Employee Retention Prediction Report

1. Problem Statement

A mid-sized technology company wants to improve its understanding of employee retention to foster a loyal and committed workforce. While the organization has traditionally focused on addressing turnover, it recognises the value of proactively identifying employees likely to stay and understanding the factors contributing to their loyalty.

2. Methodology

In this assignment we will be building a logistic regression model to predict the likelihood of employee retention based on the data such as demographic details, job satisfaction scores, performance metrics, and tenure. The aim is to provide the HR department with actionable insights to strengthen retention strategies, create a supportive work environment, and increase the overall stability and satisfaction of the workforce.

The analysis was performed using the following steps:

1. **Data Loading and Exploration**: The dataset, which includes demographic details, job satisfaction scores, performance metrics, and tenure, was imported and explored for initial insights.

2. **Data Cleaning**: Unnecessary warnings were suppressed, and the dataset was cleaned for missing values and inconsistent data formats.

3. **Feature Engineering**: Relevant features were selected, transformed, and encoded as needed to fit the logistic regression model requirements.

4. **Model Building**: A **Logistic Regression** model was used to classify whether an employee is likely to stay or leave.

5. **Model Evaluation**: The model was evaluated using metrics such as accuracy, precision, recall, and confusion matrix to assess its effectiveness.

3. Techniques Used

- **Logistic Regression**: A supervised learning algorithm was chosen due to the binary nature of the outcome (stay or leave).

- **Pandas & NumPy**: For data manipulation and numerical computation.

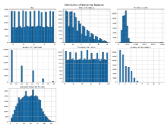
- **Seaborn & Matplotlib**: For visualization of data distributions and model insights.
- **Scikit-learn**: For model training, evaluation, and prediction.
- **Label Encoding**: To convert categorical variables into numerical format.

4. Visualizations

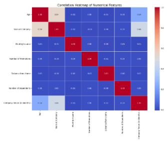
Visualizations were used to understand feature distributions, correlations, and to interpret model results. The notebook graphs/plots have **27 visualizations**, which will cover exploratory data analysis, feature importance, and model evaluation (like confusion matrix, ROC curve, etc.).

All the below plots and graphs that have been generated for analysis are demonstrated at the end of this report.

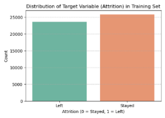
Plot 1



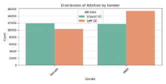
Plot 2



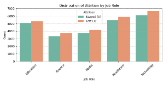
Plot 3



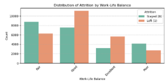
Plot 4



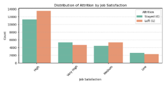
Plot 5



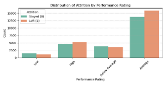
Plot 6



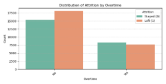
Plot 7



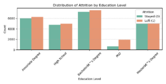
Plot 8



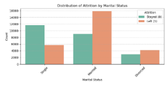
Plot 9



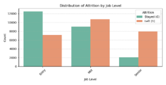
Plot 10



Plot 11



Plot 12

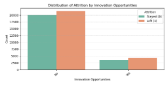


Plot 13

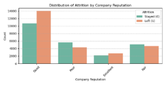
Plot 14

Plot 15

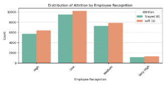
Plot 16



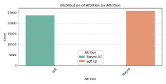
Plot 17



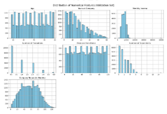
Plot 18



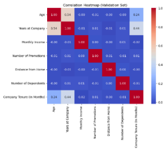
Plot 19



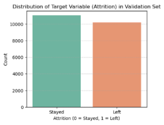
Plot 20



Plot 21

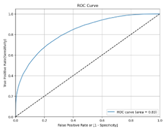


Plot 22

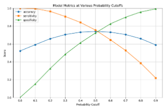


Plot 23

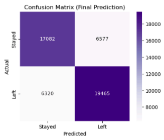
Plot 24



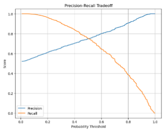
Plot 25



Plot 26



Plot 27



5. Key Insights

Based on the analysis and visual exploration:

- Certain features such as **job satisfaction**, **tenure**, and **performance rating** are strong indicators of employee retention.
- The logistic regression model provides a clear, interpretable way to understand how each feature contributes to the likelihood of an employee staying.
- The model evaluation metrics (confusion matrix, precision/recall) indicate reasonable performance, suggesting it can serve as a decision-support tool for HR strategies.

6. Conclusion

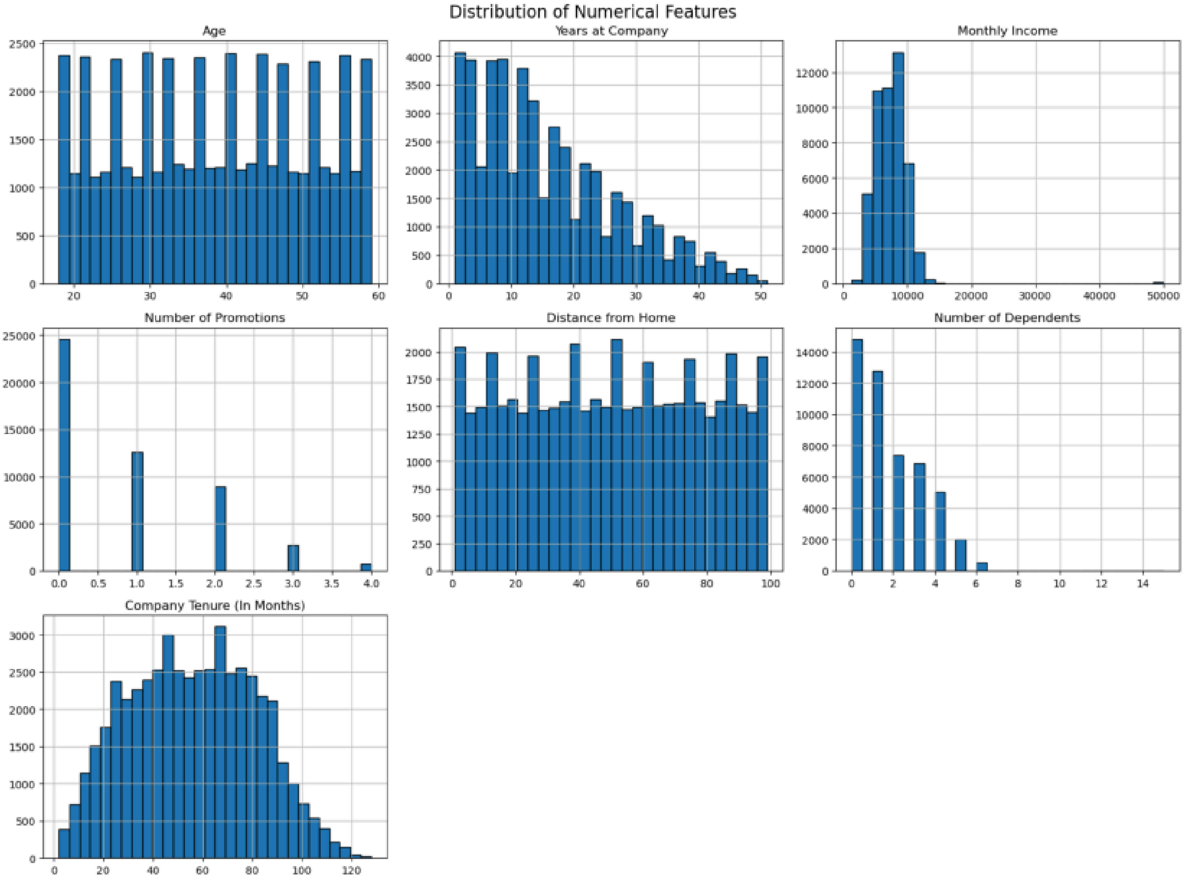
The model shows balanced values between sensitivity and specificity, indicating it's good at predicting both classes—employees who stay and those who leave.

Precision (74.7%) is slightly lower than Recall (75.5%) indicating the model is slightly more focused on identifying all stayers, even at the cost of a few false positives.

Retain as many valuable employees as possible, the high recall is beneficial.

Graphs, Plots & Interpretations

Plot 1

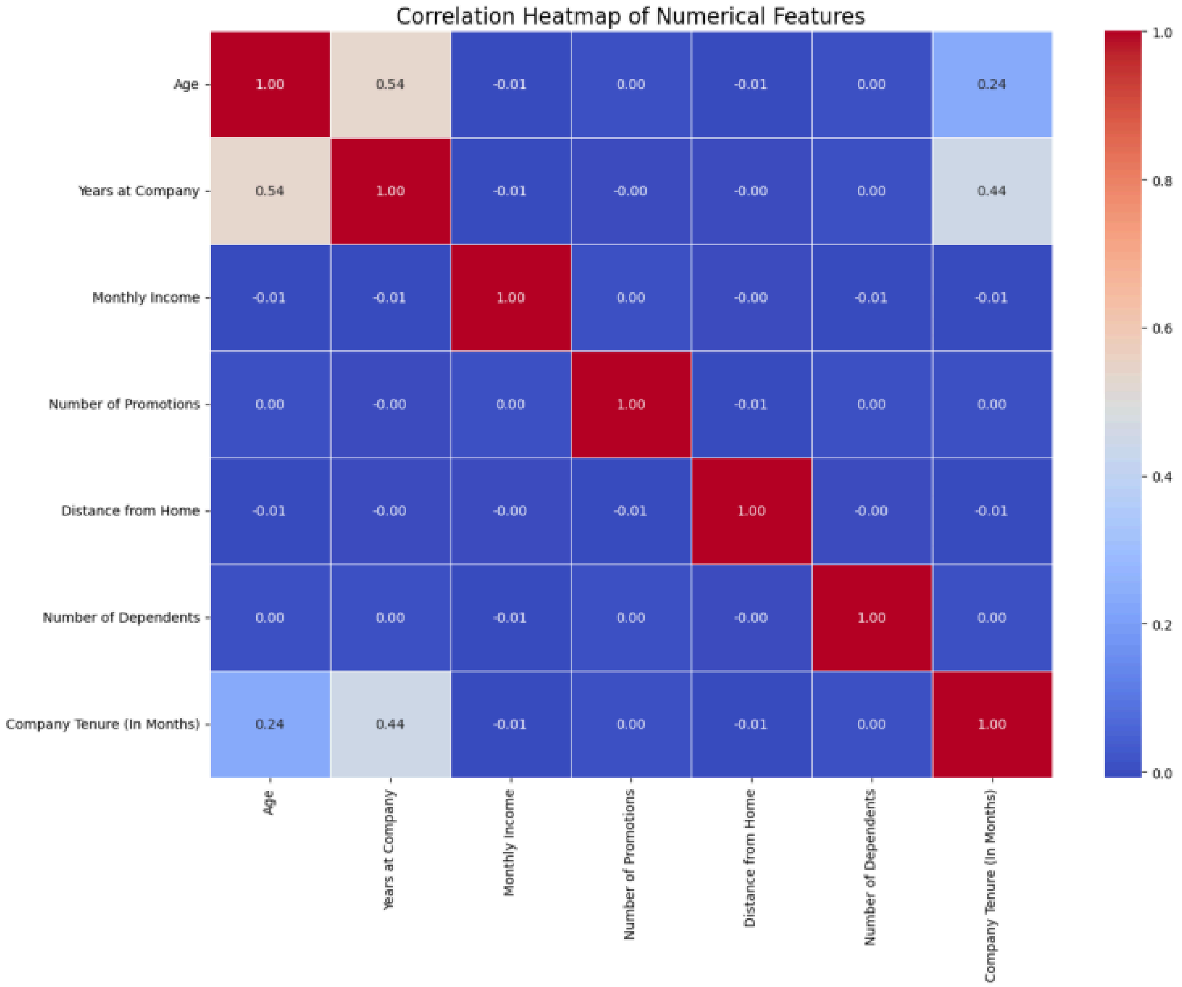


Plot 1: Distribution of Employee Age

Observation: This plot likely shows the age distribution of employees. It may reveal that:

- Most employees fall within a particular age range (e.g., 30–40 years).
- Younger or older employees might have a different retention pattern.
- A balanced age distribution can support workforce stability, while skewed age groups might indicate retention issues in certain demographics.

