IBM HR Employee Attrition Descriptive Analysis

Statistical Computing WiSe '25 - Group Project

Amaan Shaikh | Danish Ali Shaikh | Rajat Pandey

1 Introduction

The IBM HR Employee Attrition workforce data [IBM Watson Analytics, 2017] is organized as a data matrix of observations and qualitative and quantitative variables. This report performs a comprehensive descriptive statistical analysis to summarize the distribution of individual variables using frequencies, measure of location, and measure of dispersion. Bivariate techniques, including contingency tables and correlation coefficients, are applied to investigate the joint behavior and co-dependency between employee attributes and attrition. Furthermore, simple linear regression is utilized to model the linear relationship between tenure and monthly income by estimating the regression coefficients. The analysis employs visualization tools from **ggplot2** [Wickham et al., 2024] and **graphics** [R Core Team, 2024] packages, with tabular summaries generated using **knitr** [Xie, 2024].

1.1 Data Loading & Feature Selection

We analyze **seven** key variables classified into the following datatypes:

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

1.2 Data Summary

The dataset contains **1470** observations and **7** variables *(Selected from 35 variables)*.

Table 1: First six observations (Transposed View)

	1	2	3	4	5	6
Age	41	49	37	33	27	32
Attrition	Yes	No	Yes	No	No	No
JobLevel	Junior	Junior	Entry Level	Entry Level	Entry Level	Entry Level
JobSatisfaction	Very High	Medium	High	High	Medium	Very High
MonthlyIncome	5993	5130	2090	2909	3468	3068
OverTime	Yes	No	Yes	Yes	No	No
YearsAtCompany	6	10	0	8	2	7

2 Univariate Analysis

This section characterizes the workforce structure by analyzing the frequency distributions, measure of location, and dispersion of key **sentiment**, **financial**, and **demographic** variables.

2.1 Target & Sentiment

This section performs a uni-variate analysis to characterize employee **sentiment** and **professional status**, examining the frequency distributions of key qualitative variables.

2.1.1 Attrition Status (Binary Nominal)

Table 2: Attrition Status Frequency Table

Attrition Status	Frequency	Proportion
No	1233	83.88
Yes	237	16.12

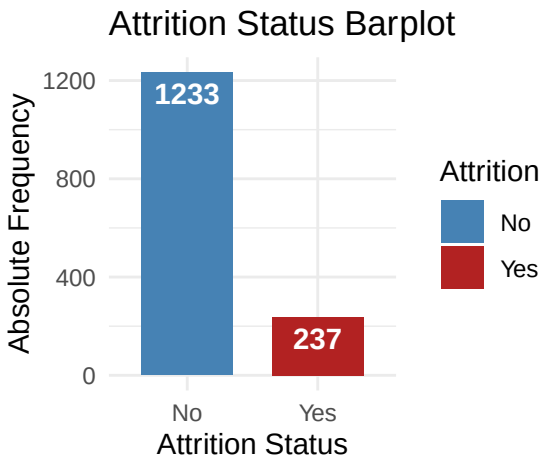

Figure 1: Bar plot of Attrition Status illustrating a significant class imbalance with high employee retention

Interpretation: The analysis reveals a significant **class imbalance**, with **1,233** *(83.88%)* active employees versus **237** *(16.12%)* attrition cases. This distribution confirms a stable workforce and establishes a realistic baseline turnover rate for subsequent bivariate comparisons.

2.1.2 Job Satisfaction (Ordinal)

Table 3: Job Satisfaction Frequency Table

Job Satisfaction	Frequency	Proportion	Cum. Freq.	Cum. %
Low	289	19.66	289	19.66
Medium	280	19.05	569	38.71
High	442	30.07	1011	68.78
Very High (Mode)	459	31.22	1470	100.00

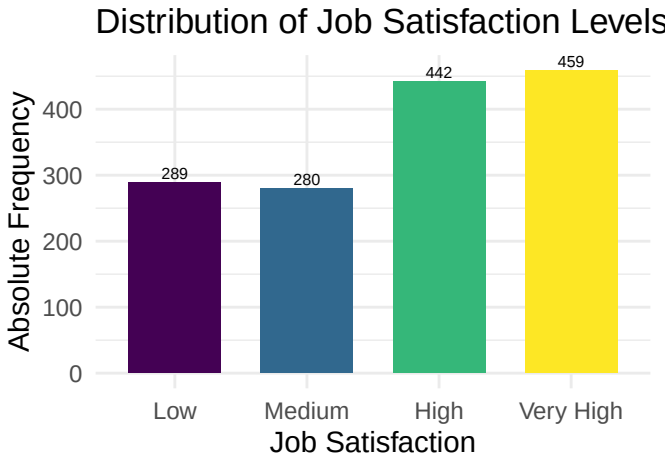

Figure 2: Distribution of Job Satisfaction Levels

Interpretation: The data indicates **high** employee morale, with **"Very High"** *(31.22%, n=459)* as the **modal** category. Cumulatively, over **61.29%** of the workforce reports **positive satisfaction** *(High or Very High)*, significantly outweighing the **19.66%** who report **low** satisfaction.

