Università degli Studi di Padova

DIPARTIMENTO MATEMATICA

DIPARTIMENTO DI MATEMATICA "TULLIO LEVI-CIVITA"

Statistical Learning Final Project

Employee Attrition Classification

Authors

Zeynep TUTAR - 2106038

Aysenur Oya ÖZEN - 0000000

Supervisor

Prof. Alberto ROVERATO

Academic Year:
2023/2024

Contents

Introduction to Dataset

The aim of this project is to develop two predictive models to determine employee attrition of a company. The dataset[1] 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. Each record includes a unique Employee ID and features that influence employee attrition. The goal is to understand the factors contributing to attrition and develop predictive models to identify at-risk employees.

The dataset is already split into train and test but in order to better understand the data, it is crucial to analyse the dataset as a whole.

```r
# import the train and test datasets
data_train <- read.csv("data/train.csv", stringsAsFactors = TRUE)
data_test <- read.csv("data/test.csv", stringsAsFactors = TRUE)

# merge the datasets
data <- rbind(data_train, data_test)
attach(data)
```

Description of the Features

The features of the dataset are presented below:

- **Employee ID:** A unique identifier assigned to each employee.
- **Age:** The age of the employee, ranging from 18 to 60 years.
- **Gender:** The gender of the employee
- **Years at Company:** The number of years the employee has been working at the company.
- **Monthly Income:** The monthly salary of the employee, in dollars.
- **Job Role:** The department or role the employee works in, encoded into categories such as Finance, Healthcare, Technology, Education, and Media.
- **Work-Life Balance:** The employee's perceived balance between work and personal life, (Poor, Below Average, Good, Excellent)
- **Job Satisfaction:** The employee's satisfaction with their job: (Very Low, Low, Medium, High)
- **Performance Rating:** The employee's performance rating: (Low, Below Average, Average, High)
- **Number of Promotions:** The total number of promotions the employee has received.
- **Distance from Home:** The distance between the employee's home and workplace, in miles.
- **Education Level:** The highest education level attained by the employee: (High School, Associate Degree, Bachelor's Degree, Master's Degree, PhD)
- **Marital Status:** The marital status of the employee: (Divorced, Married, Single)

[1] https://www.kaggle.com/datasets/stealthtechnologies/employee-attrition-dataset/data

- **Job Level:** The job level of the employee: (Entry, Mid, Senior)
- **Company Size:** The size of the company the employee works for: (Small,Medium,Large)
- **Company Tenure:** The total number of years the employee has been working in the industry.
- **Remote Work:** Whether the employee works remotely: (Yes or No)
- **Leadership Opportunities:** Whether the employee has leadership opportunities: (Yes or No)
- **Innovation Opportunities:** Whether the employee has opportunities for innovation: (Yes or No)
- **Company Reputation:** The employee's perception of the company's reputation: (Very Poor, Poor,Good, Excellent)
- **Employee Recognition:** The level of recognition the employee receives:(Very Low, Low, Medium, High)
- **Attrition:** Whether the employee has left the company, encoded as 0 (stayed) and 1 (Left).

Data Analysis

In order to develop predictive models, first it is necessary to perform exploratory data analysis (EDA) and modify the format of the data if necessary.

```r
# installing required libraries
library(car)
library(dplyr)
library(corrplot)
```

```r
# Descriptive statistics of DataFrame
summary(data)
```

```
  Employee.ID          Age            Gender      Years.at.Company
 Min.   :    1   Min.   :18.00   Female:33672   Min.   : 1.00
 1st Qu.:18625   1st Qu.:28.00   Male  :40826   1st Qu.: 7.00
 Median :37250   Median :39.00                  Median :13.00
 Mean   :37250   Mean   :38.53                  Mean   :15.72
 3rd Qu.:55874   3rd Qu.:49.00                  3rd Qu.:23.00
 Max.   :74498   Max.   :59.00                  Max.   :51.00
       Job.Role      Monthly.Income  Work.Life.Balance  Job.Satisfaction
 Education :15658   Min.   : 1226   Excellent:13432   High     :37245
 Finance   :10448   1st Qu.: 5652   Fair     :22529   Low      : 7457
 Healthcare:17074   Median : 7348   Good     :28158   Medium   :14717
 Media     :11996   Mean   : 7299   Poor     :10379   Very High:15079
 Technology:19322   3rd Qu.: 8876
                    Max.   :16149
     Performance.Rating Number.of.Promotions Overtime     Distance.from.Home
 Average      :44719    Min.   :0.0000       No :50157    Min.   : 1.00
 Below Average:11139    1st Qu.:0.0000       Yes:24341    1st Qu.:25.00
 High         :14910    Median :1.0000                    Median :50.00
```

```
 Low          : 3730    Mean   :0.8329                   Mean   :49.99  
                        3rd Qu.:2.0000                   3rd Qu.:75.00  
                        Max.   :4.0000                   Max.   :99.00  
          Education.Level   Marital.Status  Number.of.Dependents  Job.Level    
 Associate Degree :18649   Divorced:11078  Min.   :0.00          Entry :29780  
 Bachelor’s Degree:22331   Married :37419  1st Qu.:0.00          Mid   :29678  
 High School      :14680   Single  :26001  Median :1.00          Senior:15040  
 Master’s Degree  :15021                   Mean   :1.65                        
 PhD              : 3817                   3rd Qu.:3.00                        
                                           Max.   :6.00                        
 Company.Size   Company.Tenure   Remote.Work Leadership.Opportunities
 Large :14912   Min.   :  2.00   No :60300   No :70845               
 Medium:37231   1st Qu.: 36.00   Yes:14198   Yes: 3653               
 Small :22355   Median : 56.00                                       
                Mean   : 55.73                                       
                3rd Qu.: 76.00                                       
                Max.   :128.00                                       
 Innovation.Opportunities Company.Reputation Employee.Recognition
 No :62394                Excellent: 7414    High     :18550     
 Yes:12104                Fair     :14786    Low      :29620     
                          Good     :37182    Medium   :22657     
                          Poor     :15116    Very High: 3671     


  Attrition    
 Left  :35370  
 Stayed:39128  
```

```r
# Data types of columns
str(data)
```

```
'data.frame':	74498 obs. of  24 variables:
 $ Employee.ID             : int  8410 64756 30257 65791 65026 24368 64970 36999 32714 15944 ...
 $ Age                     : int  31 59 24 36 56 38 47 48 57 24 ...
 $ Gender                  : Factor w/ 2 levels "Female","Male": 2 1 1 1 2 1 2 2 2 1 ...
 $ Years.at.Company        : int  19 4 10 7 41 3 23 16 44 1 ...
 $ Job.Role                : Factor w/ 5 levels "Education","Finance",..: 1 4 3 1 1 5 1 2 1 3 ...
 $ Monthly.Income          : int  5390 5534 8159 3989 4821 9977 3681 11223 3773 7319 ...
 $ Work.Life.Balance       : Factor w/ 4 levels "Excellent","Fair",..: 1 4 3 3 2 2 2 1 3 4 ...
 $ Job.Satisfaction        : Factor w/ 4 levels "High","Low","Medium",..: 3 1 1 1 4 1 1 4 3 1 ...
 $ Performance.Rating      : Factor w/ 4 levels "Average","Below Average",..: 1 4 4 3 1 2 3 3 3 1 ...
 $ Number.of.Promotions    : int  2 3 0 1 0 3 1 2 1 1 ...
 $ Overtime                : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 1 2 2 ...
 $ Distance.from.Home      : int  22 21 11 27 71 37 75 5 39 57 ...
 $ Education.Level         : Factor w/ 5 levels "Associate Degree",..: 1 4 2 3 3 2 3 4 3 5 ...
```

```
 $ Marital.Status          : Factor w/ 3 levels "Divorced","Married",..: 2 1 2 3 1 2 1 2 2 3 ...
 $ Number.of.Dependents    : int  0 3 3 2 0 0 3 4 4 4 ...
 $ Job.Level               : Factor w/ 3 levels "Entry","Mid",..: 2 2 2 2 3 2 1 1 1 1 ...
 $ Company.Size            : Factor w/ 3 levels "Large","Medium",..: 2 2 2 3 2 2 3 2 2 1 ...
 $ Company.Tenure          : int  89 21 74 50 68 47 93 88 75 45 ...
 $ Remote.Work             : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
 $ Leadership.Opportunities: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Innovation.Opportunities: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 2 ...
 $ Company.Reputation      : Factor w/ 4 levels "Excellent","Fair",..: 1 2 4 3 2 2 3 1 2 3 ...
 $ Employee.Recognition    : Factor w/ 4 levels "High","Low","Medium",..: 3 2 2 3 3 1 3 2 3 2 ...
 $ Attrition               : Factor w/ 2 levels "Left","Stayed": 2 2 2 2 2 1 1 2 2 1 ...
```

Data Preprocessing

To prepare the dataset for further analysis, several data preprocessing steps are performed:

1. Removing features

```r
# first column contains Employee IDs, so not necessary for analysis
data <- data[, !names(data) %in% "Employee.ID"]
```

