Employee Attrition Classification

Aysenur Oya OZEN - 2107501

Zeynep TUTAR - 2106038

Introduction to Dataset

The objective of this project is to develop at least two predictive models to determine employee attrition. Additionally, it aims to identify and understand the key factors that contribute to employee turnover.

The dataset used for this project is a simulated dataset designed for the analysis and prediction of employee attrition.

It contains detailed information about various aspects of an employee's profile, including demographics, job-related features, and personal circumstances.

The dataset contains 74,498 samples and 22 features. Each record includes a unique Employee ID and features that influence employee attrition.

Description of the Features

	Variable	Description
1	Employee ID	A unique identifier assigned to each employee
2	Age	The age of the employee, ranging from 18 to 60 years
3	Gender	The gender of the employee
4	Years at Company	The number of years the employee has been working at the company.
5	Monthly Income	The monthly salary of the employee, in dollars
6	Job Role	The department or role the employee works in, encoded into categories such as Finance, Healthcare, Technology, Education, and Media
7	Work-Life Balance	The employee's perceived balance between work and personal life, (Poor, Below Average, Good, Excellent)
8	Job Satisfactio	The employee's satisfaction with their job: (Very Low, Low, Medium, High)
9	Performance Rating	The employee's performance rating: (Low, Below Average, Average, High)
10	Number of Promotions	The total number of promotions the employee has received.
11	Distance from Home	The distance between the employee's home and workplace, in miles.

	Variable	Description
12	Education Level	The highest education level attained by the employee: (High School, Associate Degree, Bachelor’s Degree, Master’s Degree, PhD)
13	Marital Status	The marital status of the employee: (Divorced, Married, Single)
14	Job Level	The job level of the employee: (Entry, Mid, Senior)
15	Company Size	The job level of the employee: (Entry, Mid, Senior)
16	Company Tenure	The total number of years the employee has been working in the industry
17	Remote Work	Whether the employee works remotely: (Yes or No)
18	Leadership Opportunities	Whether the employee has leadership opportunities: (Yes or No)
19	Innovation Opportunities	Whether the employee has opportunities for innovation: (Yes or No)
20	Company Reputation:	The employee's perception of the company's reputation: (Very Poor, Poor, Good, Excellent)
21	Employee Recognition	The level of recognition the employee receives:(Very Low, Low, Medium, High)
22	Attrition	Whether the employee has left the company, encoded as 0 (stayed) and 1 (Left)

Exploraty Data Anlaysis (EDA)

Step by Step

Removing Columns

Employee.ID and Company.Tenure dropped as they are not useful for predictive modeling.Company.Tenure column gives logically incorrect numerical values.

Changing Features Type

Some numeric columns are converted that represent categories (Number.of.Promotions, Number.of.Dependents) to factors to treat them appropriately in analyses and visualizations.

Numerical and Categorical Variables Separation

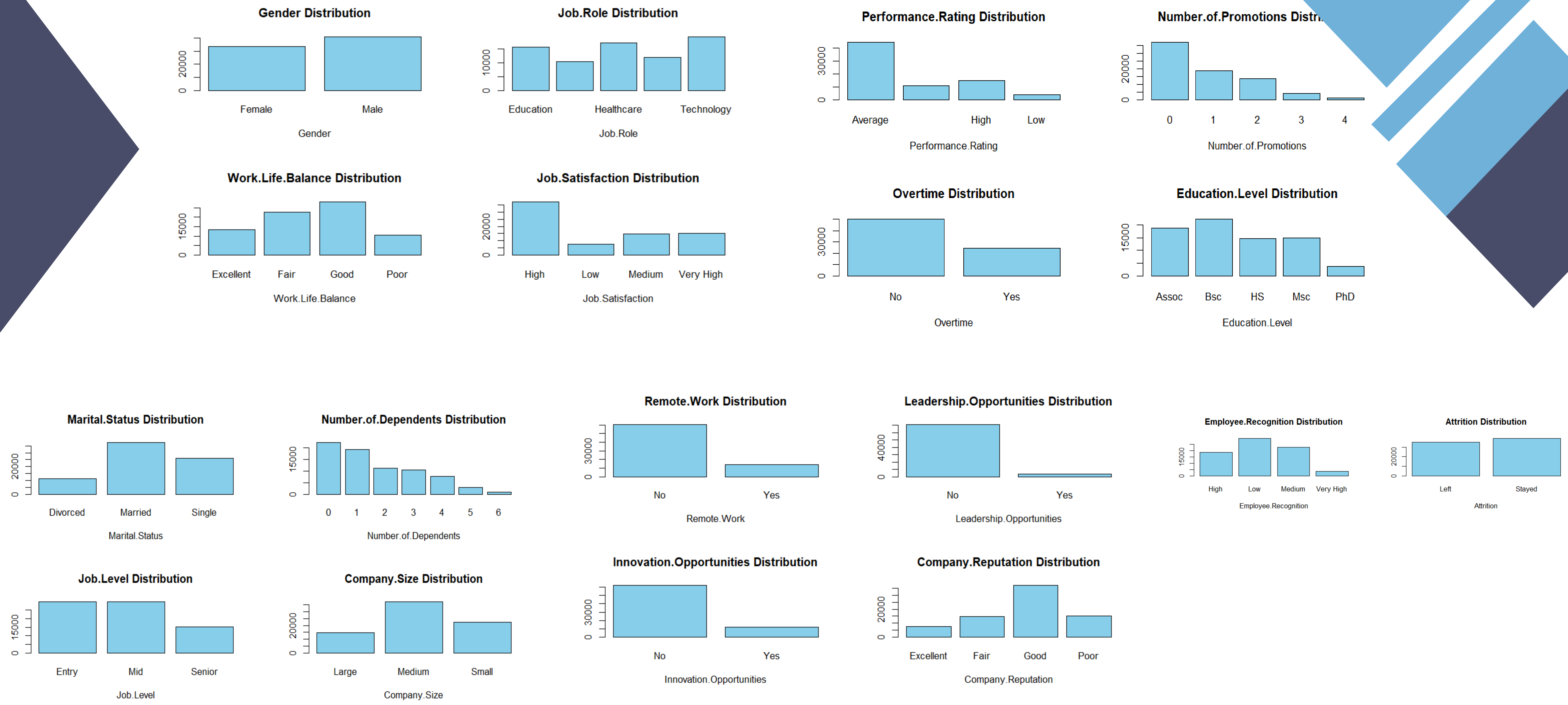
Numerical and categorical variables were separated for targeted analysis. Summary statistics were then used to provide a quick overview of the distribution of numerical features.