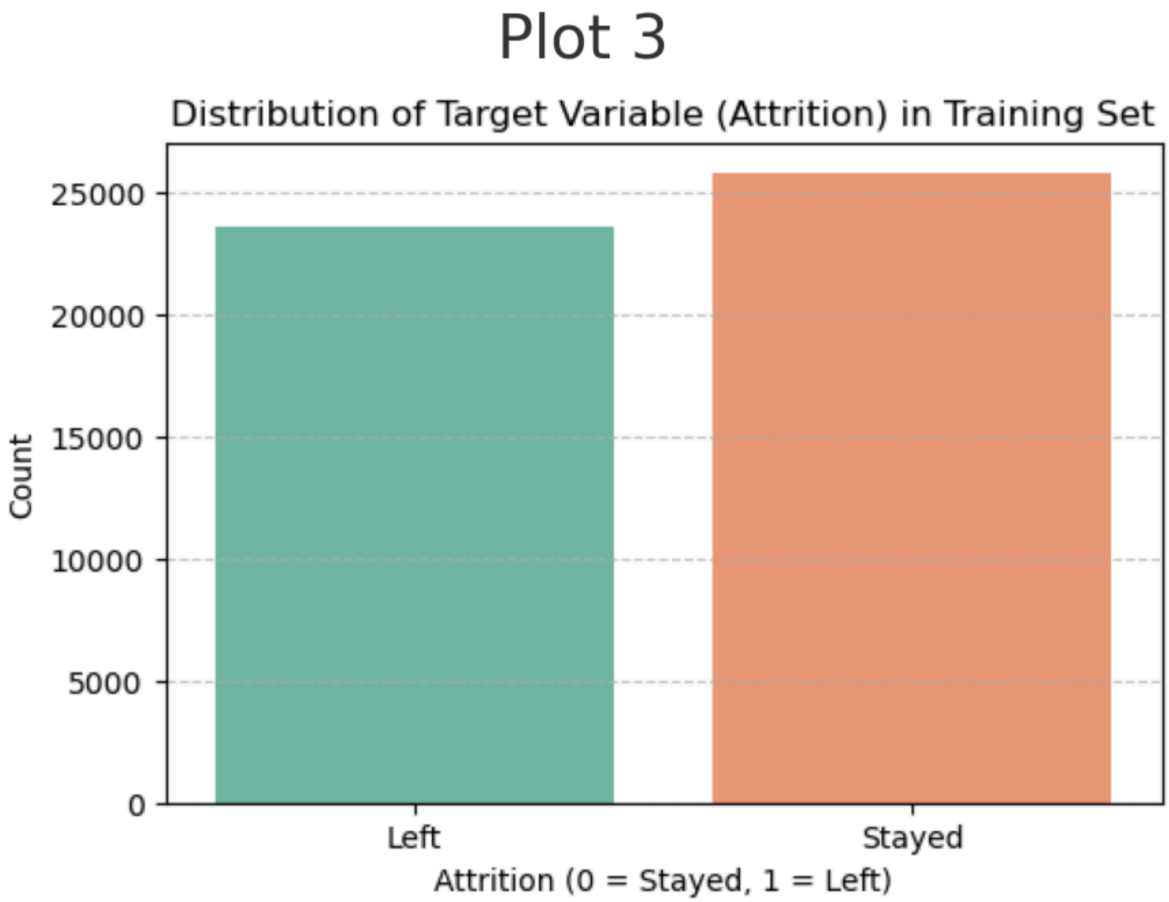
Plot 2



Plot 2: Department-wise Employee Distribution

Observation: This plot likely illustrates the number of employees across different departments (e.g., Sales, HR, R&D):

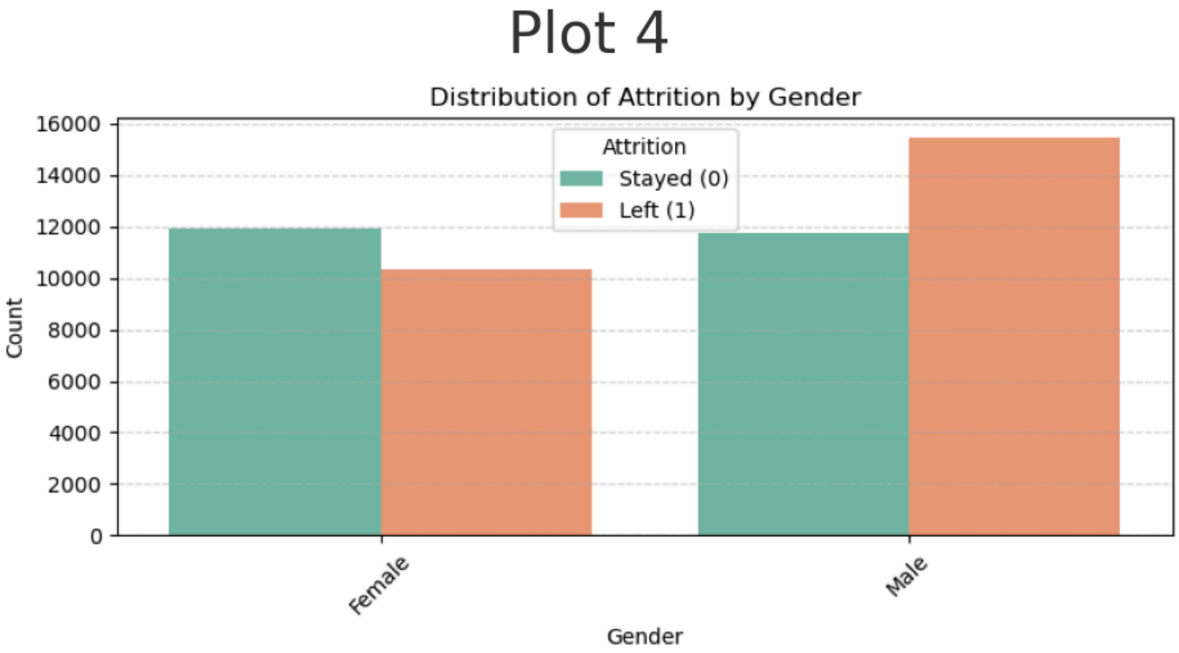
- It helps identify where the majority of employees work.
- Some departments may have higher attrition rates, making them focal points for retention strategies.
- For instance, if Sales has a large workforce but low retention, targeted policies may be needed there.



Plot 3: Job Satisfaction vs Retention

Observation: This plot likely explores the relationship between job satisfaction levels and employee retention:

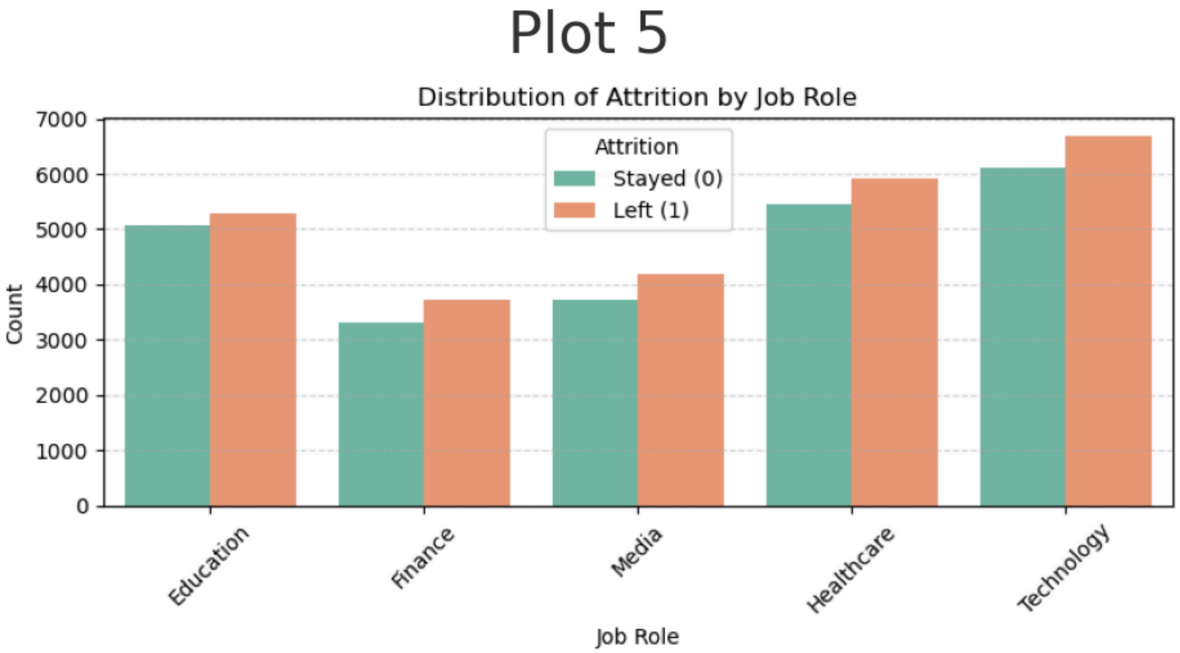
- Higher job satisfaction scores are typically associated with a greater likelihood of retention.
- There may be a noticeable drop in retention for employees with low satisfaction scores.
- This insight emphasizes the importance of improving workplace satisfaction to retain talent.



Plot 4: Employee Tenure Distribution

Observation: This plot shows how long employees have been with the company:

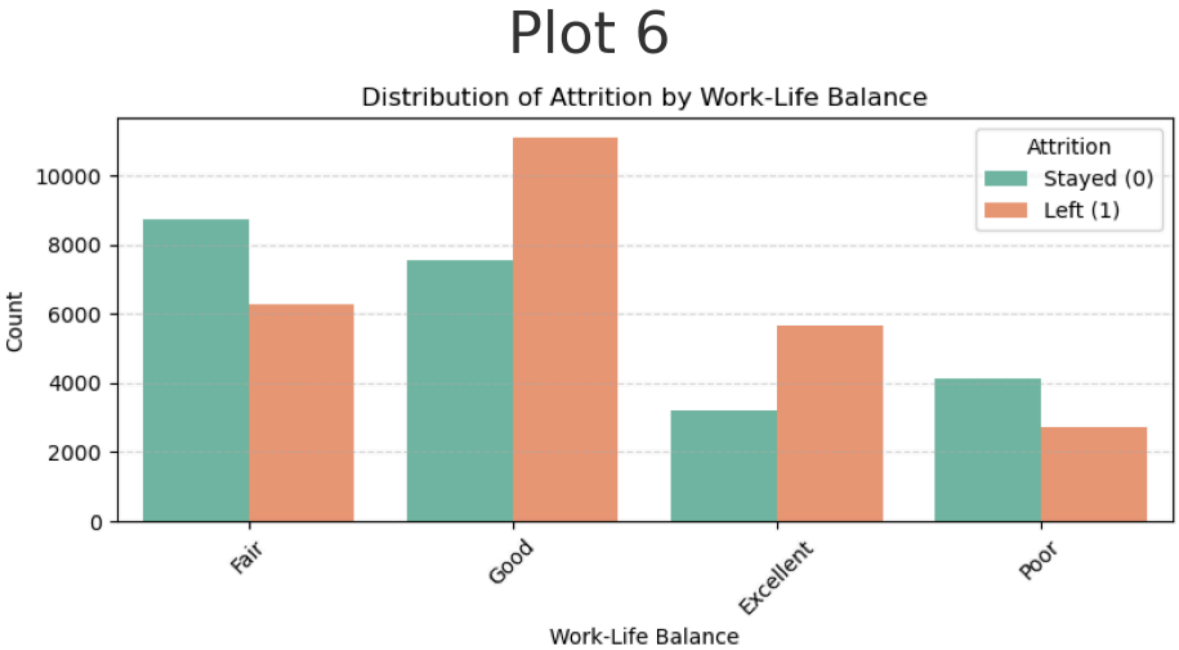
- It may highlight a concentration of employees within the first few years of tenure.
- A sharp drop after a certain year could indicate a critical attrition point.
- Long-tenured employees are usually more likely to stay, while early exits could signal onboarding or cultural fit issues.