2. Numeric and categorical value separation

```r
numeric_vars <- sapply(data, is.numeric)
categoric_vars <- sapply(data, function(x) is.factor(x) || is.character(x))

# Taking names of features
categoric_var_names <- names(data)[categoric_vars]
numeric_var_names <- names(data)[numeric_vars]

# Numeric val. summary
summary(data[, numeric_vars])
```

```
      Age         Years.at.Company Monthly.Income  Number.of.Promotions
 Min.   :18.00   Min.   : 1.00    Min.   : 1226   Min.   :0.0000
 1st Qu.:28.00   1st Qu.: 7.00    1st Qu.: 5652   1st Qu.:0.0000
 Median :39.00   Median :13.00    Median : 7348   Median :1.0000
 Mean   :38.53   Mean   :15.72    Mean   : 7299   Mean   :0.8329
 3rd Qu.:49.00   3rd Qu.:23.00    3rd Qu.: 8876   3rd Qu.:2.0000
 Max.   :59.00   Max.   :51.00    Max.   :16149   Max.   :4.0000
 Distance.from.Home Number.of.Dependents Company.Tenure
 Min.   : 1.00      Min.   :0.00         Min.   :  2.00
 1st Qu.:25.00      1st Qu.:0.00         1st Qu.: 36.00
 Median :50.00      Median :1.00         Median : 56.00
 Mean   :49.99      Mean   :1.65         Mean   : 55.73
 3rd Qu.:75.00      3rd Qu.:3.00         3rd Qu.: 76.00
 Max.   :99.00      Max.   :6.00         Max.   :128.00
```

3. Handling missing values

```r
# Missing Values --- No null Values
na_summary <- sapply(data, function(x) sum(is.na(x)))
na_summary
```

```
                     Age                   Gender        Years.at.Company
                       0                        0                       0
                Job.Role           Monthly.Income       Work.Life.Balance
                       0                        0                       0
        Job.Satisfaction       Performance.Rating    Number.of.Promotions
                       0                        0                       0
                Overtime       Distance.from.Home         Education.Level
                       0                        0                       0
          Marital.Status     Number.of.Dependents               Job.Level
                       0                        0                       0
            Company.Size          Company.Tenure             Remote.Work
                       0                        0                       0
Leadership.Opportunities Innovation.Opportunities      Company.Reputation
                       0                        0                       0
    Employee.Recognition                Attrition
                       0                        0
```

Categorical Features

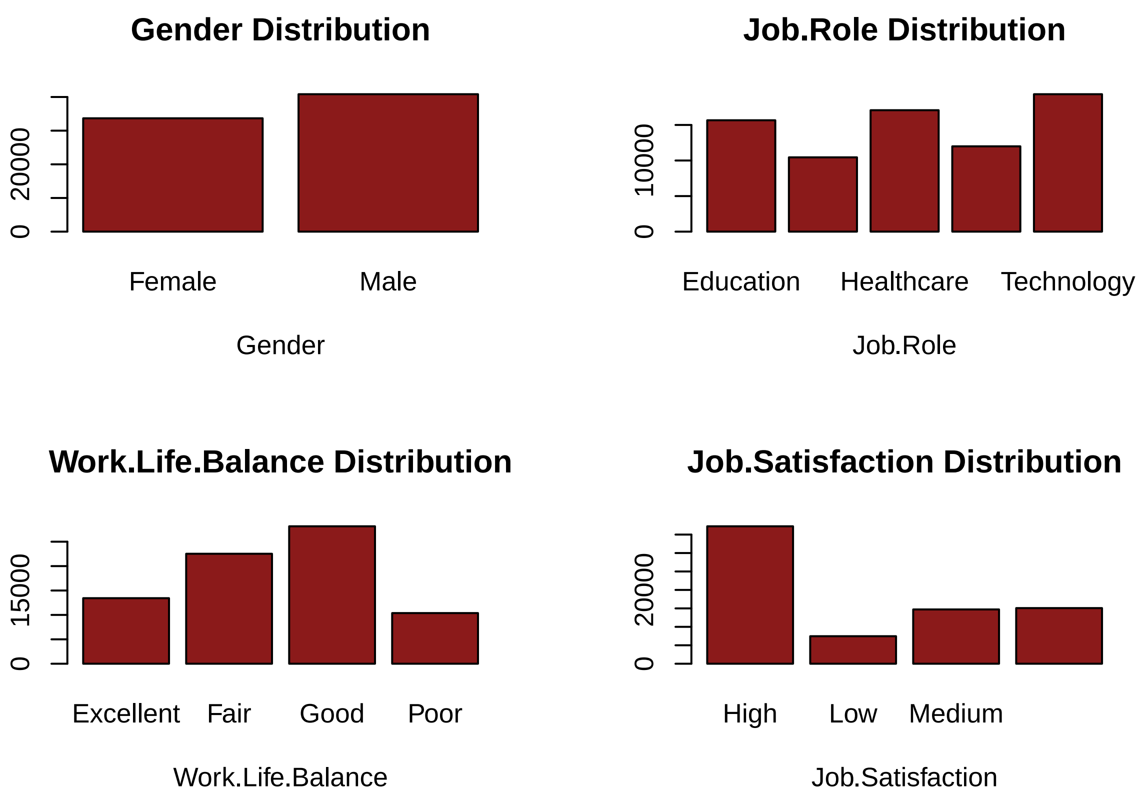
```r
# Categorical val. dist.
categoric_var_names <- names(data)[categoric_vars]
for (var in categoric_var_names) {
    cat("\nDistribution of", var, ":\n")
    print(table(data[[var]]))
}
```

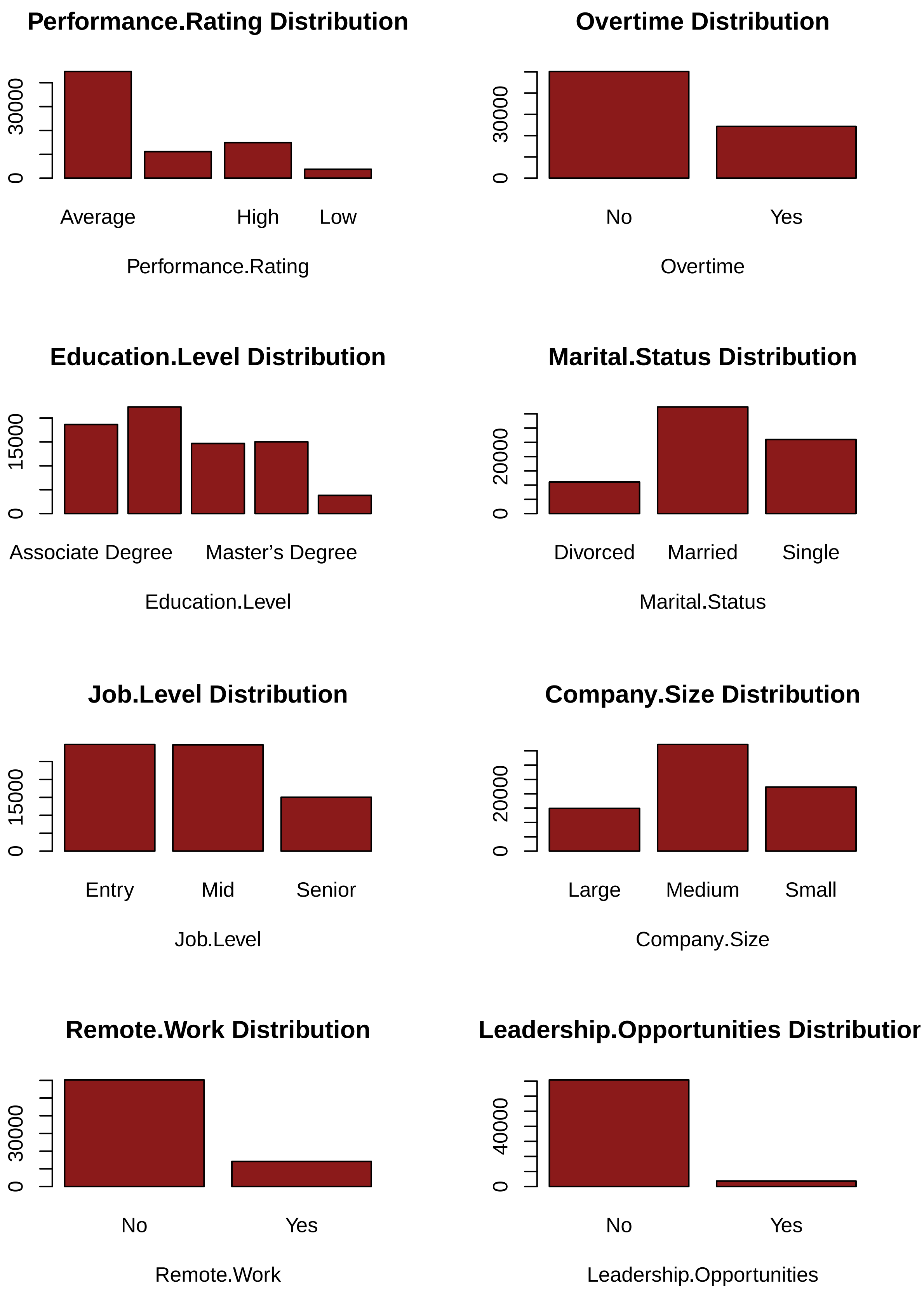
```

Distribution of Gender :

Female   Male
 33672  40826

Distribution of Job.Role :

 Education    Finance Healthcare      Media Technology
     15658      10448      17074      11996      19322

Distribution of Work.Life.Balance :

Excellent      Fair      Good      Poor
    13432     22529     28158     10379
```