Variable	Type
Age	Numeric
Years.at.Company	Numeric
Monthly.Income	Numeric
Distance.from.Home	Numeric
Gender	Categoric
Job.Role	Categoric
Work.Life.Balance	Categoric
Job.Satisfaction	Categoric
Performance.Rating	Categoric
Number.of.Promotions	Categoric
Overtime	Categoric
Education.Level	Categoric
Marital.Status	Categoric
Number.of.Dependents	Categoric
Job.Level	Categoric
Company.Size	Categoric
Remote.Work	Categoric
Leadership.Opportunities	Categoric
Innovation.Opportunities	Categoric
Company.Reputation	Categoric
Employee.Recognition	Categoric
Attrition	Categoric

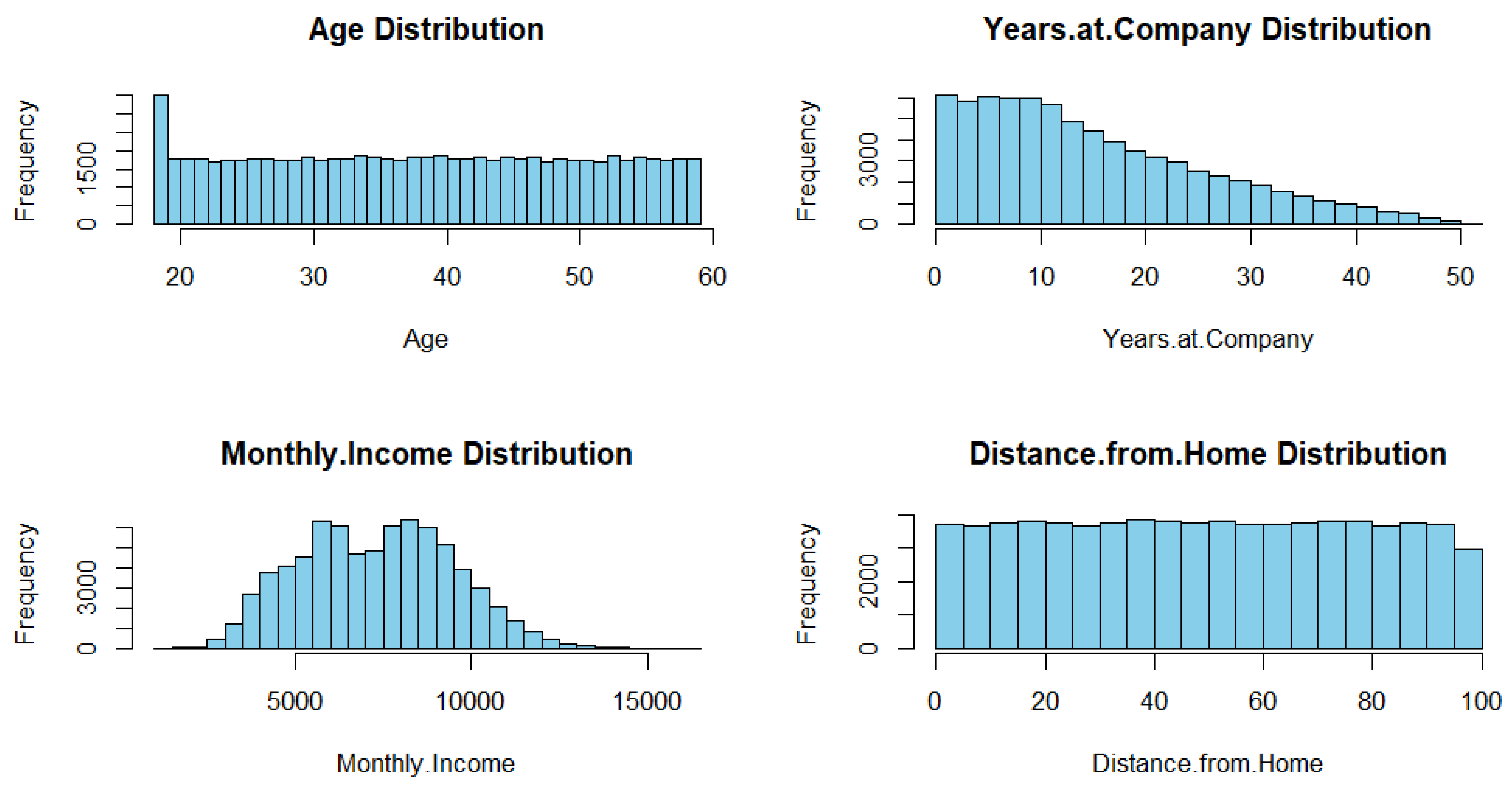
Summary Statistics for Numerical Variables

```
##       Age        Years.at.Company Monthly.Income  Distance.from.Home
##  Min.   :18.00   Min.   : 1.00    Min.   : 1226   Min.   : 1.00
##  1st Qu.:28.00   1st Qu.: 7.00    1st Qu.: 5652   1st Qu.:25.00
##  Median :39.00   Median :13.00    Median : 7348   Median :50.00
##  Mean   :38.53   Mean   :15.72    Mean   : 7299   Mean   :49.99
##  3rd Qu.:49.00   3rd Qu.:23.00    3rd Qu.: 8876   3rd Qu.:75.00
##  Max.   :59.00   Max.   :51.00    Max.   :16149   Max.   :99.00
```