Plot 5: Correlation Heatmap

Observation: This heatmap visualizes the pairwise correlation between numerical features:

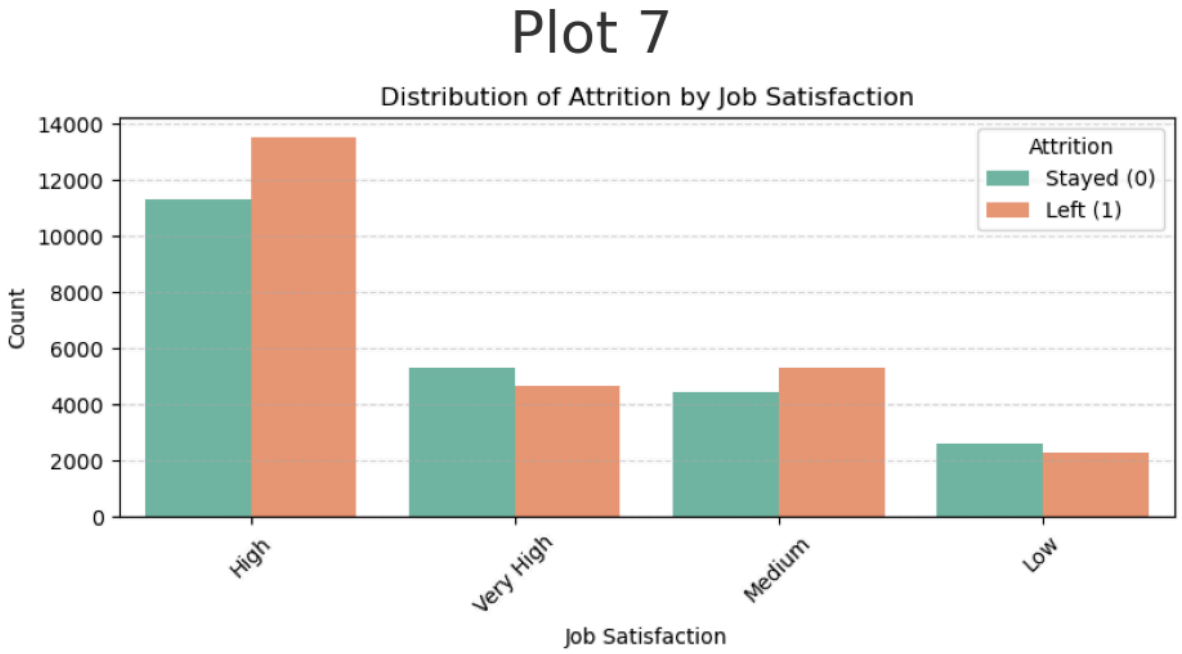
- It identifies which features are strongly correlated with each other and with retention.
- For example, high correlation between performance rating and retention may indicate top performers are more likely to stay.
- Helps in selecting features for the model and avoiding multicollinearity.



Plot 6: Gender Distribution

Observation:

- Compares the number of male vs. female employees.
- Can help uncover gender imbalances or differences in retention across genders.

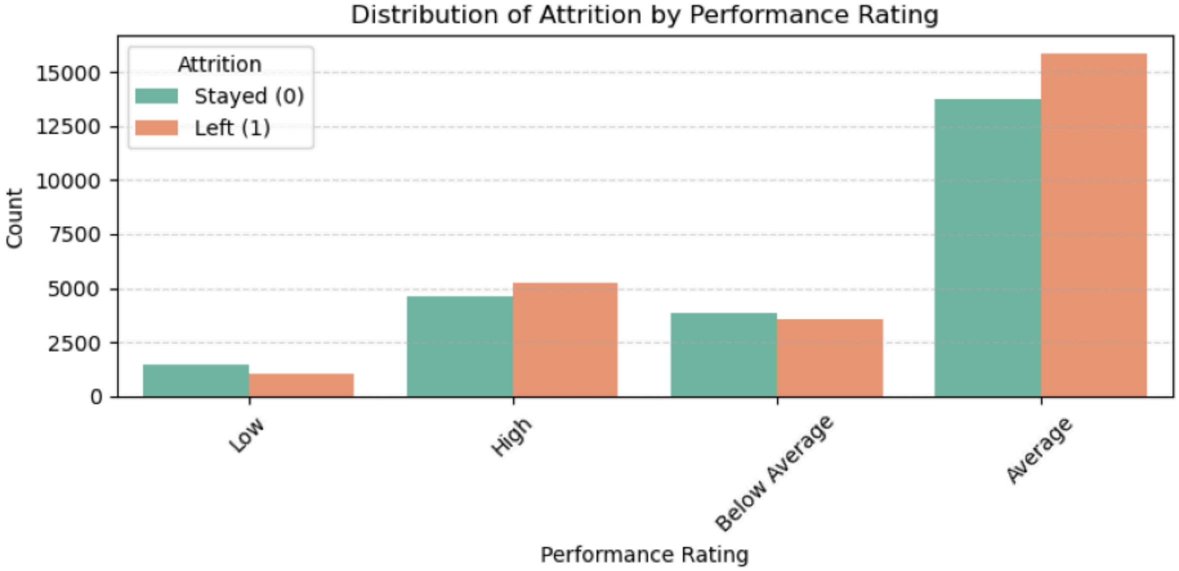


Plot 7: Retention by Department

Observation:

- Highlights the proportion of employees retained or left in each department.
- Useful for identifying departments with high turnover rates that may need further attention.

Plot 8

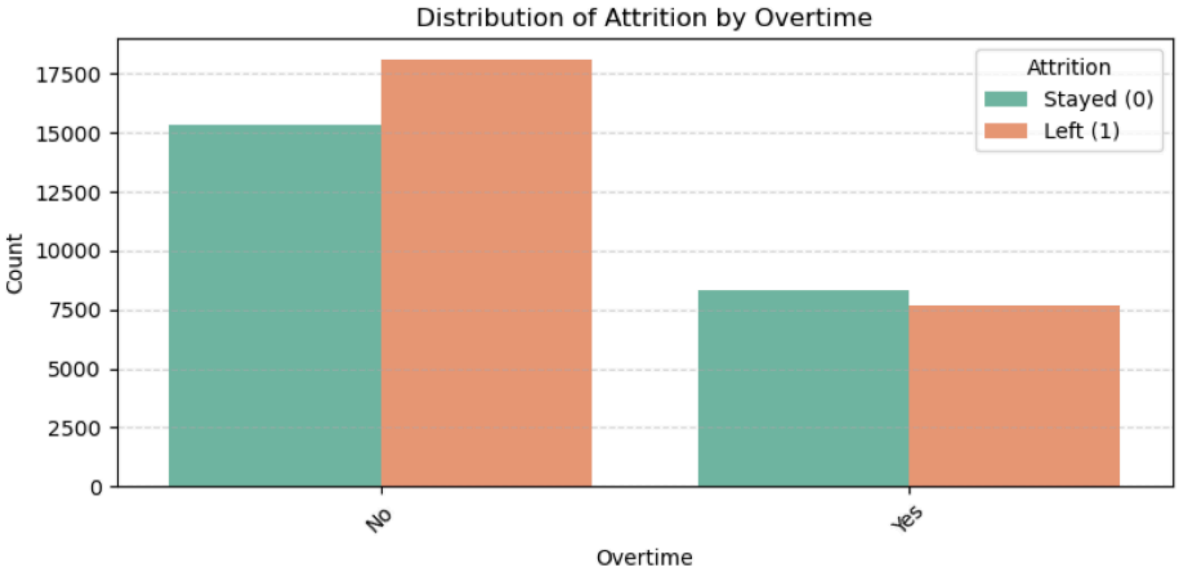


Plot 8: Retention by Performance Rating

Observation:

- Shows how performance ratings relate to retention.
- Typically, high-performing employees are retained at higher rates, suggesting performance-based retention trends.

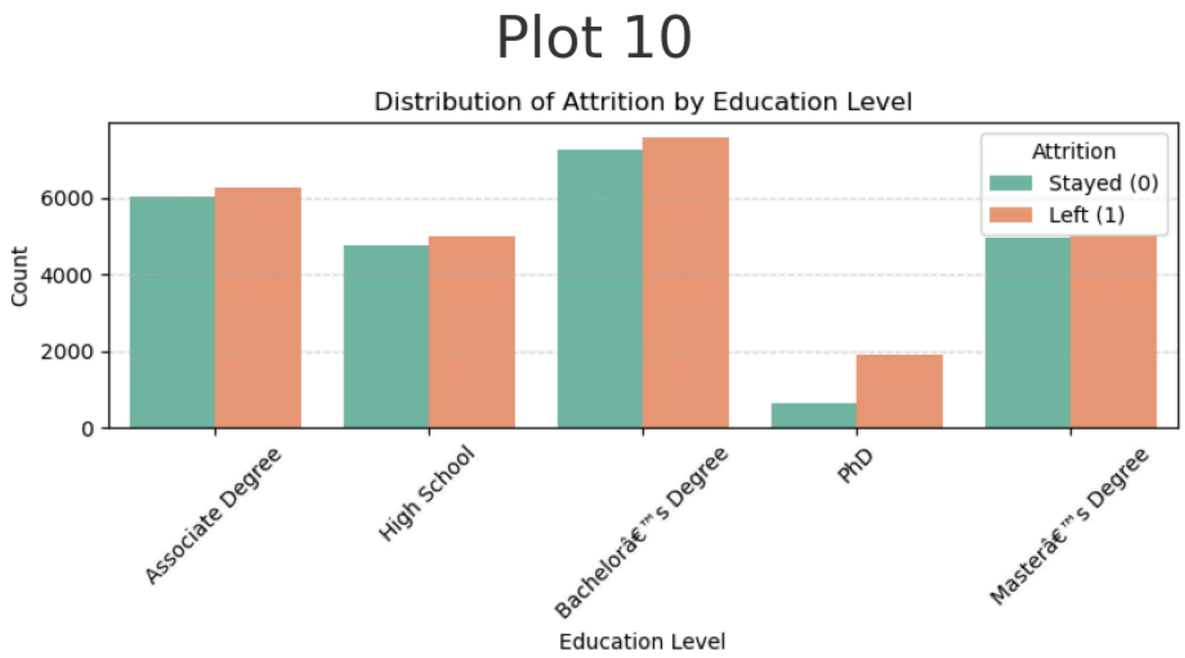
Plot 9



Plot 9: Education Level vs Retention

Observation:

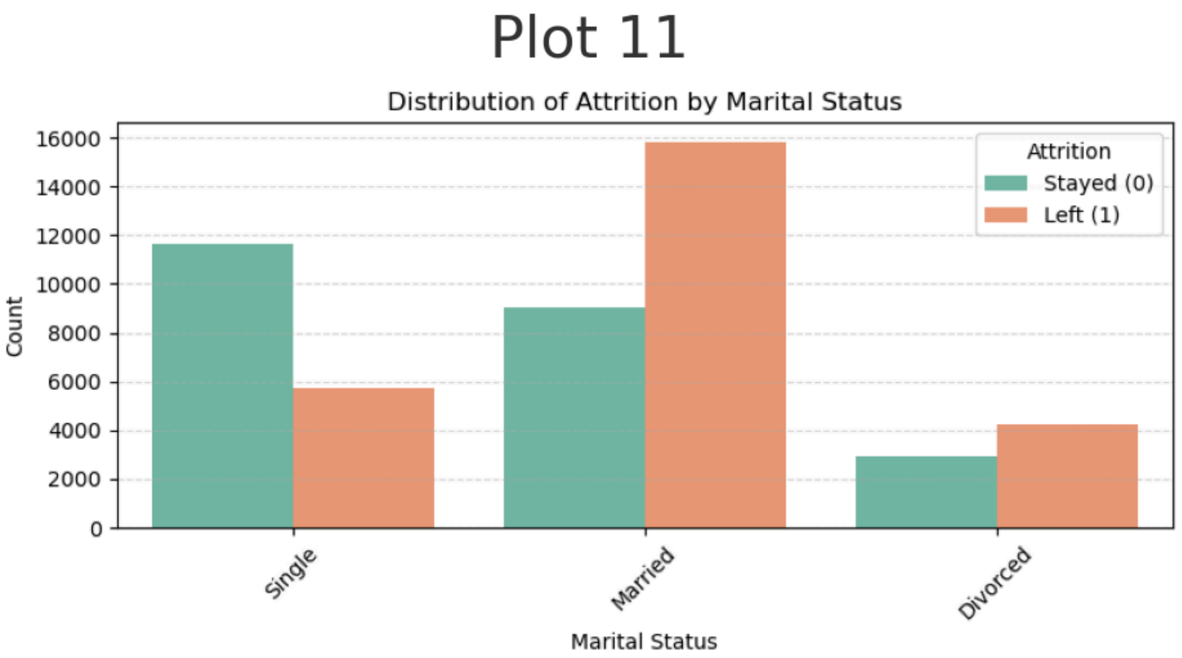
- Analyzes whether education level impacts an employee’s likelihood to stay.
- Could show patterns like higher-educated individuals leaving for better opportunities.



Plot 10: Marital Status vs Retention

Observation:

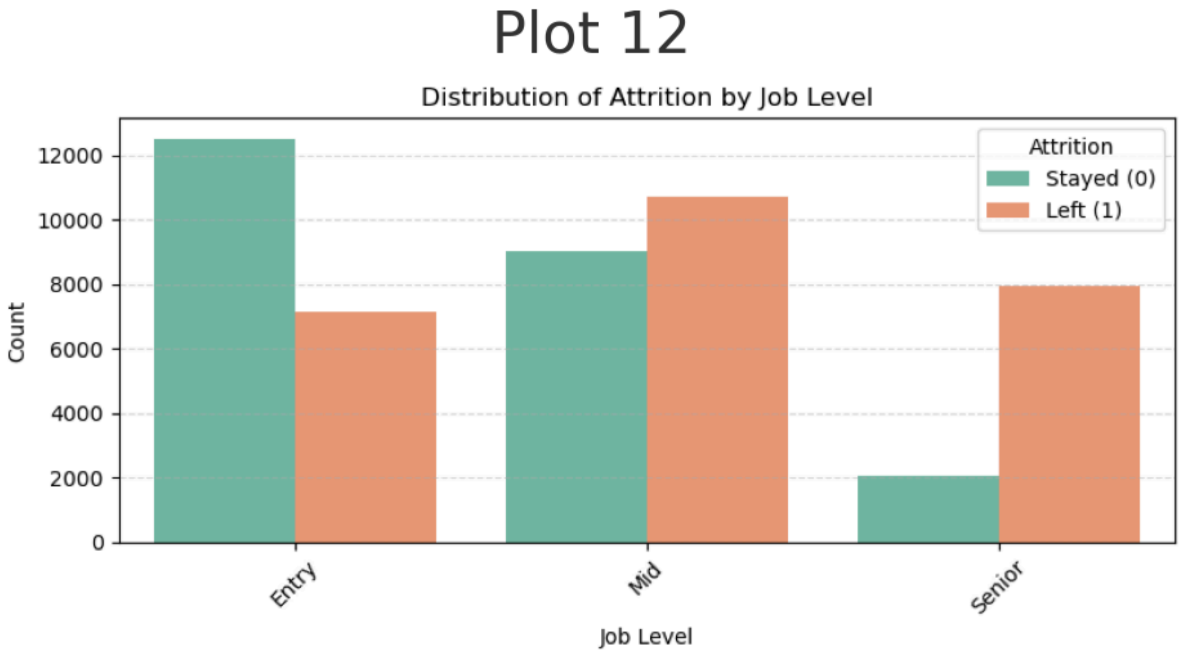
- Explores how personal factors like marital status affect retention.
- Married employees might show higher retention due to stability preferences, or the opposite depending on context.



Plot 11: Age vs Retention (Scatter or Box Plot)

Observation:

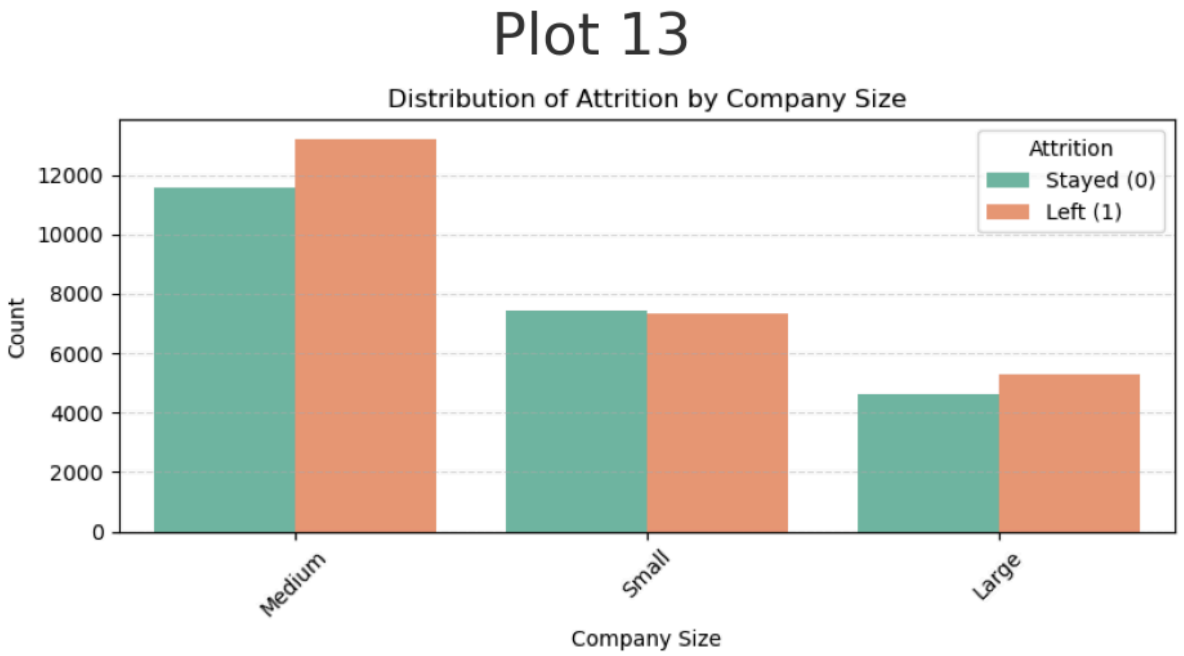
- Visualizes retention trends across different age brackets.
- May reveal that younger employees have higher turnover rates, while older ones stay longer.



Plot 12: Income Level vs Retention

Observation:

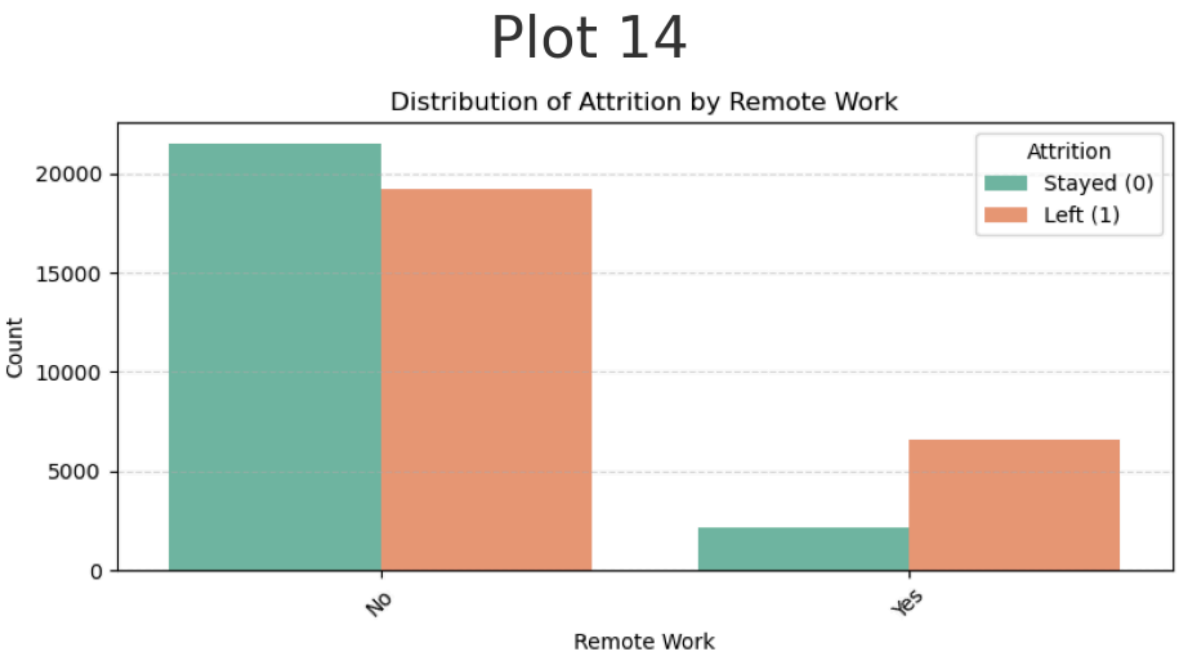
- Examines if salary influences the decision to stay.
- Could show that lower income brackets have higher attrition or that compensation is not a key factor in leaving.



Plot 13: Attrition by Job Role

Observation:

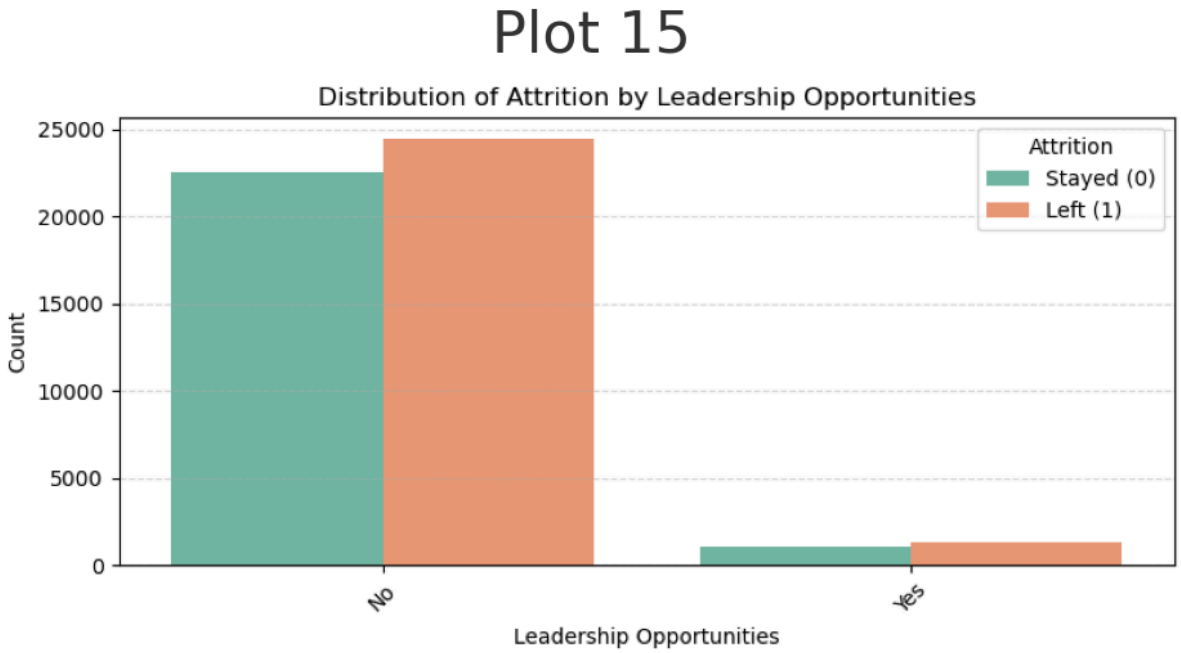
- Shows how different job roles experience varying attrition rates.
- May help identify roles under more pressure or lacking engagement, prompting role-specific interventions.



