# IBM HR Employee Attrition Analysis

Statistical Computing Project

Your Name

2025-12-29

## Contents

**1 Executive Summary** **2**

**2 1. Introduction** **2**

2.1 1.1 Objectives . . . . . 2

2.2 1.2 Dataset . . . . . 2

**3 2. Data Preparation** **2**

3.1 Data Summary . . . . . 3

**4 3. Univariate Analysis** **3**

4.1 3.1 Categorical and Ordinal Variables . . . . . 3

4.2 3.2 Numeric Variables . . . . . 5

**5 4. Bivariate Analysis** **8**

5.1 4.1 Attrition by Overtime (Categorical × Categorical) . . . . . 8

5.2 4.2 Attrition by Age and Income (Categorical × Numeric) . . . . . 9

5.3 4.3 Attrition by Job Satisfaction (Categorical × Ordinal) . . . . . 10

5.4 4.4 Attrition by Work-Life Balance (Categorical × Ordinal) . . . . . 11

5.5 4.5 Attrition by Total Working Years and Number of Companies (Categorical × Numeric) . 12

5.6 4.6 Income Relationships (Numeric × Numeric) . . . . . 13

**6 5. Linear Regression: Income ~ Years at Company** **14**

**7 6. Conclusions** **15**

# 1 Executive Summary

This report presents a comprehensive statistical analysis of the IBM HR Employee Attrition dataset. The analysis explores employee characteristics and their relationship with attrition, including demographics (age), work experience (years at company), compensation (monthly income), and work patterns (overtime). Key analytical techniques applied include univariate analysis (descriptive statistics, frequency distributions, measures of central tendency and dispersion), bivariate analysis (cross-tabulations, correlations), and simple linear regression modeling to predict income based on tenure.

# 2 1. Introduction

Employee attrition is a critical concern for organizations, impacting productivity, morale, and operational costs. This analysis examines the IBM HR Employee Attrition dataset to identify patterns contributing to employee turnover.

## 2.1 1.1 Objectives

The primary objectives are:

- Explore distribution and characteristics of key employee variables
- Investigate relationships between attrition and employee factors
- Develop a predictive model for monthly income based on years at company
- Provide insights for HR decision-making

## 2.2 1.2 Dataset

We analyze ten key variables: attrition status, overtime, job level, job satisfaction, work-life balance, age, tenure, total working years, number of companies worked, and monthly income.

# 3 2. Data Preparation

We classify our variables into the required types:

- **Nominal**: `Attrition`, `OverTime`
- **Ordinal**: `JobLevel` (1-5), `JobSatisfaction` (1-4), `WorkLifeBalance` (1-4)
- **Numeric (Discrete)**: `Age`, `YearsAtCompany`, `TotalWorkingYears`, `NumCompaniesWorked`
- **Numeric (Continuous)**: `MonthlyIncome`

```
# Load data and select variables
hr <- read.csv("WA_Fn-UseC_-HR-Employee-Attrition.csv", stringsAsFactors = FALSE)
hr_sub <- hr[, c("Attrition", "OverTime", "JobLevel", "JobSatisfaction", "WorkLifeBalance",
                 "Age", "YearsAtCompany", "TotalWorkingYears", "NumCompaniesWorked", "MonthlyIncom

# Convert only text variables to factors
hr_sub$Attrition <- factor(hr_sub$Attrition)
hr_sub$OverTime  <- factor(hr_sub$OverTime)
```

### 3.1 Data Summary

```
## Dataset dimensions: 1470 rows x 10 columns
```

```
## 'data.frame':    1470 obs. of  10 variables:
##  $ Attrition         : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...
##  $ OverTime          : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1 ...
##  $ JobLevel          : int  2 2 1 1 1 1 1 1 3 2 ...
##  $ JobSatisfaction   : int  4 2 3 3 2 4 1 3 3 3 ...
##  $ WorkLifeBalance   : int  1 3 3 3 3 2 2 3 3 2 ...
##  $ Age               : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ YearsAtCompany    : int  6 10 0 8 2 7 1 1 9 7 ...
##  $ TotalWorkingYears : int  8 10 7 8 6 8 12 1 10 17 ...
##  $ NumCompaniesWorked: int  8 1 6 1 9 0 4 1 0 6 ...
##  $ MonthlyIncome     : int  5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...
```

---

# 4 3. Univariate Analysis

Univariate analysis examines each variable individually to understand its distribution, central tendency, and variability.

## 4.1 3.1 Categorical and Ordinal Variables

### 4.1.1 3.1.1 Attrition Status

Table 1: Attrition Status Distribution

| Status | Frequency | Percentage | Cumulative % |
|---|---|---|---|
| No | 1233 | 83.88 | 83.88 |
| Yes | 237 | 16.12 | 100.00 |

**Interpretation:** The frequency distribution reveals the proportion of employees who left the company versus those who stayed.