Categorical Variables Distribution

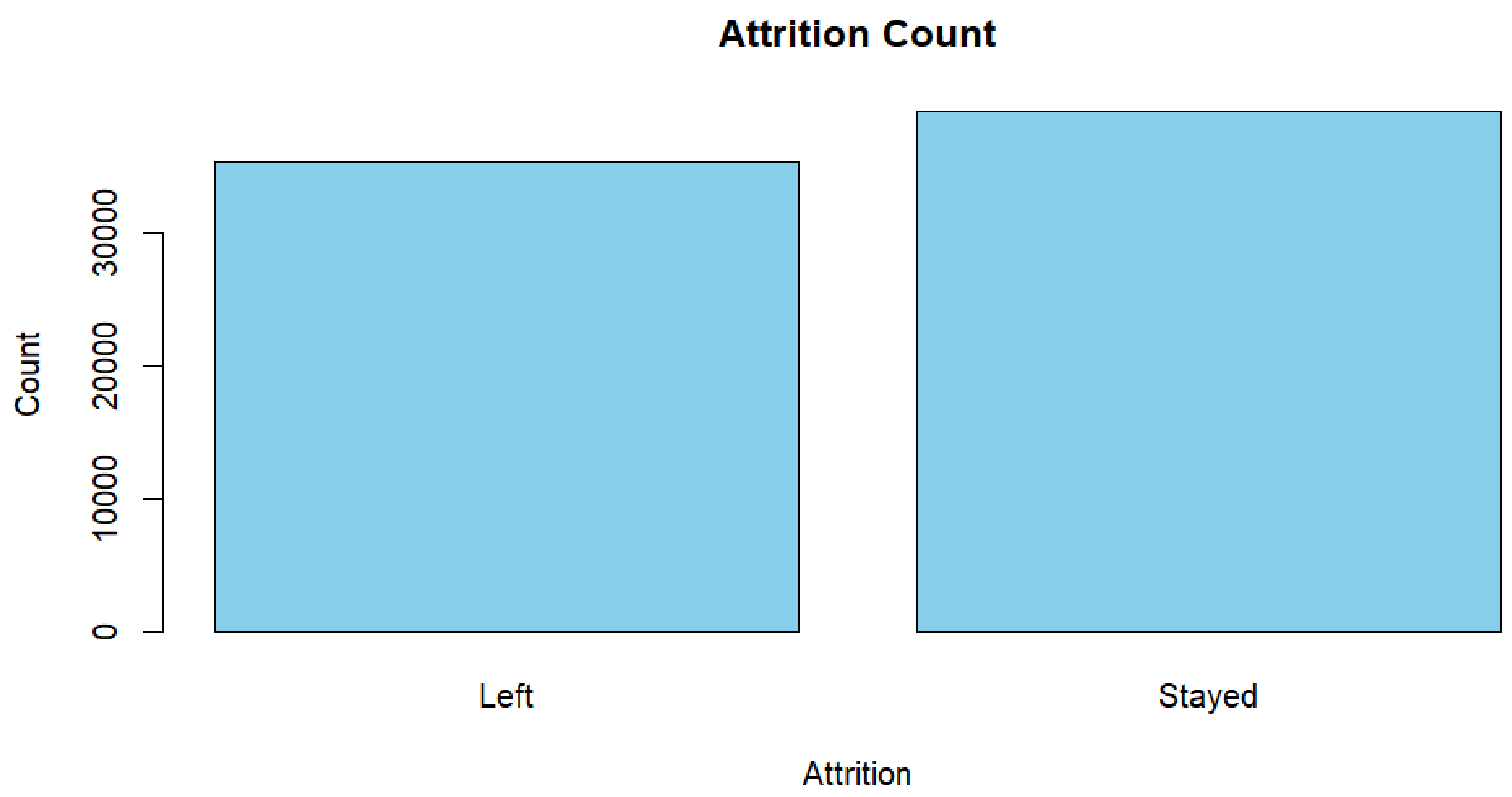


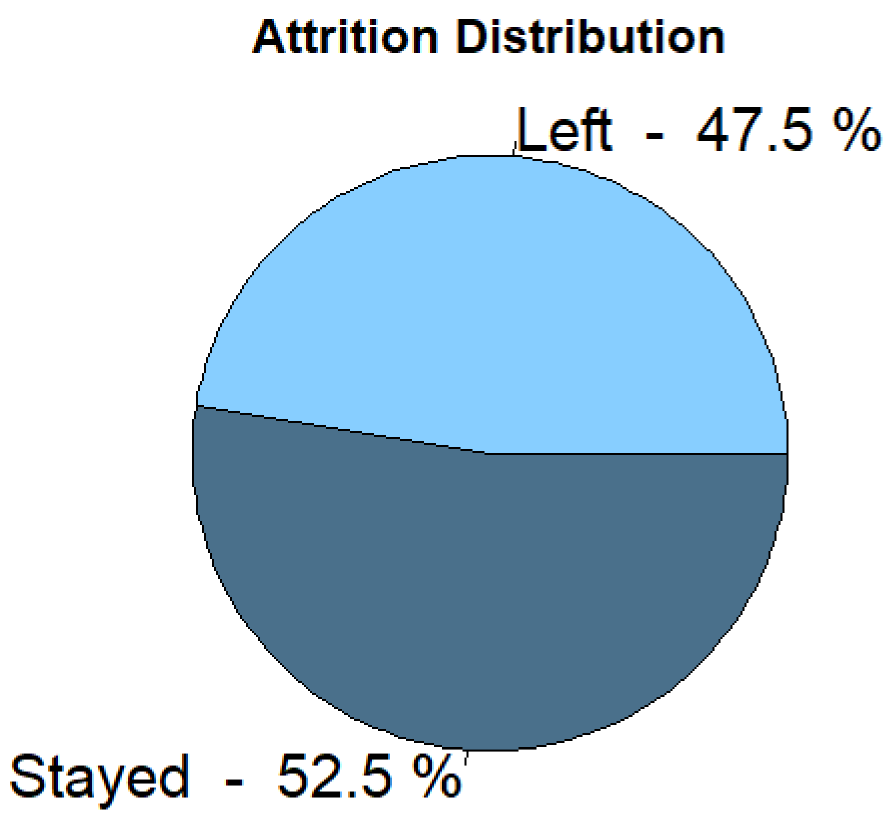
Numerical Variables Distribution



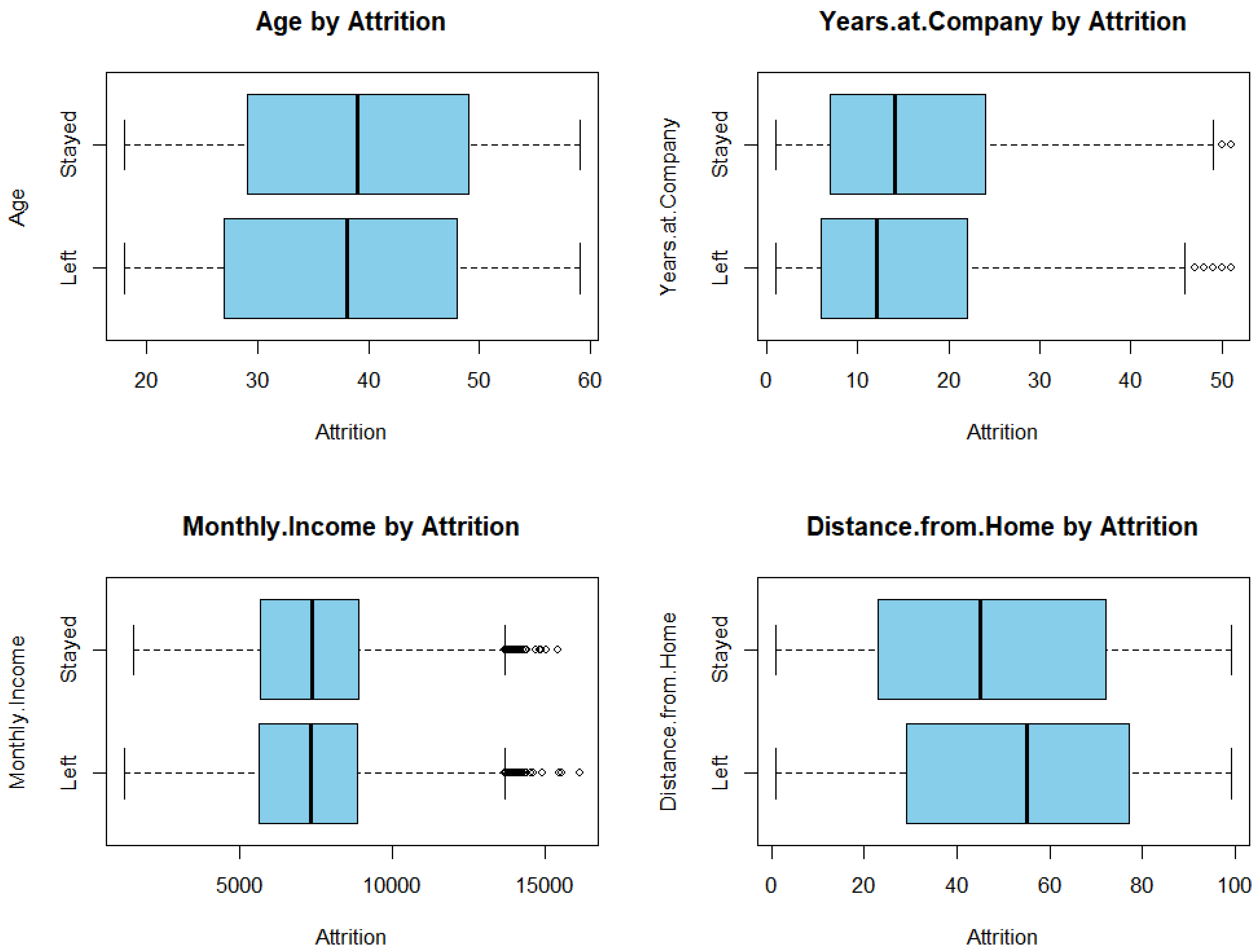
Target Value Distribution

Attrition	Count	Percentage
Left	35.370	47.50%
Stayed	39.128	52.50%

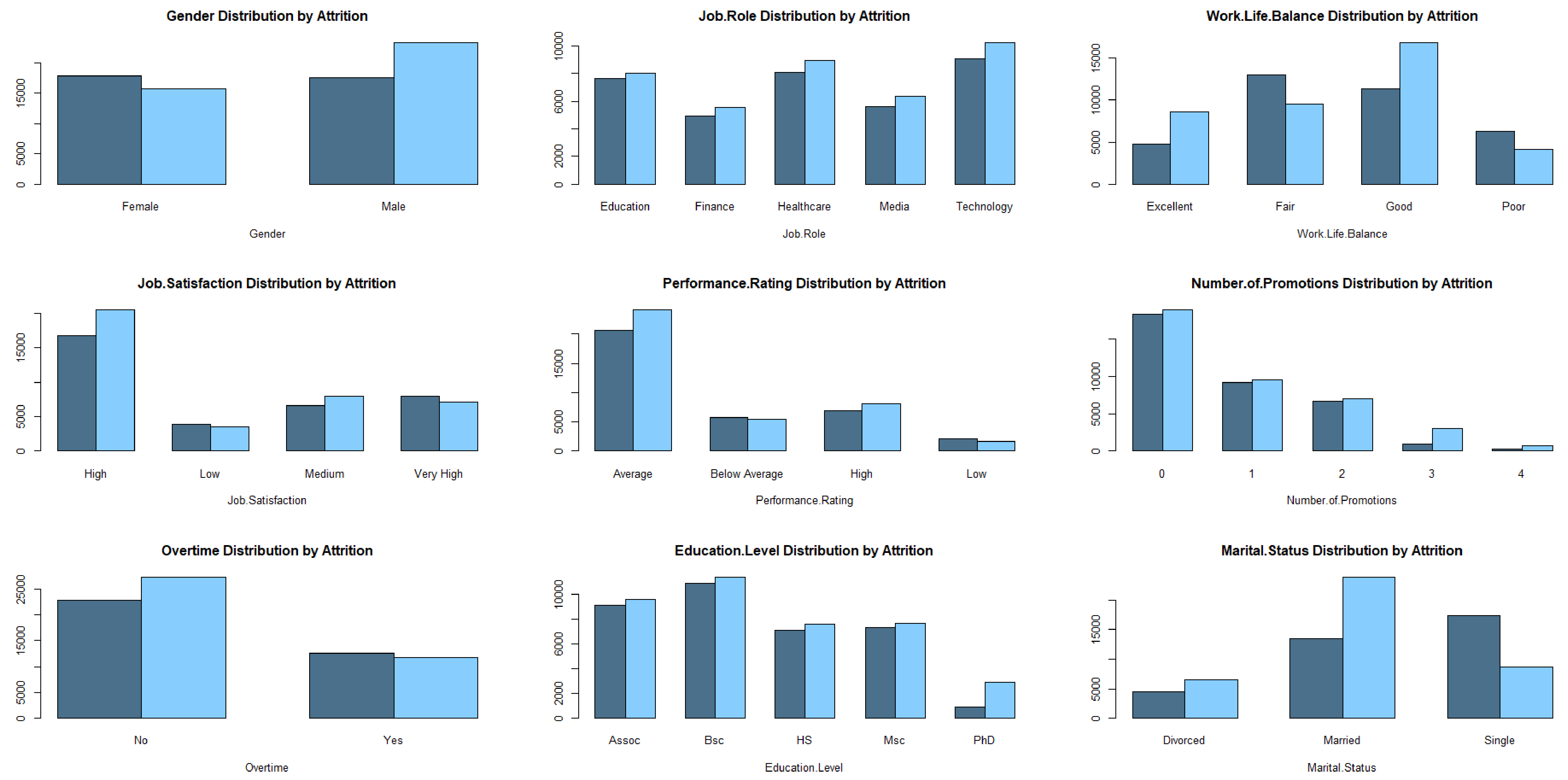




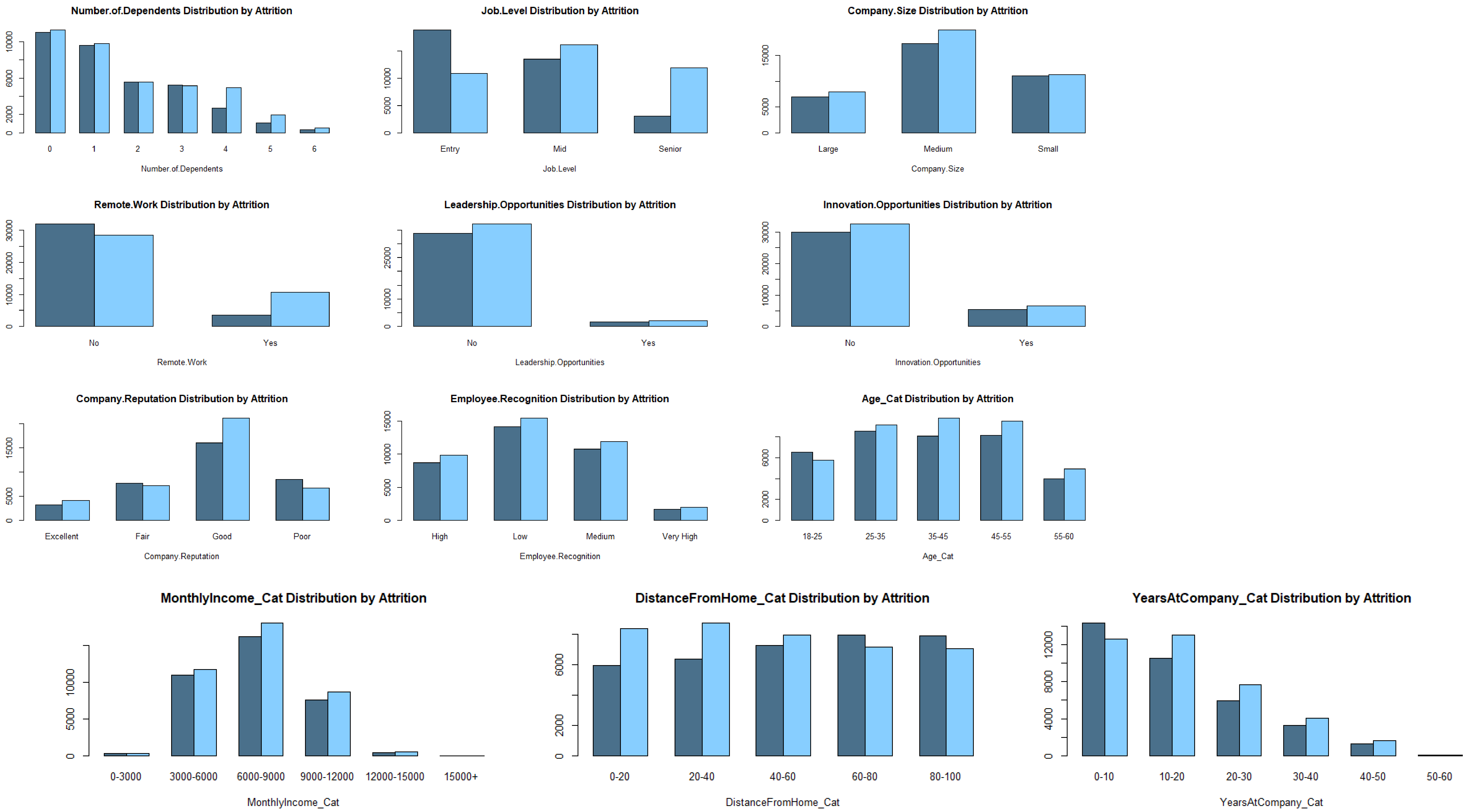
Outlier Analysis



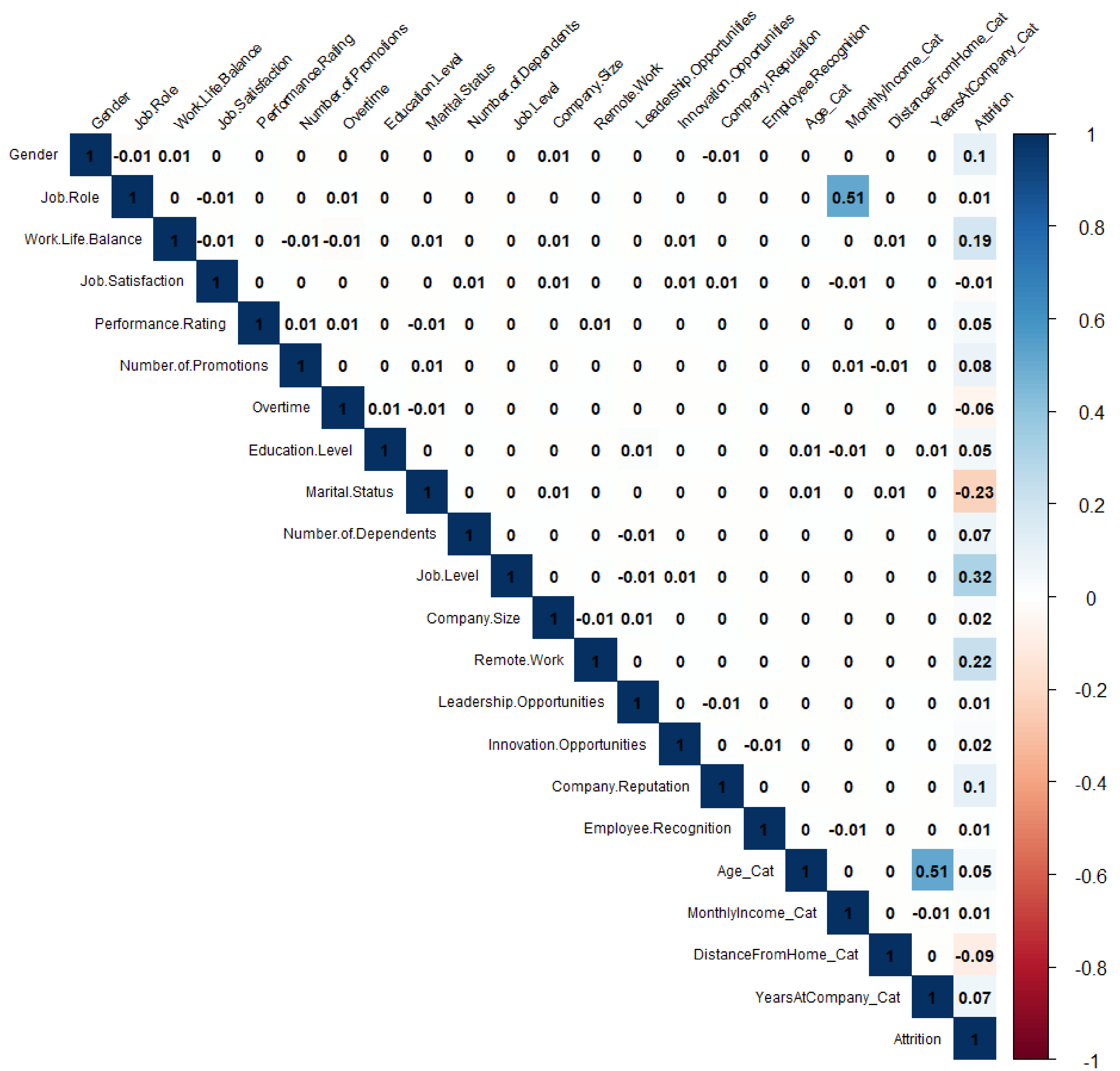
Transforming Numerical Variables into Categorical Variables



This methodology effectively transforms continuous numeric variables into categorical bins, facilitating better visualization and comparative analysis of attrition.