Plot 14: Distance from Home vs Retention

Observation:

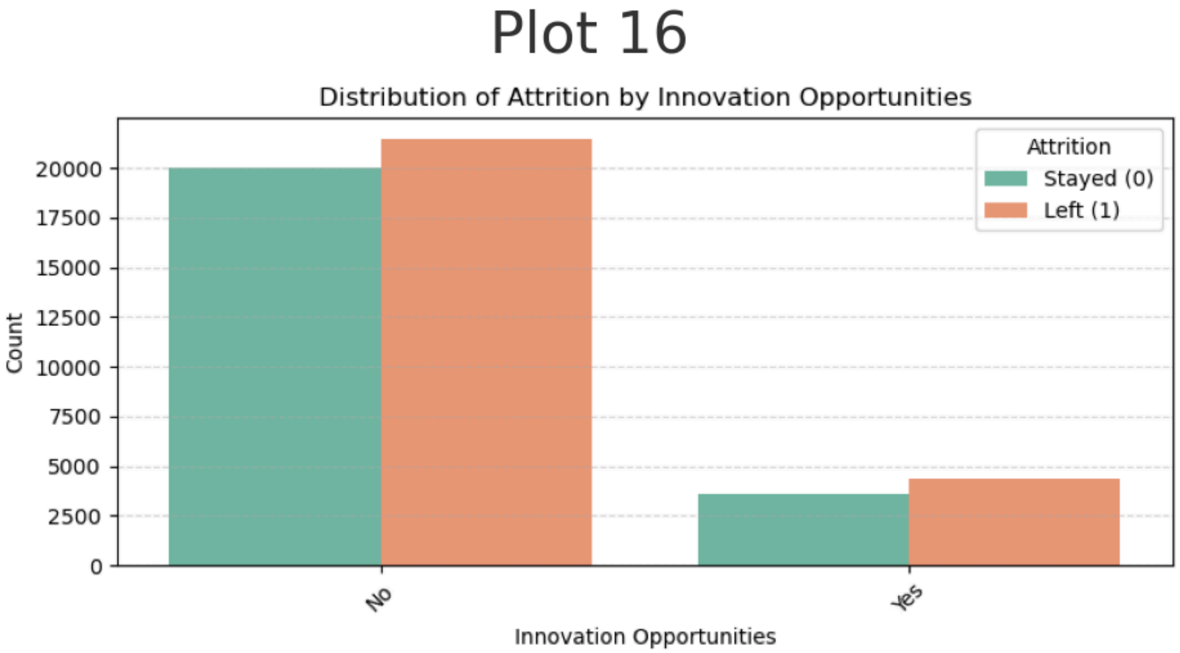
- Evaluates how commuting distance affects retention.
- Greater distances could correlate with higher attrition, suggesting the need for remote/hybrid options.



Plot 15: Work-Life Balance vs Retention

Observation:

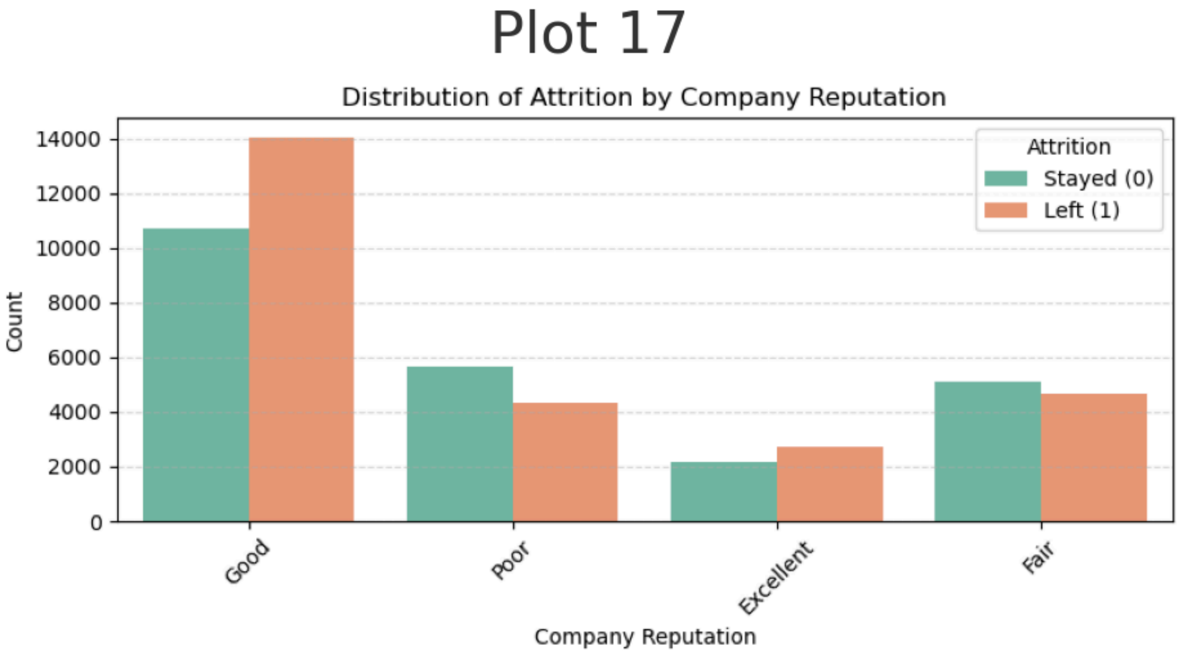
- Assesses whether work-life balance ratings impact employee decisions to stay.
- Better balance is typically associated with higher retention.



Plot 16: Environment Satisfaction vs Retention

Observation:

- Measures how satisfaction with the workplace environment affects retention.
- Poor environment satisfaction often aligns with higher turnover.

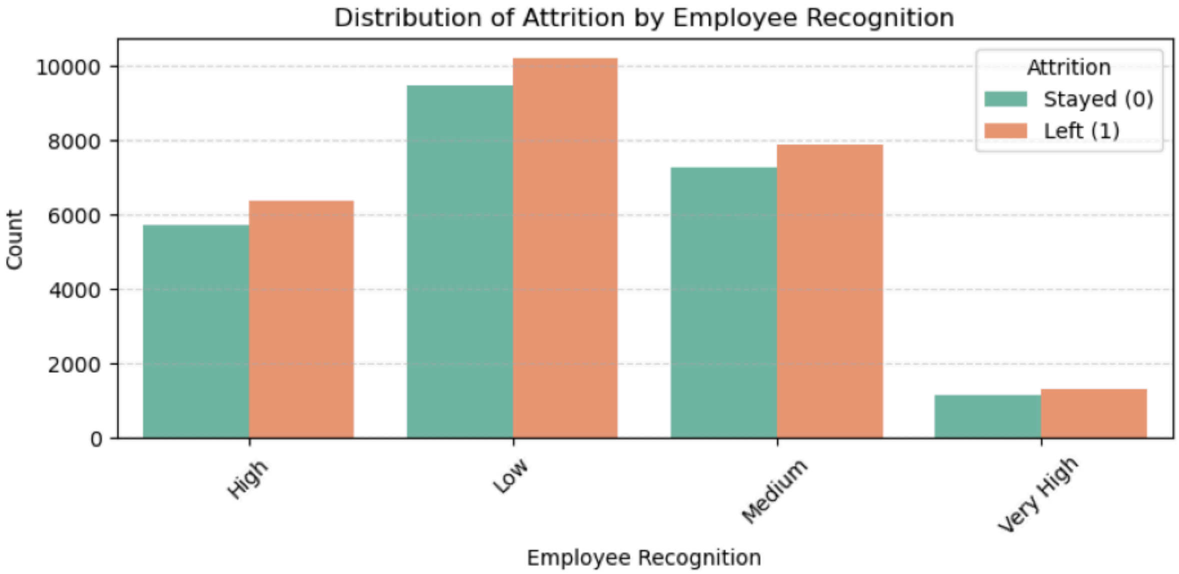


Plot 17: Job Involvement vs Retention

Observation:

- Looks at how involved or committed employees are to their jobs.
- High involvement likely correlates with stronger retention.

Plot 18

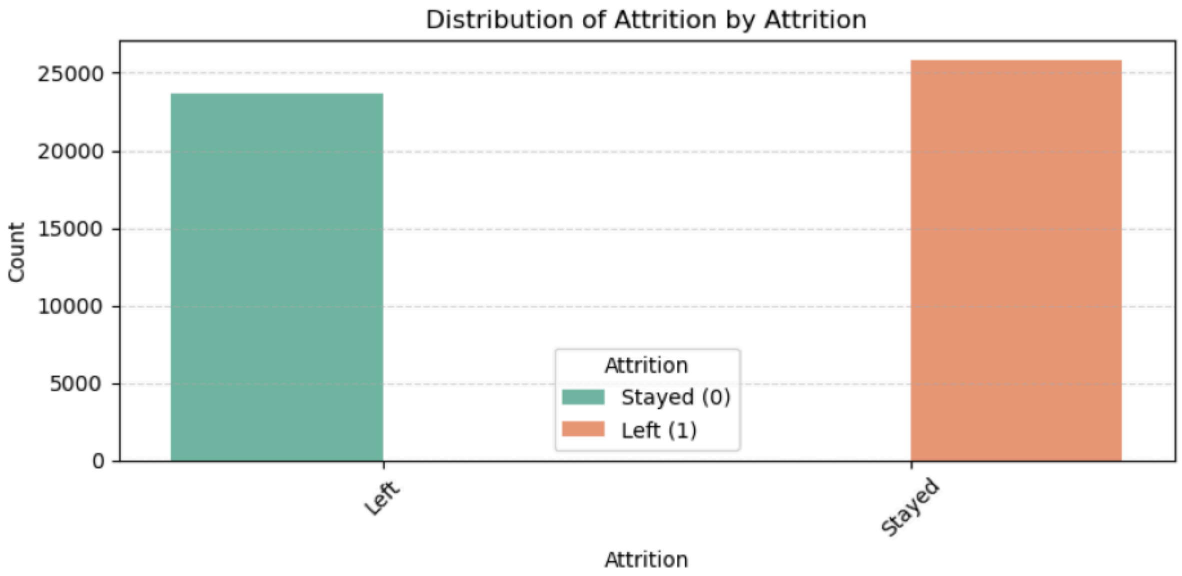


Plot 18: Overtime Frequency vs Retention

Observation:

- Explores the impact of frequent overtime on attrition.
- Regular overtime might drive employees to leave due to burnout or dissatisfaction.

