IBM HR Employee Attrition Analysis

Statistical Computing Project

Amaan Shaikh | Danish Ali Shaikh | Rajat Pandey

2025-12-29

1. Introduction

This document provides a short documentation of the IBM HR Employee Attrition dataset used for the Statistical Computing group project. Instead of just a script to load the data we thought of sharing a summary so that it is easier for you to assess the dataset.

The dataset was obtained from Kaggle and was originally created by IBM to study employee attrition patterns within an organization.

Source: IBM HR Analytics Employee Attrition & Performance
URL: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

The purpose of this documentation is to summarize the dataset, describe the selected variables and their data types.

2. Dataset Description

This is a fictional data set created by IBM data scientists. It includes demographic information, job-related attributes, compensation details, and work–life balance indicators.

The primary target variable is **Attrition**, which indicates whether an employee has left the company. Since the original dataset contains many variables, we focus on a subset of variables that are most relevant for attrition analysis and suitable for statistical exploration.

```r
# Load dataset.
hr <- read.csv("WA_Fn-UseC_-HR-Employee-Attrition.csv", stringsAsFactors = FALSE)

# Our selection of variables of interest.
data_selected <- hr[, c("Attrition", "OverTime", "JobLevel", "JobSatisfaction",
                        "WorkLifeBalance", "Age", "YearsAtCompany", "TotalWorkingYears",
                        "NumCompaniesWorked", "MonthlyIncome")]
```

3. Summary of the Dataset

3.1 Number of Rows and Columns

```
dim(data_selected)
```

```
## [1] 1470   10
```

3.2 Summary of Variables

```
summary(data_selected)
```

```
##   Attrition           OverTime            JobLevel     JobSatisfaction
##  Length:1470        Length:1470        Min.   :1.000   Min.   :1.000  
##  Class :character   Class :character   1st Qu.:1.000   1st Qu.:2.000  
##  Mode  :character   Mode  :character   Median :2.000   Median :3.000  
##                                        Mean   :2.064   Mean   :2.729  
##                                        3rd Qu.:3.000   3rd Qu.:4.000  
##                                        Max.   :5.000   Max.   :4.000  
##  WorkLifeBalance      Age        YearsAtCompany   TotalWorkingYears
##  Min.   :1.000   Min.   :18.00   Min.   : 0.000   Min.   : 0.00    
##  1st Qu.:2.000   1st Qu.:30.00   1st Qu.: 3.000   1st Qu.: 6.00    
##  Median :3.000   Median :36.00   Median : 5.000   Median :10.00    
##  Mean   :2.761   Mean   :36.92   Mean   : 7.008   Mean   :11.28    
##  3rd Qu.:3.000   3rd Qu.:43.00   3rd Qu.: 9.000   3rd Qu.:15.00    
##  Max.   :4.000   Max.   :60.00   Max.   :40.000   Max.   :40.00    
##  NumCompaniesWorked MonthlyIncome  
##  Min.   :0.000      Min.   : 1009  
##  1st Qu.:1.000      1st Qu.: 2911  
##  Median :2.000      Median : 4919  
##  Mean   :2.693      Mean   : 6503  
##  3rd Qu.:4.000      3rd Qu.: 8379  
##  Max.   :9.000      Max.   :19999  
```

4. Variable Description

Variable Name	Data Type	Scale of Measure	Description
Attrition	Binary	Nominal	Whether the employee left the company (Yes / No)
OverTime	Binary	Nominal	Whether the employee works overtime

Variable Name	Data Type	Scale of Measure	Description
JobLevel	Categorical	Ordinal	Job level on a scale from 1 (lowest) to 5 (highest)
JobSatisfaction	Categorical	Ordinal	Job satisfaction level (1 = Low, 4 = Very High)
WorkLifeBalance	Categorical	Ordinal	Work-life balance rating (1 = Bad, 4 = Best)
Age	Numeric	Discrete	Age of the employee (in years)
YearsAtCompany	Numeric	Discrete	Number of years the employee has worked at the company
TotalWorkingYears	Numeric	Discrete	Total number of years of professional experience
NumCompaniesWorked	Numeric	Discrete	Number of companies the employee has previously worked for
MonthlyIncome	Numeric	Continuous	Monthly salary of the employee (in USD)

5. Example Data (First Few Observations)

```r
head(data_selected, 5)
```

```
##   Attrition OverTime JobLevel JobSatisfaction WorkLifeBalance Age
## 1       Yes      Yes        2               4               1  41
## 2        No       No        2               2               3  49
## 3       Yes      Yes        1               3               3  37
## 4        No      Yes        1               3               3  33
## 5        No       No        1               2               3  27
##   YearsAtCompany TotalWorkingYears NumCompaniesWorked MonthlyIncome
## 1              6                 8                  8          5993
## 2             10                10                  1          5130
## 3              0                 7                  6          2090
## 4              8                 8                  1          2909
## 5              2                 6                  9          3468
```

6. Conclusion

We found the IBM HR employee attrition dataset suitable for the given task as it has simple yet sufficient variables to perform descriptive statistical analysis.