Number.of.Dependents Distribution by Attrition
0 2000 6000 10000
0 1 2 3 4 5 6
Number.of.Dependents
Job.Level Distribution by Attrition
0 5000 10000
Entry Mid Senior
Job.Level
Company.Size Distribution by Attrition
0 5000 15000
Large Medium Small
Company.Size
Remote.Work Distribution by Attrition
0 10000 20000 30000
No Yes
Remote.Work
Leadership.Opportunities Distribution by Attrition
0 10000 25000
No Yes
Leadership.Opportunities
Innovation.Opportunities Distribution by Attrition
0 10000 20000 30000
No Yes
Innovation.Opportunities
Company.Reputation Distribution by Attrition
0 5000 15000
Excellent Fair Good Poor
Company.Reputation
Employee.Recognition Distribution by Attrition
0 5000 10000 15000
High Low Medium Very High
Employee.Recognition
Age_Cat Distribution by Attrition
0 2000 6000
18-25 25-35 35-45 45-55 55-60
Age_Cat
MonthlyIncome_Cat Distribution by Attrition
0 5000 10000
0-3000 3000-6000 6000-9000 9000-12000 12000-15000 15000+
MonthlyIncome_Cat
DistanceFromHome_Cat Distribution by Attrition
0 2000 6000
0-20 20-40 40-60 60-80 80-100
DistanceFromHome_Cat
YearsAtCompany_Cat Distribution by Attrition
0 4000 8000 12000
0-10 10-20 20-30 30-40 40-50 50-60
YearsAtCompany_Cat

Correlation



Data Preparation

Handling Categorical Features

In order to use the categorical features in the model,we need to convert categorical features to numeric (ordinal or nominal) representations.

