HUMAN CAPITAL ANALYTICS

Project 1 – 6010 Practicum Analytics

Subash Yadav

By: Subash Yadav | TO: JP WANG

Table of Contents

Executive Summary

This project focuses on understanding and addressing employee turnover through advanced data-driven approaches. By analyzing key factors influencing turnover and segmenting employees into actionable groups, the study provides valuable insights for retention strategies. The primary business goals included identifying key drivers of turnover, such as job dissatisfaction, workload imbalance, and lack of promotions, and proactively retaining at-risk employees through targeted interventions. A comprehensive dataset of 14,999 employees was analyzed using methods like hypothesis testing, feature engineering, predictive modeling, and clustering.

For classification, models such as Logistic Regression, Decision Tree, and Random Forest were utilized to predict turnover likelihood. Random Forest achieved the highest accuracy (97.71%), identifying job satisfaction, tenure, workload, and salary as critical predictors. Clustering with K-means segmented stayed employees into three groups based on satisfaction, tenure, and workload. Cluster 3, with low satisfaction and high tenure, represented high-risk employees requiring immediate retention efforts.

Retention strategies were tailored to cluster insights: high-risk employees need workload redistribution and career growth opportunities, medium-risk employees require engagement programs, and low-risk employees should be incentivized to sustain their satisfaction. The action plan emphasizes data-driven interventions, workload monitoring, and professional development to improve engagement and reduce turnover.

This study provides a structured framework to foster a committed workforce, reduce turnover costs, and enhance productivity by addressing employee-specific needs effectively.

Business Goal

Primary Goal 1: Identify and Address Key Drivers of Employee Turnover

Objective: Use data-driven methods to uncover and address the primary factors influencing employee turnover. This will involve:

1. **Identifying Key Drivers of Turnover:**
    - Analyzing employee data to pinpoint critical turnover factors, such as job dissatisfaction, salary disparities, limited promotion opportunities, and tenure-related challenges.
    - Evaluating the direct and indirect impacts of these factors to derive actionable insights.
2. **Developing Actionable Strategies:**
    - Formulating specific retention initiatives to mitigate identified turnover causes.
    - Improving job satisfaction, defining clear career paths, and ensuring competitive and equitable compensation.

Primary Goal 2: Proactively Retain At-Risk Employees

Objective: Leverage predictive analytics to identify and support employees at high risk of leaving, enabling early interventions. This will involve:

1. **Predicting Turnover Risk:**
    - Building predictive models to identify employees most likely to leave based on key factors such as job satisfaction, promotion history, and workload.
    - Facilitating timely identification of at-risk employees.
2. **Implementing Retention Strategies:**
    - Empowering HR and managers with actionable insights to engage and support at-risk employees.
    - Developing tailored intervention plans to improve satisfaction, engagement, and retention.

By achieving these goals, the organization aims to foster a committed workforce, reduce turnover costs, and enhance long-term organizational productivity.

Analytical Goal

Primary Analytical Goal: Utilize Data Analytics to Address Key Factor for Employee Turnover and Enhance Retention

Objective: Apply advanced data-driven techniques to understand, predict, and address employee turnover effectively. This will involve:

1. **Understanding Key Turnover Drivers:**
    - Perform exploratory data analysis (EDA) to uncover patterns and relationships in employee data, focusing on factors like job satisfaction, salary, promotions, and workload.
    - Validate hypotheses using statistical methods to confirm the significance of identified turnover drivers and their impact.
2. **Developing Predictive Models:**
    - Build and refine classification models (e.g., logistic regression, decision trees, random forests) to predict the likelihood of employee turnover.
    - Use model evaluation metrics (accuracy, precision, recall, F1-score, and AUC-ROC) to ensure reliable identification of at-risk employees.

3. **Feature Engineering and Optimization:**
    - Create derived variables (e.g., workload balance, satisfaction-evaluation interaction) to capture complex relationships and improve model performance.
    - Apply feature selection and dimensionality reduction techniques (e.g., LASSO) to enhance model interpretability and efficiency.
4. **Clustering and Segmentation:**
    - Segment employees based on shared characteristics (e.g., satisfaction levels, tenure, performance) using clustering methods to tailor retention strategies.

5. **Actionable Insights and Visualization:**
    - Generate clear, actionable insights from the data analysis, linking key findings to business goals.
    - Visualize results through dashboards and reports to enable HR and managers to implement targeted interventions.

By achieving these analytical goals, the organization will gain a comprehensive understanding of turnover dynamics, improve retention strategies, and make data-informed decisions to maintain a motivated and productive workforce.

Analytical Approaches

Data Exploration and Preparation

The first step involves understanding the dataset's structure, identifying potential quality issues, and preparing the data for subsequent analysis.

- **Exploratory Data Analysis (EDA):** Techniques such as summary statistics, correlation matrices, and data visualizations will be used to explore the distribution and relationships among variables.
- **Data Cleaning and Transformation:** Missing values and outliers will be addressed to ensure data integrity. Categorical variables (e.g., salary levels, department) will be encoded appropriately, while new features will be derived to better capture employee dynamics, such as "workload balance" and "engagement index."

Hypothesis Testing and Key Driver Analysis

A hypothesis-driven approach will validate assumptions about turnover factors and uncover actionable insights.

- **Testing Key Hypotheses:** Statistical tests (e.g., t-tests, chi-square tests) will be used to confirm the significance of variables such as salary, job satisfaction, and promotion history.
- **Identifying Key Drivers:** Correlation analysis and feature importance methods will quantify the impact of these variables on turnover, enabling the identification of critical factors that influence employee decisions.

Predictive Model Development

Predictive models will be designed to identify employees at high risk of leaving, providing a basis for early interventions.

- **Model Selection:** A combination of machine learning models (e.g., logistic regression, decision trees, random forests) will be developed to classify employees based on turnover likelihood.

- **Evaluation Metrics:** Models will be assessed using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to ensure reliable predictions.
- **Feature Optimization:** Techniques such as LASSO regression and recursive feature elimination will be applied to enhance model interpretability and performance.

Employee Segmentation
To tailor retention strategies, employees will be segmented into distinct groups based on shared characteristics.

- **Clustering Analysis:** Methods like K-means clustering will identify natural groupings of employees based on factors such as tenure, satisfaction levels, and workload.
- **Segment Insights:** Each segment will be analyzed to understand unique challenges and opportunities for targeted interventions.

Actionable Insights and Visualization
The final step focuses on deriving meaningful insights and communicating results effectively.

- **Data Visualization:** Visualizations will highlight key patterns and predictive outcomes, allowing stakeholders to interpret findings intuitively.
- **Strategic Recommendations:** Insights from the analysis will be translated into specific, actionable retention strategies, such as improving compensation for underpaid employees or addressing workload imbalances.

This analytical approach ensures a comprehensive understanding of turnover dynamics, enabling the organization to make informed decisions that align with business goals.

Dataset Explanation
The analysis is based on the dataset "Employee.csv," which contains data on 14,999 employees. The dataset provides various attributes related to employees' job satisfaction, compensation, career growth, and other factors that could influence their decision to stay with or leave the company. Below is a detailed explanation of the dataset's key variables:

Variable	Description	Type	Importance
satisfaction_level	Employee's job satisfaction level (0 to 1 scale)	Numerical	Key for understanding job satisfaction's impact on turnover
last_evaluation	Score of the employee’s most recent performance evaluation (0 to 1 scale)	Numerical	Assesses if performance outcomes correlate with turnover
number_project	Total number of projects the employee has worked on	Numerical	Insight into workload and job engagement's influence on turnover
average_montly_hours	Average number of hours worked per month	Numerical	Analyzes whether workload intensity affects turnover

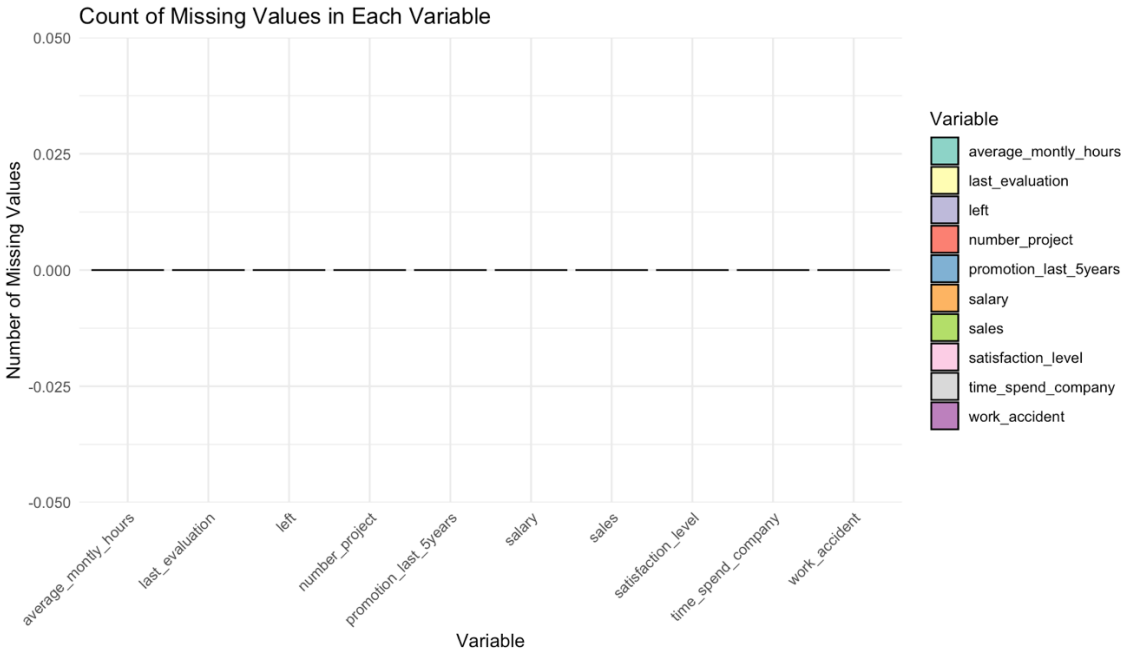
Variable	Description	Type	Importance
time_spent_company	Number of years the employee has been with the company	Numerical	Identifies tenure-related trends in turnover
work_accident	Binary indicator if the employee experienced a work accident (1 = Yes, 0 = No)	Binary	Examines if workplace safety issues contribute to turnover
left	Binary indicator if the employee has left the company (1 = Yes, 0 = No)	Binary	Target variable for predicting employee turnover
promotion_last_5years	Binary indicator if the employee was promoted in the last 5 years (1 = Yes, 0 = No)	Binary	Examines correlation between promotions and turnover
department	Department the employee works in (e.g., sales, technical, HR)	Categorical	Identifies departmental differences in turnover rates
salary	Employee's salary level (low, medium, high)	Categorical	Assesses whether compensation levels influence turnover

(Table. 1 – Dataset Variables)

The table gives a comprehensive overview of each variable's role in the dataset and its relevance to the analysis.

Data Exploration Summary

1. **Data Completeness:**
    - No missing values found in the dataset, ensuring full data coverage for analysis.



(Fig. 1 – Missing Value Plot)

The dataset was evaluated for missing values across all variables to ensure data integrity. As depicted in the plot, no missing values were identified in any of the variables. This confirms that the dataset is complete and does not require imputation or additional preprocessing steps for missing data handling. This completeness ensures reliable analysis and modeling, as no biases are introduced due to missing data imputation.

2. **Summary Statistics:**

The summary statistics provide an overview of the key numerical variables in the dataset, highlighting their distribution, central tendencies, and range:

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Satisfaction Level	0.09	0.44	0.64	0.61	0.82	1.00
Last Evaluation	0.36	0.56	0.72	0.71	0.87	1.00
Number of Projects	2.00	3.00	4.00	3.80	5.00	7.00
Average Monthly Hours	96.00	156.00	200.00	201.10	245.00	310.00
Time Spent at Company	2.00	3.00	3.00	3.49	4.00	10.00

(Table.2 – Summary Statistics)

- **Satisfaction Level**: Employee satisfaction ranges from 0.09 to 1.00, with a median value of 0.64 and an average of 0.61. This indicates that satisfaction levels are generally moderate to high among employees.
- **Last Evaluation**: Scores of employees' last performance evaluations range from 0.36 to 1.00, with a median of 0.72. The mean score of 0.71 suggests consistent performance evaluations across employees.
- **Number of Projects**: Employees have worked on a minimum of 2 and a maximum of 7 projects. The median value of 4 and the mean of 3.80 suggest that most employees handle between 3 and 5 projects.