2.2 Financial Structure

This section summarizes the **job hierarchy**, **income distribution**, and **overtime** status of employees using univariate descriptive statistics and graphical representations to describe the compensation structure and workload intensity.

2.2.1 Job Level (Ordinal)

Table 4: Job Level Frequency Table

Job Level	Frequency	Proportion	Cum. Freq.	Cum. %
Entry Level (Mode)	543	36.94	543	36.94
Junior	534	36.33	1077	73.27
Mid Level	218	14.83	1295	88.10
Senior	106	7.21	1401	95.31
Executive	69	4.69	1470	100.00

Interpretation: The workforce structure exhibits a distinct pyramidal hierarchy, heavily concentrated in early-career roles. **Entry Level** *(36.94%, n=543)* and **Junior** *(36.33%, n=534)* positions **dominate**, cumulatively accounting for **73.27%** of the organization. In contrast, upper management is exclusive, with **Executives** representing only **4.69%** of the total population.

2.2.2 Monthly Income (Continuous)

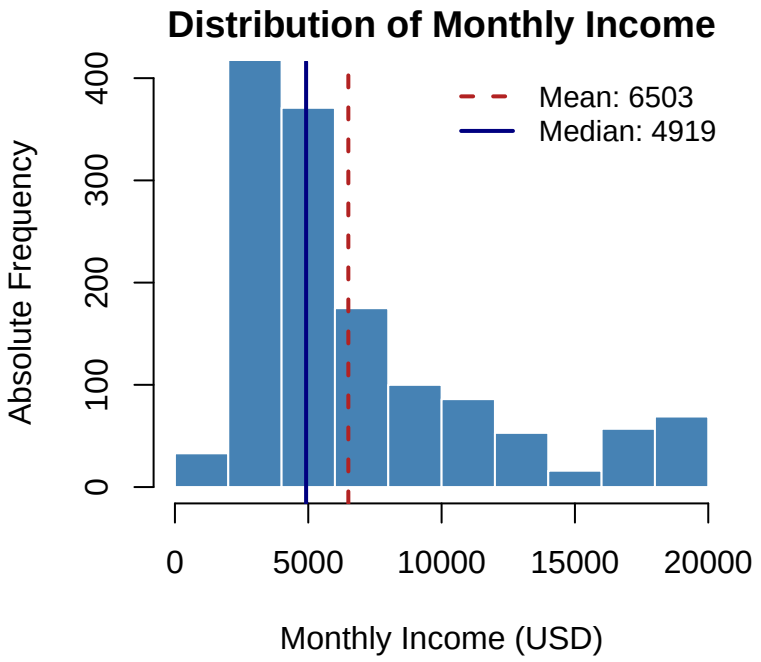

Figure 3: Histogram of Monthly Income. The distribution is right-skewed, indicated by the Mean (Red) being higher than the Median (Blue)

Table 5: Descriptive Statistics: Monthly Income (Location & Dispersion)

Mean	1st Q	Median	3rd Q	Variance	SD	Range	IQR
6502.93	2911	4919	8379	22164857	4707.96	1009 – 19999	5468

Interpretation: The distribution of monthly income is **strongly right-skewed**, evidenced by the **Mean** *($6,503)* significantly exceeding the **Median** *($4,919)*. While the majority of employees cluster in the lower income brackets *(approx. $2,500–$5,000)*, a distinct **right tail** of **high earners** extends to nearly **$20,000**. This indicates that a small cohort of highly paid executives disproportionately inflates the average, resulting in **high variability** *(SD = $4,708)* across the workforce.

2.2.3 Over Time (Binary Nominal)

Table 6: Over Time Frequency Table

Over Time	Frequency	Proportion
No	1054	71.7
Yes	416	28.3

Interpretation: The majority of the workforce *(71.7%, n=1,054)* does **not work overtime**. However, a significant minority of **28.3%** *(n=416)* actively engages in overtime, representing a distinct subgroup that may be more susceptible to burnout or attrition.