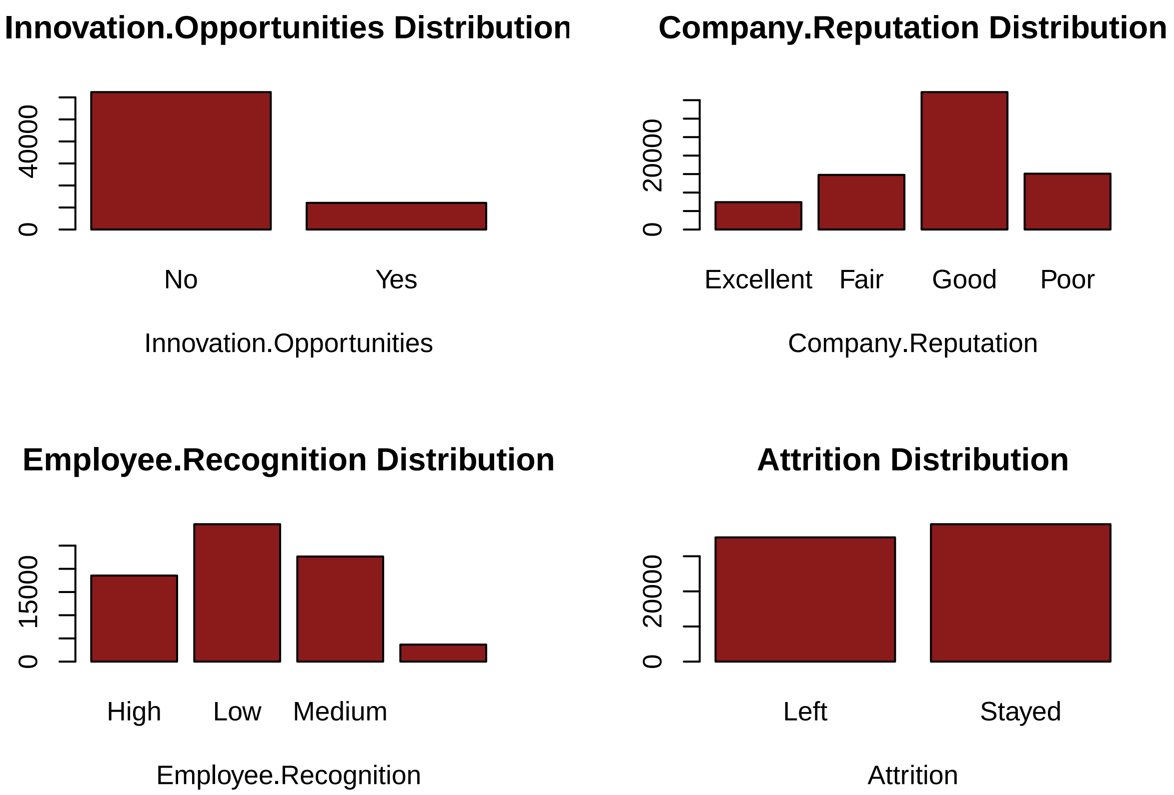
```
Distribution of Job.Satisfaction :

     High       Low    Medium Very High 
    37245      7457     14717     15079 

Distribution of Performance.Rating :

      Average Below Average          High           Low 
        44719         11139         14910          3730 

Distribution of Overtime :

   No   Yes 
50157 24341 

Distribution of Education.Level :

 Associate Degree Bachelor’s Degree       High School   Master’s Degree 
            18649             22331             14680             15021 
              PhD 
             3817 

Distribution of Marital.Status :

Divorced  Married   Single 
   11078    37419    26001 

Distribution of Job.Level :

 Entry    Mid Senior 
 29780  29678  15040 

Distribution of Company.Size :

 Large Medium  Small 
 14912  37231  22355 

Distribution of Remote.Work :

   No   Yes 
60300 14198 

Distribution of Leadership.Opportunities :

   No   Yes 
70845  3653 

Distribution of Innovation.Opportunities :
```

```
   No   Yes
62394 12104

Distribution of Company.Reputation :

Excellent      Fair      Good      Poor
     7414     14786     37182     15116

Distribution of Employee.Recognition :

     High       Low    Medium Very High
    18550     29620     22657      3671

Distribution of Attrition :

  Left Stayed
 35370  39128
```

```r
# Categorical val. dist.--barplot
par(mfrow = c(2, 2))
for (cat_var in categoric_var_names) {
    barplot(table(data[[cat_var]]), main = paste(cat_var, "Distribution"),
        xlab = cat_var, col = "firebrick4")
}
```

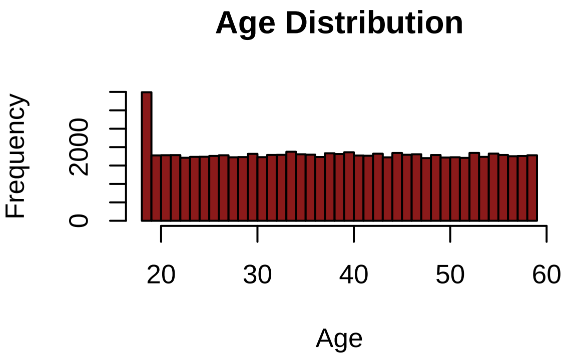


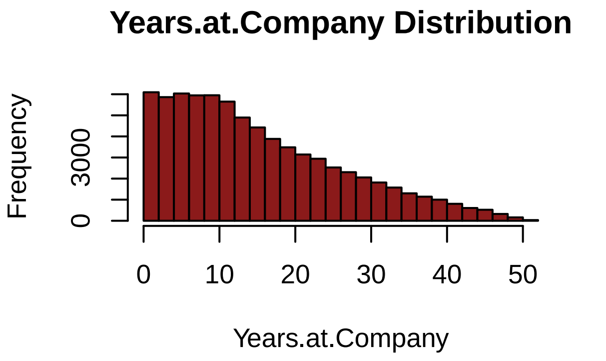

Performance.Rating Distribution
0
30000
Average
High
Low
Performance.Rating
Overtime Distribution
0
30000
No
Yes
Overtime
Education.Level Distribution
0
15000
Associate Degree
Master’s Degree
Education.Level
Marital.Status Distribution
0
20000
Divorced
Married
Single
Marital.Status
Job.Level Distribution
0
15000
Entry
Mid
Senior
Job.Level
Company.Size Distribution
0
20000
Large
Medium
Small
Company.Size
Remote.Work Distribution
0
30000
No
Yes
Remote.Work
Leadership.Opportunities Distributior
0
40000
No
Yes
Leadership.Opportunities

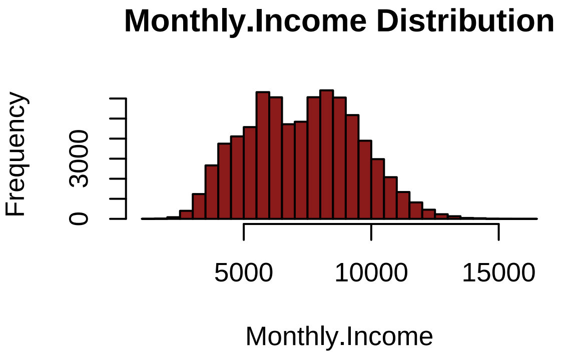


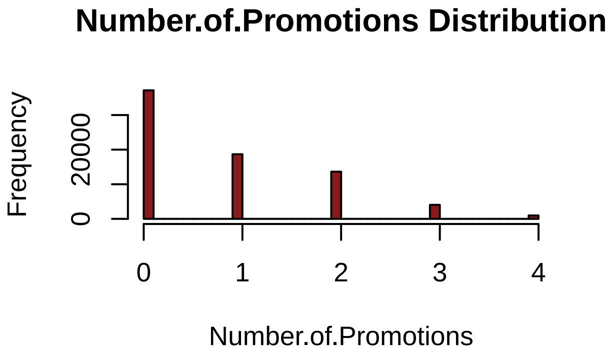
Numeric Features

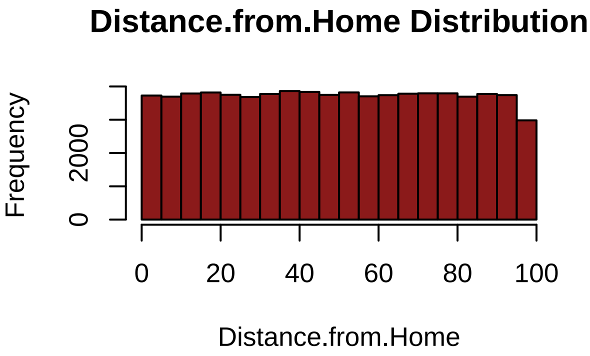
```r
# Numeric features--hist graph
plots_per_page <- 4
num_plots <- length(numeric_var_names)
num_pages <- ceiling(num_plots/plots_per_page)

plot_index <- 1
for (page in 1:num_pages) {
    par(mfrow = c(2, 2))
    for (i in 1:plots_per_page) {
        if (plot_index > num_plots)
            break
        num_var <- numeric_var_names[plot_index]
        hist(data[[num_var]], main = paste(num_var, "Distribution"), xlab = num_var,
            col = "firebrick4", breaks = 30)

        plot_index <- plot_index + 1
    }
}
```