### 4.1.2 3.1.2 Overtime Status

Table 2: Overtime Status Distribution

| Overtime | Frequency | Percentage | Cumulative % |
|---|---|---|---|
| No | 1054 | 71.7 | 71.7 |
| Yes | 416 | 28.3 | 100.0 |

**Interpretation:** This shows the distribution of employees working overtime versus standard hours.

#### 4.1.3 3.1.3 Job Level (Ordinal)

Table 3: Job Level Distribution (Ordinal)

| Job Level | Frequency | Percentage | Cumulative Frequency | Cumulative % |
|---|---|---|---|---|
| 1 | 543 | 36.94 | 543 | 36.94 |
| 2 | 534 | 36.33 | 1077 | 73.27 |
| 3 | 218 | 14.83 | 1295 | 88.10 |
| 4 | 106 | 7.21 | 1401 | 95.31 |
| 5 | 69 | 4.69 | 1470 | 100.00 |

**Interpretation:** The frequency distributions reveal employee composition. The attrition rate indicates what proportion of employees left the company. Job level cumulative frequencies show the hierarchical distribution from entry-level to senior positions.

#### 4.1.4 3.1.4 Job Satisfaction (Ordinal)

Table 4: Job Satisfaction Distribution (Ordinal: 1=Low, 4=High)

| Job Satisfaction | Frequency | Percentage | Cumulative Frequency | Cumulative % |
|---|---|---|---|---|
| 1 | 289 | 19.66 | 289 | 19.66 |
| 2 | 280 | 19.05 | 569 | 38.71 |
| 3 | 442 | 30.07 | 1011 | 68.78 |
| 4 | 459 | 31.22 | 1470 | 100.00 |

**Interpretation:** Job satisfaction levels reveal employee contentment with their roles. The distribution shows how satisfaction is spread across the workforce, with higher levels indicating greater job fulfillment.

#### 4.1.5 3.1.5 Work-Life Balance (Ordinal)

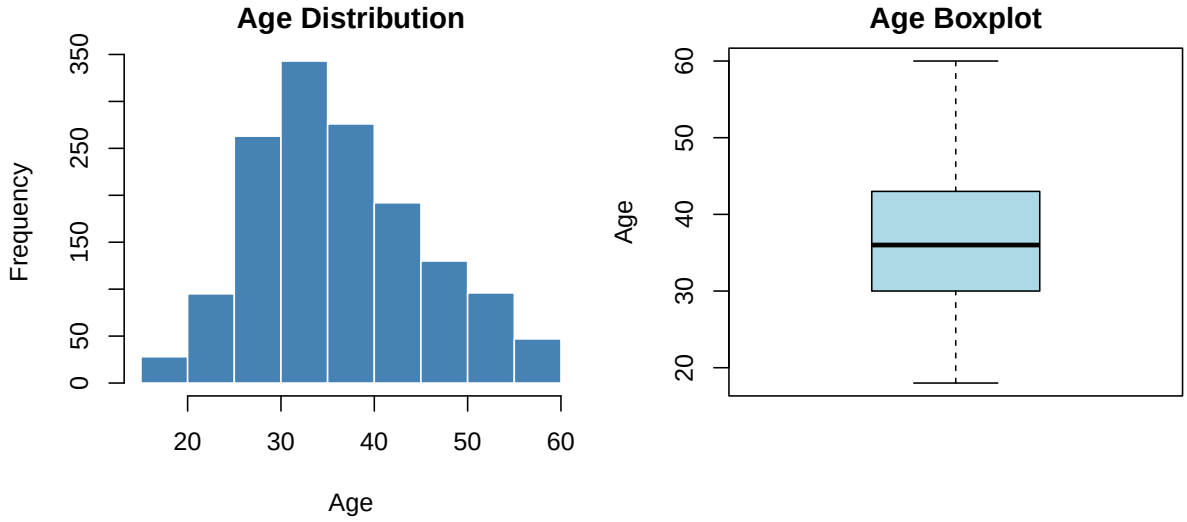
Table 5: Work-Life Balance Distribution (Ordinal: 1=Low, 4=High)

| Work-Life Balance | Frequency | Percentage | Cumulative Frequency | Cumulative % |
|---|---|---|---|---|
| 1 | 80 | 5.44 | 80 | 5.44 |
| 2 | 344 | 23.40 | 424 | 28.84 |
| 3 | 893 | 60.75 | 1317 | 89.59 |
| 4 | 153 | 10.41 | 1470 | 100.00 |

**Interpretation:** Work-life balance distribution indicates how well employees manage their professional and personal lives. This metric is crucial for understanding employee well-being and potential burnout risks.