2.3 Demographics

This section analyzes **Age** and **Years at Company** using descriptive statistics and distributional plots to characterize employee demographic composition and tenure.

2.2.1 Age (Continuous)

Table 7: Descriptive Statistics: Age (Location & Dispersion)

Mean	1st Q	Median	3rd Q	Variance	SD	Range	IQR
36.924	30	36	43	83.455	9.135	18 – 60	13

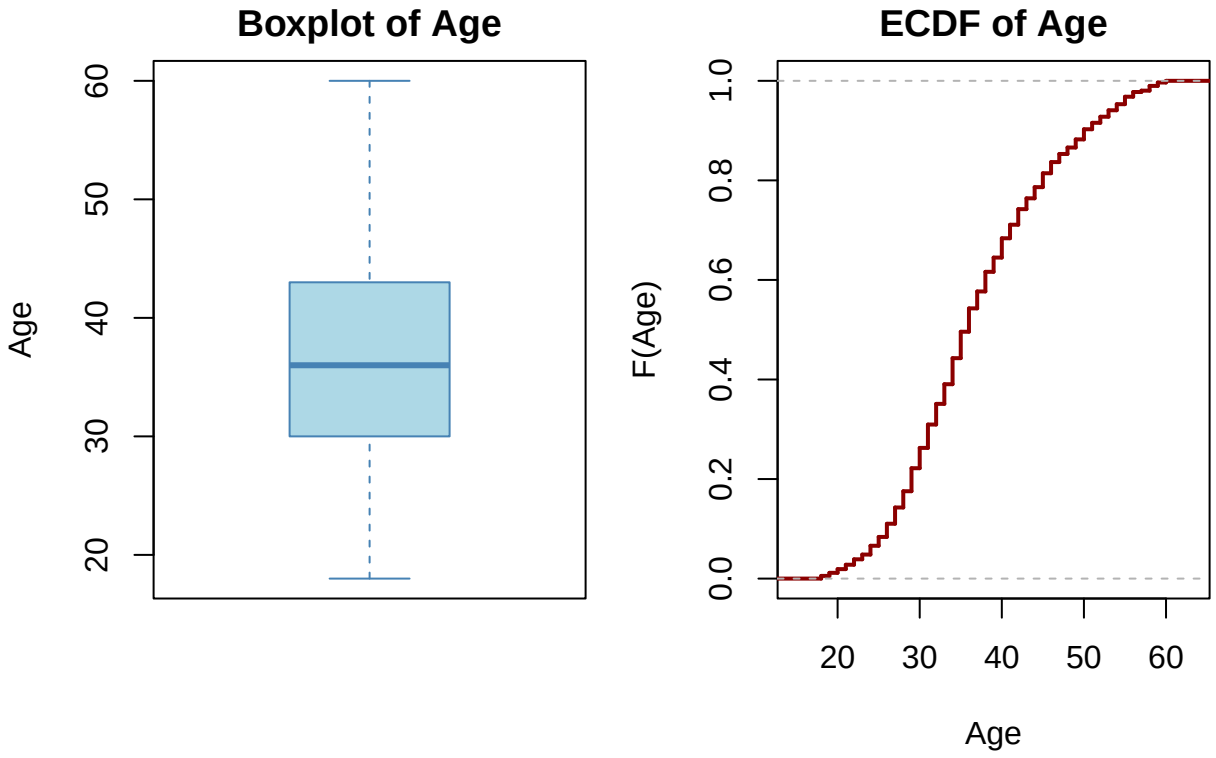

Figure 4: Distribution of employee age. The boxplot (left) summarizes dispersion and outliers, and the empirical cumulative distribution function (right) shows the cumulative age distribution

Interpretation: The age distribution is approximately symmetric and bell-shaped, spanning from **18** to **60** years with a standard deviation of **9.14**. The moderate variability **(IQR = 13)** indicates that the middle 50% of the workforce is concentrated within a relatively narrow age band, suggesting a demographically balanced employee base without extreme skewness. The ECDF reinforces this concentration with a steep incline between ages **30** and **45**, indicating that the majority of the workforce density accumulates in this mid-career bracket. The curve crosses the 0.5 cumulative probability threshold at approximately **36 years**, confirming the median, before tapering off near age **60**.