- **Average Monthly Hours**: Monthly hours worked range from 96 to 310, with a median of 200 and an average of 201.1. This reflects a reasonable workload for most employees but also points to some outliers working very high or low hours.
- **Time Spent at the Company**: Employee tenure ranges from 2 to 10 years, with a median of 3 years and an average of 3.49 years. Most employees have relatively short tenures, suggesting potential turnover trends or rapid employee cycling.

Variable	Levels	Count per Level
Work Accident	0, 1	0: No, 1: Yes
Left (Turnover)	0, 1	0: Employees stayed, 1: Employees left
Promotion Last 5 Years	0, 1	0: No promotion, 1: Promoted
Department	Sales, Technical, HR, etc.	Department (e.g., sales: majority)
Salary	Low, Medium, High	Low: Count (majority), Medium: Count, High: Count

(Table.2 – Frequency Distribution Statistics)

Work Accident:
- **Description**: Indicates whether an employee experienced a work accident.
- **Levels**:
  - 0: No work accident (majority).
  - 1: Experienced a work accident (minority).
- **Relevance**: Helps evaluate if workplace safety contributes to employee turnover.

Left (Turnover):
- **Description**: Indicates whether the employee has left the company.
- **Levels**:
  - 0: Employee stayed.
  - 1: Employee left.
- **Relevance**: Serves as the target variable for predictive modeling.

Promotion Last 5 Years:
- **Description**: Indicates whether the employee was promoted in the last 5 years.
- **Levels**:
  - 0: Not promoted (majority).
  - 1: Promoted (minority).
- **Relevance**: Evaluates the impact of promotions on employee retention.

Department (Sales):
- **Description**: Represents the department or role of the employee (e.g., sales, technical, HR).
- **Levels**: Includes categories like Sales, Technical, HR, and others.
- **Relevance**: Analyzes turnover trends across departments.

Salary:
- **Description**: Indicates the salary level of the employee.
- **Levels**:
  - Low: Majority of employees.
  - Medium: Moderately represented.

    - High: Minority of employees.
- **Relevance**: Examines the correlation between compensation and turnover.

These statistics establish a baseline understanding of the dataset's numerical variables and will guide further exploratory and predictive analyses.

Hypothesis Testing

Visual Insights for hypothesis:

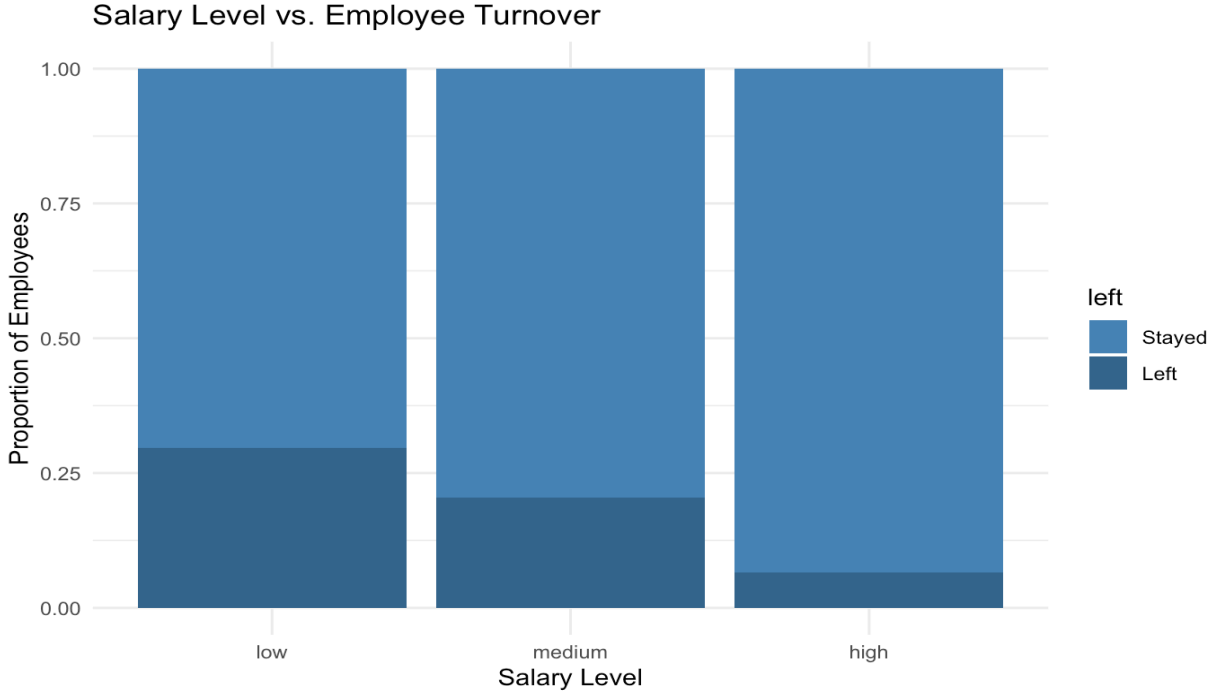
Hypothesis 1: Salary Is the Reason Why Employees Left the Company

- To test the hypothesis that salary plays a significant role in employee turnover, a Pearson's Chi-Square test was conducted to examine the association between salary levels (low, medium, high) and whether employees left the company (left = 1) or stayed (left = 0).

Chi-Square Test Results:

- **Chi-Square Statistic (X-squared):** 381.23
- **Degrees of Freedom (df):** 2
- **p-value:** < 2.2e-16

The Chi-Square test reveals a **highly significant** p-value (less than 0.05), meaning that the observed differences in employee turnover across different salary levels are **statistically significant**. This supports the hypothesis that salary levels are strongly associated with employee turnover.



(Fig. 2 – Salary Level v/s Turnover)

Salary	Total_Employees	Employees_left	Proportion_left
Low	7316	2172	0.297
Medium	6446	1317	0.204
High	1237	82	0.0663

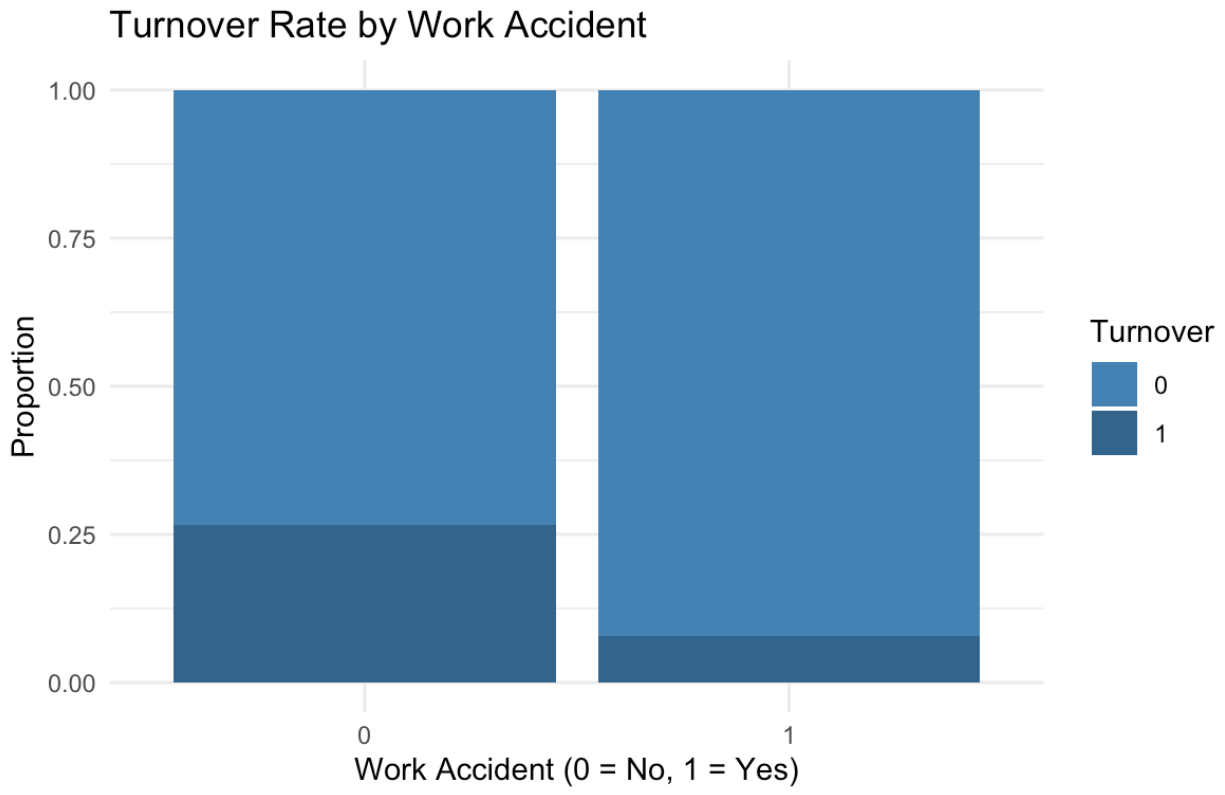
(Table 4 – Salary Level v/s Turnover)

Interpretation:

- The bar plot shows that a significantly higher proportion of employees with a **low salary** have left the company compared to those with **medium** or **high salaries**. This suggests that employees with lower salaries are more likely to leave, supporting the hypothesis that salary is a significant factor in turnover.
- Conversely, almost all employees with **high salaries** have stayed, indicating that higher compensation is associated with better retention.
- Both the **graphical representation** and the **Chi-Square test** strongly support the hypothesis that salary is a key factor in employee turnover. Employees in the low-salary category are much more likely to leave, while those in higher salary brackets are more likely to stay, reinforcing the importance of competitive compensation in retaining talent.

Hypothesis 2: Employees Leave the Company Because Work Is Not Safe

- To test this hypothesis, used a stacked bar plot (Figure 2) to show the proportion of employees who left and stayed based on whether they experienced a work accident (work_accident).



(Fig. 3 – Work Accident v/s Turnover)

Work_accident	Total_Employees	Employee_Left	Employee_Stayed	Proportion_Left
0	12830	3402	9428	0.265
1	2169	169	2000	0.0779

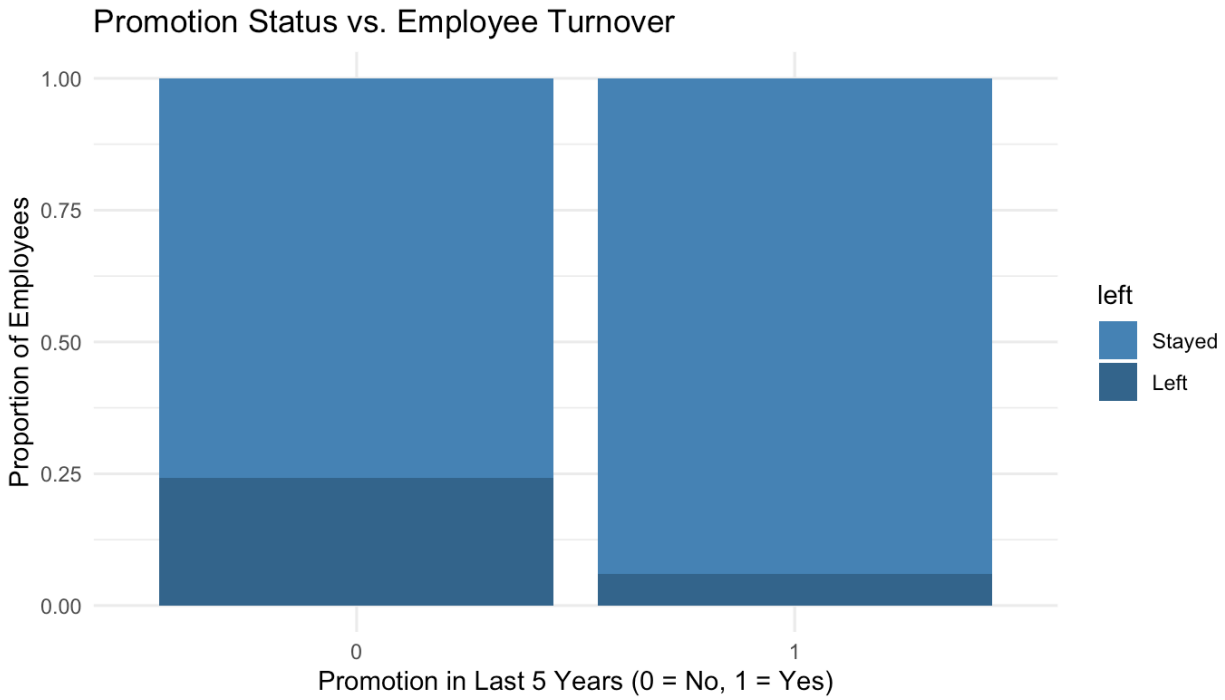
(Table 5 – Work Accident v/s Turnover)

Interpretation:

- The data shows that **employees who experienced a work accident (work_accident = 1) have a relatively low proportion of turnover**, with only **7.79%** of them leaving the company. In contrast, **26.5%** of employees who did not experience a work accident (work_accident = 0) have left the company.
- The plot shows that the proportion of employees who left after experiencing a **work accident** (work_accident = 1) is only slightly higher than those who stayed, and most employees who left did not experience a work accident (work_accident = 0).
- This indicates that the **majority of employees who left did not experience a work accident**, and the proportion of employees leaving after a work accident is only marginally higher than those who stayed.
- Therefore, the data **does not provide strong evidence** to support the hypothesis that **unsafe work conditions** are a significant driver of turnover. While safety concerns may have some influence, they are not a major factor contributing to employees leaving the company. Most employees who left had not experienced a work-related accident, suggesting other factors are more likely to be the primary reasons for turnover.