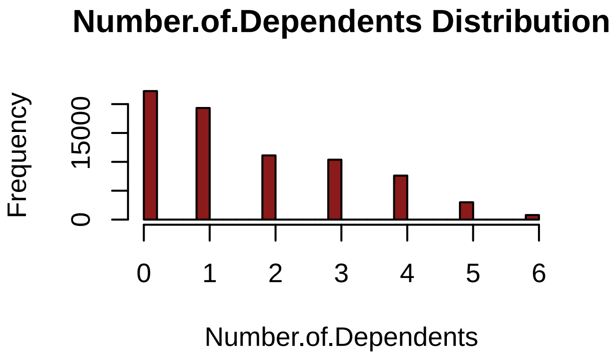

Age Distribution
Frequency
0
2000
20
30
40
50
60
Age

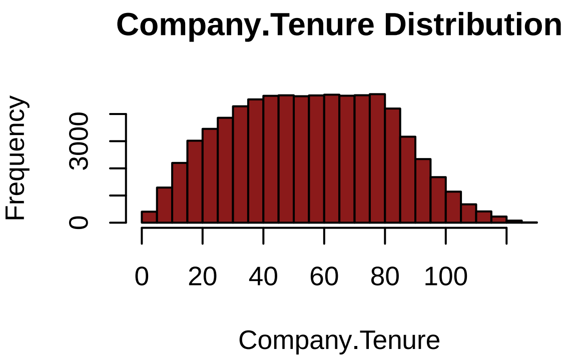

Years.at.Company Distribution
Frequency
0
3000
0
10
20
30
40
50
Years.at.Company


Monthly.Income Distribution
Frequency
0
3000
5000
10000
15000
Monthly.Income


Number.of.Promotions Distribution
Frequency
0
20000
0
1
2
3
4
Number.of.Promotions

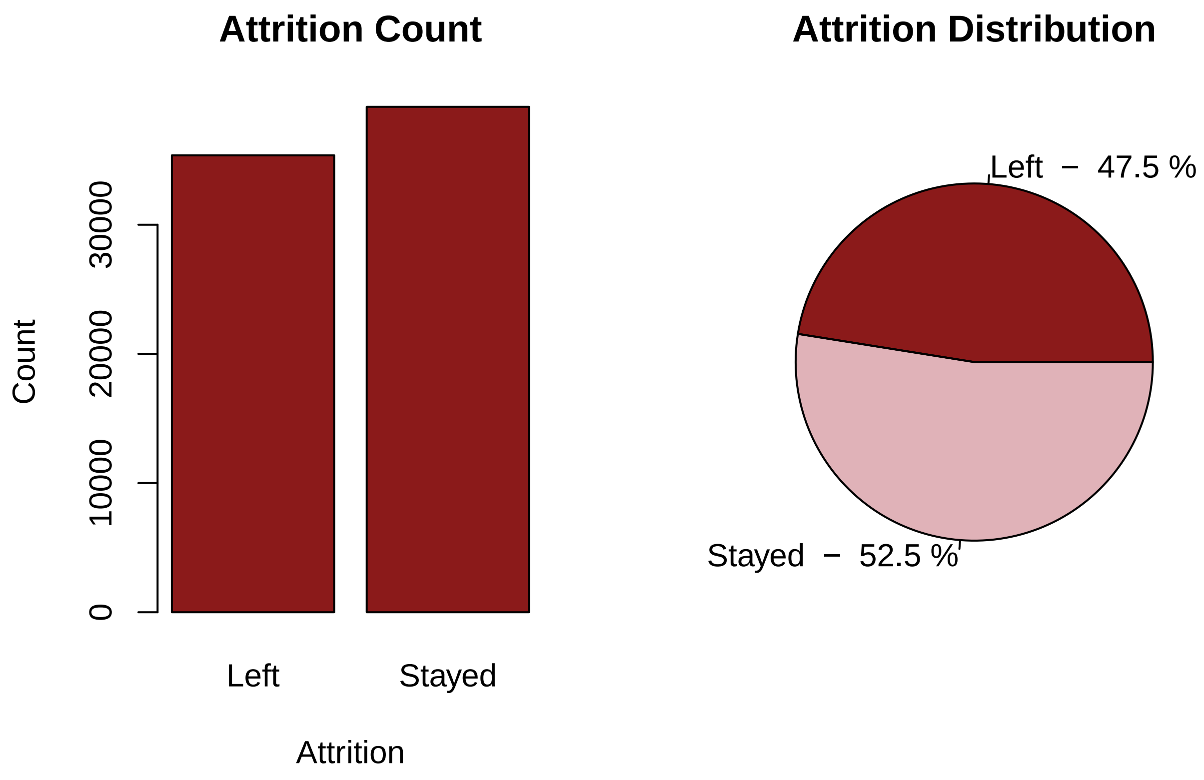

Distance.from.Home Distribution
Frequency
0
2000
0
20
40
60
80
100
Distance.from.Home


Number.of.Dependents Distribution
Frequency
0
15000
0
1
2
3
4
5
6
Number.of.Dependents


Company.Tenure Distribution
Frequency
0
3000
0
20
40
60
80
100
Company.Tenure

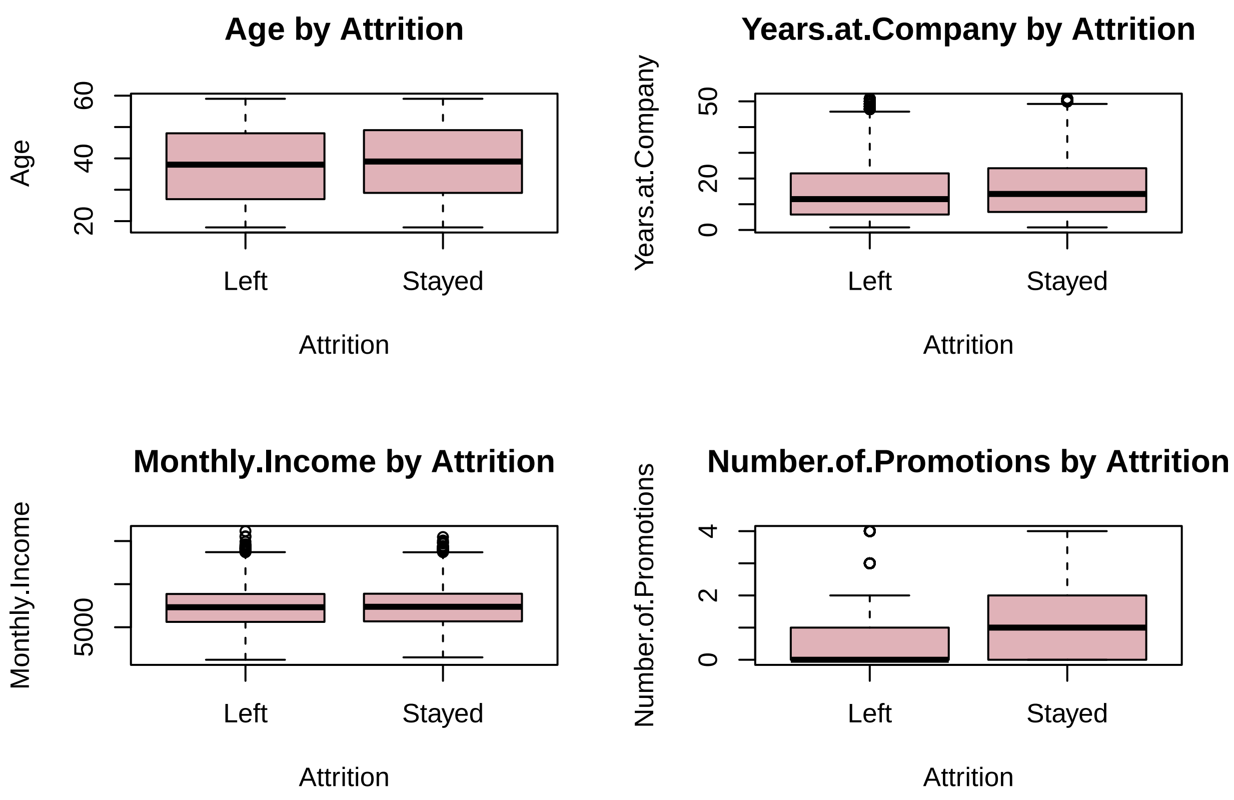
Target Values

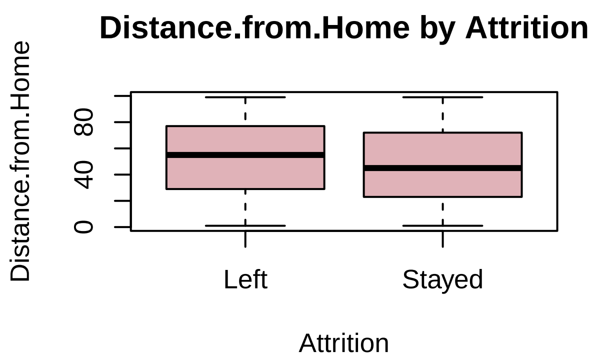
```
# Target values
par(mfrow = c(1, 2))

barplot(table(data$Attrition), main = "Attrition Count", xlab = "Attrition",
    ylab = "Count", col = "firebrick4")

# Target dist - Pie chart
attrition_table <- table(data$Attrition)
attrition_df <- as.data.frame(attrition_table)
colnames(attrition_df) <- c("Attrition", "Count")
attrition_df$Percentage <- round(100 * attrition_df$Count/sum(attrition_df$Count),
    1)

pie(attrition_df$Count, labels = paste(attrition_df$Attrition, " - ",
 ↪  attrition_df$Percentage,
    "%"), col = c("firebrick4", rgb(red = 155/255, green = 0/255, blue = 20/255,
    alpha = 0.3)), main = "Attrition Distribution", cex = 1, radius = 1)
```