2.3.2 Years At Company (Discrete)

Table 8: Descriptive Statistics: Years at Company (Location & Dispersion)

Mean	1st Q	Median	3rd Q	Variance	SD	Range	IQR
7.008	3	5	9	37.534	6.127	0 – 40	6

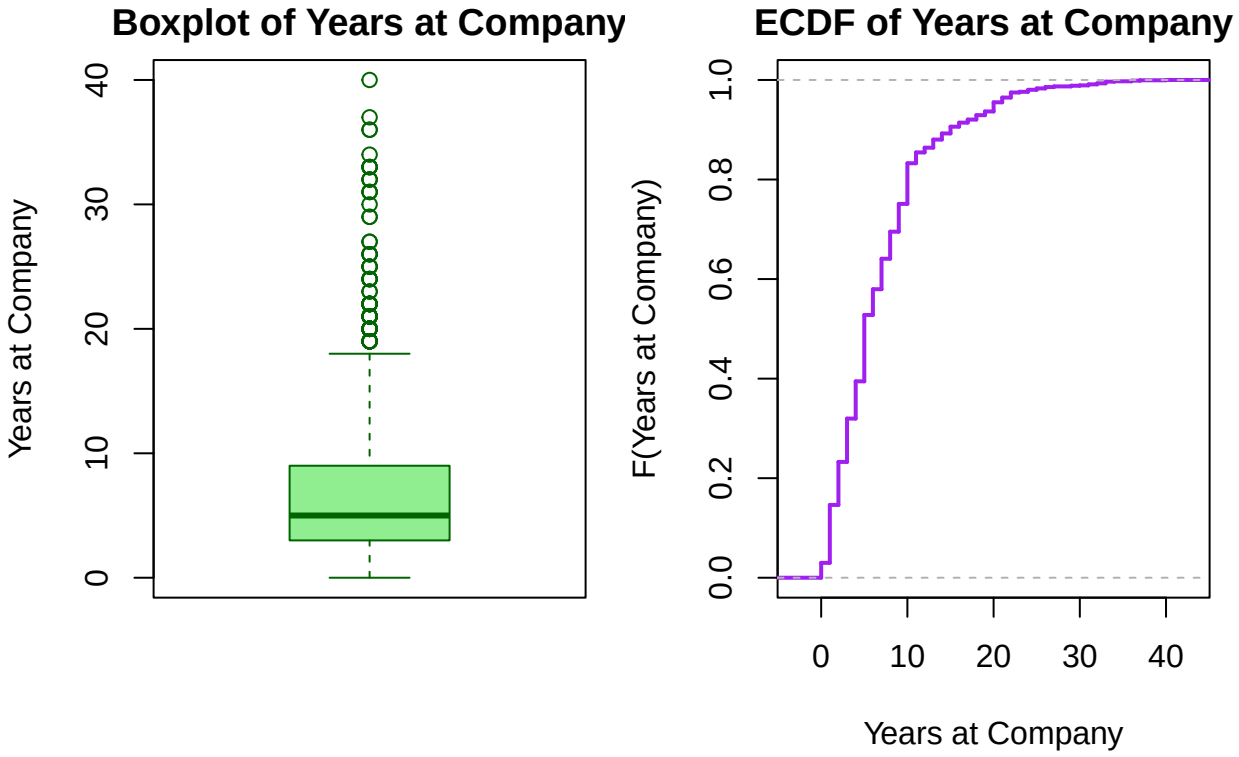

Figure 5: Distribution of years at company showing low median tenure and a long right tail with high-tenure outliers

Interpretation: The distribution of tenure is **strongly right-skewed**, evidenced by the **Mean** *(7.01)* exceeding the **Median** *(5.0)*. The boxplot reveals significant **upper outliers** *(up to 40 years)*, indicating that while the typical employee has a tenure of 5 years or less, a distinct cohort of long-term veterans remains, heavily influencing the average. The ECDF confirms this early-career concentration with a sharp vertical rise between 0 and 10 years, showing that approximately **75%** of the workforce has been with the company for **9 years or less** *(3rd Quartile)*, before the curve flattens out for the remaining long-term employees.

3 Bivariate Analysis

This section investigates the relationships between pairs of variables to identify potential drivers of attrition. We employ contingency tables for categorical data and correlation matrices for numerical attributes.

3.1 Impact of Sentiment on Attrition (Categorical vs. Categorical)

We analyze the relationship between **Job Satisfaction** and **Attrition** to determine if lower employee morale correlates with higher turnover.