Hypothesis 3: this company is a good place to grow professionally.

- To test this hypothesis, created a bar plot (Figure 6) showing the proportion of employees who left and stayed based on their promotion status over the last five years (promotion_last_5years) and provided with tabular data to prove the hypothesis.



(Fig. 4 – Promotion Status v/s Employee turnover)

Promotion_last_5years	Total_Employees	Employee_Left	Employee_Stayed	Proportion_Left
0	14680	3552	11128	0.242
1	319	19	300	0.0596

(Table 6 – Promotion Status v/s Employee Turnover)

Interpretation:

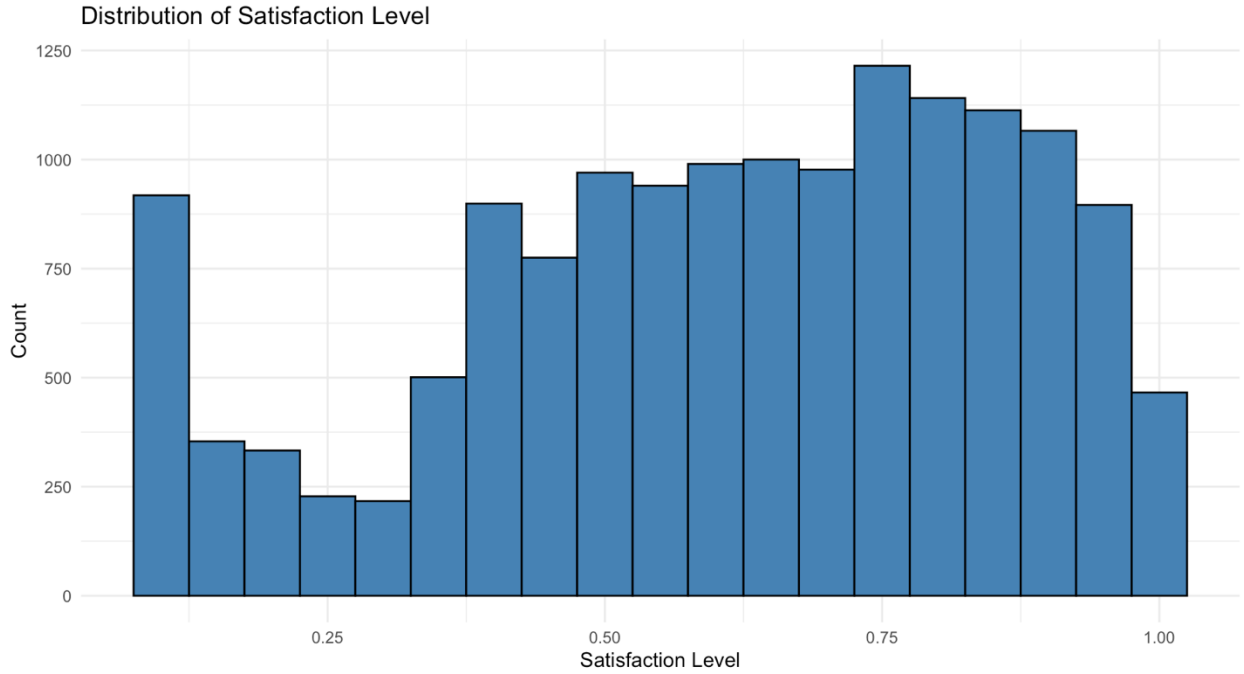
- The data in the table shows that **24.2% of employees who did not receive a promotion** in the last five years (promotion_last_5years = 0) have left the company, whereas only **5.96% of employees who received a promotion** (promotion_last_5years = 1) have left.
- The plot indicates that a higher proportion of employees who **did not receive a promotion** in the last five years (promotion_last_5years = 0) left the company compared to those who received a promotion (promotion_last_5years = 1).
- This supports the hypothesis that a **lack of promotions and career advancement opportunities** is a key factor in employee turnover. Employees who perceive limited opportunities for professional growth are more likely to leave the company.

Data Visualization

Distribution of Satisfaction Level

The distribution of employee satisfaction levels was analyzed using a histogram. The data was divided into bins with an interval of 0.05 to observe the spread and concentration of satisfaction levels across employees.

The histogram below represents the satisfaction levels of employees in the dataset.



(Fig. 5 – Distribution of Satisfaction Level)

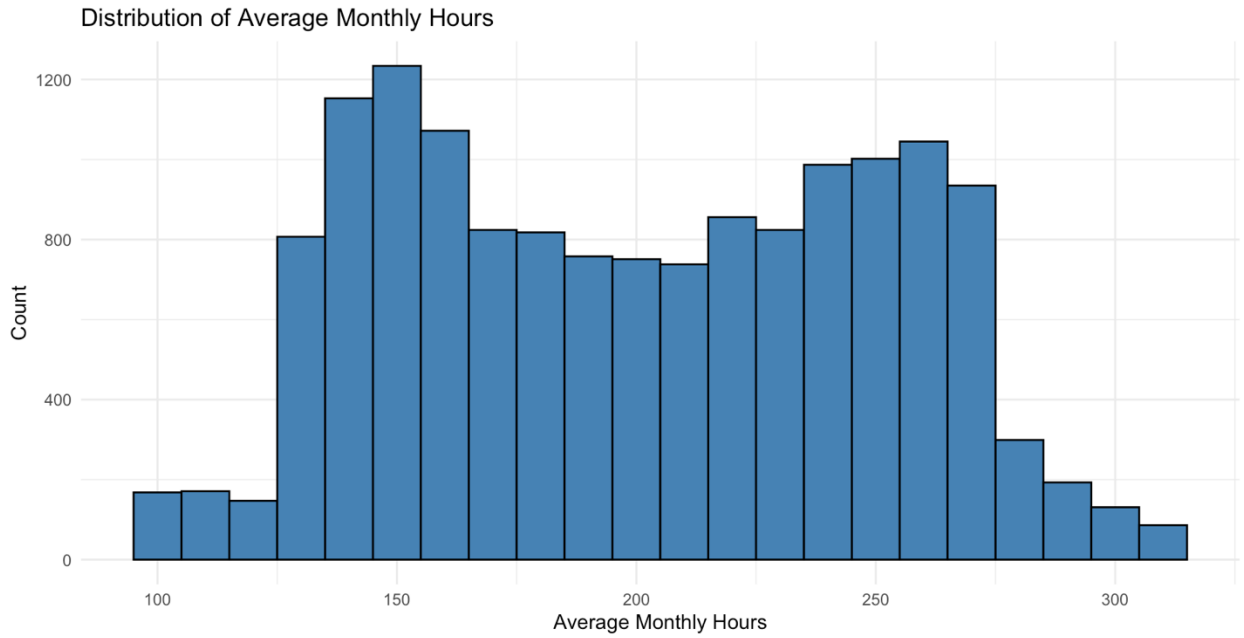
Interpretation:

1. **Low Satisfaction (0.0 - 0.4)**:
    - A notable portion of employees falls into this range, indicating dissatisfaction or disengagement.
2. **Moderate Satisfaction (0.4 - 0.7)**:
    - This range has a consistent number of employees, representing moderately satisfied workers.
3. **High Satisfaction (0.7 - 1.0)**:
    - A significant peak is observed around 0.75, showing that many employees report high satisfaction.

The presence of both highly satisfied and dissatisfied groups highlights areas to focus on—either to sustain satisfaction or address disengagement.

Distribution of Average Monthly Hours

The distribution of average monthly hours worked by employees was analyzed using a histogram.

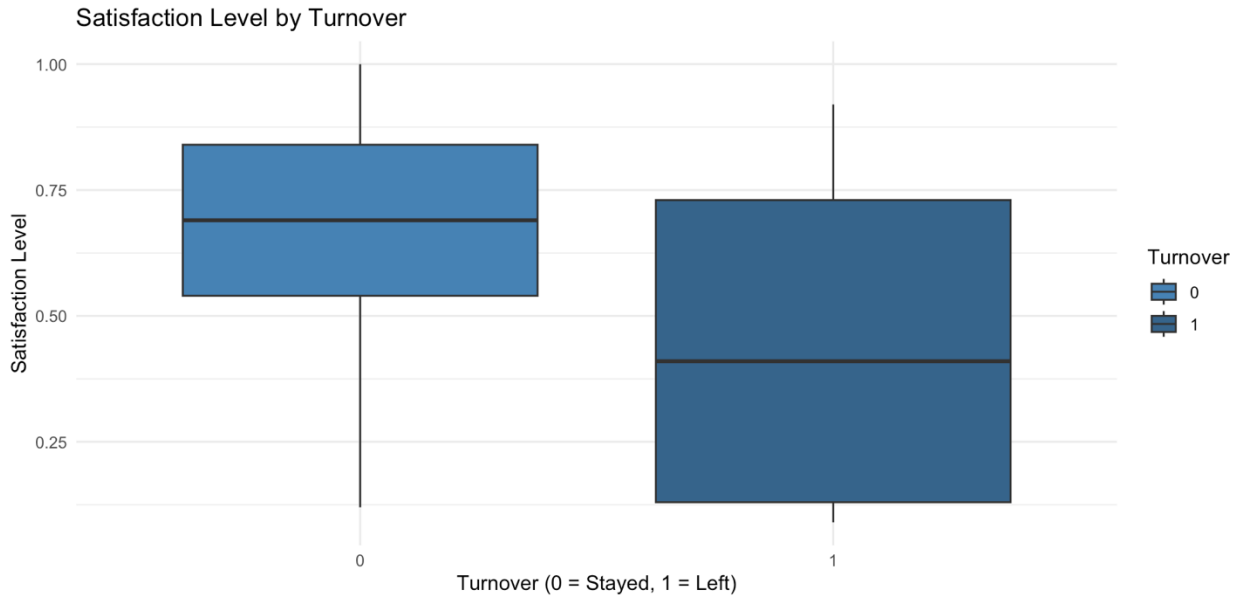


(Fig. 6 – Distribution of Average Monthly Hours)

The plot shows the distribution of average monthly hours worked by employees. The distribution is **right-skewed**, with a noticeable concentration of employees in the range of **150-250 hours**, indicating this as the standard workload.
Fewer employees work **less than 150 hours**, suggesting part-time roles or reduced workloads, while a significant number of employees work **more than 250 hours**, highlighting a subset with high workloads or consistent overtime.

Satisfaction Level by Turnover

A **boxplot** was created to compare the distribution of satisfaction levels between employees who stayed (left = 0) and those who left (left = 1). This analysis helps to identify differences in satisfaction levels between the two groups.