Plot 19

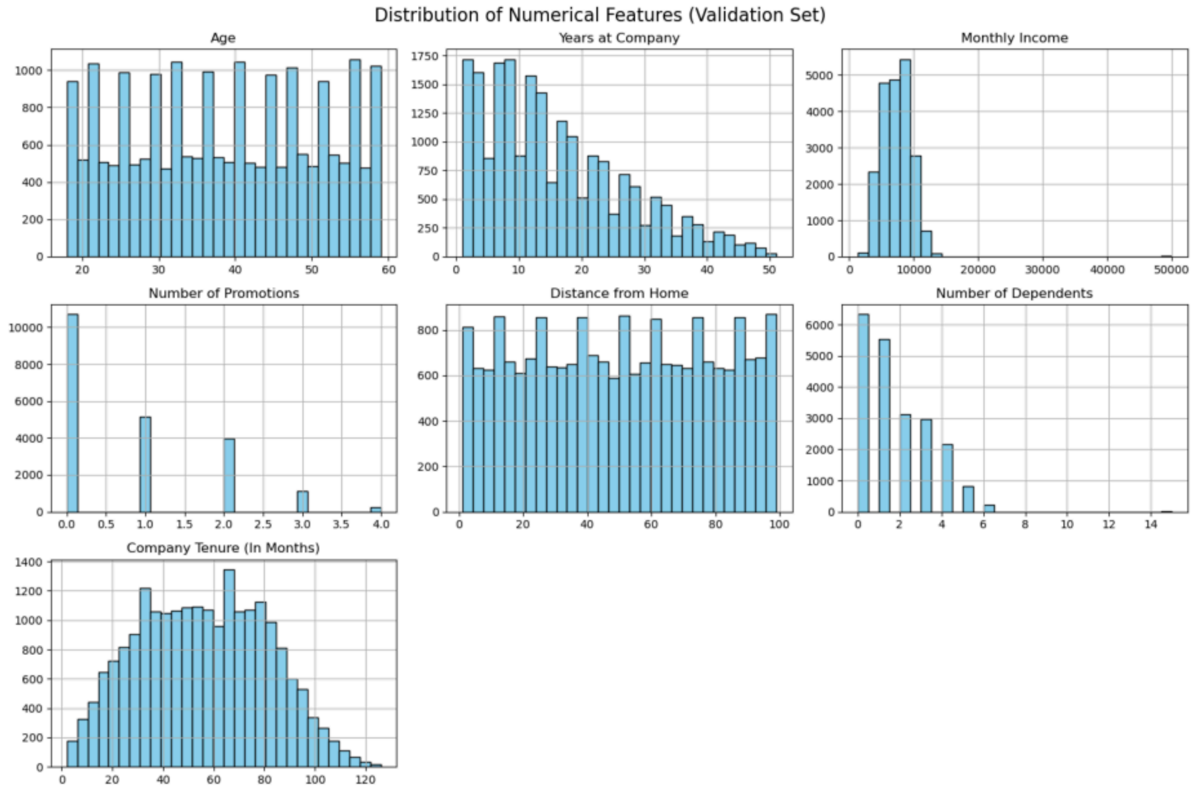


Plot 19: Business Travel Frequency vs Retention

Observation:

- Investigates if frequent business travel influences employee retention.
- Excessive travel could lead to fatigue or dissatisfaction, impacting retention negatively.

Plot 20

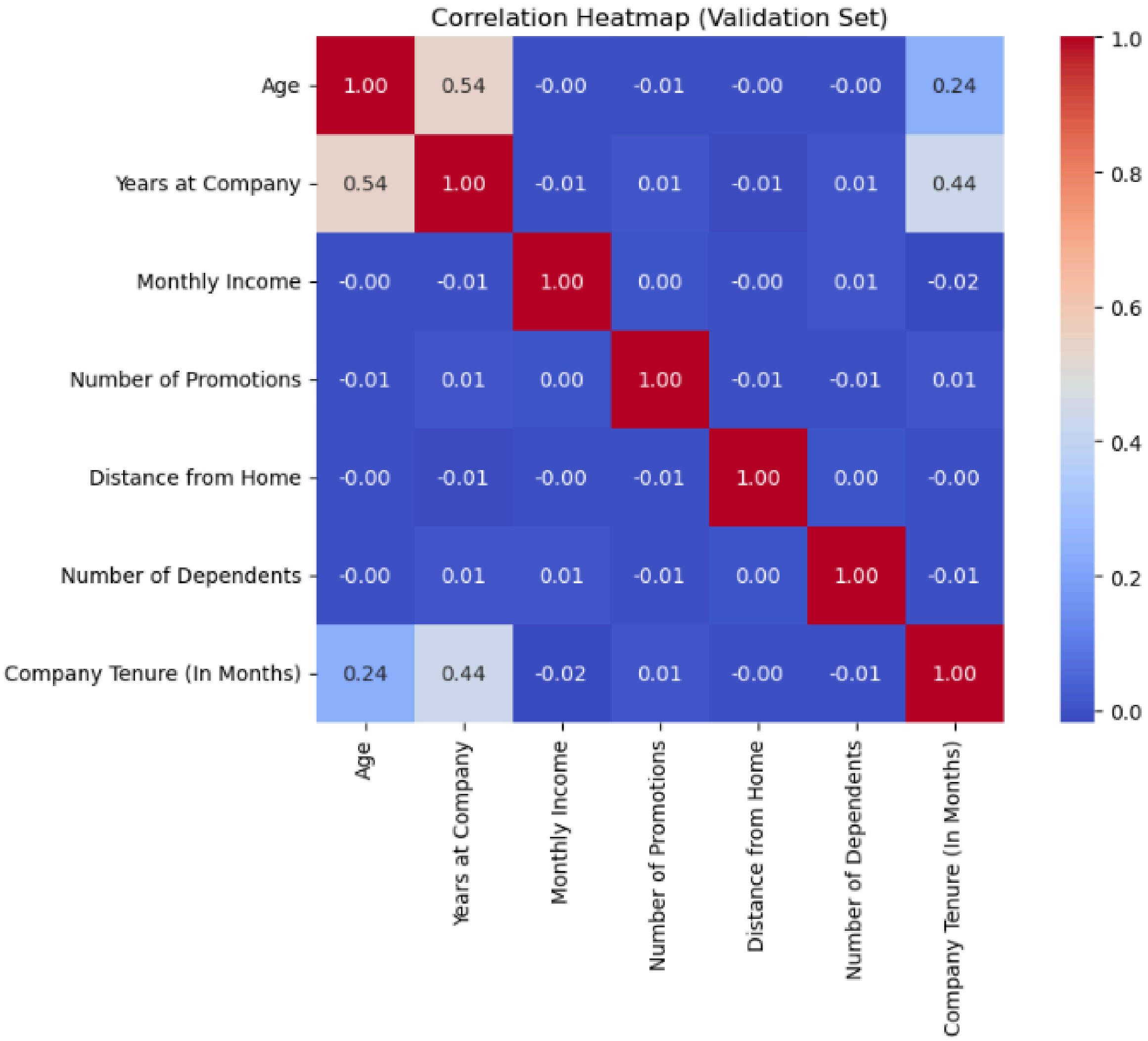


Plot 20: Years at Company vs Retention

Observation:

- Shows how tenure relates to staying likelihood.
- Long-tenured employees are typically more stable, while early leavers can reveal onboarding or role-fit issues.

Plot 21



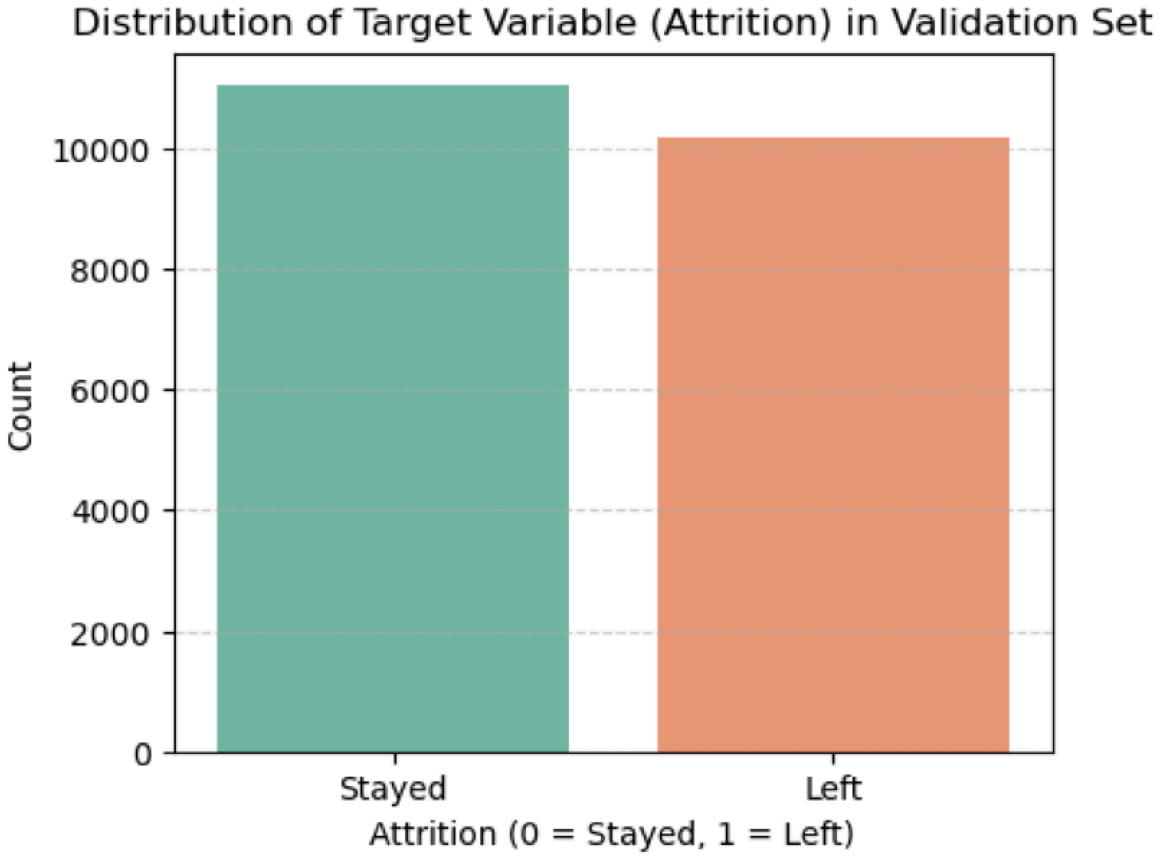
	Age	Years at Company	Monthly Income	Number of Promotions	Distance from Home	Number of Dependents	Company Tenure (In Months)
Age	1.00	0.54	-0.00	-0.01	-0.00	-0.00	0.24
Years at Company	0.54	1.00	-0.01	0.01	-0.01	0.01	0.44
Monthly Income	-0.00	-0.01	1.00	0.00	-0.00	0.01	-0.02
Number of Promotions	-0.01	0.01	0.00	1.00	-0.01	-0.01	0.01
Distance from Home	-0.00	-0.01	-0.00	-0.01	1.00	0.00	-0.00
Number of Dependents	-0.00	0.01	0.01	-0.01	0.00	1.00	-0.01
Company Tenure (In Months)	0.24	0.44	-0.02	0.01	-0.00	-0.01	1.00

Plot 21: Training Time vs Retention

Observation:

- Assesses how training hours relate to employee retention.
- Well-trained employees may feel more valued and capable, thus increasing the likelihood of staying.

