The clustering results provide a segmentation perspective that HR managers can use to develop targeted policies for specific employee groups.

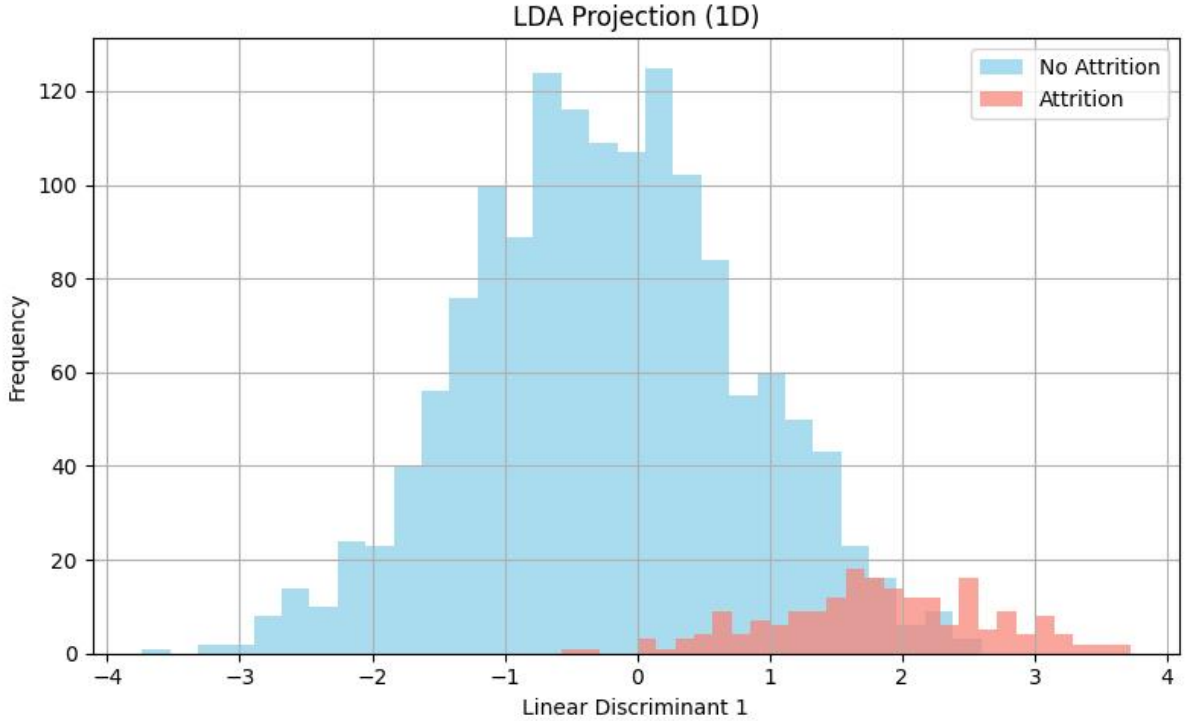


7.Feature Selection and Dimensionality Reduction

Multiple feature selection techniques were applied to identify variables most predictive of attrition. **ANOVA F-tests** revealed that Age, TotalWorkingYears, JobLevel, and YearsAtCompany were statistically significant in distinguishing between employees who stay versus those who leave. These findings were corroborated by **Mutual Information scores**, where MonthlyIncome and YearsWithCurrManager also ranked highly.
In addition, **Recursive Feature Elimination (RFE)** selected Age, DistanceFromHome, JobInvolvement, Shift, and YearsInCurrentRole as the top five predictors of attrition. This combination of statistical and model-based selection provided a reliable set of features for future modeling efforts. For visual interpretability, **Principal Component Analysis (PCA)** and **Linear Discriminant Analysis (LDA)** were employed. PCA showed limited class separability, while LDA demonstrated clear class division based on PerformanceRating, validating the effectiveness of supervised dimensionality reduction in class distinction.




PCA Projection (First 2 Components)
Attrition
0
1
Principal Component 2
Principal Component 1
4
2
0
−2
−4
−4
−2
0
2
4
6
8