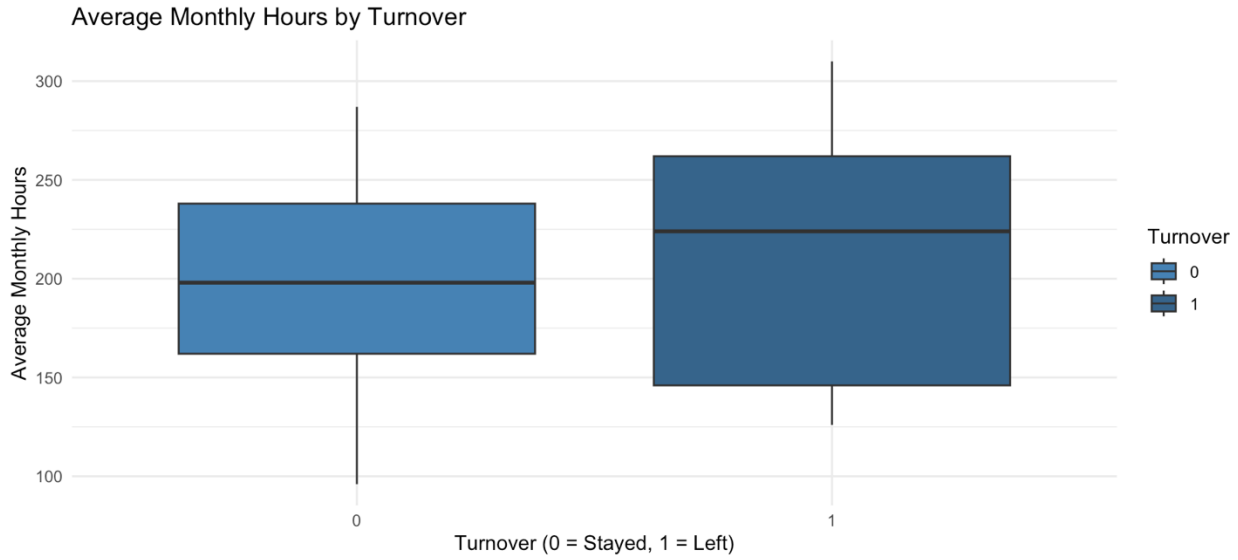
(Fig. 7 – Satisfaction level by turnover)

1. **Stayed (Turnover = 0):**
    - The median satisfaction level is relatively high, around 0.75.
    - Most satisfaction levels are concentrated between 0.65 and 0.85, with fewer outliers below 0.5.
2. **Left (Turnover = 1):**
    - The median satisfaction level is significantly lower, around 0.45.
    - A wider range of satisfaction levels is observed, with many employees reporting satisfaction levels as low as 0.2 and a few as high as 0.8.

Employees who left the company generally had lower satisfaction levels compared to those who stayed. This suggests that dissatisfaction may play a key role in turnover decisions.

Average Monthly Hours by Turnover

A **boxplot** was created to examine the distribution of average monthly hours worked by employees who stayed (left = 0) versus those who left (left = 1). This visualizes whether workload differences correlate with turnover.



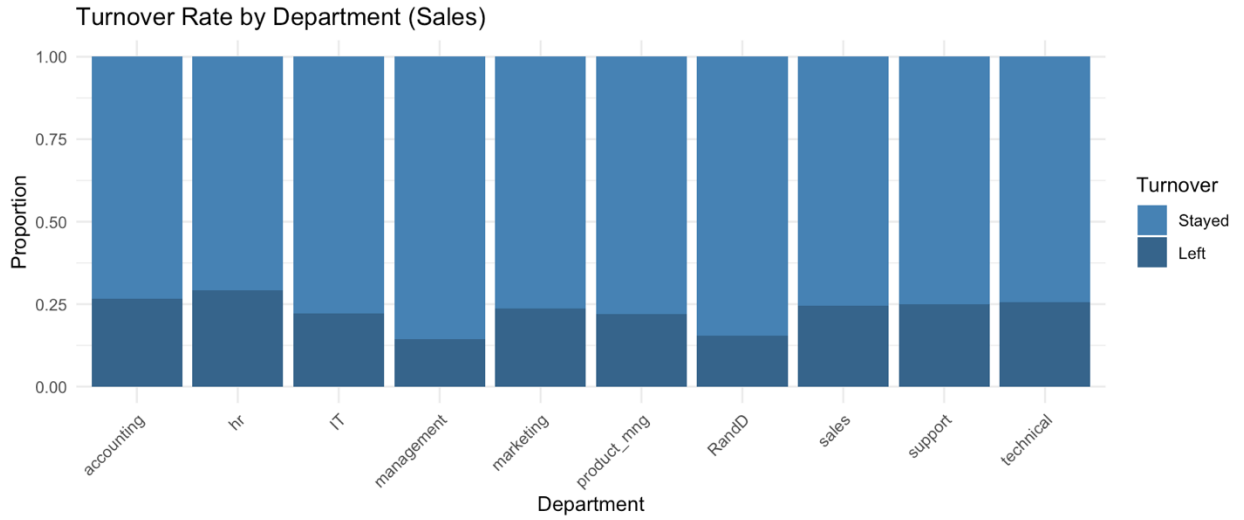
(Fig. 8 – Average Monthly Hours by turnover)

1. **Stayed (Turnover = 0):**
    - The median average monthly hours are approximately 200.
    - Most employees worked between 150 and 250 hours per month.
    - There are very few outliers outside this range.
2. **Left (Turnover = 1):**
    - The median average monthly hours are higher, around 250.
    - A broader range is observed, with many employees working more than 250 hours and some close to the maximum of 300.
    - Higher workloads are evident among employees who left compared to those who stayed.

Employees who left the company tended to work significantly more hours on average compared to those who stayed. This suggests that excessive workload may be a potential driver of turnover.

Turnover Rate by Sales (Department)

A **stacked bar chart** was created to compare the **turnover rates** across different departments. The proportions of employees who stayed (Turnover = Stayed) and left (Turnover = Left) are visualized for each department, showing department-specific trends in turnover.



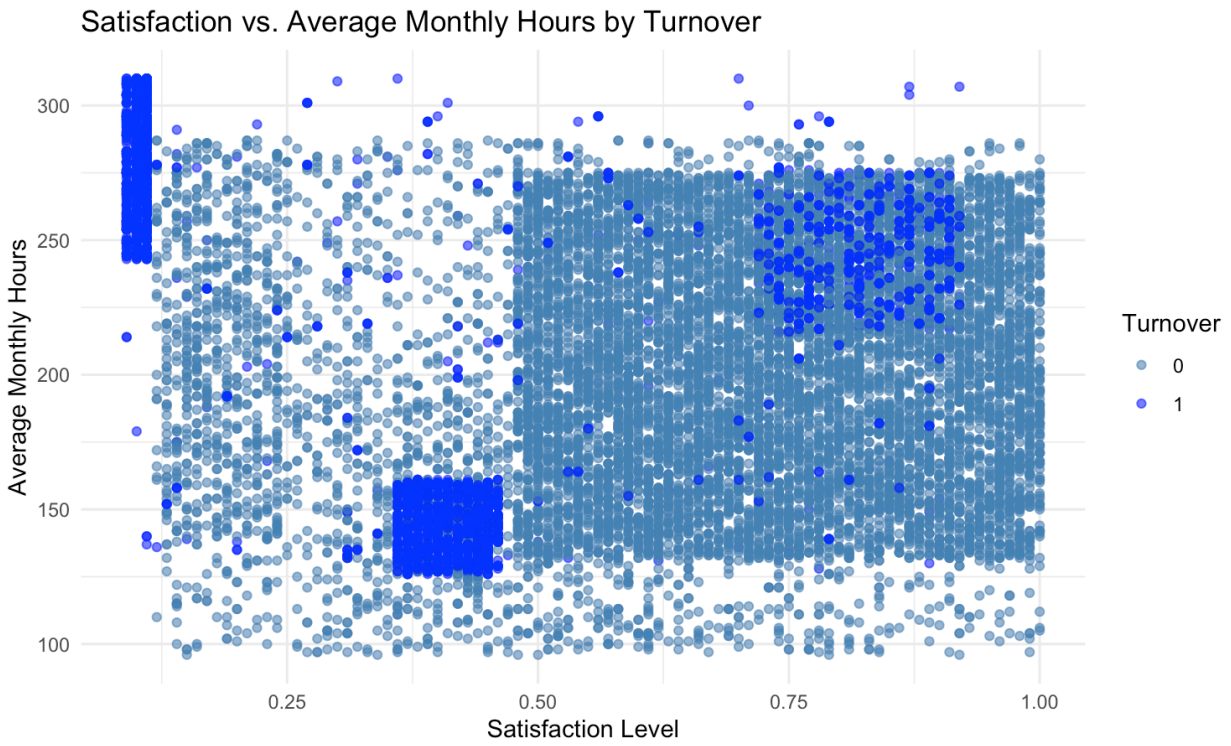
(Fig. 9 – Sales (Department) by turnover)

1. **High Turnover Departments:**
   - Some departments, such as **sales** and **support**, exhibit higher proportions of employees leaving compared to other departments.
   - Sales appears to have a relatively higher turnover rate than technical or management.
2. **Low Turnover Departments:**
   - Departments such as **management** and **R&D** have lower proportions of employees leaving, indicating better retention.

Certain departments, especially sales and support, seem to struggle with employee retention. Targeted interventions for these departments may help reduce turnover.

Satisfaction Vs. Average Monthly Hours by Turnover

A **scatterplot** was created to examine the relationship between **satisfaction level** (X-axis) and **average monthly hours** (Y-axis), with turnover (0 = stayed, 1 = left) differentiated by color. Each dot represents an employee.



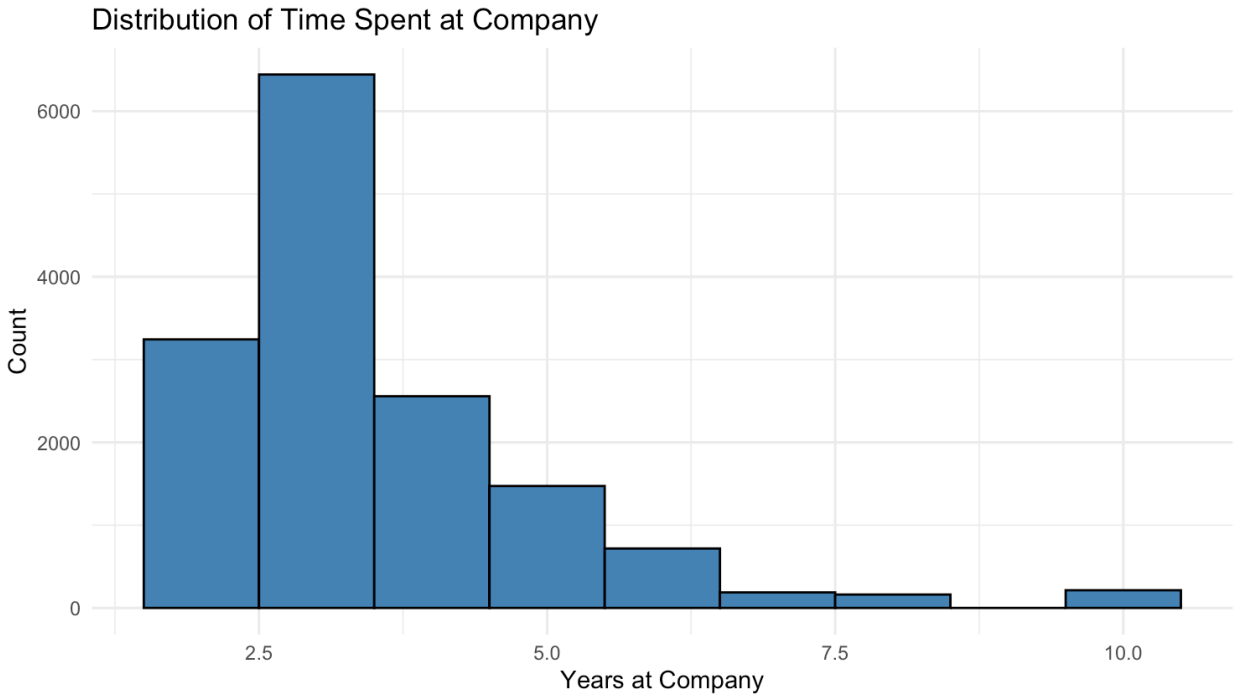
(Fig. 10 – Satisfaction Vs. Average Monthly Hours by Turnover)

1. **High Turnover Zones**:
    - Employees with **low satisfaction levels (<0.4)** and **high monthly hours (>250)** exhibit higher turnover.
    - A cluster of turnover is also visible for **low satisfaction and low monthly hours**.
2. **Stable Zones**:
    - Employees with **moderate to high satisfaction (≥0.6)** generally have a lower turnover rate, irrespective of monthly hours.

Turnover risk is notably higher for employees who feel dissatisfied and either overworked or underutilized, indicating these factors play a crucial role in employee retention.

Distribution of Time Spent at Company

A **histogram** was plotted to visualize the distribution of the number of years employees have spent at the company.