Normalization

Normalization scales the data values to a range between 0 and 1, which helps improve the performance and training stability.

Normalization Function
A custom function normalize is defined to scale numerical values. The function takes a numeric vector x and returns a normalized vector where each value is scaled between 0 and 1.

$$\text{normalized\_value}$$
$$(x - \min(x))/(\max(x) - \min(x))$$

Train-Test Split

To evaluate the performance of the predictive models, the dataset is split into training and testing sets. 80% of the rows in data as train and the remaining 20% of the data is used for the testing set.

Checking Dataset Dimensions and Balance

Before proceeding to the modeling step, it is essential to examine the dimensions and balance of the datasets. This helps ensure that the training and testing sets are appropriately sized and balanced for effective model training and evaluation.

Number of samples in training data: **59,598**

Number of samples in testing data: **14,900**

Attrition	Percentage
Left	47.33%
Stayed	52.67%

y_train

Attrition	Percentage
Left	48.06%
Stayed	51.94%

y_test

Predictive Classification Models

Logistic Regression

Logistic regression is a model that estimates the odds of the dependent variable occurring and applies the logit (log-odds) transformation to express this relationship.

```
Call:
glm(formula = y.train ~ ., family = binomial, data = X.train)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      -0.553730   0.064437  -8.593  < 2e-16 ***
Age                               0.229757   0.039315   5.844 5.10e-09 ***
GenderMale                        0.543694   0.019788  27.476  < 2e-16 ***
Years.at.Company                  0.654574   0.051922  12.607  < 2e-16 ***
Job.RoleFinance                   0.109910   0.046309   2.373 0.017625 *  
Job.RoleHealthcare                0.061875   0.040506   1.528 0.126628    
Job.RoleMedia                     0.106027   0.034483   3.075 0.002107 ** 
Job.RoleTechnology                0.098439   0.046401   2.121 0.033880 *  
Monthly.Income                    0.018438   0.117791   0.157 0.875614    
Work.Life.Balance                -0.565409   0.031321 -18.052  < 2e-16 ***
Job.Satisfaction                 -0.365624   0.024033 -15.213  < 2e-16 ***
Performance.Rating               -0.289214   0.030814  -9.386  < 2e-16 ***
Number.of.Promotions              0.928255   0.039924  23.250  < 2e-16 ***
OvertimeYes                      -0.336407   0.020902 -16.095  < 2e-16 ***
Distance.from.Home               -0.886487   0.033969 -26.097  < 2e-16 ***
Education.LevelBachelor.s.Degree -0.035818   0.026276  -1.363 0.172835    
Education.LevelHigh.School        0.001582   0.029370   0.054 0.957050    
Education.LevelMaster.s.Degree   -0.006862   0.029087  -0.236 0.813509    
Education.LevelPhD                1.506973   0.053149  28.354  < 2e-16 ***
Marital.StatusMarried             0.257001   0.028268   9.092  < 2e-16 ***
Marital.StatusSingle             -1.409863   0.030515 -46.203  < 2e-16 ***
Number.of.Dependents              0.842765   0.038204  22.060  < 2e-16 ***
Job.Level                         2.292477   0.028636  80.056  < 2e-16 ***
Company.Size                     -0.193325   0.027989  -6.907 4.94e-12 ***
Remote.WorkYes                    1.635385   0.027848  58.726  < 2e-16 ***
Leadership.OpportunitiesYes       0.162824   0.045057   3.614 0.000302 ***
Innovation.OpportunitiesYes       0.127904   0.026574   4.813 1.49e-06 ***
Company.Reputation               -0.342783   0.033757 -10.154  < 2e-16 ***
Employee.Recognition             -0.003910   0.034468  -0.113 0.909677    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82451  on 59597  degrees of freedom
Residual deviance: 62455  on 59569  degrees of freedom
AIC: 62513

Number of Fisher Scoring iterations: 4
```

Basic Logistic Classifier

Understanding Model Fit Using R^2 Statistics

R^2 provides an indication of how well the independent variables in the model explain the variability of the dependent variable.
A higher R^2 value indicates a better fit of the model to the data.

With the full model the value of R^2 **24.25%** indicates that approximately 24.25% of the variance in the target can be explained by the features in the model.

It suggests that the model is not capturing much of the underlying pattern in the data.

Evaluating Multicollinearity Using VIF Values

	features	VIF
7	Job.RoleTechnology	4.317840
5	Job.RoleHealthcare	3.028354
8	Monthly.Income	3.014476
4	Job.RoleFinance	2.699408
20	Marital.StatusSingle	2.164088
19	Marital.StatusMarried	2.085912
6	Job.RoleMedia	1.682568
15	Education.LevelBachelor.s.Degree	1.528098
17	Education.LevelMaster.s.Degree	1.434283
16	Education.LevelHigh.School	1.424799
3	Years.at.Company	1.405468
1	Age	1.403168
18	Education.LevelPhD	1.124677
22	Job.Level	1.098028
24	Remote.WorkYes	1.060436
2	GenderMale	1.013155
14	Distance.from.Home	1.011754
12	Number.of.Promotions	1.009919
21	Number.of.Dependents	1.009231
9	Work.Life.Balance	1.007059
10	Job.Satisfaction	1.005314
13	OvertimeYes	1.005171
27	Company.Reputation	1.002671
11	Performance.Rating	1.001850
23	Company.Size	1.001579
25	Leadership.OpportunitiesYes	1.000932
28	Employee.Recognition	1.000568
26	Innovation.OpportunitiesYes	1.000566

A VIF value of 1 indicates no correlation, values between 1 and 5 indicate moderate correlation, and values above 5 suggest significant multicollinearity, leading to unreliable coefficient estimates.