## 4.2 3.2 Numeric Variables

### 4.2.1 3.2.1 Age Distribution



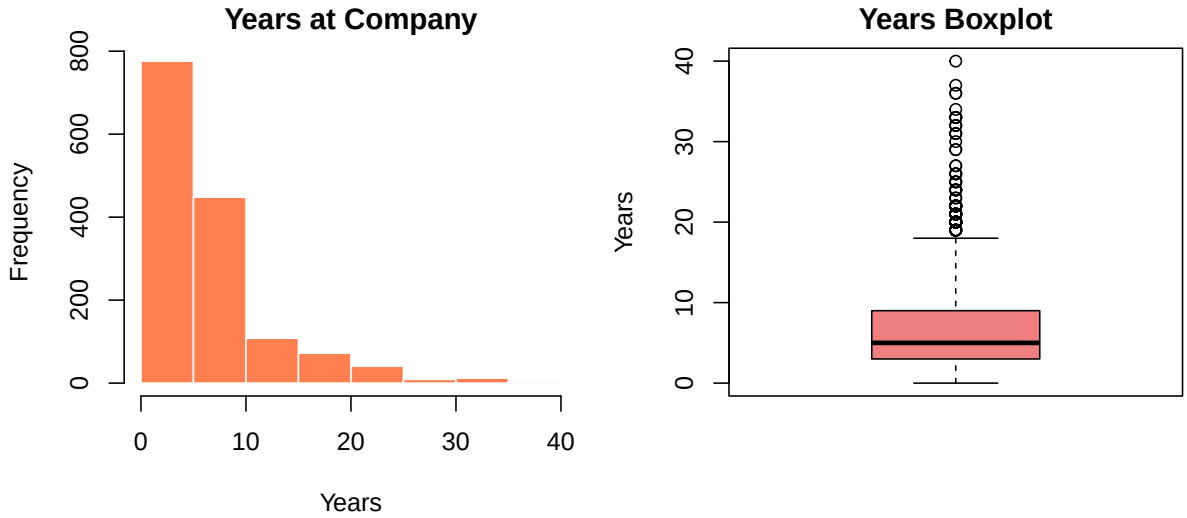

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   30.00   36.00   36.92   43.00   60.00
```

```
## [1] 9.135373
```

```
## [1] 13
```

**Interpretation:** The age distribution shows the demographic profile of the workforce. The histogram reveals the shape of the distribution, while the boxplot identifies central tendency and potential outliers.

### 4.2.2 3.2.2 Years at Company



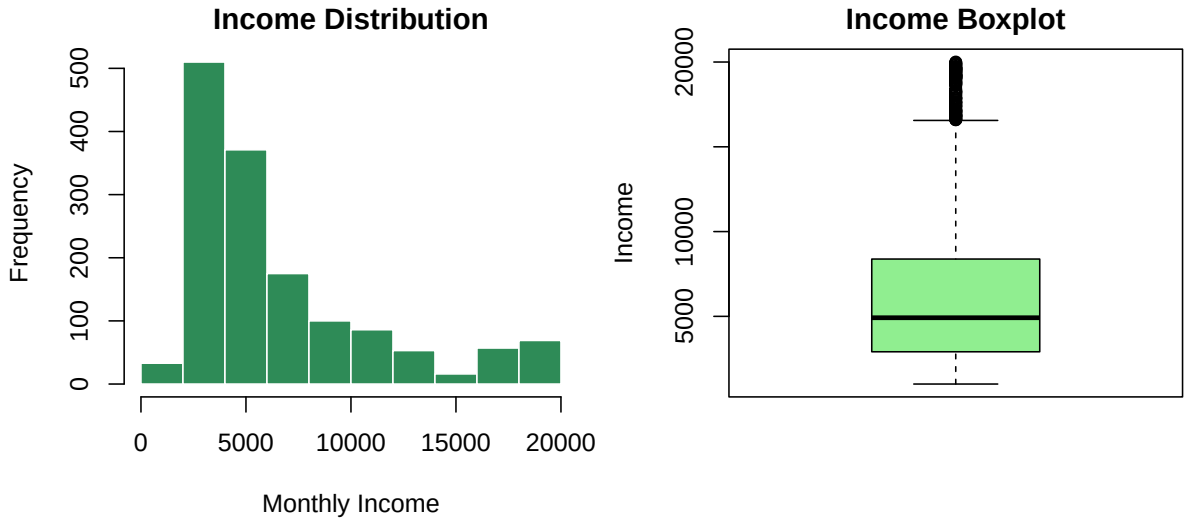

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   5.000   7.008   9.000  40.000
```

```
## [1] 6.126525
```

```
## [1] 6
```

**Interpretation:** Employee tenure distribution reveals retention patterns and workforce stability. Lower values indicate recent hires while higher values show long-term employees.