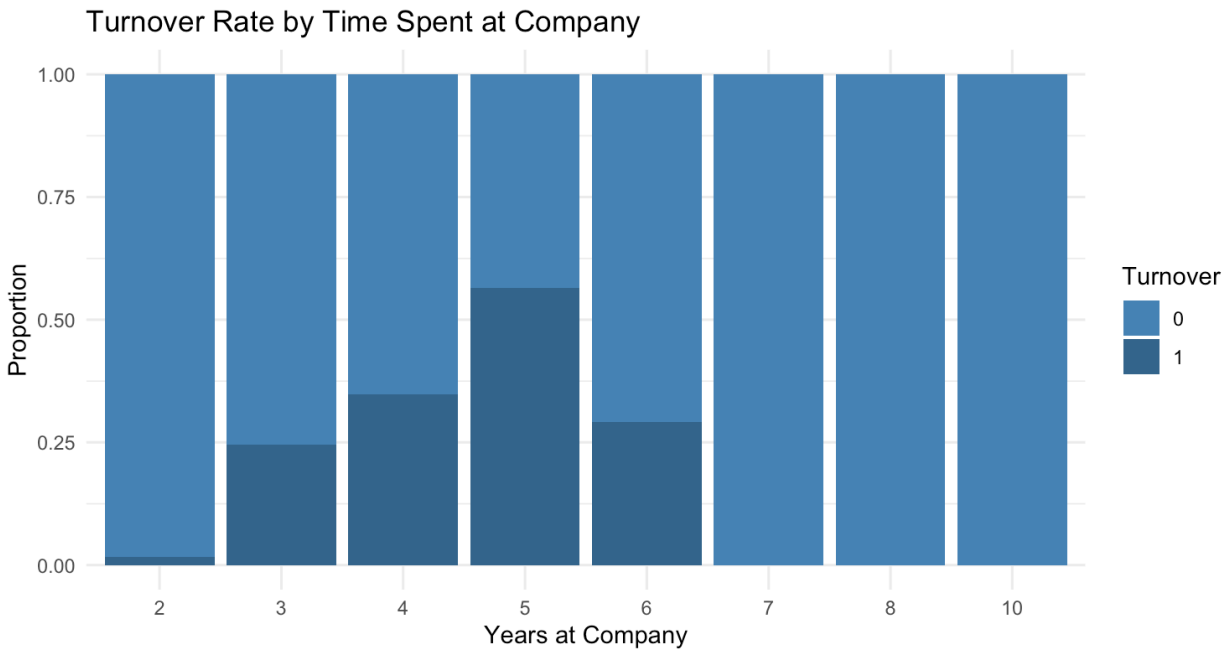
(Fig. 11 – Distribution of time Spent at Company)

1. **Short Tenure Dominates**:
    - Most employees have spent **2 to 3 years** at the company, with the highest peak at 3 years.
    - A rapid decline in employee counts is observed after 5 years of tenure.
2. **Long-Tenure Employees Are Rare**:
    - Few employees have stayed longer than **7 years**, indicating potential issues with long-term retention.

The company has a workforce primarily composed of short-tenured employees. The declining counts for longer tenures suggest potential turnover issues for employees staying beyond 5 years. Strategies to retain experienced employees may be necessary.

Time Spent at Company Vs Turnover

A **stacked bar plot** was created to display the proportion of employees who stayed versus those who left based on the number of years spent at the company.



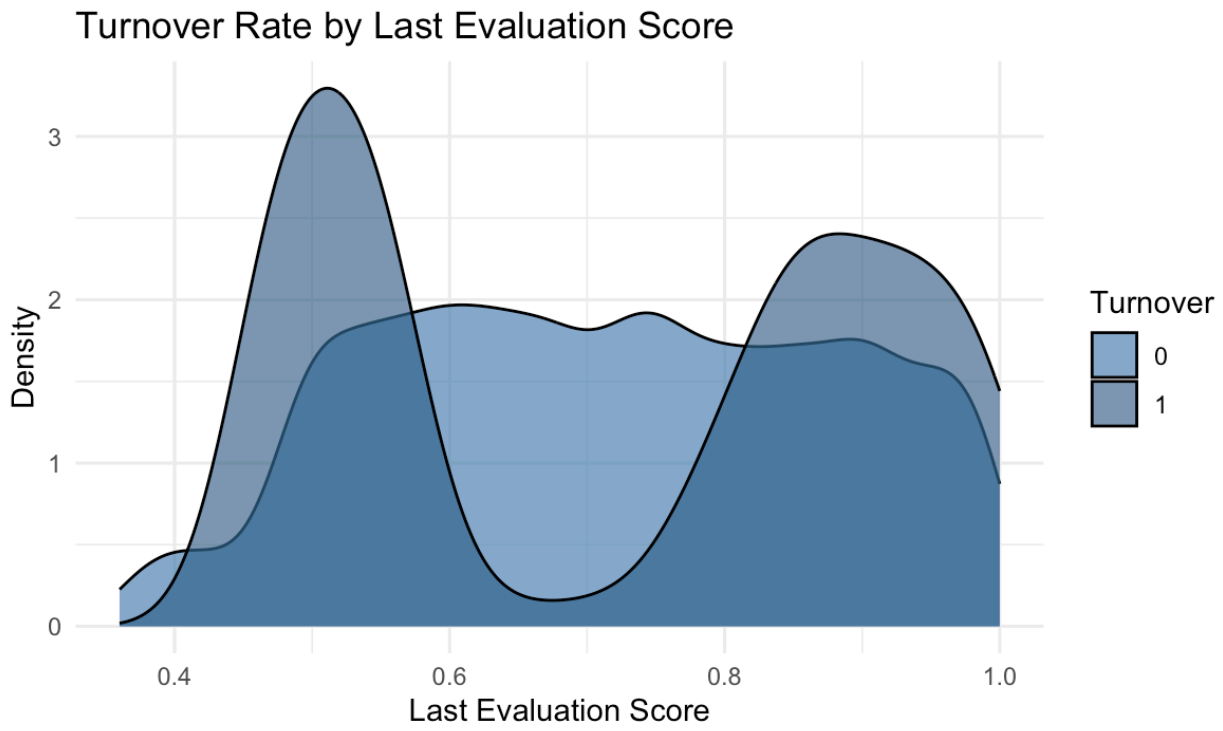
(Fig. 12 – Time Spent at Company VS Turnover)

1. **Short Tenure (2-3 years)**:
    - Employees with short tenure exhibit a low turnover rate, as most of them stay with the company.
2. **Mid-Tenure (5-6 years)**:
    - A significant spike in turnover is observed for employees with **5 years** at the company, showing the highest proportion of employees leaving.
    - Turnover rates remain high for **6 years**, highlighting a critical retention challenge.
3. **Long Tenure (>7 years)**:
    - Employees with tenure beyond **7 years** have very low turnover, suggesting greater retention among long-tenured staff.

Retention efforts need to focus on mid-tenure employees (5-6 years), as this group has the highest turnover rates. This might be due to stagnation, limited growth opportunities, or burnout.

Turnover Rate by Last Evaluation Score

A **density plot** was created to visualize the turnover rate based on employees' last evaluation scores. The plot shows the distribution of evaluation scores for employees who stayed versus those who left.



(Fig. 13 – Turnover by Last Evaluation Score)

1. **High Evaluation Scores (0.8-1.0)**:
    - Employees with very high evaluation scores show a **higher density of turnover**, possibly due to feeling overburdened, undervalued despite high performance, or burnout.
2. **Moderate Scores (0.5-0.7)**:
    - Employees in this range have the **lowest turnover rates**, suggesting that moderate evaluations might reflect balanced workloads and satisfaction.
3. **Low Evaluation Scores (0.3-0.4)**:
    - A secondary peak in turnover is observed for employees with low evaluation scores, potentially due to performance issues or dissatisfaction with roles.

Turnover is **bimodal**: Employees with **low** and **high evaluation scores** are more likely to leave, whereas employees with **moderate scores** are more likely to stay. Retention strategies should focus on both high performers (to address burnout) and low performers (to improve satisfaction or fit).

Correlation

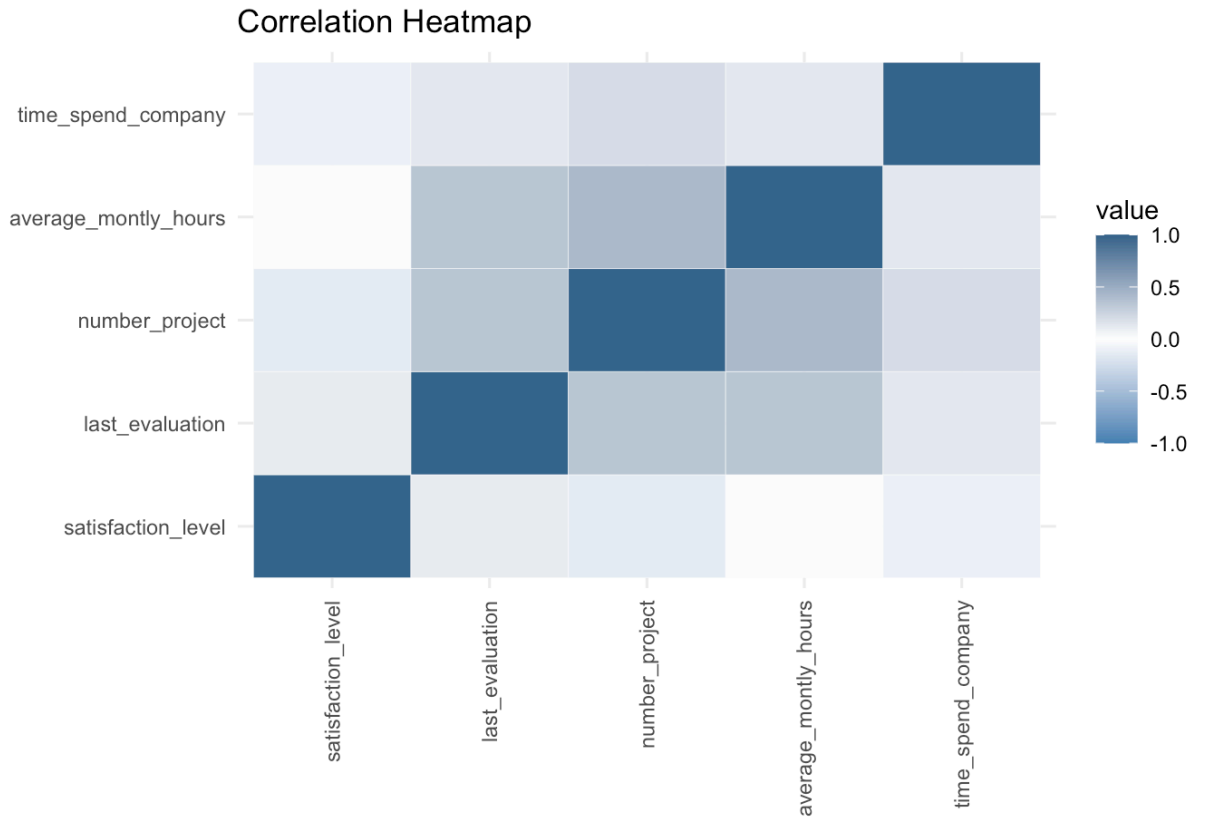
A **correlation heatmap** was generated to analyze relationships between numerical variables in the dataset, providing insights into how these variables are interrelated. The correlation matrix shows values ranging from -1 to 1, where:

- **+1** indicates a perfect positive correlation.
- **0** indicates no correlation.
- **-1** indicates a perfect negative correlation.

Variable	Satisfaction Level	Last Evaluation	Number of Projects	Average Monthly Hours	Time Spent at Company
Satisfaction Level	**1.00**	**0.11**	**-0.14**	**-0.02**	**-0.10**
Last Evaluation	**0.11**	**1.00**	**0.35**	**0.34**	**0.13**
Number of Projects	**-0.14**	**0.35**	**1.00**	**0.42**	**0.20**
Average Monthly Hours	**-0.02**	**0.34**	**0.42**	**1.00**	**0.13**
Time Spent at Company	**-0.10**	**0.13**	**0.20**	**0.13**	**1.00**

(Table. 7 – Correlation Table)

1. **Satisfaction Level**:
   - Negatively correlated with the **number of projects (-0.14)** and **time spent at the company (-0.10)**, indicating that lower satisfaction might be associated with higher workloads or longer tenure.
2. **Last Evaluation**:
   - Positively correlated with **number of projects (0.35)** and **average monthly hours (0.34)**, suggesting that higher evaluations are linked to higher workloads.
3. **Number of Projects**:
   - Strongly correlated with **average monthly hours (0.42)**, as expected since more projects typically involve longer working hours.
4. **Average Monthly Hours**:
   - Positively correlated with **time spent at the company (0.13)**, indicating that longer tenure might result in slightly higher working hours.
5. **Time Spent at Company**:
   - Weakly correlated with other variables, showing no strong association except a modest link with **number of projects (0.20)**.