Logistic Regression with Backward Stepwise Search

```
Call:
glm(formula = y.train ~ Age + GenderMale + Years.at.Company +
    Work.Life.Balance + Job.Satisfaction + Performance.Rating +
    Number.of.Promotions + OvertimeYes + Distance.from.Home +
    Education.LevelPhD + Marital.StatusMarried + Marital.StatusSingle +
    Number.of.Dependents + Job.Level + Company.Size + Remote.WorkYes +
    Leadership.OpportunitiesYes + Innovation.OpportunitiesYes +
    Company.Reputation, family = binomial, data = X.train)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -0.48883    0.05177  -9.442  < 2e-16 ***
Age                           0.22990    0.03930   5.849 4.94e-09 ***
GenderMale                    0.54243    0.01978  27.423  < 2e-16 ***
Years.at.Company              0.65402    0.05191  12.599  < 2e-16 ***
Work.Life.Balance            -0.56426    0.03131 -18.022  < 2e-16 ***
Job.Satisfaction             -0.36546    0.02403 -15.210  < 2e-16 ***
Performance.Rating           -0.28983    0.03080  -9.409  < 2e-16 ***
Number.of.Promotions          0.92821    0.03991  23.256  < 2e-16 ***
OvertimeYes                  -0.33570    0.02089 -16.068  < 2e-16 ***
Distance.from.Home           -0.88551    0.03396 -26.076  < 2e-16 ***
Education.LevelPhD            1.51895    0.05049  30.086  < 2e-16 ***
Marital.StatusMarried         0.25788    0.02826   9.126  < 2e-16 ***
Marital.StatusSingle         -1.40864    0.03050 -46.183  < 2e-16 ***
Number.of.Dependents          0.84234    0.03820  22.052  < 2e-16 ***
Job.Level                     2.29026    0.02862  80.023  < 2e-16 ***
Company.Size                 -0.19248    0.02798  -6.879 6.02e-12 ***
Remote.WorkYes                1.63481    0.02784  58.729  < 2e-16 ***
Leadership.OpportunitiesYes   0.16201    0.04504   3.597 0.000322 ***
Innovation.OpportunitiesYes   0.12815    0.02657   4.823 1.41e-06 ***
Company.Reputation           -0.34247    0.03374 -10.149  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82451  on 59597  degrees of freedom
Residual deviance: 62476  on 59578  degrees of freedom
AIC: 62516

Number of Fisher Scoring iterations: 4
```

Backward stepwise selection is a greedy algorithm that iteratively removes the least significant features to develop a predictive model. The goal is to find the best subset of features that improve model performance.

Backward Stepwise Search Results

	features	VIF
12	Marital.StatusSingle	2.163022
11	Marital.StatusMarried	2.085498
3	Years.at.Company	1.405299
1	Age	1.402928
14	Job.Level	1.097342
16	Remote.WorkYes	1.060083
10	Education.LevelPhD	1.015134
2	GenderMale	1.012783
9	Distance.from.Home	1.011547
7	Number.of.Promotions	1.009742

This method decreased VIF scores, p-values, and the AIC score. Although $R2$ slightly decreased to **24.23%**, this is expected with fewer features. Most importantly, reducing the variance inflation factor makes the model more stable.

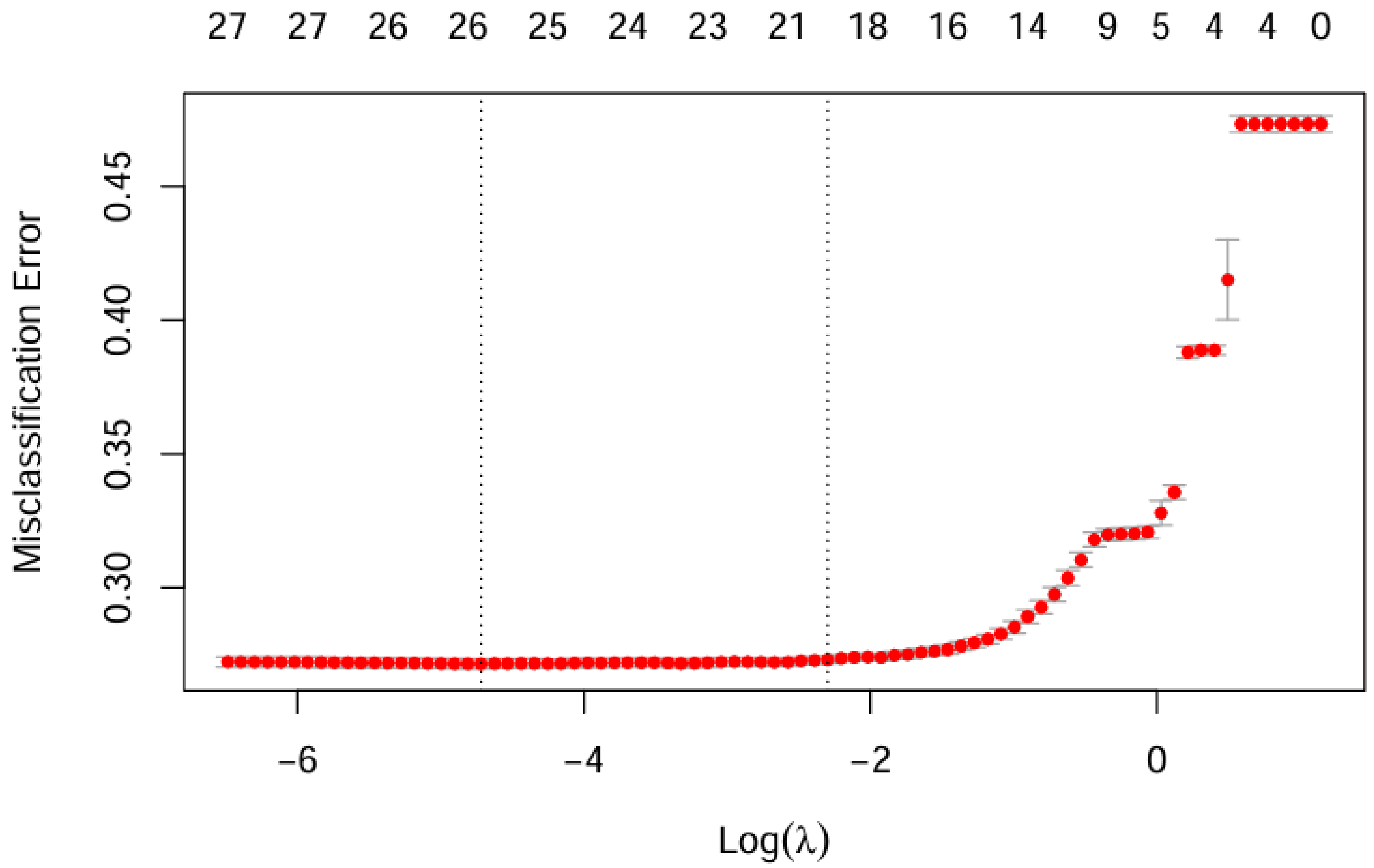
Logistic Regression with Shrinkage Methods

Shrinkage methods prevent overfitting by adding a penalty for large coefficients, thus reducing their variance and enhancing generalizability.