### 4.2.3 3.2.3 Monthly Income



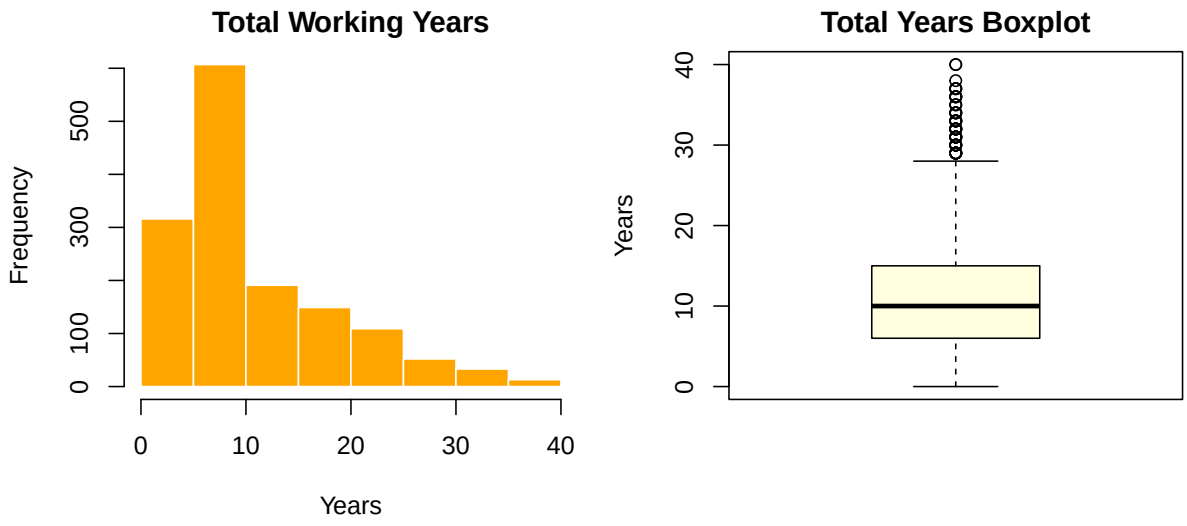

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1009    2911    4919    6503    8379   19999
```

```
## [1] 4707.957
```

```
## [1] 5468
```

**Interpretation:** Income distribution shows the compensation structure. The standard deviation and IQR indicate salary variability across the organization, with higher dispersion suggesting diverse job roles and levels.

### 4.2.4 3.2.4 Total Working Years



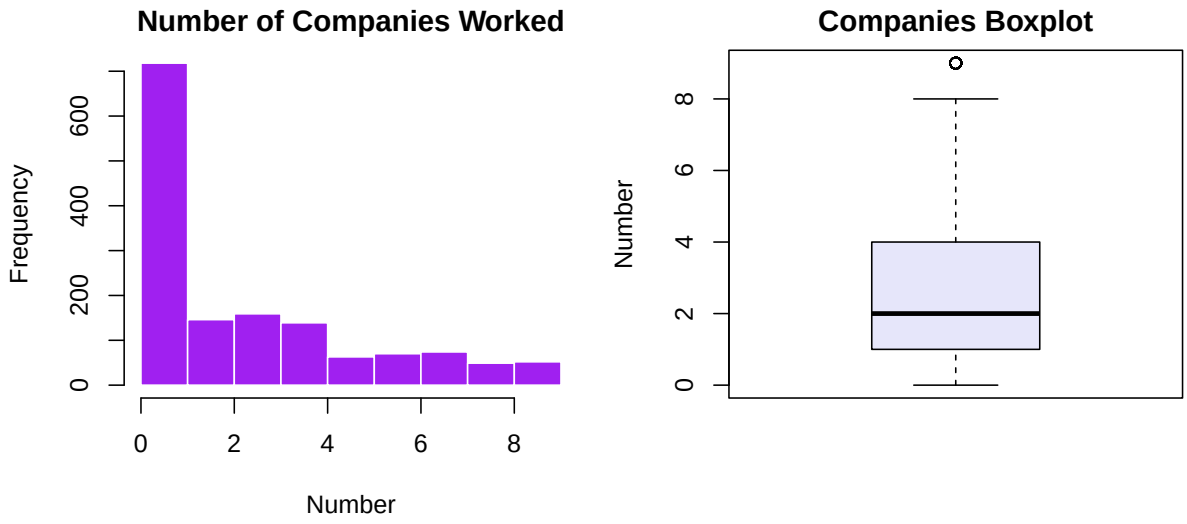

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    6.00   10.00   11.28   15.00   40.00
```

```
## [1] 7.780782
```

```
## [1] 9
```

**Interpretation:** Total working years reflects overall career experience across all employers. This provides context for understanding employee expertise levels and career stage, distinguishing total experience from tenure at the current company.