(Fig. 14 – Correlation of numeric variables)

- The diagonal values are all **1**, representing perfect self-correlation.
- Significant correlations include:
    - **Number of Projects** and **Average Monthly Hours (0.42)**: Employees working on more projects also tend to have higher monthly hours.
    - **Last Evaluation** and **Number of Projects (0.35)**: Better evaluations are tied to a higher number of projects.

The heatmap highlights relationships between workload, performance, and satisfaction, providing a foundation for deeper analysis of employee behavior and potential turnover factors. These correlations can help prioritize key variables in predictive models and retention strategies.

Data Transformation

1. Binning Average Monthly Hours

The average monthly hours worked by employees were binned into three categories:

- **Low**: Hours less than 150.
- **Moderate**: Hours between 150 and 250.
- **High**: Hours greater than 250.

The original average_montly_hours variable was removed after the transformation.

- **Purpose**: Simplify analysis by categorizing employees into workload categories, allowing for easier comparisons of turnover rates.

Summary Table for Turnover by Hours Category

Hours Category	Total Employees	Employees Left	Employees Stayed	Proportion Left
High	3,202	1,242	1,960	0.388
Low	2,945	1,059	1,886	0.360
Moderate	8,852	1,270	7,582	0.143

(Table. 8 –Summary Table of Turnover by Hours Category)

(Fig. 15 – Prop. Of Employees who left by hours category)

- **Interpretation of Plot**:
  - **High Hours**: Highest turnover proportion (38.8%)—indicates overworked employees are more likely to leave.
  - **Low Hours**: Turnover proportion (36.0%) suggests underutilized employees also tend to leave.
  - **Moderate Hours**: Lowest turnover proportion (14.3%)—indicates employees with balanced workloads are less likely to leave.

Employees with high or low hours show a higher likelihood of turnover compared to those with moderate workloads.

2. Satisfaction-to-Performance Score

- **Formula Used**:

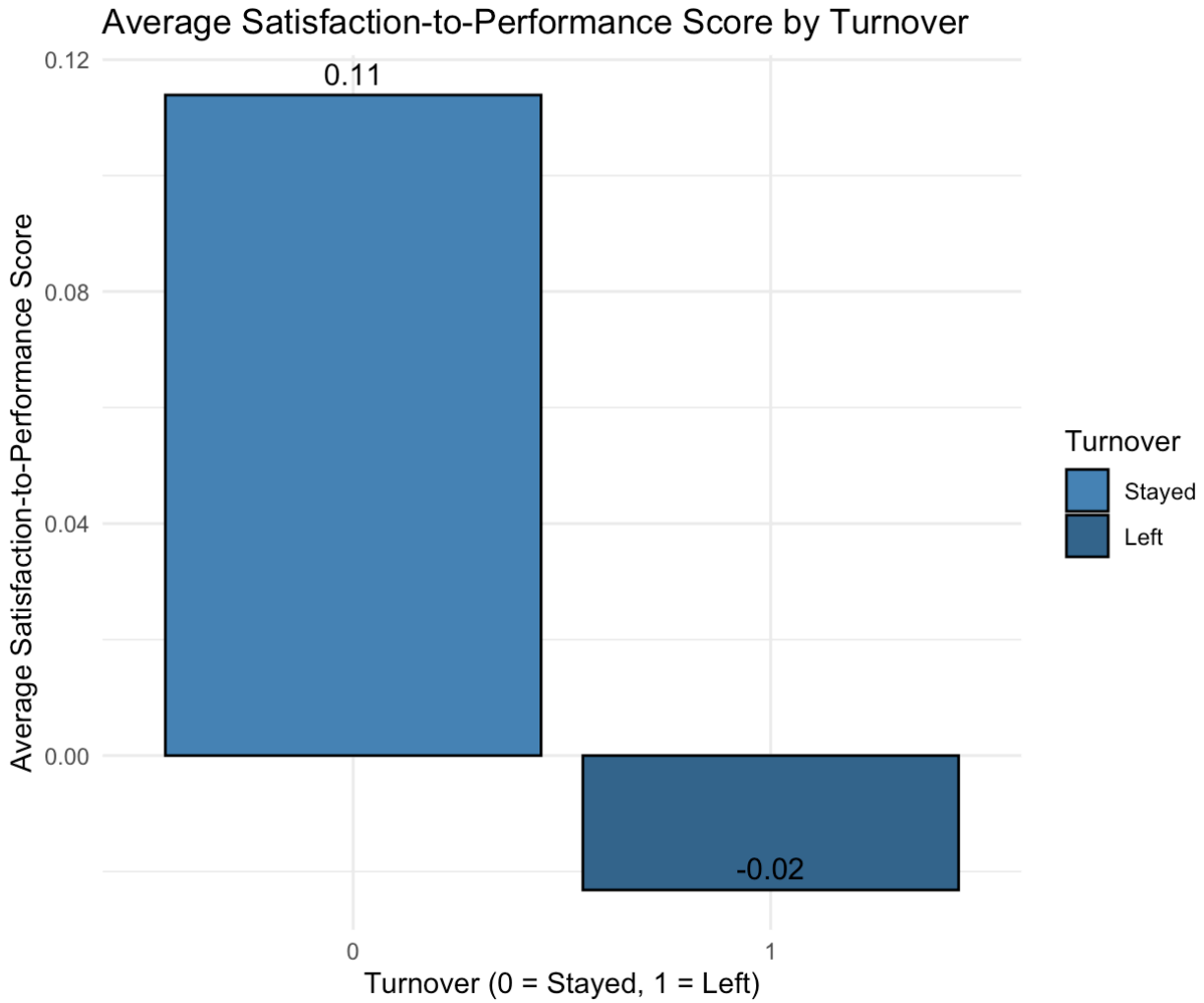
$$Satisfaction\_To\_Performance\_Score = 0.6 * Satisfaction_{level} - 0.4 * last\_evaluation$$

  - This formula assigns more weight to satisfaction levels compared to last evaluation scores.
- **Purpose**: Create a derived metric to analyze the combined effect of satisfaction and evaluation on turnover.

Summary Table for Satisfaction-to-Performance by Turnover

Turnover (Left)	Average Score	Minimum Score	Maximum Score
Stayed (0)	0.11	-0.15	0.55
Left (1)	-0.02	-0.45	0.30

(Table. 9 –Summary Table of Satisfaction_to_performance by Turnover)



(Fig. 16 – Average Satisfaction to performance score by turnover)

- **Interpretation of Plot**:
    - Employees who stayed have a significantly higher average satisfaction-to-performance score (0.11) than those who left (-0.02).
    - Lower scores for employees who left indicate dissatisfaction or poor performance evaluations may have contributed to turnover.

Employees who left the company tend to have a lower satisfaction-to-performance score, suggesting interventions to improve satisfaction could reduce turnover.

LASSO

The objective was to apply LASSO (Least Absolute Shrinkage and Selection Operator) regression for feature selection and predictive modeling. LASSO helps in penalizing irrelevant features, allowing the identification of significant predictors that influence employee turnover. The procedure involved the following steps:

1. **Data Preparation**:
    - Converted non-numeric columns to numeric or factor-encoded variables (e.g., salary, sales, work_accident, etc.).
    - Prepared the predictor matrix X and the target variable Y.
    - Removed the intercept column to avoid redundancy in the model.matrix.
2. **LASSO Model Training**:
    - Fit a LASSO regression model with cross-validation (cv.glmnet) to optimize the penalty parameter λ (lambda) for the binomial family.
    - Identified the best value of λ (lambda.min) using the model's cross-validation results.
3. **LASSO Coefficients**:
    - Extracted the coefficients of the LASSO model using the optimal λ. This step highlights the importance of each feature in predicting turnover, where non-zero coefficients indicate significant variables.
4. **Plot Cross-Validation Results**:
    - Visualized the relationship between Log(λ) and binomial deviance, allowing the identification of the penalty that minimizes prediction error.

Key Results

- **Best Lambda Value**:
    - The optimal lambda value (lambda.min) was found to be 0.000567.
- **LASSO Coefficients**:
    - The table below summarizes the coefficients obtained for each variable:

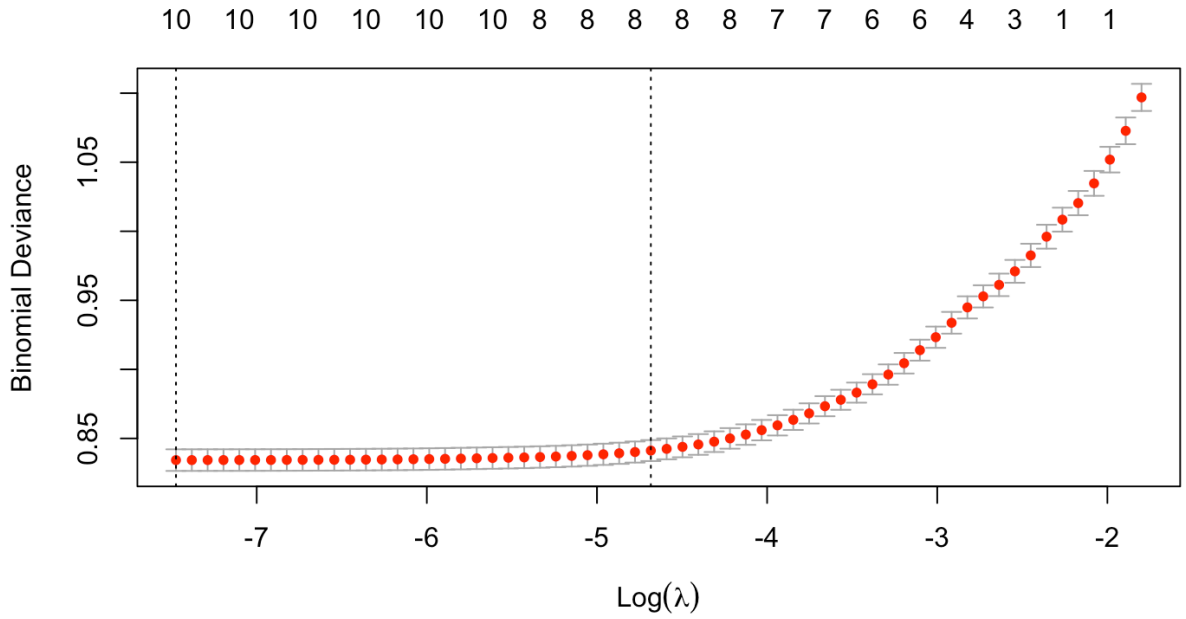
Feature	**Coefficient**
Intercept	1.989884
Satisfaction Level	-2.140467
Last Evaluation	0 (eliminated)

Feature	Coefficient
Number of Projects	-0.257999
Time Spent at Company	0.250880
Work Accident	-1.511600
Promotion in Last 5 Years	-1.408971
Sales	0.009575
Salary	-0.691590
Hours Category (Low)	-0.174644
Hours Category (Moderate)	-1.110976
Satisfaction-Performance Score	-2.694608

(Table. 10 –Summary of LASSO Result)

Interpretation of Results

1. **Important Predictors**:
   - Variables with large magnitude coefficients include:
     - Satisfaction Level (-2.140) and Satisfaction-Performance Score (-2.695): Strong negative impact on the likelihood of staying, indicating dissatisfied employees are more likely to leave.
     - Work Accident (-1.512) and Promotion Last 5 Years (-1.409): Highlight poor workplace safety and lack of growth opportunities as significant drivers of turnover.
2. **Eliminated Features**:
   - Last Evaluation had a coefficient of 0, suggesting it does not significantly influence turnover.
3. **Feature Insights**:
   - Higher satisfaction scores (both satisfaction_level and satisfaction_performance_score) are negatively correlated with turnover.
   - Employees with fewer work accidents or recent promotions are less likely to leave, as indicated by their negative coefficients.



(Fig. 17 – LASSO binomial deviance plot)

- As λ increases (moves to the right on the x-axis), more coefficients are penalized to zero, simplifying the model.
- Around Log(λ) ≈ -7, the binomial deviance is minimized, signifying the optimal balance between model complexity and performance.

This analysis indicates that a subset of features (satisfaction levels, accidents, promotions, etc.) play a critical role in predicting employee turnover. The LASSO model provides a clear pathway for actionable insights by focusing on these significant variables.

Data Preparation and Partitioning

Dataset Creation

The dataset was streamlined using predictors identified as significant through the LASSO regularization method. These variables were carefully chosen for their contribution to the prediction of employee turnover. Below is a brief description of the included variables:

- **satisfaction_level**: Measures employee satisfaction on a scale from 0 to 1.
- **satisfaction_performance_score**: A derived variable capturing the weighted relationship between satisfaction and performance evaluation.
- **work_accident**: Binary indicator (0 = No, 1 = Yes) of whether the employee experienced a workplace accident.
- **promotion_last_5years**: Binary indicator (0 = No, 1 = Yes) of whether the employee received a promotion in the last five years.
- **salary**: Categorical variable (low, medium, high), converted to a numeric factor for modeling purposes.
- **hours_category**: Categorized ranges of average monthly hours (Low, Moderate, High), replacing the original continuous average_montly_hours variable.
- **time_spend_company**: The number of years the employee has been with the company.