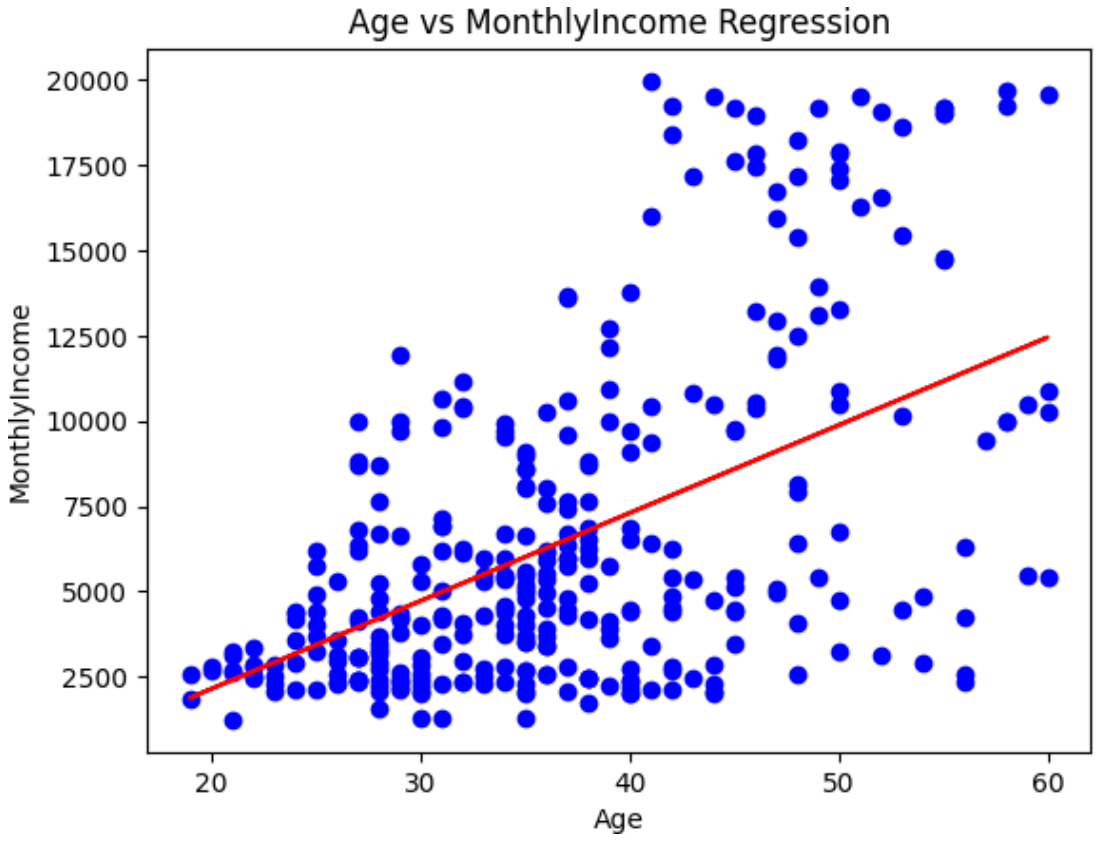

LDA Projection (1D)
No Attrition
Attrition
Frequency
Linear Discriminant 1
120
100
80
60
40
20
0
−4
−3
−2
−1
0
1
2
3
4

```python
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA

lda = LDA(n_components=1)
X_lda = lda.fit_transform(X_scaled, y)

print("LDA reduced shape:", X_lda.shape)
print("First 5 LDA values:", X_lda[:5].flatten())
```

```
LDA reduced shape: (1676, 1)
First 5 LDA values: [ 2.07607521 -1.08639521  1.68610818  1.08080548  0.99330297]
```

8.Regression Modeling

To explore the direct relationship between Age and MonthlyIncome, a **simple linear regression model** was built. The results showed a positive trend, suggesting that older employees tend to earn more. However, the model produced a relatively low **R^2 score of 0.31**, indicating that Age alone explains only 31% of the variance in income. The regression plot, characterized by a high spread of data points, demonstrated that other factors—such as education, department, job level, and tenure—play significant roles in determining employee income. This insight highlights the limitations of univariate models and underscores the need for multivariate approaches when dealing with compensation modeling.



9.Frequent Pattern Mining

Using the **Apriori algorithm**, frequent itemsets and association rules were mined from categorical features like Department, JobRole, and MaritalStatus. The analysis revealed strong relationships such as the one between Marketing and Cardiology, and between Nursing and Cardiology, with **lift values greater than 3**. This indicates strong associations that go beyond random chance. These patterns suggest interdepartmental links that may influence team dynamics, career mobility, or collaborative structures. Such insights are valuable for organizational restructuring or HR strategy formulation.

10.Conclusion and Future Scope

This EDA project demonstrated the power of data science in understanding complex human resource phenomena. The results provided clear evidence that age, tenure, job level, and satisfaction scores are critical in understanding attrition and compensation trends. Outlier detection methods highlighted potential HR policy gaps, while clustering and feature selection helped segment the workforce for targeted intervention. Although the linear regression model was simplistic, it revealed the multifactorial nature of income prediction.

For future work, the project can be extended by implementing **supervised learning models** like logistic regression, decision trees, or random forests for attrition classification. **Cross-validation** and **hyperparameter tuning** can further improve model robustness. Integration with external datasets, such as regional economic indicators or employee satisfaction surveys, could provide a more holistic view. Finally, deploying advanced techniques like **time-series forecasting**, **XGBoost**, or **neural networks (LSTM, GRU)** would add scalability and predictive power to the analysis, transforming it from a diagnostic tool into a strategic decision-making engine for HR departments.