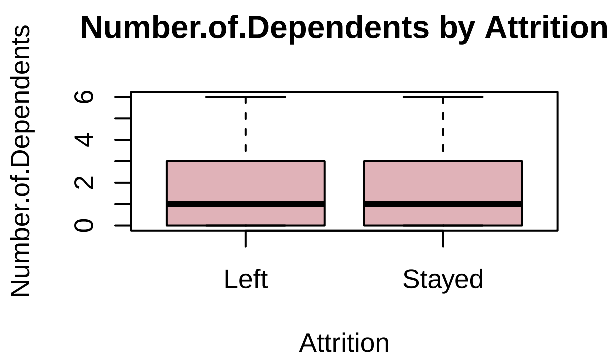


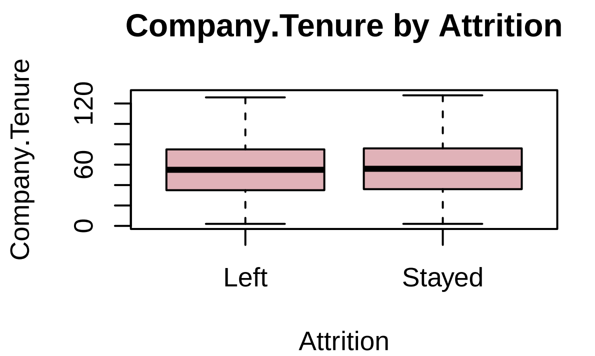

```
# Target Visualization with Numeric features Outlier Check --boxplot
cat_var <- "Attrition"
```

```r
plot_index <- 1
for (page in 1:num_pages) {
    par(mfrow = c(2, 2))
    for (i in 1:plots_per_page) {
        if (plot_index > num_plots)
            break
        num_var <- numeric_var_names[plot_index]
        boxplot(data[[num_var]] ~ data[[cat_var]], main = paste(num_var,
            "by", cat_var), xlab = cat_var, ylab = num_var, col = rgb(red = 155/255,
            green = 0/255, blue = 20/255, alpha = 0.3))
        plot_index <- plot_index + 1
    }
}
```