- **number_project**: The total number of projects the employee has worked on.

The target variable **left** (indicating turnover) was converted into a factor to facilitate classification tasks in all subsequent modeling.

Data Partitioning

To build and evaluate predictive models effectively, the dataset was split into two subsets:

- **Training Data (70%)**: Used for model building and parameter tuning.
- **Testing Data (30%)**: Used for evaluating model performance on unseen data.

The **createDataPartition() function** from the caret package was employed to split the dataset. This approach ensured **stratified sampling**, preserving the class distribution of the target variable (left) across both subsets. For example, if 20% of employees in the dataset had left the company, the same proportion was maintained in both the training and testing sets.

Key Aspects of the Partitioning Process:

1. **Reproducibility**:
   - A random seed (set.seed(123)) was set to ensure consistent partitioning across runs, enabling reproducible results.
2. **Stratification**:
   - The function stratified the data based on the target variable (left). This prevented imbalances in the training or testing sets, especially critical for datasets with an imbalanced target variable.
3. **Benefits**:
   - Ensured that both training and testing sets reflected the overall dataset's distribution.
   - Enabled fair evaluation of model performance on the testing data without bias introduced by skewed sampling.

The result of this partitioning process was a well-balanced dataset that facilitated robust model training and accurate evaluation. This step ensured that models were trained on a representative sample of the data while also being tested on a subset reflective of real-world scenarios.

Modelling: Classification and Clustering

Classification

For the classification task, models were implemented to predict employee turnover and identify key drivers contributing to it. The following models were used:

- Logistic Regression
- Decision Tree
- Random Forest

Clustering

For clustering, the **K-Means** algorithm was applied to group employees into distinct segments based on shared characteristics such as satisfaction level, tenure, and workload, enabling deeper insights into employee profiles.

Logistic Regression Analysis

Logistic Regression Overview: Logistic regression was implemented to model the probability of employee turnover (left), using LASSO-selected predictors such as satisfaction level, satisfaction-to-performance score, and other key features. The model estimates the likelihood of an employee leaving and outputs probabilities that were converted to binary classes using a threshold of 0.5.

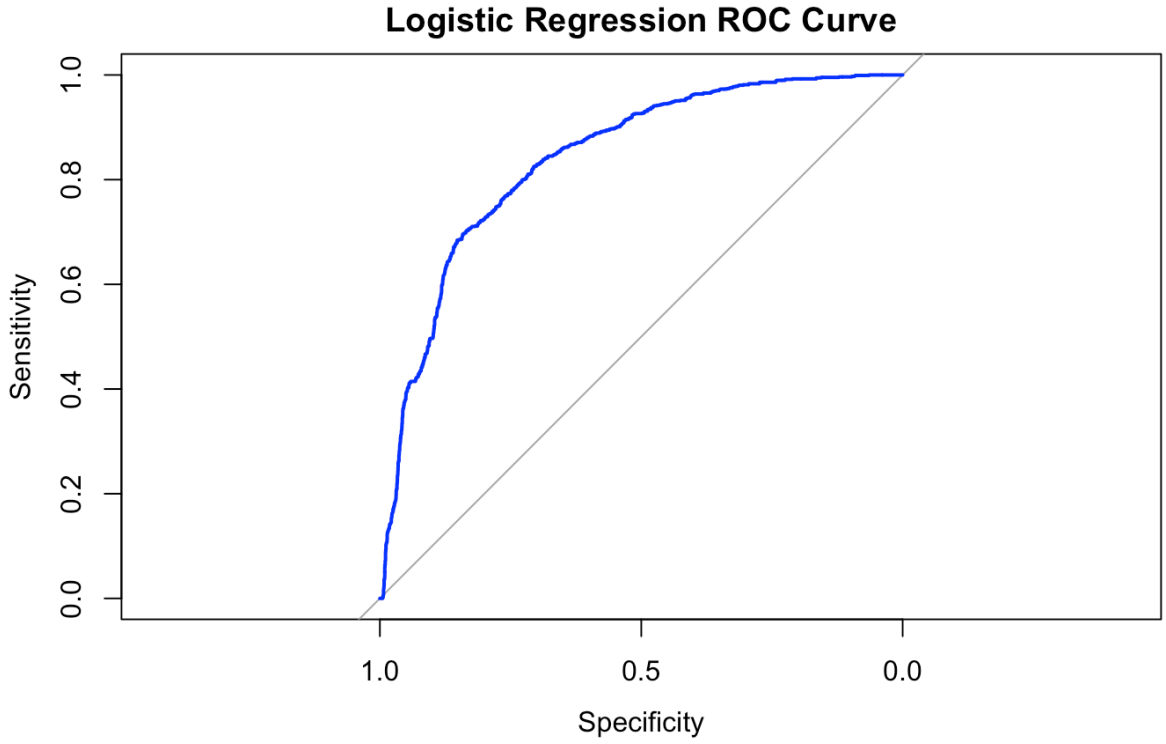
Tabular Results: Logistic Regression Metrics

Metric	Value
Accuracy	0.8082
Sensitivity (Recall)	0.9288
Specificity	0.4220
AUC (Area Under the ROC Curve)	0.84

(Table. 11 – Logistic Regression Metrics)

Key Metrics:

1. **Accuracy (0.8082):** The overall proportion of correctly classified observations (both stayed and left) in the testing dataset.
2. **Sensitivity/Recall (0.9288):** The model's ability to correctly identify employees who left. High sensitivity indicates strong performance in identifying true positives.
3. **Specificity (0.4220):** The model's ability to correctly identify employees who stayed. The lower specificity suggests challenges in identifying true negatives.
4. **AUC (0.84):** The area under the ROC curve indicates a good balance between sensitivity and specificity, demonstrating that the model performs well in distinguishing between employees who stayed and those who left.



(Fig. 18 – ROC Curve)

ROC Curve Interpretation: The ROC curve demonstrates the trade-off between sensitivity and specificity at various thresholds. The curve rises sharply, indicating strong classification performance, particularly for employees who left. The AUC value of 0.84 further validates the model's robustness in predicting turnover.

Decision Tree

Overview of Decision Tree Model: A decision tree model was employed to classify employees into "left" (employees who left) and "stayed" (employees who remained) categories. This model uses a series of hierarchical decisions based on the most critical predictors to classify data. The visualization highlights the decision-making process, with branches representing different splits based on predictor thresholds.

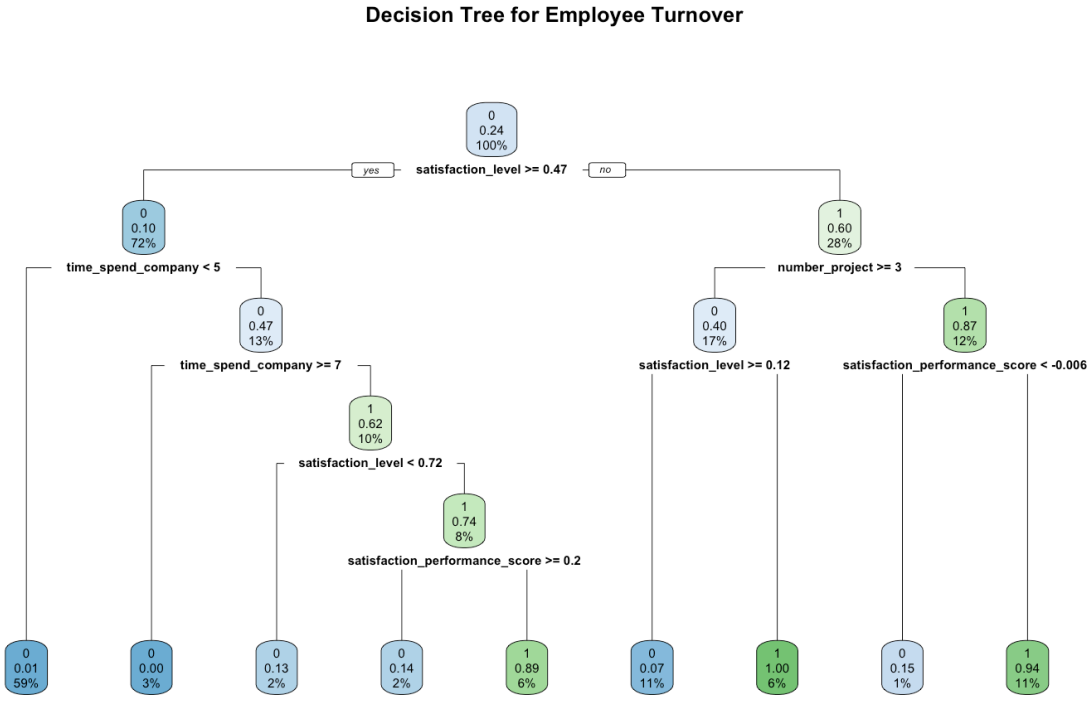
Key Metrics of Decision Tree Performance:

Metric	Value
Accuracy	0.9633
Sensitivity (Recall)	0.9851
Specificity	0.8936

(Table. 12 – Decision Tree Metrics)

Key Metrics:

1. **Accuracy (96.33%):** Indicates that the decision tree correctly classified 96.33% of the employees in the testing dataset.
2. **Sensitivity/Recall (98.51%):** Demonstrates the model's excellent ability to identify employees who left.
3. **Specificity (89.36%):** Indicates that the model effectively identified employees who stayed, albeit slightly lower than its ability to identify those who left.



(Fig. 19 – Decision Tree Plot)

Decision Tree Interpretation:

1. The root node splits based on the **satisfaction_level**. Employees with a satisfaction level below a certain threshold (0.47) are more likely to leave.
2. Subsequent splits occur on factors such as **time_spend_company**, **satisfaction_performance_score**, and **number_project**, showcasing their importance in determining turnover.
3. The leaf nodes provide the final classification (stayed or left), with associated probabilities and proportions.

Insights Derived from the Decision Tree:

- **Satisfaction Level** emerged as the most significant factor influencing turnover, appearing as the first split in the tree.
- **Tenure (time_spend_company)** and **project involvement (number_project)** further refined the predictions, highlighting their importance in turnover decisions.
- The decision tree's interpretability makes it a useful tool for explaining turnover predictions to stakeholders and for identifying actionable areas for intervention.

This model demonstrates robust performance, especially in identifying employees likely to leave, while maintaining high accuracy and specificity levels.

Random Forest

Overview of Random Forest Model: The Random Forest model, a robust ensemble learning method, was utilized to classify employee turnover. It builds multiple decision trees during training

and outputs the mode of their predictions (for classification tasks). This method enhances accuracy and reduces overfitting, providing both high performance and feature importance insights.

