Employee Attrition Analysis

Milestone 1: Data Collection, Exploration, and Preprocessing

Executive Summary

This report presents the findings from the initial phase of the Employee Attrition Analysis project,
covering data collection, exploration, and preprocessing. The analysis was performed on the HR-Employee-Attrition
dataset containing 1,470 employee records with 35 features. Through comprehensive exploration and preprocessing,
we identified key patterns in employee attrition and prepared a clean dataset for predictive modeling.

Key Findings:

- Dataset contains 1,470 employees with 16.1% attrition rate (237 employees)
- No missing values or duplicate records detected - excellent data quality
- Identified and removed 4 redundant features (EmployeeCount, StandardHours, Over18, EmployeeNumber)
- Successfully transformed 8 ordinal categorical features into meaningful labels
- Final dataset: 26 numerical and 9 categorical features

1. Data Collection

1.1 Dataset Overview

Attribute	Details
Source	HR-Employee-Attrition.csv (Kaggle)
Format	CSV (Comma-Separated Values
Total Records	1,470 employees
Total Features	35 columns
Target Variable	Attrition (Yes/No)
Data Quality	High - No missing values or duplicates

1.2 Feature Categories

The dataset comprises multiple feature categories essential for attrition analysis:

- **Demographics:**
  - Age, Gender, MaritalStatus, Education, EducationField
- **Job Information:**
  - Department, JobRole, JobLevel, BusinessTravel
- **Compensation:**
  - MonthlyIncome, MonthlyRate, DailyRate, HourlyRate, PercentSalaryHike, StockOptionLevel
- **Satisfaction Metrics:**
  - JobSatisfaction, EnvironmentSatisfaction, RelationshipSatisfaction, WorkLifeBalance
- **Performance:**
  - PerformanceRating, JobInvolvement, TrainingTimesLastYear
- **Tenure:**
  - YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, TotalWorkingYears

2. Initial Data Exploration

2.1 Data Quality Assessment

Quality Metric	Result	Action Taken
Missing Values	0	None required
Duplicate Records	0	None required
Total Rows	1,470	All retained
Total Columns	35	4 removed

2.2 Target Variable Distribution

The target variable 'Attrition' shows class imbalance:

Attrition Status	Count	Percentage
No (Stayed)	1,233	83.9%
Yes (Left)	237	16.1%

Key Insight: The dataset shows a typical imbalanced classification problem with an 84:16 ratio. This will require appropriate handling during model training (e.g., SMOTE, class weights).

2.3 Data Type Distribution

Feature distribution by data type:

- Numerical Features: 26 columns (integers and floats)
- Categorical Features: 9 columns (object type)
- Binary Features: Attrition, Gender, OverTime

3. Data Preprocessing

3.1 Removal of Redundant Features

Four features were identified as redundant and removed:

Feature	Reason	Impact
EmployeeCount	All values = 1 (constant)	No variance, no predictive value
StandardHours	All values = 80 (constant)	No variance, no predictive value
Over18	All values = "Y" (constant)	Minimum age is 18, no variance
EmployeeNumber	Unique identifier	Not relevant for prediction

Result: Dataset reduced from 35 to 31 features, improving model efficiency without losing information.

3.2 Feature Engineering - Label Transformation

Eight ordinal categorical features were transformed from numeric codes to meaningful labels for better interpretability and analysis:

Feature	Original Range	Transformed Labels
Education	1-5	High School → Doctorate
EnvironmentSatisfaction	1-4	Low → Very High Satisfaction
JobInvolvement	1-4	Low → Very High Involvement
JobLevel	1-5	Entry → Executive Level
JobSatisfaction	1-4	Low → Very High Satisfaction
PerformanceRating	1-4	Poor → Outstanding Performance
RelationshipSatisfaction	1-4	Low → Very High Satisfaction
WorkLifeBalance	1-4	Poor → Excellent Balance

Benefits: Enhanced interpretability for stakeholders, easier visualization, and maintained ordinal relationships for modeling.

4. Exploratory Data Analysis (EDA)

4.1 Numerical Features Summary Statistics

Key statistical insights from numerical features:

- Average employee age: 37 years (range: 18-60)
- Average monthly income: $6,503 (range: $1,009-$19,999)
- Average tenure: 7 years at company, 11 years total working experience
- Average distance from home: 9 miles (median: 7 miles)
- Training times last year: average 2.8 sessions

4.2 Categorical Features Distribution

Analysis of categorical variable distributions revealed:

- **BusinessTravel:** Travel_Rarely (71%), Travel_Frequently (19%), Non-Travel (10%)
- **Department:** R&D (65%), Sales (30%), HR (4%)
- **Gender:** Male (60%), Female (40%)
- **MaritalStatus:** Married (46%), Single (32%), Divorced (22%)
- **OverTime:** No (72%), Yes (28%)
- **JobRole:** 9 unique roles, most common: Sales Executive (22%), Research Scientist (20%)

5. Key Patterns and Relationships

Initial exploration revealed several noteworthy patterns in the data:

5.1 Attrition-Related Observations

- Higher attrition observed among employees working overtime
- Entry-level and junior-level positions show higher turnover rates
- Single employees exhibit higher attrition compared to married employees
- Employees with frequent business travel show elevated attrition
- Lower job satisfaction and work-life balance scores correlate with leaving

5.2 Compensation and Tenure Patterns

- Strong positive correlation between job level and monthly income
- Years at company shows right-skewed distribution (many recent hires)
- Performance ratings show limited variance (most employees rated "Excellent")
- Stock option levels vary, with many employees at level 0 or 1

6. Data Quality and Validation

6.1 Validation Checks Performed

Check	Result	Status
Missing Values	None detected	✓ Pass
Duplicate Records	None detected	✓ Pass
Data Type Consistency	All appropriate	✓ Pass
Value Ranges	All within expected bounds	✓ Pass
Categorical Integrity	All valid categories	✓ Pass

6.2 Outlier Analysis

Outlier detection revealed:

- Age: No significant outliers (all values reasonable for workforce)
- Monthly Income: Some high earners (executives) - legitimate outliers retained
- Distance from Home: Max 29 miles - within reasonable commuting distance
- Years at Company: One employee with 40 years - legitimate long-tenure employee

Decision: All outliers retained as they represent legitimate business scenarios.

7. Conclusions and Next Steps

7.1 Summary

The data collection, exploration, and preprocessing phase has been successfully completed. The HR-Employee-Attrition dataset has been thoroughly examined, cleaned, and prepared for advanced analysis and modeling. The data quality is excellent with no missing values or duplicates, and meaningful transformations have
been applied to enhance interpretability.

.2 Recommendations for Milestone 2

- Conduct chi-square tests for categorical features (already initiated)
- Perform correlation analysis for numerical features
- Apply feature selection techniques (RFE, feature importance)
- Create derived features (e.g., salary-to-tenure ratio, promotion rate)
- Develop interactive visualizations and dashboards
- Address class imbalance in target variable