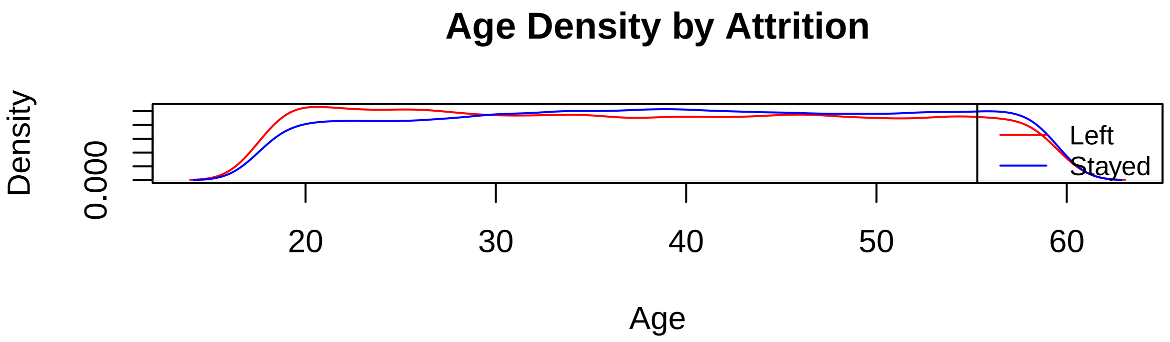


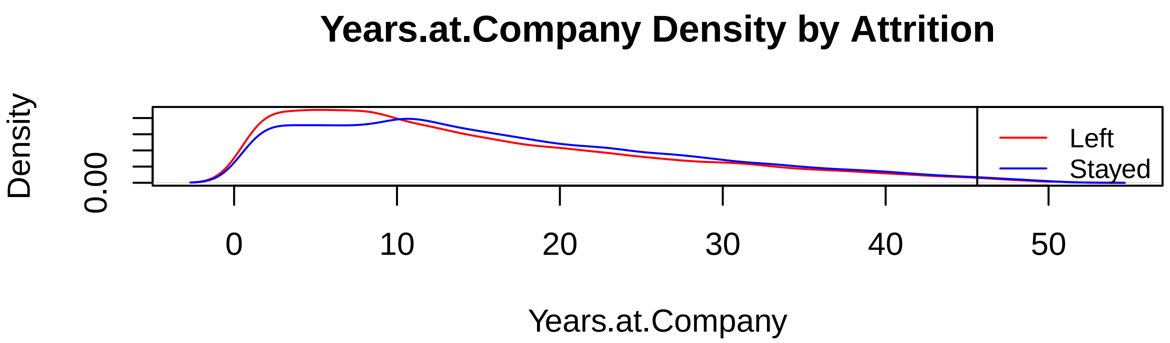


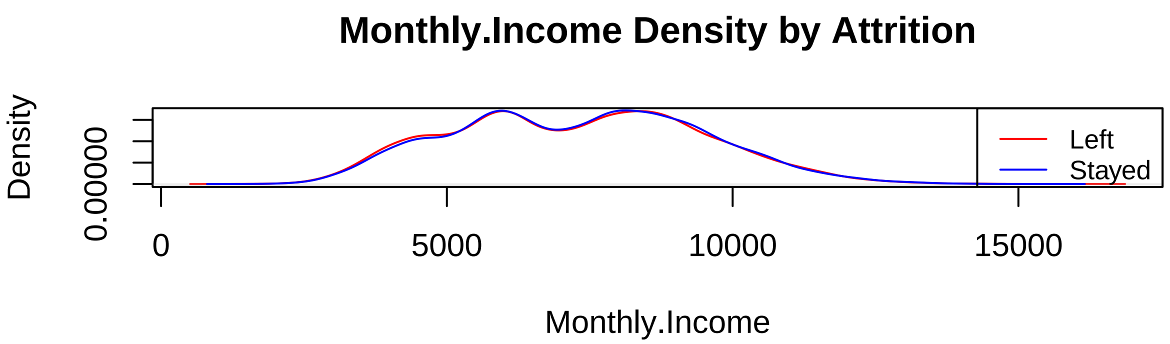


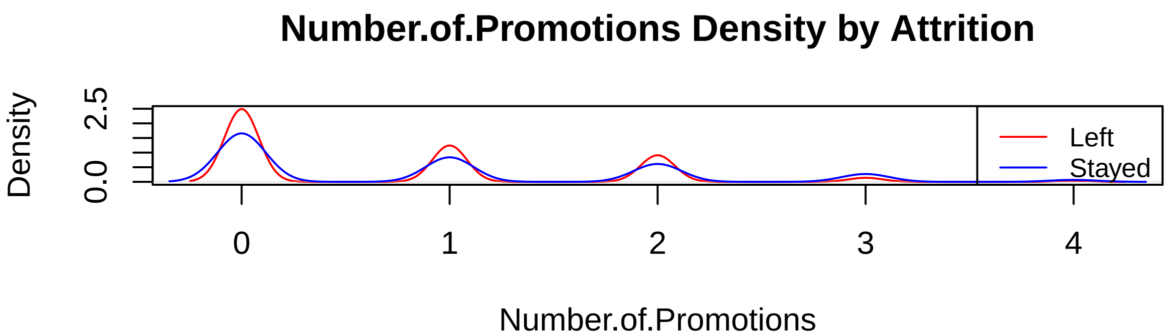


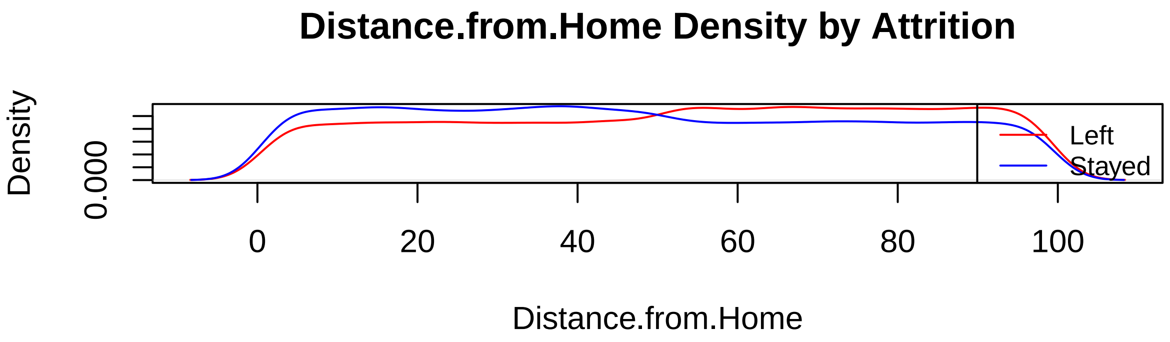

```r
# Density plots
cat_var <- "Attrition"

plot_index <- 1
for (page in 1:num_pages) {
    par(mfrow = c(2, 1))
    for (i in 1:plots_per_page) {
        if (plot_index > num_plots)
            break
        num_var <- numeric_var_names[plot_index]
        plot(density(data[[num_var]][data[[cat_var]] == "Left"], na.rm = TRUE),
            col = "red", main = paste(num_var, "Density by", cat_var), xlab = num_var,
            ylab = "Density")
        lines(density(data[[num_var]][data[[cat_var]] == "Stayed"], na.rm = TRUE),
            col = "blue")
        legend("topright", legend = c("Left", "Stayed"), col = c("red", "blue"),
            lty = 1, cex = 0.8)
        plot_index <- plot_index + 1
    }
}
```

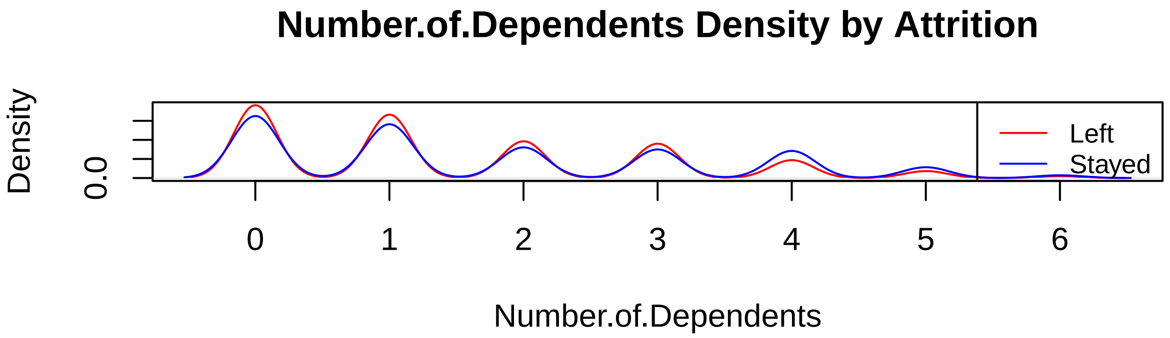

Age Density by Attrition
Density
0.000
20
30
40
50
60
Age
Left
Stayed


Years.at.Company Density by Attrition
Density
0.00
0
10
20
30
40
50
Years.at.Company
Left
Stayed

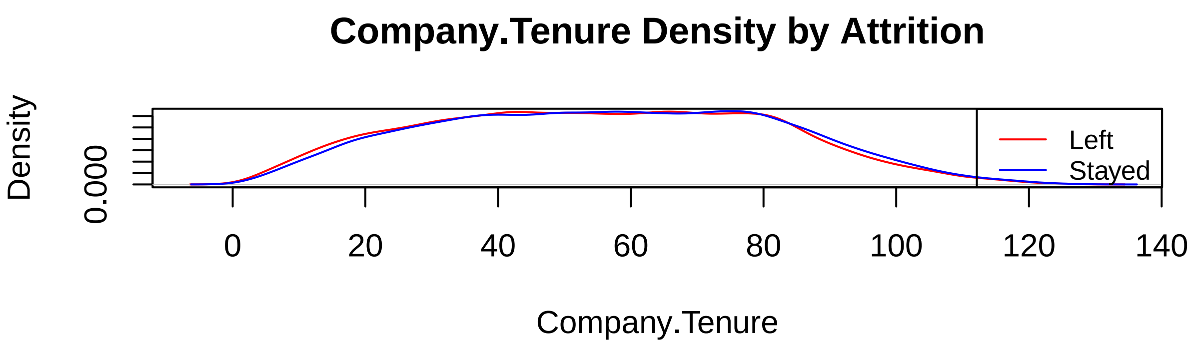

Monthly.Income Density by Attrition
Density
0.00000
0
5000
10000
15000
Monthly.Income
Left
Stayed


Number.of.Promotions Density by Attrition
Density
0.0
2.5
0
1
2
3
4
Number.of.Promotions
Left
Stayed



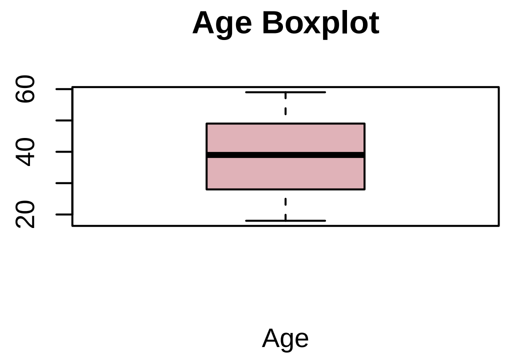


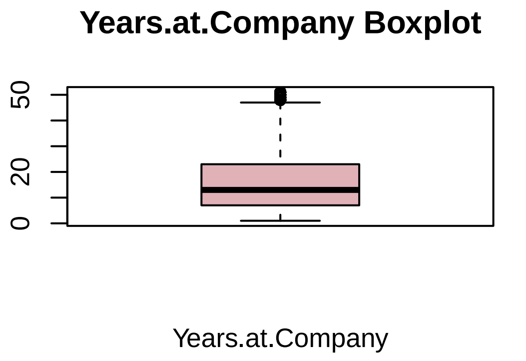

```r
par(mfrow = c(1, 1))
```

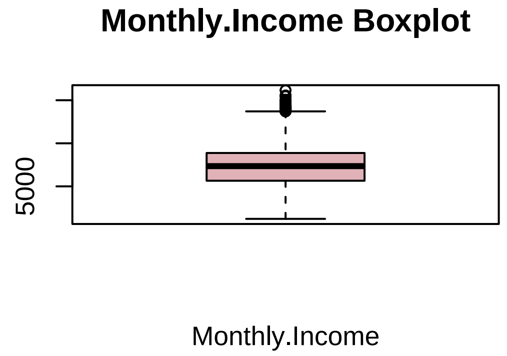


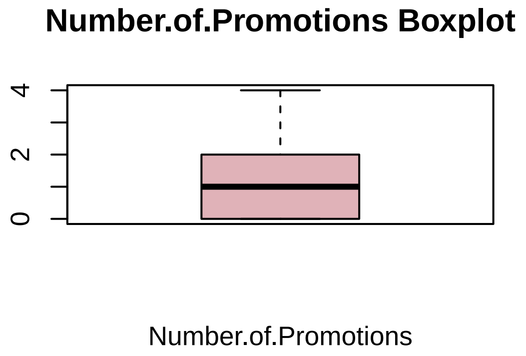
Outliers

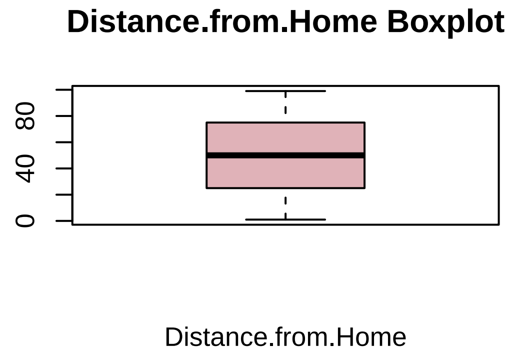
```
# Outlier Analysis
par(mfrow = c(2, 2))
for (num_var in numeric_var_names) {
    boxplot(data[[num_var]], main = paste(num_var, "Boxplot"), xlab = num_var,
        col = rgb(red = 155/255, green = 0/255, blue = 20/255, alpha = 0.3))
}
```