Table 9: Attrition by Job Satisfaction Level (with Marginal Frequencies)

	Low	Medium	High	Very High	Sum
No	223	234	369	407	1233
Yes	66	46	73	52	237
Sum	289	280	442	459	1470

Table 10: Row Relative Frequency: Job Satisfaction within Attrition Groups

	Low	Medium	High	Very High
No	0.18	0.19	0.30	0.33
Yes	0.28	0.19	0.31	0.22

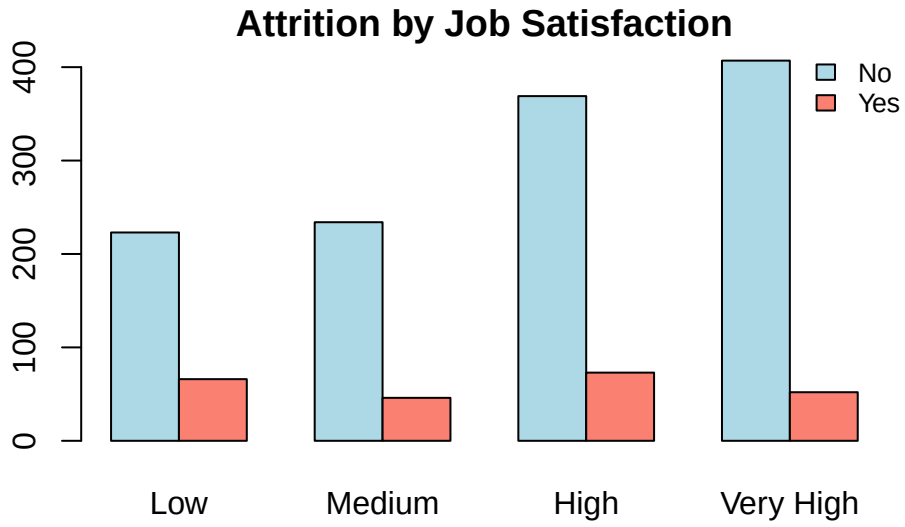

Figure 6: Attrition counts across Job Satisfaction levels. Employees reporting 'Low' satisfaction show a higher relative frequency of attrition compared to those in the 'High' and 'Very High' categories

Interpretation: The analysis shows a clear association between job satisfaction and attrition. Among employees who left the organization, **28%** reported Low satisfaction, compared to 18% among those who stayed, while Very High satisfaction is more common among retained employees. The bar chart visually confirms this trend: as **satisfaction levels rise**, the volume of **retained employees** *(Blue)* **increases** substantially,although attrition does not disappear entirely at higher satisfaction levels, indicating that job satisfaction alone does not fully explain attrition.

3.2 Financial Factors & Attrition (Numeric vs. Categorical)

This section examines the relationship between **Monthly Income** and employee **Attrition** to assess whether compensation levels differ between employees who stay and those who leave the organization. Group-wise summary statistics and correlation coefficients are used to describe the association between income and attrition status.

Table 11: Monthly Income Summary by Attrition (Mean, Median, SD)