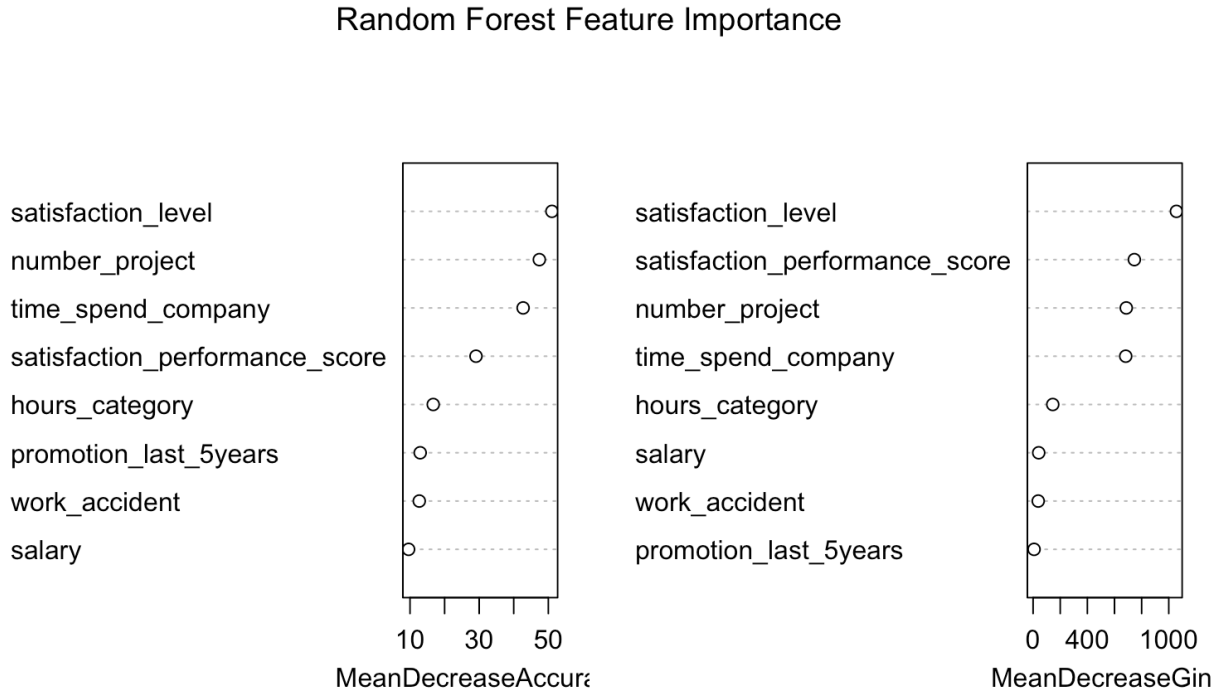
Key Metrics of Random Forest Performance:

Metric	Value
Accuracy	0.9771
Sensitivity (Recall)	0.9945
Specificity	0.9216

(Table. 13 – Random Forest Metrics)

Metric Explanation:

1. **Accuracy (97.71%):** Reflects the overall correctness of the model in predicting employee turnover.
2. **Sensitivity/Recall (99.45%):** Demonstrates the model's strong ability to correctly identify employees who left.
3. **Specificity (92.16%):** Indicates the model's ability to accurately classify employees who stayed.



(Fig. 20 – Random Forest Feature Importance Plot)

Feature Importance Analysis: The feature importance plot ranks the predictors based on their influence on the model's decision-making process. Key features and their relevance:

- **Satisfaction Level:** Most influential variable, reaffirming its critical role in turnover decisions.

- **Satisfaction Performance Score:** Indicates the interaction between satisfaction and performance evaluations.
- **Number of Projects and Time Spent at the Company:** Highlight the importance of workload and tenure.
- **Hours Category:** Reflects workload intensity, grouped into low, moderate, and high.
- **Salary and Promotion Last 5 Years:** Moderate influence, indicating compensation and career advancement impact turnover.
- **Work Accident:** Low importance, showing minimal correlation with turnover in this dataset.

Insights Derived:

1. **Satisfaction Level** remains the most critical determinant of turnover, similar to trends seen in other models.
2. Work-related factors such as tenure, workload (number of projects), and workload intensity significantly contribute to turnover.
3. Compensation (salary) and career progression (promotion history) moderately influence the likelihood of turnover.

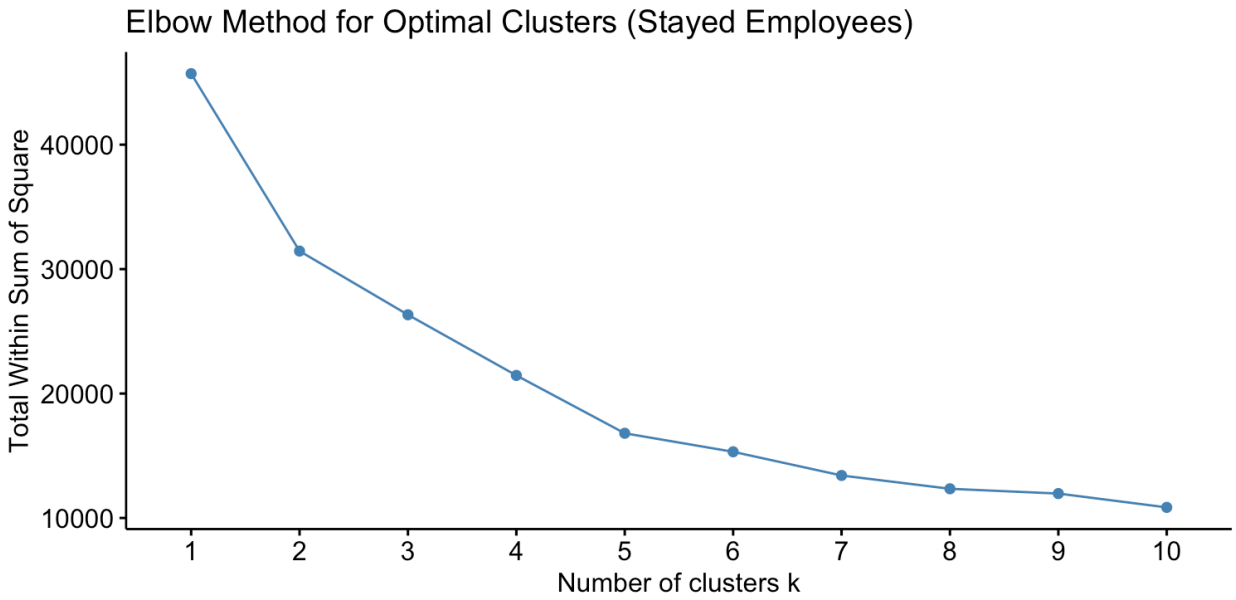
The Random Forest model's high accuracy and recall make it highly effective for predicting turnover. Its feature importance insights are invaluable for targeting retention strategies effectively.

Clustering – K-means

The clustering aimed to segment employees who **stayed** in the organization based on features like job satisfaction, tenure, project involvement, and satisfaction-to-performance balance. This analysis identifies distinct groups and helps target specific retention strategies.

Elbow Methods

1. **Elbow Method**:
    - **Objective**: Determine the optimal number of clusters by minimizing within-cluster sum of squares (WSS).
    - **Observation**: The plot showed a clear "elbow" at 3 clusters, indicating diminishing returns in WSS reduction beyond this point.



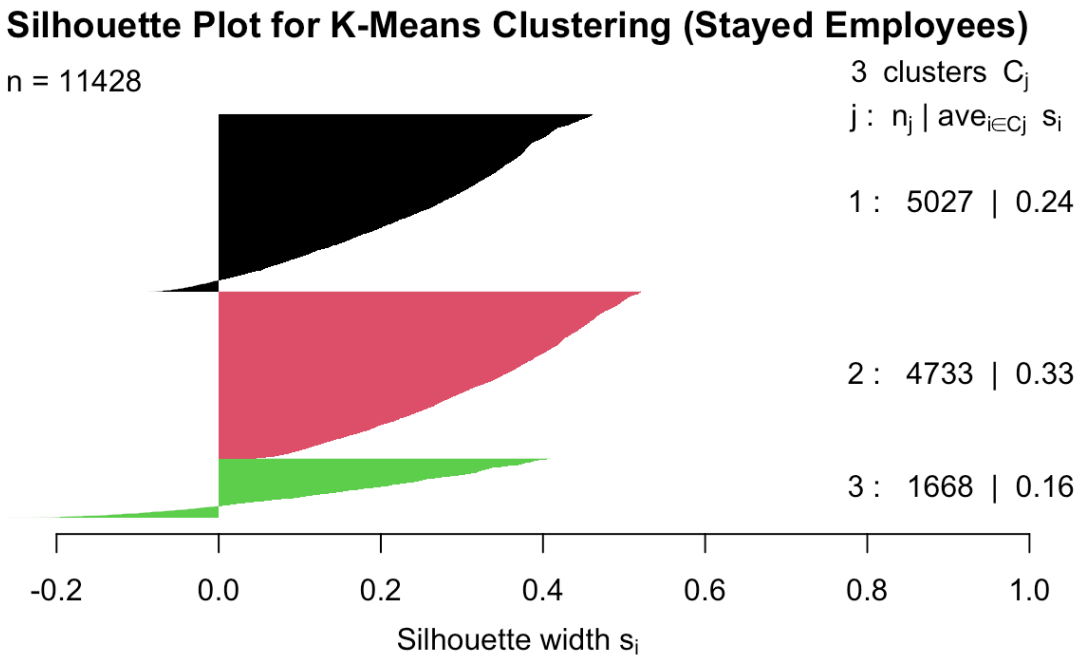
(Fig. 21 – Elbow Method Plot)

Cluster Characteristics

Clusters were defined using features: satisfaction_level, time_spend_company, number_project, and satisfaction_performance_score. The summary metrics are:

Cluster	Avg. Satisfaction	Avg. Tenure	Avg. Projects	Avg. Satisfaction-to-Performance
1	0.852	3.25	3.89	0.232
2	0.586	2.83	3.42	0.0598
3	0.339	5.34	4.51	-0.0874

(Table. 14 – Cluster Stats)



(Fig. 22 – Silhouette Plot)

- **Cluster 1**:
  - High satisfaction and balanced workload.
  - Short to moderate tenure.
  - Likely represents employees with strong organizational alignment.
- **Cluster 2**:
  - Moderate satisfaction and workload.
  - Employees likely experiencing medium engagement.
- **Cluster 3**:
  - Low satisfaction and high tenure.
  - Higher project workload but poor satisfaction-to-performance balance. These employees are at the **highest risk of disengagement** despite their experience.

Cluster Centers Plot

The bar plot highlights the feature averages for each cluster:

(Fig. 23 – Cluster Centers by Feature)

- **Observation**:
  - Cluster 3 exhibits the lowest satisfaction and satisfaction-performance score but the highest tenure.
  - Cluster 1 shows balanced metrics, signifying a content workforce.

Insights and Retention Strategies

Key Insights

Employee Turnover Drivers:

- **Job Satisfaction**: A significant factor influencing turnover. Employees with lower satisfaction levels are more likely to leave.
- **Tenure**: Employees with shorter tenure (2–3 years) or very long tenure (5+ years) display distinct behaviors that require tailored strategies.
- **Workload and Project Involvement**: Employees with higher workloads or project counts (Cluster 3 in clustering results) report lower satisfaction and are at higher risk of disengagement.
- **Promotion and Compensation**: Lack of promotions in the last five years and lower salary levels are key contributors to dissatisfaction.
- **Satisfaction-to-Performance Balance**: Poor satisfaction-performance alignment in certain employee groups highlights the need for balance between employee effort and recognition.

Retention Strategies

1. **For Cluster 3 (High Risk Employees)**:
    - **Address Dissatisfaction**:
        - Conduct targeted surveys to identify specific dissatisfaction sources (e.g., workload, lack of recognition).
        - Introduce flexible work arrangements or redistribute workloads to reduce pressure.
    - **Recognition and Career Growth**:
        - Offer career progression opportunities or additional training to reignite motivation.
        - Implement mentorship programs to leverage their experience while improving engagement.
2. **For Cluster 2 (Medium Risk Employees)**:
    - **Boost Engagement**:
        - Introduce team-building activities and incentive programs to foster collaboration.
        - Monitor project assignments to ensure balanced workloads and align tasks with employee interests.
    - **Preemptive Measures**:
        - Provide periodic feedback sessions and reward consistent performance to prevent dissatisfaction.
3. **For Cluster 1 (Engaged Employees)**:
    - **Maintain Satisfaction**:
        - Offer professional development opportunities to sustain their enthusiasm and loyalty.
        - Ensure competitive compensation and recognition for high-performing employees.
    - **Encourage Advocacy**:
        - Empower these employees to act as mentors or champions within the organization to influence others positively.

Action Plan

1. **Data-Driven Interventions**:
    - Use clustering insights to prioritize retention strategies for high-risk groups.
    - Implement a continuous feedback loop to track satisfaction and engagement levels across all clusters.
2. **Performance and Workload Balance**:
    - Establish clearer career paths and align responsibilities with employee capacity.
    - Create automated dashboards to monitor workload distribution and identify burnout risks early.
3. **Employee Engagement Programs**:
    - Offer targeted workshops, upskilling programs, and mental health support initiatives.
    - Develop personalized incentive structures to align with individual and group motivations.

Conclusion

This project successfully identified the primary drivers of employee turnover and provided actionable strategies to retain at-risk employees. Through robust predictive modeling and clustering analysis, we found that job satisfaction, workload balance, and career advancement opportunities are critical to retention. Employees with low satisfaction and high workloads were at the greatest risk, necessitating targeted interventions.

Our clustering analysis revealed three distinct employee segments, each with unique characteristics and needs. Tailored retention strategies, including workload redistribution, promotion opportunities, and engagement initiatives, are essential for addressing these challenges. Additionally, fostering professional development and maintaining satisfaction levels for engaged employees will ensure long-term workforce stability.

By implementing the proposed action plan, the organization can proactively manage turnover risks, improve employee satisfaction, and align its workforce strategies with business objectives. This approach not only enhances organizational productivity but also builds a stronger and more committed workforce.