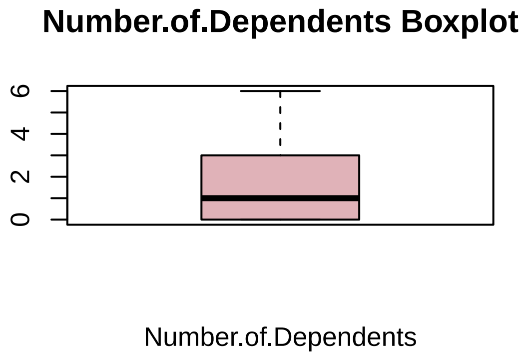

Age Boxplot
20
40
60
Age

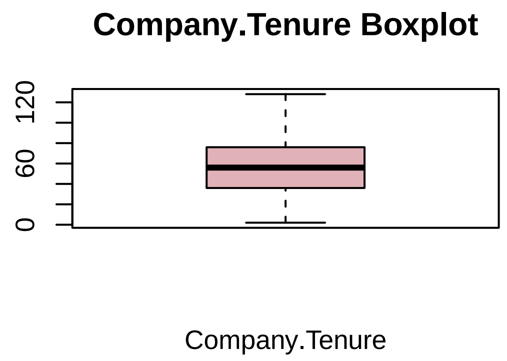

Years.at.Company Boxplot
0
20
50
Years.at.Company


Monthly.Income Boxplot
5000
Monthly.Income


Number.of.Promotions Boxplot
0
2
4
Number.of.Promotions


Distance.from.Home Boxplot
0
40
80
Distance.from.Home


Number.of.Dependents Boxplot
0
2
4
6
Number.of.Dependents

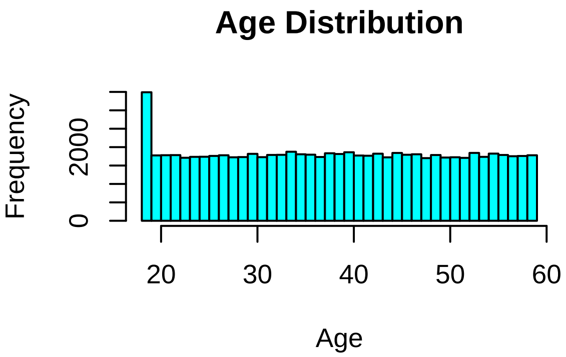

Company.Tenure Boxplot
0
60
120
Company.Tenure

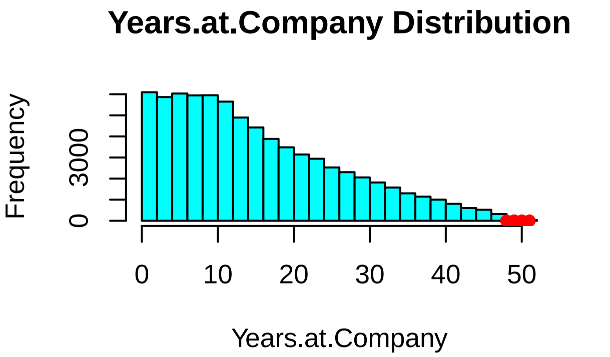
```r
# Function to identify outliers using IQR
identify_outliers <- function(x) {
    Q1 <- quantile(x, 0.25, na.rm = TRUE)
    Q3 <- quantile(x, 0.75, na.rm = TRUE)
    IQR <- Q3 - Q1
    outliers <- x[x < (Q1 - 1.5 * IQR) | x > (Q3 + 1.5 * IQR)]
    return(outliers)
}

# Identify and show outliers for each numeric variable
outliers_list <- list()
for (var in numeric_var_names) {
    outliers <- identify_outliers(data[[var]])
    outliers_list[[var]] <- outliers
    cat("\nOutliers in", var, ":\n")
    print(outliers)
}
```

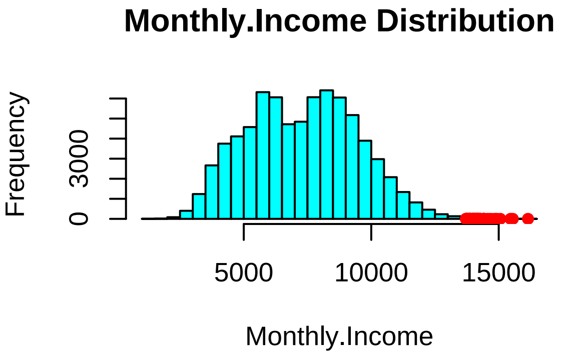
```
Outliers in Age :
integer(0)

Outliers in Years.at.Company :
  [1] 48 49 49 48 48 49 50 48 50 48 48 49 48 51 48 48 51 48 48 48 48 49 49 50 51
 [26] 48 48 50 48 49 48 49 48 48 49 50 48 48 48 49 49 49 48 49 50 51 49 48 51 49
 [51] 49 48 50 50 49 49 49 50 48 48 48 48 48 49 48 51 48 49 50 49 48 48 48 49 51
 [76] 48 50 50 50 50 50 48 49 48 49 49 50 49 51 48 50 49 48 48 50 48 49 48 48 48
[101] 48 50 51 49 49 49 48 51 48 49 49 50 50 48 51 49 49 48 48 48 48 49 50 49 48
[126] 49 50 48 48 49 48 49 48 48 51 49 50 48 48 48 50 48 51 50 48 49 49 49 51 49
[151] 48 49 51 48 50 50 49 48 48 48 49 48 48 51 48 49 48 48 48 49 48 51 49 49 48
[176] 48 50 48 48 49 48 48 49 49 50 50 49 48 49 48 48 48 50 51 50 49 48 50 50 48
[201] 48 50 49 49 48 49 48 48 48 49 49 48 48 49 49 49 51 48 51 48 49 49 48 50 50
[226] 51 49 49 48 48 51 49 48 49 48 51 49 48 49 49 49 50 49 49 50 51 49 50 49 49
[251] 50 48 49 48 50 51 50 50 49 50 49 50 48 49 48 48 48 51 48 48 50 50 49 48 48
[276] 49 50 48 48 49 49 51 48 48 48 51 48 48 48 49 48 49 49 51 48 49 48 50 48 48
[301] 50 48 49 49 48 48 49 48 51 50 48 48 50 49 48 49 51 48 49 49 48 48 49 48 49
[326] 48 49 51 49 48 49 50 48 48 48 51 48 48

Outliers in Monthly.Income :
 [1] 15495 13961 14014 14016 14176 13962 14276 14066 13876 14421 13959 13722
[13] 13747 13768 14622 13739 14163 16149 13833 14271 14235 13800 14226 13988
[25] 14147 14286 14885 13859 14396 14210 13715 14127 13793 14002 14185 14076
[37] 14067 13875 14398 14137 14103 14924 13728 13713 14405 13877 15464 15552
[49] 14839 14406 14110 13840 14412 13896 14021 14181 14292 13893 13830 13764
[61] 14707 14433 14028 14547 15063

Outliers in Number.of.Promotions :
integer(0)
```

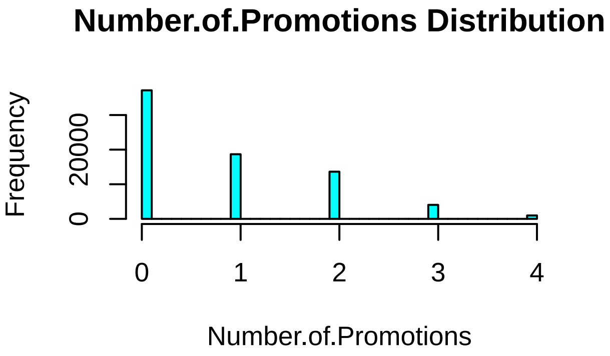
```
Outliers in Distance.from.Home :
integer(0)

Outliers in Number.of.Dependents :
integer(0)

Outliers in Company.Tenure :
integer(0)
```

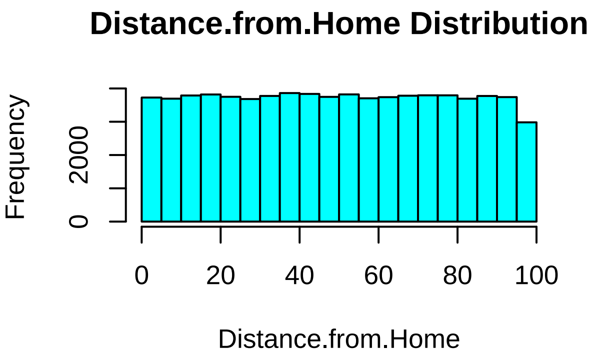
```r
# Plot histograms and highlight outliers

plot_index <- 1
for (page in 1:num_pages) {
    par(mfrow = c(2, 2))
    for (i in 1:plots_per_page) {
        if (plot_index > num_plots)
            break
        num_var <- numeric_var_names[plot_index]

        hist(data[[num_var]], main = paste(num_var, "Distribution"), xlab = num_var,
            col = "cyan", breaks = 30)
        outliers <- outliers_list[[num_var]]
        if (length(outliers) > 0) {
            points(outliers, rep(0, length(outliers)), col = "red", pch = 16)
        }

        plot_index <- plot_index + 1
    }
}
```

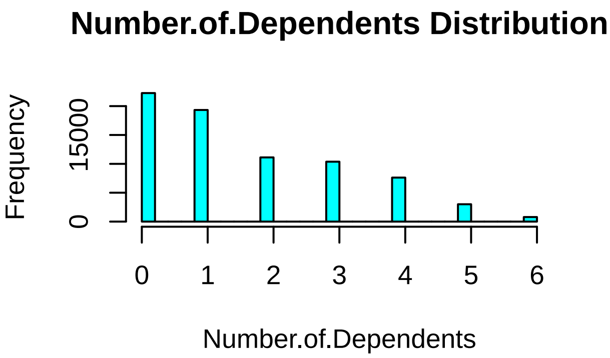

Age Distribution
Frequency
0
2000
20
30
40
50
60
Age

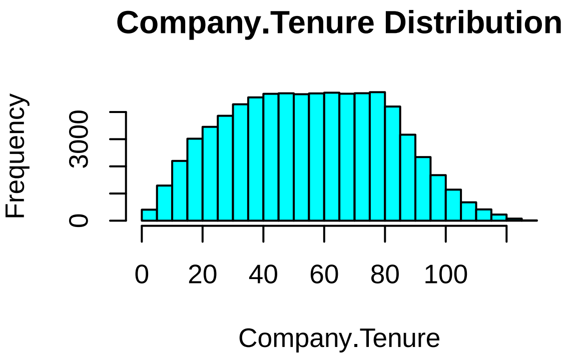

Years.at.Company Distribution
Frequency
0
3000
0
10
20
30
40
50
Years.at.Company


Monthly.Income Distribution
Frequency
0
3000
5000
10000
15000
Monthly.Income


Number.of.Promotions Distribution
Frequency
0
20000
0
1
2
3
4
Number.of.Promotions