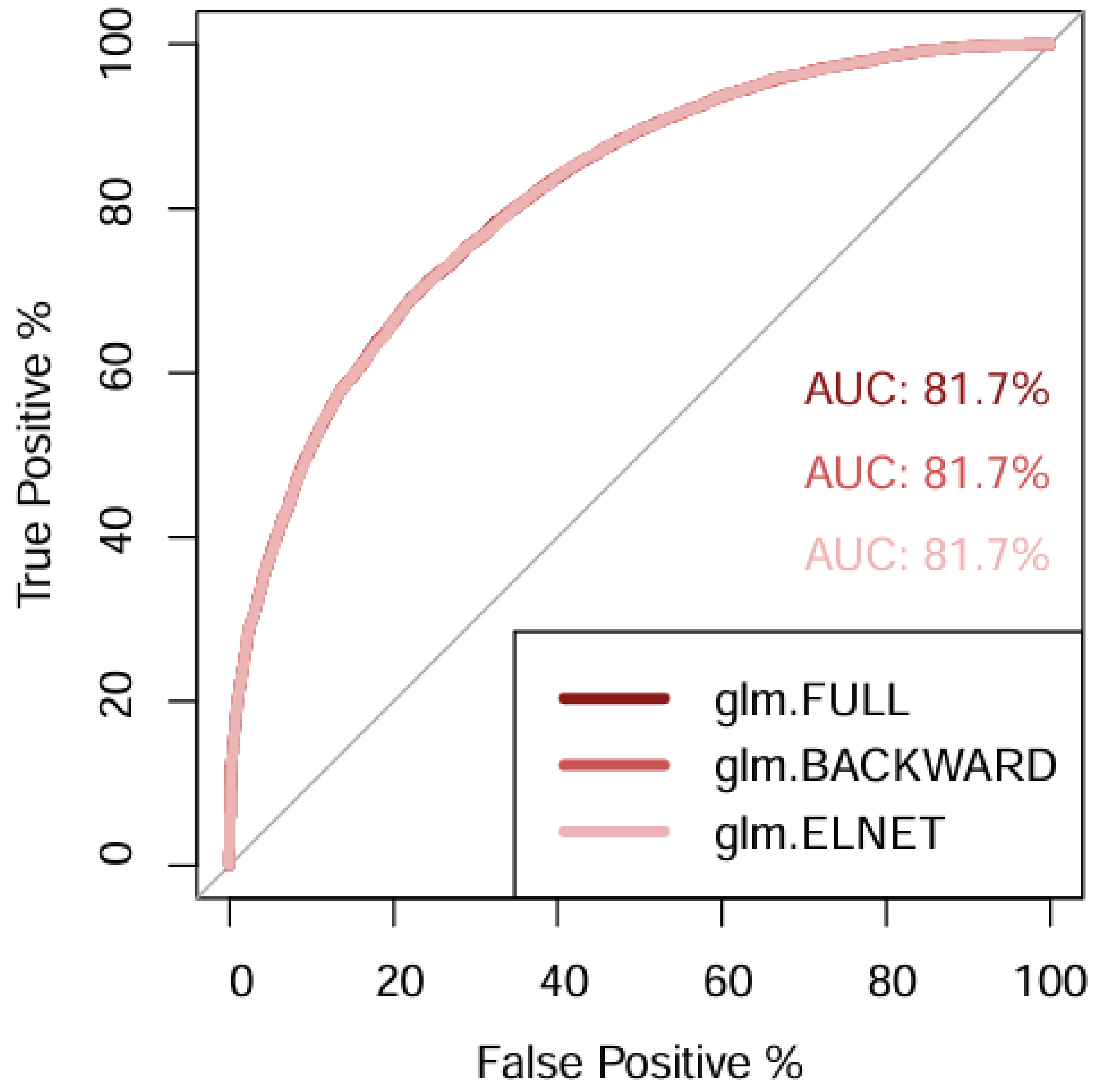
Ridge Regression: Adds a penalty to shrink all coefficients, useful for keeping all features in the model.

Lasso Regression: Introduces a penalty that can shrink some coefficients to zero, aiding in feature selection.

Elastic Net: Combines Ridge and Lasso penalties, allowing for feature selection and handling multicollinearity, leveraging the strengths of both methods.



Comparison of Logistic Classifiers



The ROC curve is a tool for assessing the performance of binary classification models, plotting true positive rate against false positive rate at various thresholds.

The Area Under the Curve (AUC) provides a measure of the model's ability to predict the target values, with higher values indicating better performance.

Logistic Model Performance Metrics at Different Thresholds

Threshold	Accuracy	F1.Score	Precision	Recall
0.2	0.6401	0.7367	0.5941	0.9692
0.3	0.6912	0.7554	0.6417	0.9179
0.4	0.7215	0.7593	0.6889	0.8456
0.5	0.7313	0.7456	0.7337	0.7578
0.6	0.7263	0.7120	0.7850	0.6515

Basic Logistic Classifier

Threshold	Accuracy	F1.Score	Precision	Recall
0.2	0.6395	0.7363	0.5937	0.9691
0.3	0.6907	0.7549	0.6414	0.9170
0.4	0.7226	0.7601	0.6901	0.8458
0.5	0.7323	0.7467	0.7343	0.7595
0.6	0.7270	0.7129	0.7855	0.6525

Feature Selection with Backward Stepwise Search

Threshold	Accuracy	F1.Score	Precision	Recall
0.2	0.6266	0.7306	0.5842	0.9748
0.3	0.6829	0.7524	0.6328	0.9278
0.4	0.7214	0.7609	0.6864	0.8536
0.5	0.7316	0.7462	0.7332	0.7598
0.6	0.7247	0.7083	0.7876	0.6435

Elastic Net Shrinkage Method

Linear Discriminant Analysis (LDA)

LDA is a classification algorithm that finds a linear combination of features that best separates two or more classes.
It assumes that the features follow a multivariate normal distribution with a common mean and variance for all classes.

Accuracy	F1.Score	Precision	Recall
0.7328	0.7462	0.7365	0.756

LDA Results

Quadratic Discriminant Analysis (QDA)

QDA is similar to LDA but allows for different mean and variance for each class. This results in quadratic decision boundaries.

Accuracy	F1.Score	Precision	Recall
0.7121	0.6958	0.7712	0.6338

QDA Results

Conclusion

This study compares the performance of multiple predictive models for employee attrition classification, focusing on Logistic Regression , LDA, and QDA. Each model was evaluated using Accuracy, F1 Score, Precision, and Recall.

Three variations were examined: Basic Logistic Classifier, Logistic Regression with Backward Stepwise Search, and Logistic Regression with Elastic Net Regularization.

In conclusion, **the Elastic Net regularization approach in logistic regression (gml.ELNET)** emerged as the best performer due to its optimal balance of generalization, precision, and recall. This model is recommended for practical applications in predicting employee attrition and devising targeted retention strategies.

Model	Accuracy	F1_Score	Precision	Recall
Basic Logistic Classifier	0.7215	0.7593	0.6889	0.8456
Logistic Regression with Backward Stepwise Search	0.7226	0.7601	0.6901	0.8458
Logistic Regression with Elastic Net	0.7214	0.7609	0.6864	0.8536
Linear Discriminant Analysis	0.7336	0.7471	0.7315	0.7633
Quadratic Discriminant Analysis	0.7130	0.6968	0.6923	0.7012

Thanks!

Do you have any questions?