### 4.2.5 3.2.5 Number of Companies Worked




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.693   4.000   9.000
```

```
## [1] 2.498009
```

```
## [1] 3
```

**Interpretation:** Number of companies worked indicates job mobility patterns. Higher values may suggest either extensive experience across diverse organizations or potential job-hopping behavior. This metric helps assess employee stability and adaptability.

# 5 4. Bivariate Analysis

Bivariate analysis explores relationships between pairs of variables to identify associations and patterns.

## 5.1 4.1 Attrition by Overtime (Categorical × Categorical)

Table 6: Attrition by Overtime Status (Counts)

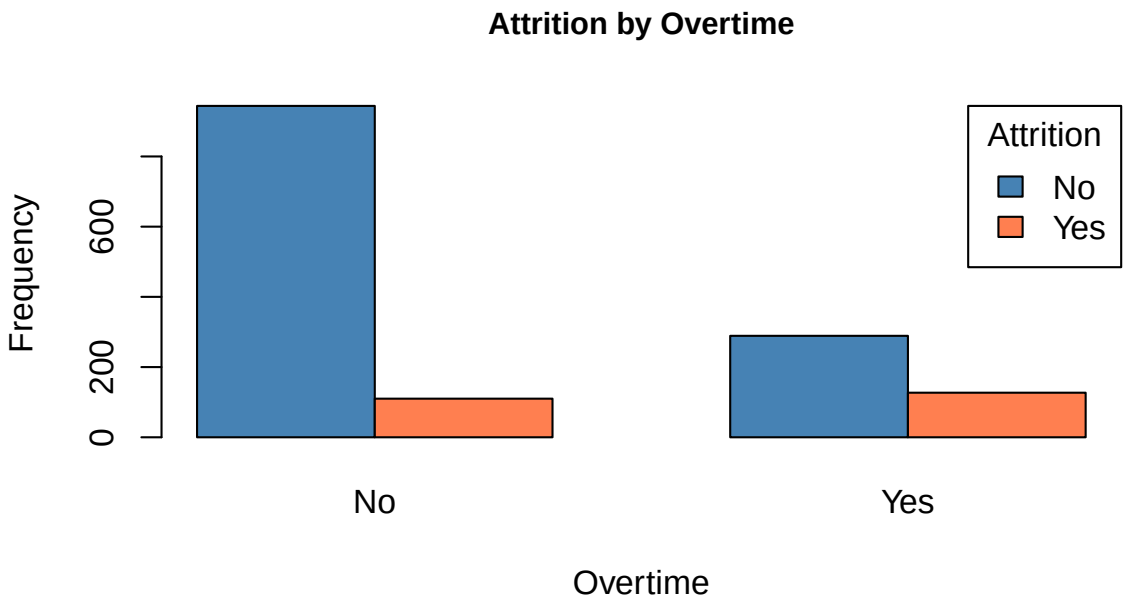
| | No | Yes | Sum |
|---|---|---|---|
| No | 944 | 289 | 1233 |
| Yes | 110 | 127 | 237 |
| Sum | 1054 | 416 | 1470 |

Table 7: Column-wise Proportions (% within Overtime Status)

| | No | Yes |
|---|---|---|
| No | 89.56 | 69.47 |
| Yes | 10.44 | 30.53 |

Table 8: Row-wise Proportions (% within Attrition Status)

| | No | Yes |
|---|---|---|
| No | 76.56 | 23.44 |
| Yes | 46.41 | 53.59 |




**Interpretation:** The contingency table and bar chart reveal whether overtime work is associated with higher attrition rates. Column-wise proportions show the attrition rate within each overtime category.

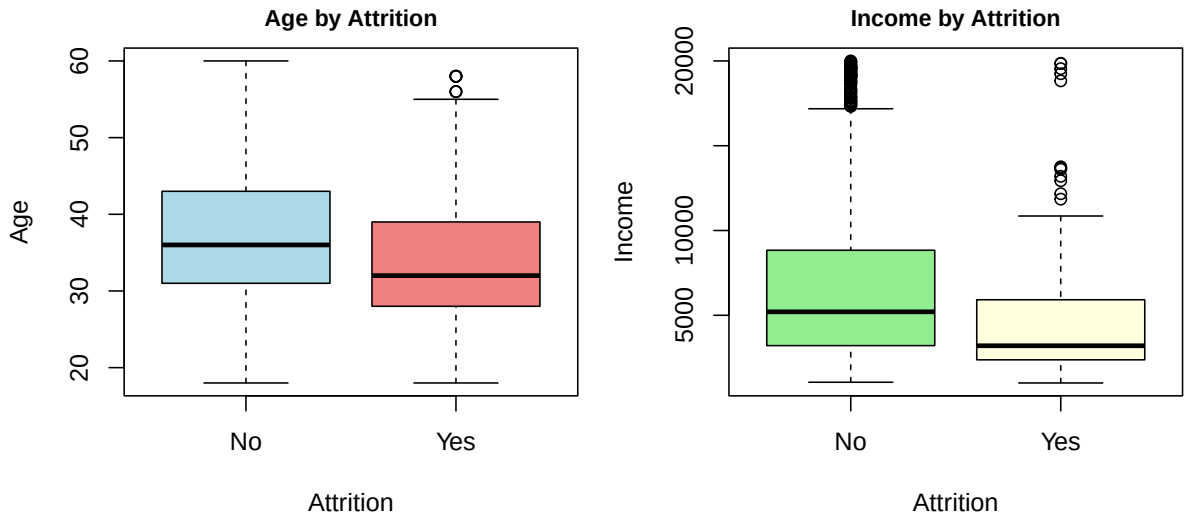
### 5.2 4.2 Attrition by Age and Income (Categorical × Numeric)