	Mean	Median	SD
No	6832.74	5204	4818.21
Yes	4787.09	3202	3640.21

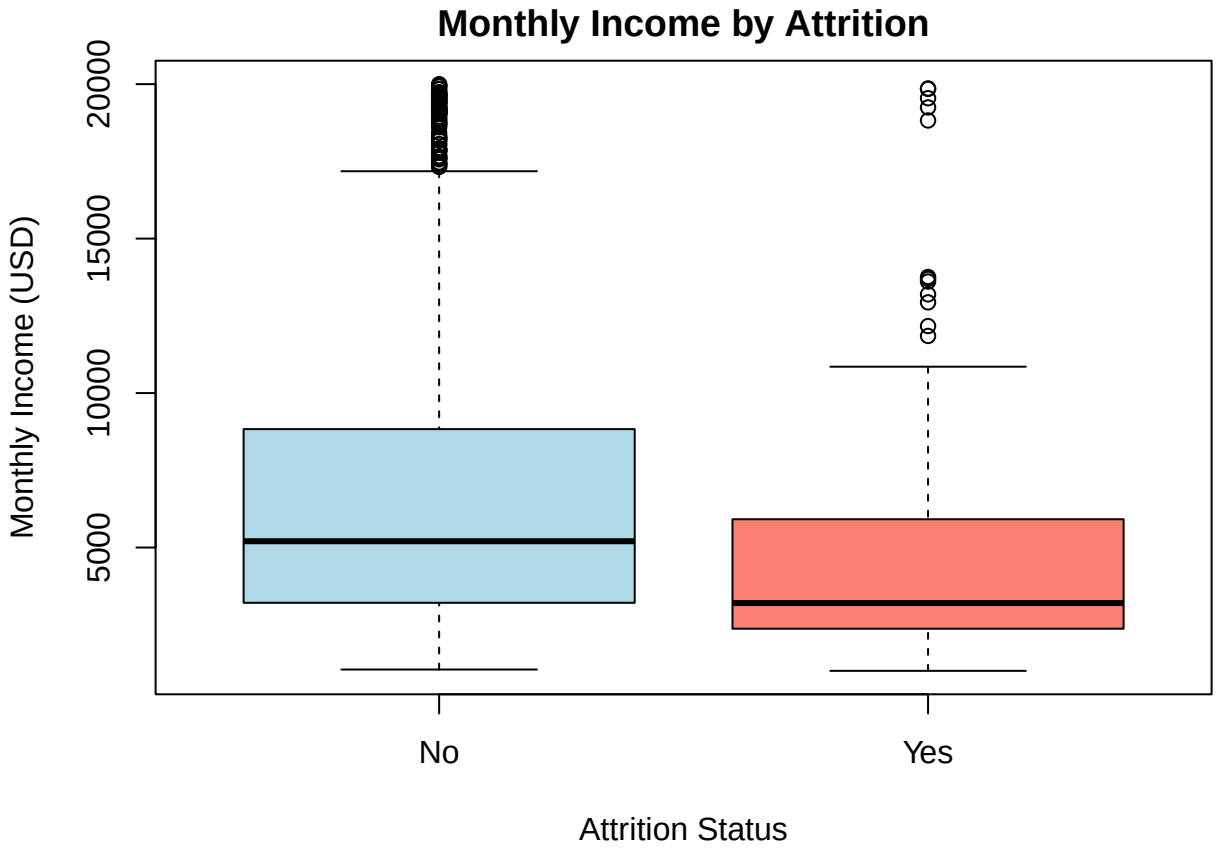

Figure 7: Boxplot of monthly income by attrition status, showing lower median income and reduced income dispersion among employees who left the organization compared to those who stayed.

Interpretation: Employees who left the organization earn **lower monthly incomes** than those who stayed, as reflected in both mean and median values.

3.3 Demographics & Attrition (Discrete / Continuous vs. Nominal)

We examine the demographic profile of the workforce, specifically focusing on **Age** and **Years at Company** *(Tenure)*. We compare the distributions of these variables across attrition groups to determine if younger or newer employees are at higher risk of leaving.