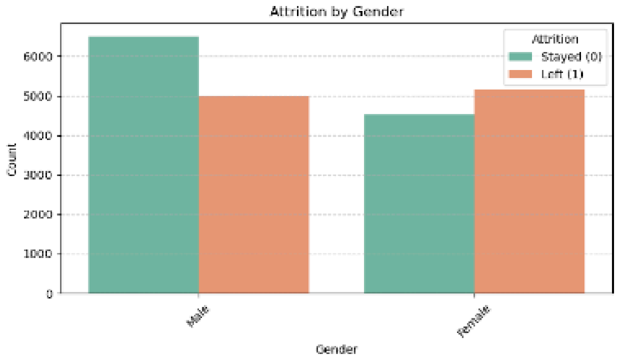
Plot 22

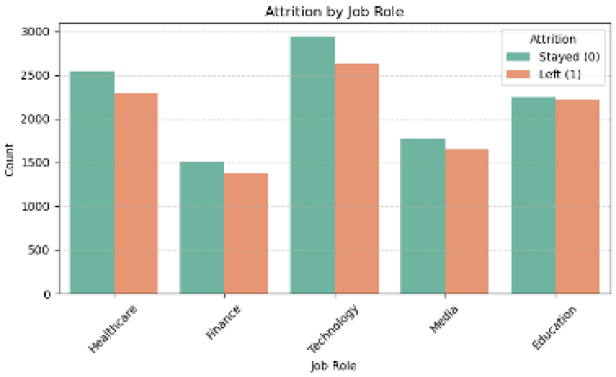


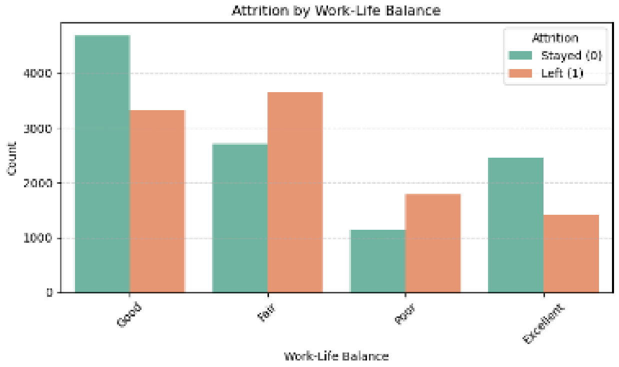
Plot 22: Histogram of Monthly Income

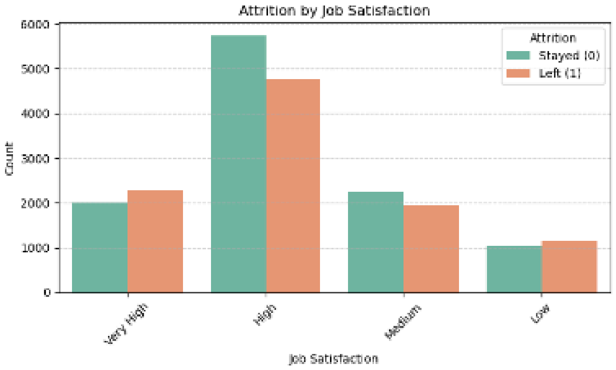
Observation:

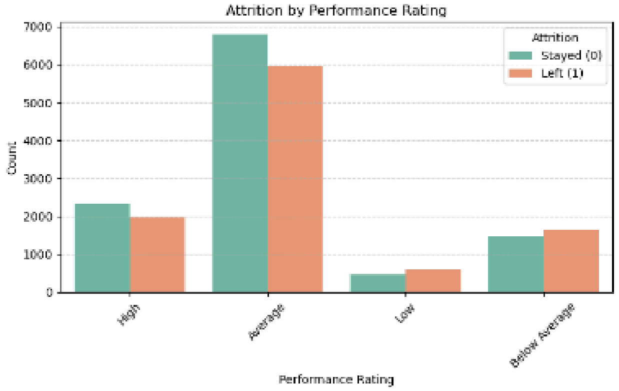
- Displays income distribution across the workforce.
- Helps understand pay scale spreads and can be analyzed against attrition rates for insights on compensation fairness.

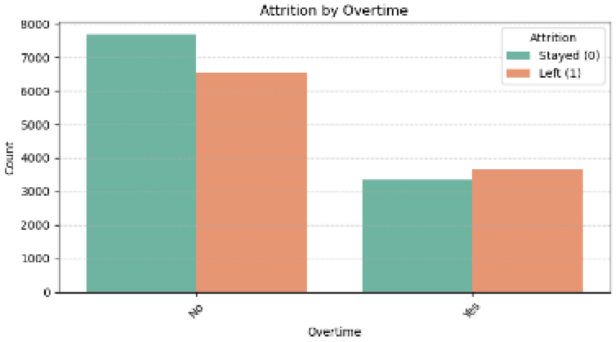

Attrition by Gender
Count
0
1000
2000
3000
4000
5000
6000
Male
Female
Gender
Attrition
Stayed (0)
Left (1)

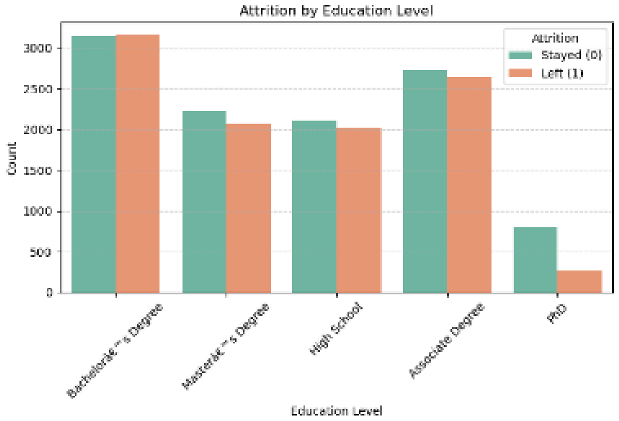

Attrition by Job Role
Count
0
500
1000
1500
2000
2500
3000
Healthcare
Finance
Technology
Media
Education
Job Role
Attrition
Stayed (0)
Left (1)

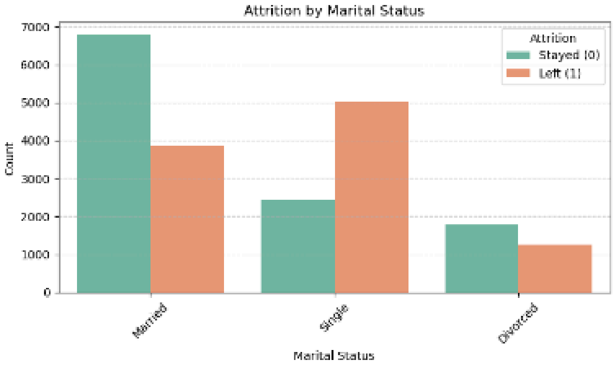

Attrition by Work-Life Balance
Count
0
1000
2000
3000
4000
Good
Fair
Poor
Excellent
Work-Life Balance
Attrition
Stayed (0)
Left (1)

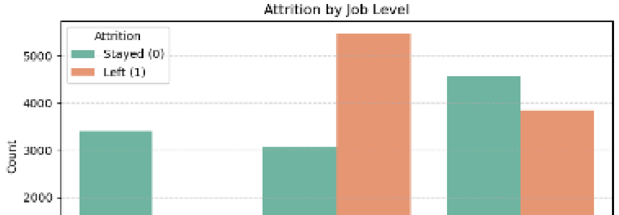

Attrition by Job Satisfaction
Count
0
1000
2000
3000
4000
5000
6000
Very High
High
Medium
Low
Job Satisfaction
Attrition
Stayed (0)
Left (1)

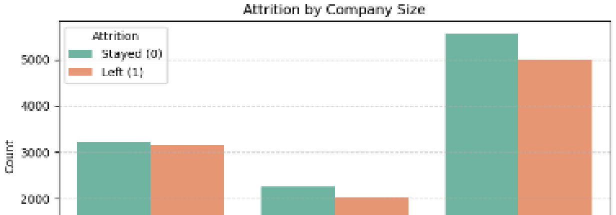

Attrition by Performance Rating
Count
0
1000
2000
3000
4000
5000
6000
7000
High
Average
Low
Below Average
Performance Rating
Attrition
Stayed (0)
Left (1)


Attrition by Overtime
Count
0
1000
2000
3000
4000
5000
6000
7000
8000
No
Yes
Overtime
Attrition
Stayed (0)
Left (1)


Attrition by Education Level
Count
0
500
1000
1500
2000
2500
3000
Bachelorâ€™s Degree
Masterâ€™s Degree
High School
Associate Degree
PhD
Education Level
Attrition
Stayed (0)
Left (1)


Attrition by Marital Status
Count
0
1000
2000
3000
4000
5000
6000
7000
Married
Single
Divorced
Marital Status
Attrition
Stayed (0)
Left (1)


Attrition by Job Level
Count
2000
3000
4000
5000
Attrition
Stayed (0)
Left (1)

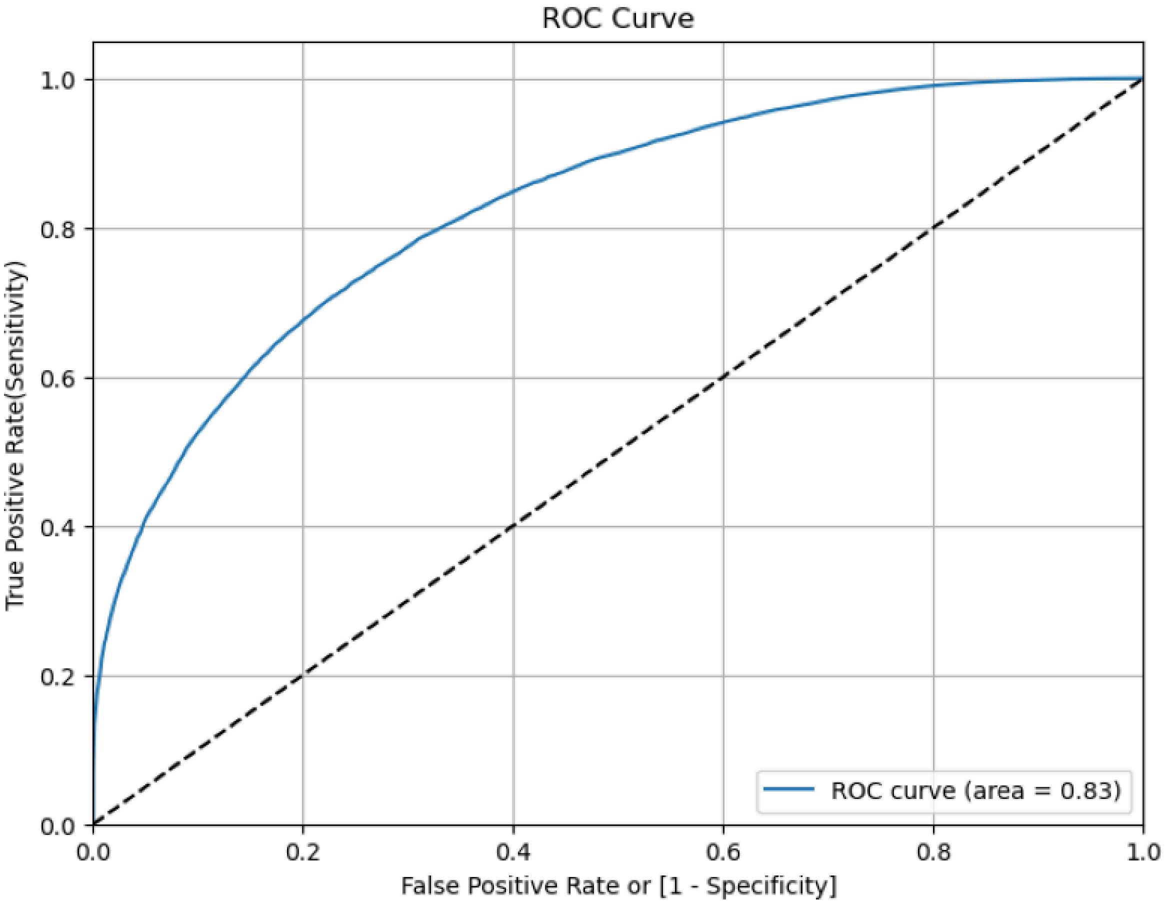

Attrition by Company Size
Count
2000
3000
4000
5000
Attrition
Stayed (0)
Left (1)

Plot 23: Histogram of Years in Current Role

Observation:

- Shows how long employees have been in their current role.
- Short stints might suggest role dissatisfaction or limited growth, contributing to higher turnover.