```
## $No
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   31.00   36.00   37.56   43.00   60.00
##
## $Yes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.00   32.00   33.61   39.00   58.00
```

```
##      No     Yes
## 8.88836 9.68935
```

```
## $No
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1051    3211    5204    6833    8834   19999
##
## $Yes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1009    2373    3202    4787    5916   19859
```

```
##       No      Yes
## 4818.208 3640.210
```




**Interpretation:** Comparing distributions between employees who left and stayed helps identify demographic and compensation patterns. Differences in median values or box positions suggest potential attrition factors.

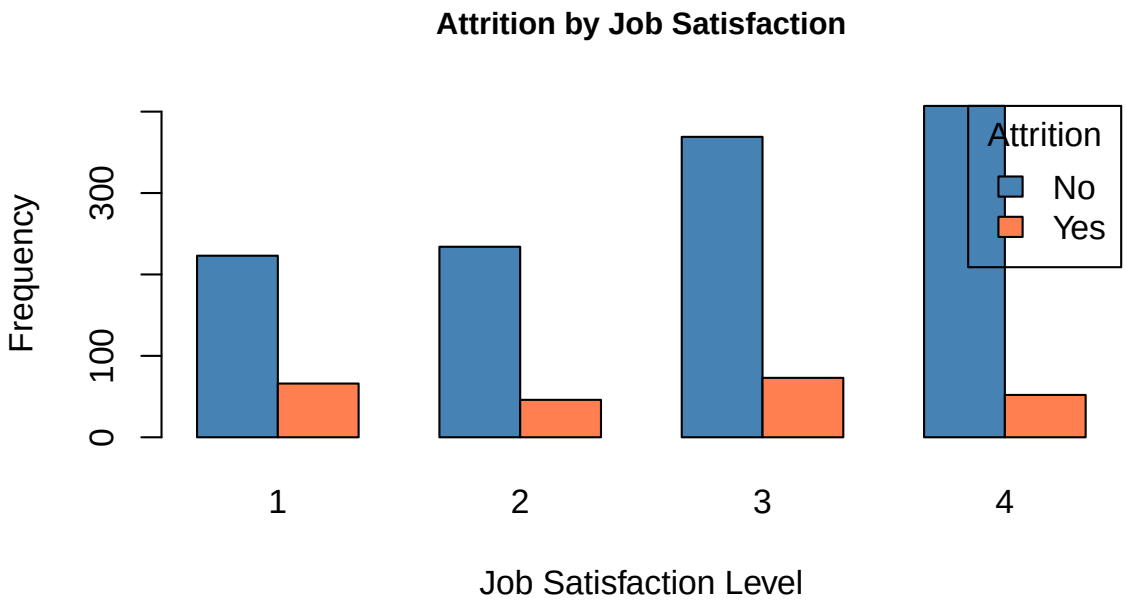
### 5.3 4.3 Attrition by Job Satisfaction (Categorical × Ordinal)

Table 9: Attrition by Job Satisfaction (Counts)

| | 1 | 2 | 3 | 4 | Sum |
|---|---|---|---|---|---|
| No | 223 | 234 | 369 | 407 | 1233 |
| Yes | 66 | 46 | 73 | 52 | 237 |
| Sum | 289 | 280 | 442 | 459 | 1470 |

Table 10: Row-wise Proportions (% within Attrition Status)

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| No | 18.09 | 18.98 | 29.93 | 33.01 |