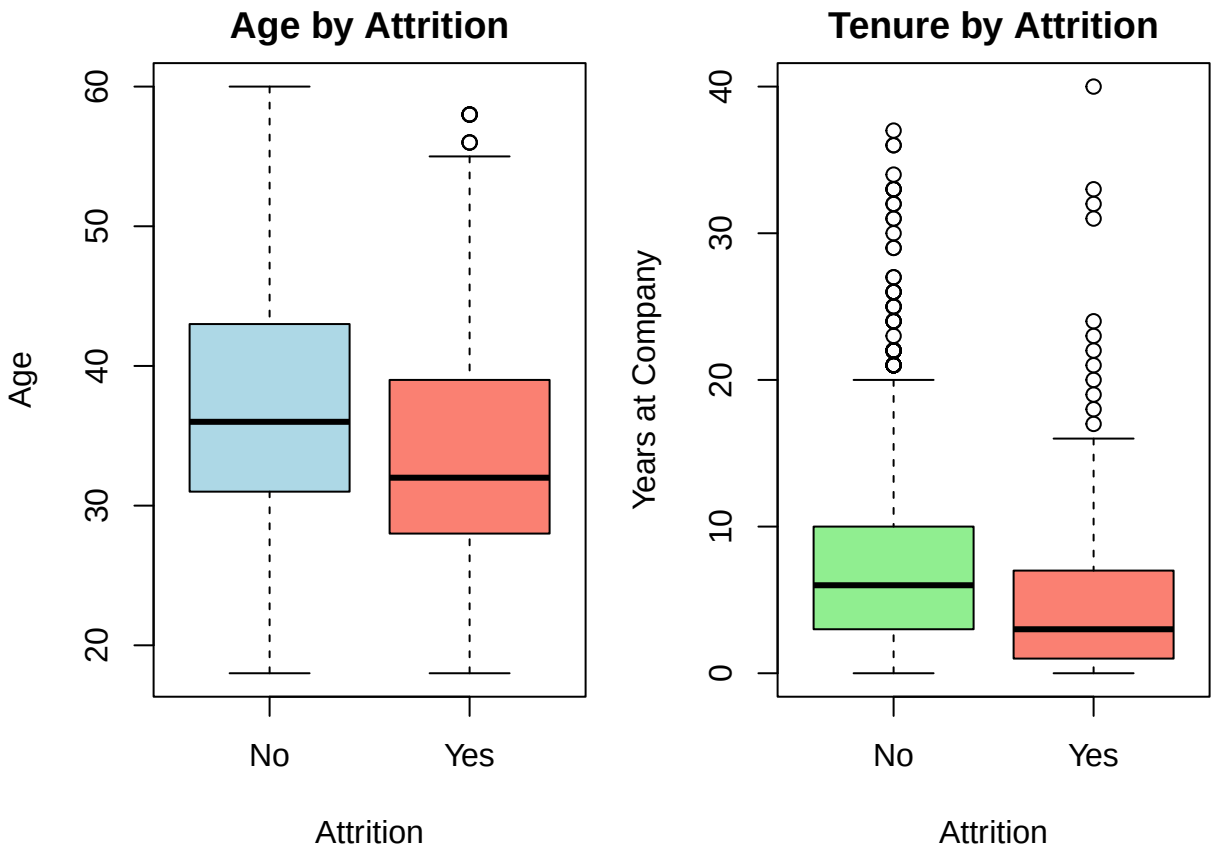

Figure 8: Demographic Drivers of Attrition: (Left) Age distribution showing a lower median age for leavers; (Right) Tenure distribution showing early-stage attrition among new hires.

Interpretation: The analysis confirms that attrition is primarily an **early-career phenomenon**. Employees who leave are generally **younger** *(median age 32 vs. 36)* and have significantly **shorter tenure** *(median 3 years vs. 6 years)* compared to those who stay. This indicates that the organization struggles most with retaining new and junior talent, rather than long-term employees.

4 Linear Regression Analysis

4.1 Correlation Analysis: Years at Company vs. Monthly Income (Discrete vs. Continuous)

Before proceeding with regression modeling, we examine the strength and nature of the relationship between **tenure** *(Years at Company)* and **compensation** *(Monthly Income)* using both **Pearson** and **Spearman** correlation coefficients. Pearson correlation measures the linear association between two continuous variables, while Spearman correlation assesses monotonic relationships based on ranked data and is more robust to outliers.

Table 12: Pearson and Spearman Correlations: Years at Company vs Monthly Income

Correlation Type	Coefficient
Pearson	0.51
Spearman	0.46

Interpretation: The correlation coefficients reveal a **positive** relationship between **tenure** and **income** that borders between **weak** and **medium** strength. The **Pearson** correlation **($r = 0.51$)** indicates a **medium** correlation *(just exceeding the 0.5 threshold)*, while the **Spearman** correlation **($\rho = 0.46$)** falls within the **weak** correlation range. The consistency of both signs confirms a positive association, justifying the exploration of a linear model, though the moderate-to-weak strength suggests variability that tenure alone cannot explain.

4.2 Linear Regression: Monthly Income vs Years at Company

We model **Monthly Income** (Y) as a linear function of **Years at Company** (X) to quantify the financial premium of tenure.

