HR Employee Attrition Data Preparation Report

Executive Summary

This report documents the initial data preparation and exploratory analysis of an HR Employee Attrition dataset containing 1,470 employee records across 35 variables. The primary objective is to understand employee attrition patterns and prepare the data for predictive modeling.

- Key Findings:
    - Dataset has no missing values and no duplicate records
    - Attrition rate: 16.1% (237 out of 1,470 employees)
    - Dataset includes 26 numerical and 9 categorical features
    - Several columns were removed due to lack of variability or relevance

Dataset Overview

Basic Statistics

Metric	Value
Total Records	1,470
Total Features	35 (reduced to 31)
Missing Values	0
Duplicate Records	0
Numerical Columns	26
Categorical Columns	9

The dataset demonstrates excellent quality with complete records and no duplications, requiring minimal cleaning efforts.

Feature Analysis

Target Variable: Attrition

- No: 1,233 employees (83.9%)
- Yes: 237 employees (16.1%)

The dataset shows a class imbalance that will need to be addressed during modeling to prevent bias toward the majority class.

Categorical Features Summary

Demographics:

- Gender: Male (882), Female (588)
- Marital Status: Married (673), Single (470), Divorced (327)
- Age Range: 18-60 years (average: 37 years)

Work-Related:

- Department: Research & Development (961), Sales (446), HR (63)
- Job Roles: 9 distinct roles, led by Sales Executive (326) and Research Scientist (292)
- Business Travel: Travel Rarely (1,043), Travel Frequently (277), Non-Travel (150)

Education:

- Education Levels: Bachelor's Degree (572), Master's Degree (398), College Graduate (282)
- Education Fields: Life Sciences (606), Medical (464), Marketing (159)

Satisfaction Metrics:

All satisfaction measures (Environment, Job, Relationship) show relatively balanced distributions across 4 levels, suggesting reasonable workplace conditions.

Numerical Features Summary

Compensation:

- Monthly Income: $1,009 - $19,999 (median: $4,919)
- Hourly Rate: $30 - $100 (median: $66)
- Percent Salary Hike: 11% - 25% (average: 15.2%)

Experience:

- Total Working Years: 0 - 40 years (average: 11.3 years)
- Years at Company: 0 - 40 years (average: 7.0 years)
- Years in Current Role: 0 - 18 years (average: 4.2 years)

Work Engagement:

- Distance from Home: 1 - 29 miles (average: 9.2 miles)
- Overtime: Yes (416), No (1,054)
- Training Times Last Year: 0 - 6 (average: 2.8)

Data Transformations

1. Encoded Feature Conversion

Eight ordinal features were converted from numeric codes to meaningful categorical labels:

- Education: 1-5 scale → Descriptive labels (High School to Doctorate)
- Satisfaction Metrics: 1-4 scale → Low to Very High Satisfaction
- Job Involvement: 1-4 scale → Low to Very High Involvement
- Job Level: 1-5 scale → Entry to Executive Level
- Performance Rating: 1-4 scale → Poor to Outstanding Performance
- Work-Life Balance: 1-4 scale → Poor to Excellent Balance

This transformation enhances interpretability and prepares features for proper encoding during modeling.

2. Column Removal

Four columns were removed due to lack of analytical value:

Column	Reason for Removal
Over18	All employees are 18+; no variation
EmployeeCount	Constant value of 1 for all records
StandardHours	Constant value of 80 for all records
EmployeeNumber	Identifier with no predictive value

Key Observations

Potential Attrition Indicators

High Cardinality Features:

- Monthly Income (1,349 unique values)
- Monthly Rate (1,427 unique values)
- Daily Rate (886 unique values)
- Employee age spans 43 years

Notable Patterns:

1. Job Level Distribution: Majority are Entry (543) and Junior level (534), suggesting retention challenges at early career stages
2. Overtime: 28.3% work overtime, which may correlate with burnout
3. Distance from Home: Wide range (1-29 miles) could impact commute satisfaction
4. Stock Options: 75% have level 0 or 1, limited equity incentives

Data Preparation Status

Completed Steps

- ✓ Data loading and initial inspection
- ✓ Missing value analysis (none found)
- ✓ Duplicate detection (none found)
- ✓ Feature type identification
- ✓ Ordinal encoding conversion
- ✓ Unnecessary column removal
- ✓ Exploratory data analysis

Conclusion

The HR Attrition dataset is clean, complete, and well-structured for predictive modeling. The 16.1% attrition rate suggests a meaningful business problem worth addressing. Key preparation work has successfully transformed encoded features into interpretable categories and removed non-informative columns.

The dataset contains a rich mix of demographic, work-related, and satisfaction metrics that should provide strong predictive signals. The primary challenge will be addressing the class imbalance to ensure the model can effectively identify at-risk employees.

The dataset is now ready for the next phase of analysis, including feature engineering, visualization, and model development.