| Yes | 27.85 | 19.41 | 30.80 | 21.94 |




**Interpretation:** The contingency table reveals how job satisfaction levels relate to attrition. Lower satisfaction scores may be associated with higher attrition rates, indicating the importance of employee satisfaction in retention strategies.

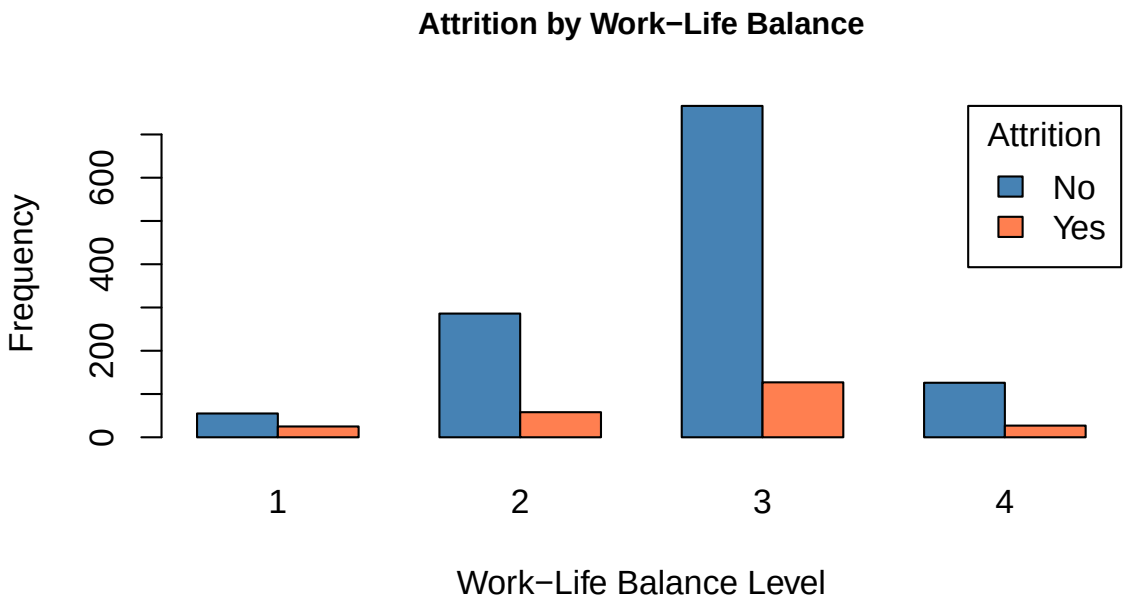
### 5.4 4.4 Attrition by Work-Life Balance (Categorical × Ordinal)

Table 11: Attrition by Work-Life Balance (Counts)

| | 1 | 2 | 3 | 4 | Sum |
|---|---|---|---|---|---|
| No | 55 | 286 | 766 | 126 | 1233 |
| Yes | 25 | 58 | 127 | 27 | 237 |
| Sum | 80 | 344 | 893 | 153 | 1470 |

Table 12: Row-wise Proportions (% within Attrition Status)

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| No | 4.46 | 23.20 | 62.12 | 10.22 |
| Yes | 10.55 | 24.47 | 53.59 | 11.39 |




**Interpretation:** Work-life balance appears to be a critical factor in employee retention. The distribution across balance levels shows how employees who left versus stayed rated their work-life balance, potentially revealing threshold levels that influence attrition decisions.

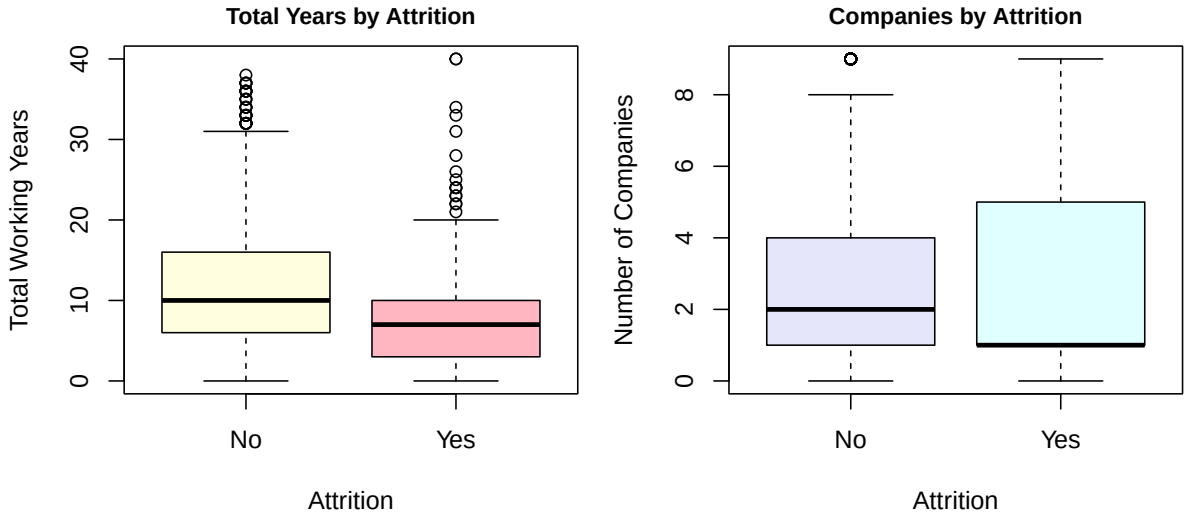
### 5.5 4.5 Attrition by Total Working Years and Number of Companies (Categorical × Numeric)