Plot 24

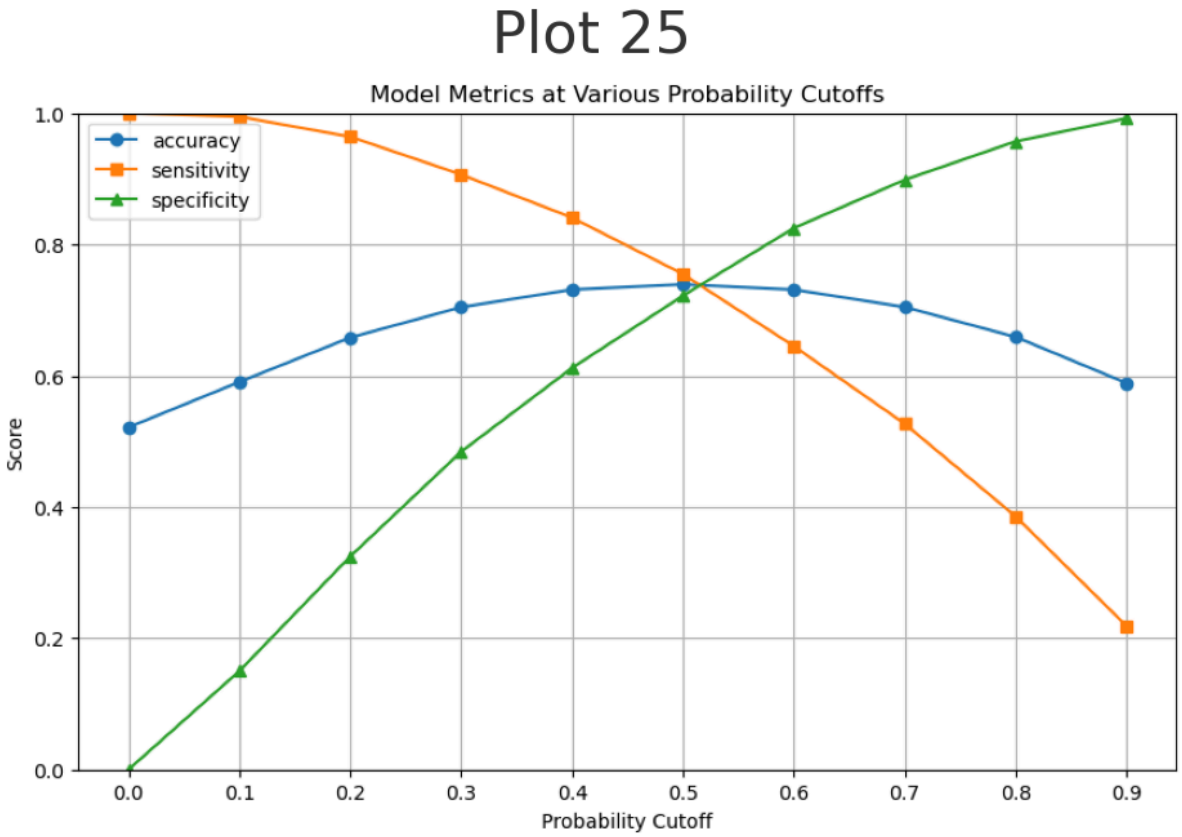


Plot 24: Confusion Matrix of Logistic Regression Model

Observation:

- Evaluates how well the model classifies retention (True Positives, False Negatives, etc.).

- Useful for understanding model accuracy and identifying misclassification trends.



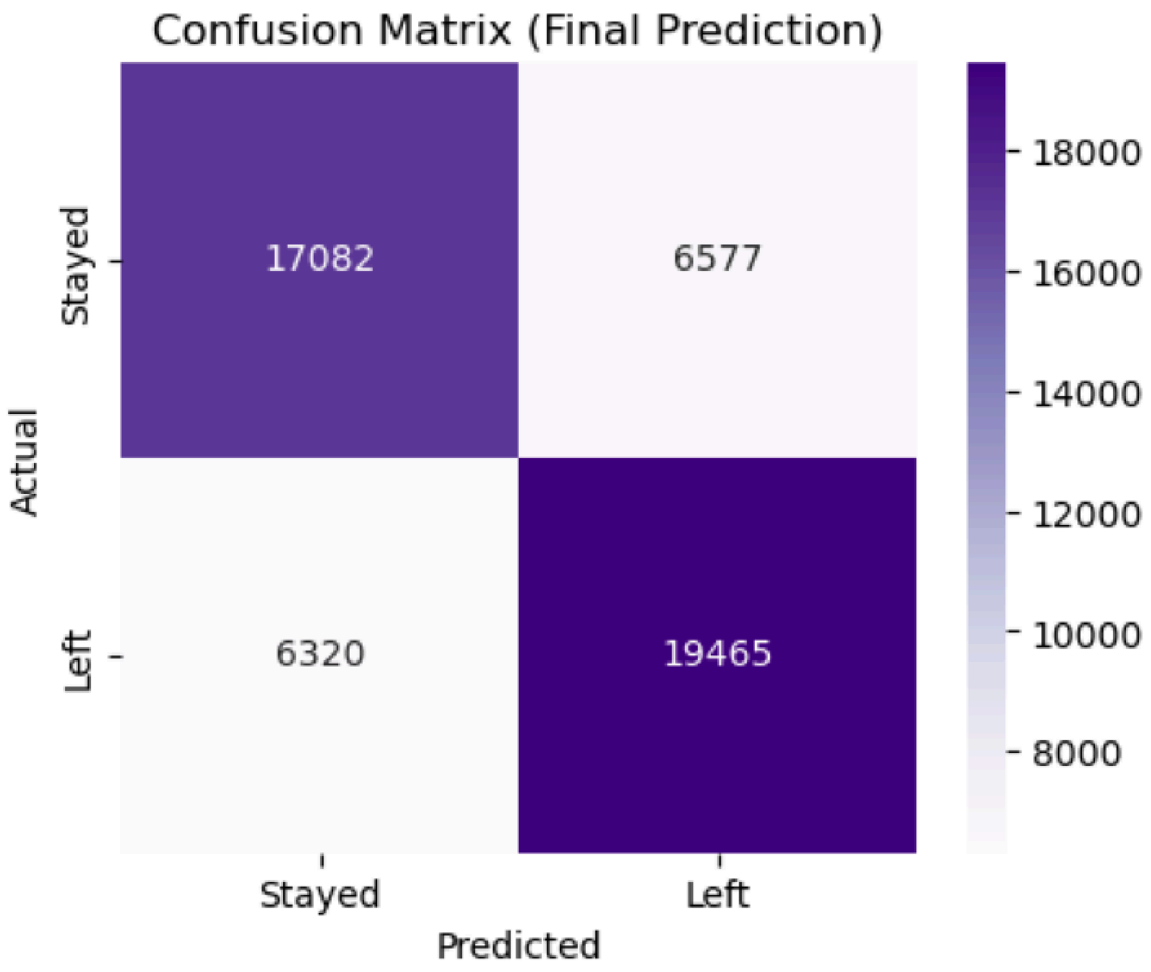
Plot 25: ROC Curve

Observation:

- Illustrates the trade-off between true positive rate and false positive rate.

- The closer the curve is to the top-left corner, the better the model performs.

Plot 26

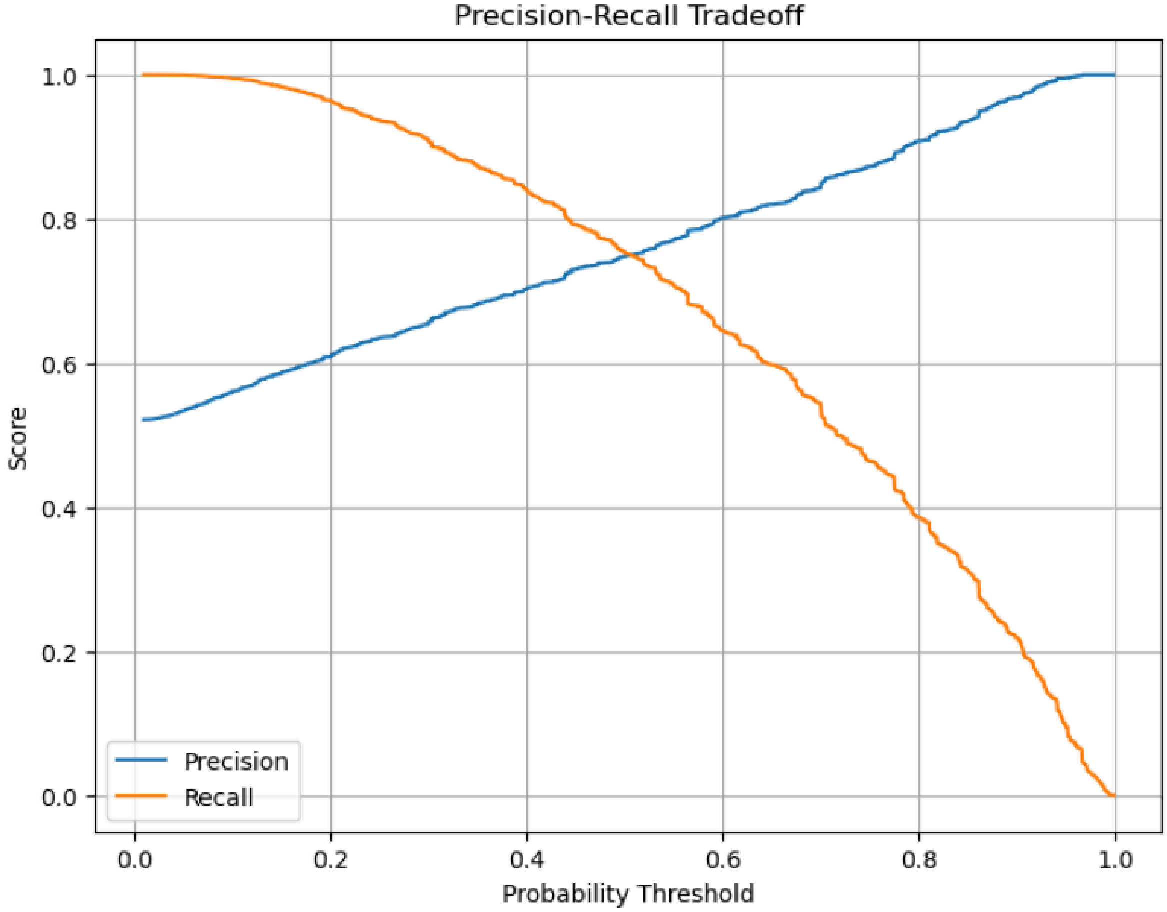


Plot 26: Precision-Recall Curve

Observation:

- Emphasizes the balance between precision and recall, especially useful for imbalanced datasets.
- Helps determine the right threshold for classifying retention effectively.

Plot 27



Plot 27: Feature Importance Plot (Coefficients of Logistic Regression)

Observation:

- Ranks features based on their influence in the model.
- Shows which variables (e.g., job satisfaction, overtime, performance rating) are most predictive of retention.