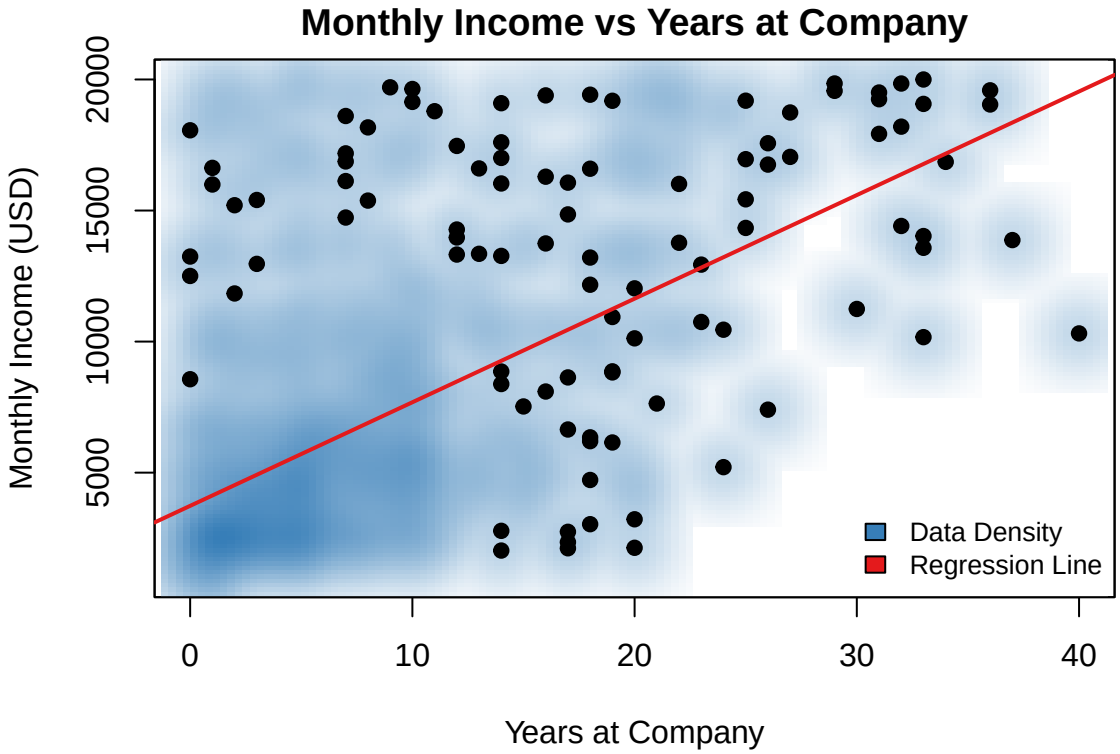

Figure 9: Scatterplot with density shading illustrating the relationship between Years at Company and Monthly Income. The red regression line indicates a positive linear trend, while the blue density shading reveals the concentration of data points, particularly in the lower tenure and income ranges.

Table 13: Linear Regression Coefficients

Predictor	Estimate	Std Error	t-Statistic	p-Value
Intercept	3733.27	160.09	23.32	< 0.001
Years at Company	395.20	17.20	22.98	< 0.001

Table 14: Model Goodness of Fit

Measure	Value
R-Squared	0.2645
Residual Std. Error (RSE)	4039.01
Observations	1,470

Regression Equation:

$$\widehat{\text{MonthlyIncome}} = 3733.27 + 395.2(\text{YearsAtCompany})$$

Interpretation: The analysis confirms a significant positive link ($p < 0.001$) where income grows by **$395/year** from a **$3,733** baseline. However, tenure explains only **26.45%** of the variance ($R^2 = 0.26$), and the **high residual error** *(4,039)* implies that compensation is largely driven by other factors like role or performance.

5 Conclusion

This report characterized the IBM workforce structure and attrition dynamics through a rigorous descriptive statistical framework. Uni-variate analysis revealed that key quantitative variables - specifically **Monthly Income** and **Years at Company** are strongly **right-skewed**, with means consistently exceeding medians. This distributional asymmetry indicates that "average" workforce metrics are inflated by a small cohort of high-tenure, high-income outliers, suggesting that median values serve as more robust central estimators for this population.

Bivariate investigation identified statistically distinct profiles for attrition. Contingency table analysis quantified the risk of low morale: employees with 'Low' job satisfaction exhibited a conditional attrition rate of **28%**, compared to only **18%** among their retained counterparts. Demographic comparisons further contextualized turnover as an early-career phenomenon, with boxplots confirming that attrition is statistically concentrated in the lower quartiles of the **Age** *(Median: 32 vs 36)* and **Tenure** distributions

Finally, the linear regression analysis established a statistically significant **positive relationship** ($p < 0.001$) between *tenure* and *compensation*. However, the model's moderate coefficient of determination ($R^2 \approx 0.26$) and **high residual standard error** ($4,039$) demonstrate that tenure is a necessary but insufficient predictor of income. The substantial unexplained variance implies that unobserved heterogeneity-likely driven by **Job Level** or some other variables we did not include - plays a dominant role in the organization's compensation structure.

References

IBM Watson Analytics. Ibm hr analytics employee attrition & performance. Kaggle, 2017. URL https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/smoothScatter.html. R package version 4.x.

H. Wickham, W. Chang, L. Henry, T. L. Pedersen, K. Takahashi, C. Wilke, K. Woo, H. Yutani, D. Dunnington, and T. van den Brand. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*, 2024. URL https://ggplot2.tidyverse.org/reference/index.html. R package.

Y. Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2024. URL https://rdocumentation.org/packages/knitr/topics/kable. R package.