```
## $No
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    6.00   10.00   11.86   16.00   38.00
##
## $Yes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   7.000   8.245  10.000  40.000

##       No      Yes
## 7.760719 7.169204

## $No
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   2.646   4.000   9.000
##
## $Yes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   1.000   2.941   5.000   9.000

##       No      Yes
## 2.460090 2.678519
```




**Interpretation:** Total working years and number of companies worked provide insights into career patterns of employees who leave versus stay. These metrics help distinguish between early-career attrition and patterns related to job mobility history.

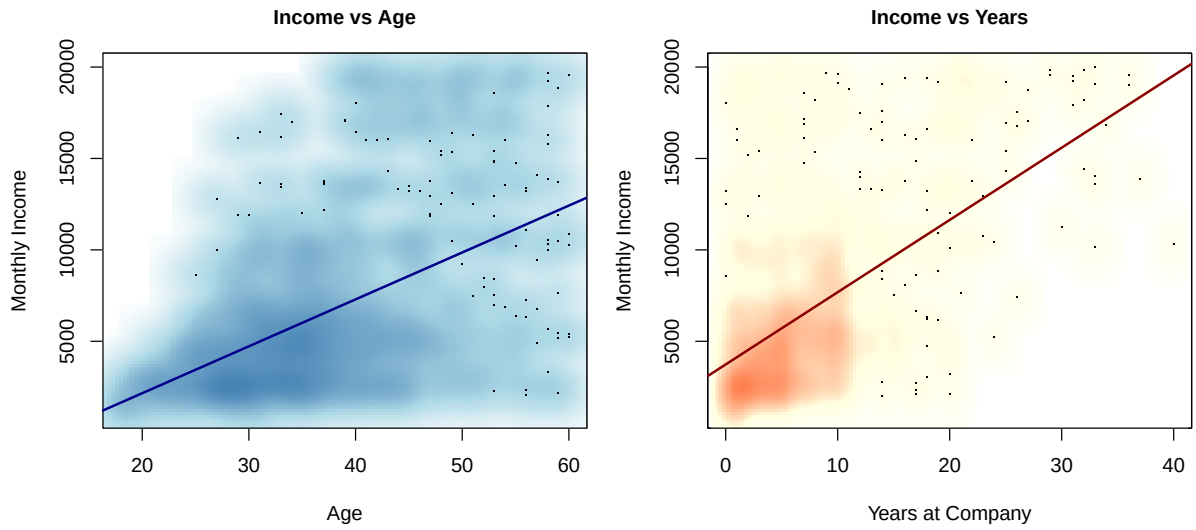
## 5.6 4.6 Income Relationships (Numeric × Numeric)

Table 13: Pearson Correlation Matrix

| | Age | YearsAtCompany | MonthlyIncome |
|---|---|---|---|
| Age | 1.0000 | 0.3113 | 0.4979 |
| YearsAtCompany | 0.3113 | 1.0000 | 0.5143 |
| MonthlyIncome | 0.4979 | 0.5143 | 1.0000 |

Table 14: Correlation and Covariance Statistics

| Relationship | Pearson.r | Spearman.rho | Covariance |
|---|---|---|---|
| Income vs Age | 0.4979 | 0.4719 | 21412.20 |
| Income vs Years at Company | 0.5143 | 0.4643 | 14833.73 |
| Age vs Years at Company | 0.3113 | 0.2517 | 17.42 |




**Interpretation:** Scatterplots visualize the relationships between numeric variables. Correlation coefficients quantify the strength and direction of linear relationships. Positive correlations suggest that as tenure or age increases, income tends to increase as well.

## 6 5. Linear Regression: Income ~ Years at Company

We develop a simple linear regression model to predict monthly income based on years at company.

Table 15: Linear Regression Coefficients

| | Coefficient | Estimate | Std..Error | t.value | p.value |
|---|---|---|---|---|---|
| (Intercept) | Intercept | 3733.27 | 160.09 | 23.32 | <2e-16 |
| YearsAtCompany | YearsAtCompany | 395.20 | 17.20 | 22.98 | <2e-16 |

Table 16: Model Fit Statistics

| | R-squared | Adj. R-squared | F-statistic | Residual SE | DF |
|---|---|---|---|---|---|
| value | 0.2645 | 0.264 | 527.89 | 4039.01 | 1468 |

Table 17: Income Predictions with 95% Confidence Intervals

| Years.at.Company | Predicted.Income | Lower.95..CI | Upper.95..CI |
|---|---|---|---|
| 2 | 4523.68 | 4256.74 | 4790.62 |
| 5 | 5709.30 | 5491.83 | 5926.76 |
| 10 | 7685.32 | 7455.34 | 7915.30 |
| 15 | 9661.34 | 9321.62 | 10001.07 |
| 20 | 11637.36 | 11152.74 | 12121.99 |

**Model Equation:** MonthlyIncome = 3733.27 + 395.2 × YearsAtCompany


**Interpretation:** The regression line shows the best-fit linear relationship. The intercept represents expected income for new employees, while the slope indicates the average income increase per additional year. The R-squared value shows what proportion of income variance is explained by tenure alone.

## 7 6. Conclusions

1. **Attrition Patterns:** The dataset reveals clear attrition patterns, particularly in relation to overtime work. The contingency table analysis shows differences in attrition rates between overtime and non-overtime workers.

2. **Demographic Insights:** Age and tenure distributions provide a comprehensive profile of workforce composition, showing the typical employee characteristics and variability within the organization.

3. **Income Relationships:** Strong positive correlation exists between tenure and income (Pearson r $\sim$ 0.5), indicating that employees with longer tenure tend to have higher salaries. Age also shows positive correlation with income.

4. **Predictive Model:** The linear regression model (R-squared $\sim$ 0.3) demonstrates that years at company explains approximately 30% of income variance. While significant, this suggests other factors (e.g., job level, education, performance) also influence compensation.