Distance.from.Home Distribution
Frequency
0
2000
0
20
40
60
80
100
Distance.from.Home


Number.of.Dependents Distribution
Frequency
0
15000
0
1
2
3
4
5
6
Number.of.Dependents


Company.Tenure Distribution
Frequency
0
3000
0
20
40
60
80
100
Company.Tenure

As a result of the analysis, the following observations were made regarding the characteristics of the data:

-
-
-
-
-
-
-

Features vs. Target

Categorical Features vs. Target

Numerical Features vs. Target

Correlation Matrix

```r
# Cor. and Cov.
cov_matrix <- cov(data[, numeric_var_names])
cor_matrix <- cor(data[, numeric_var_names])

print("Covariance Matrix:")
```

```
[1] "Covariance Matrix:"
```

```r
print(cov_matrix)
```

```
                            Age Years.at.Company Monthly.Income
Age                 146.009914270      72.87199387      -45.51579
Years.at.Company     72.871993868     125.97242911     -144.24846
Monthly.Income      -45.515785898    -144.24846489   4633293.12554
Number.of.Promotions  0.008083759       0.01048464       12.14436
Distance.from.Home   -1.579927372      -1.54734492     -117.21026
Number.of.Dependents  0.069262885       0.07649657        5.04010
Company.Tenure       72.534611568     126.16897906     -377.81514
                     Number.of.Promotions Distance.from.Home
Age                           0.008083759        -1.57992737
Years.at.Company              0.010484640        -1.54734492
Monthly.Income               12.144360519      -117.21026410
Number.of.Promotions          0.990599761        -0.19392391
Distance.from.Home           -0.193923912       813.02599733
Number.of.Dependents         -0.002255666        -0.04226003
Company.Tenure                0.130192015        -4.15337376
                     Number.of.Dependents Company.Tenure
Age                           0.069262885    72.53461157
```

```
Years.at.Company            0.076496571   126.16897906
Monthly.Income              5.040099975  -377.81513829
Number.of.Promotions       -0.002255666     0.13019201
Distance.from.Home         -0.042260030    -4.15337376
Number.of.Dependents        2.413775012     0.01663453
Company.Tenure              0.016634533   645.12692138
```

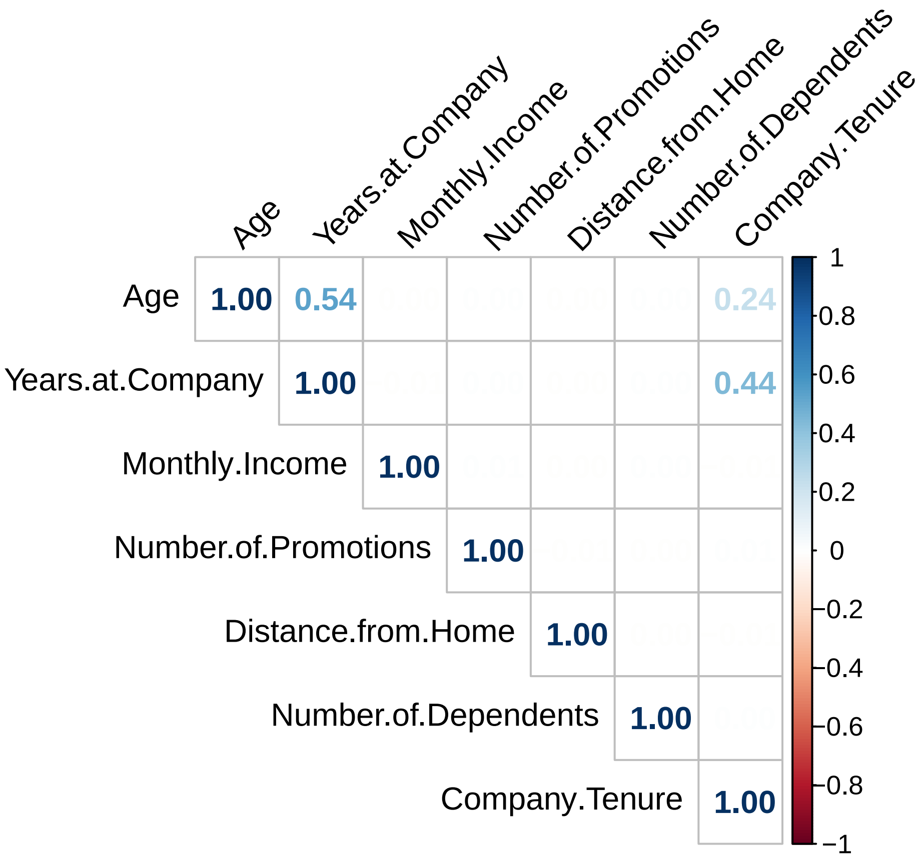
```r
print("Correlation Matrix:")
```

```
[1] "Correlation Matrix:"
```

```r
print(cor_matrix)
```

```
                               Age Years.at.Company Monthly.Income
Age                   1.0000000000      0.537318418   -0.001749951
Years.at.Company      0.5373184182      1.000000000   -0.005970745
Monthly.Income       -0.0017499514     -0.005970745    1.000000000
Number.of.Promotions  0.0006721606      0.000938570    0.005668663
Distance.from.Home   -0.0045855743     -0.004835008   -0.001909715
Number.of.Dependents  0.0036894448      0.004386881    0.001507113
Company.Tenure        0.2363368996      0.442580479   -0.006910538
                     Number.of.Promotions Distance.from.Home
Age                          0.0006721606      -0.0045855743
Years.at.Company             0.0009385700      -0.0048350077
Monthly.Income               0.0056686632      -0.0019097149
Number.of.Promotions         1.0000000000      -0.0068332929
Distance.from.Home          -0.0068332929       1.0000000000
Number.of.Dependents        -0.0014587377      -0.0009539579
Company.Tenure               0.0051500643      -0.0057349048
                     Number.of.Dependents Company.Tenure
Age                          0.0036894448   0.2363368996
Years.at.Company             0.0043868807   0.4425804786
Monthly.Income               0.0015071132  -0.0069105384
Number.of.Promotions        -0.0014587377   0.0051500643
Distance.from.Home          -0.0009539579  -0.0057349048
Number.of.Dependents         1.0000000000   0.0004215408
Company.Tenure               0.0004215408   1.0000000000
```

```r
corrplot(cor_matrix, method = "number", type = "upper", tl.col = "black",
    tl.srt = 45)
```



Partial Correlation Matrices

Data Preparation

After completing the data analysis steps, it is necessary to prepare the data for model development.

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.

```r
# Ordinal mappings:
balance.map <- c(Poor = 1, Fair = 2, Good = 3, Excellent = 4)
data$Work.Life.Balance <- balance.map[as.numeric(data$Work.Life.Balance)]

satisfaction.map <- c(Low = 1, Medium = 2, High = 3, `Very High` = 4)
data$Job.Satisfaction <- satisfaction.map[as.numeric(data$Job.Satisfaction)]

performance.map <- c(Low = 1, `Below Average` = 2, Average = 3, High = 4)
data$Performance.Rating <- performance.map[as.numeric(data$Performance.Rating)]
```

```r
education.map <- c(`High School` = 1, `Associate Degree` = 2, `Bachelor’s Degree` = 3,
    `Master’s Degree` = 4, PhD = 5)
data$Education.Level <- education.map[as.numeric(data$Education.Level)]

level.map <- c(Entry = 1, Mid = 2, Senior = 3)
data$Job.Level <- level.map[as.numeric(data$Job.Level)]

reputation.map <- c(Poor = 1, Fair = 2, Good = 3, Excellent = 4)
data$Company.Reputation <- reputation.map[as.numeric(data$Company.Reputation)]

recognition.map <- c(Low = 1, Medium = 2, High = 3, `Very High` = 4)
data$Employee.Recognition <- recognition.map[as.numeric(data$Employee.Recognition)]

size.map <- c(Small = 1, Medium = 2, Large = 3)
data$Company.Size <- size.map[as.numeric(data$Company.Size)]

# Nominal mappings: Create dummy variables for nominal data
data_numeric <- model.matrix(~., data = data)

# Convert the resulting matrix back to a data frame
data_numeric <- as.data.frame(data_numeric)[, -1]  # -1 to remove the intercept column
```

Train-Test-Split

Before splitting the data into training and test, first features and target should be defined.

```r
# Splitting data into features and target:
X <- data_numeric[, !(colnames(data_numeric) %in% c("Employee.ID", "AttritionStayed"))]

y <- data_numeric$AttritionStayed
```

Now, we can split the dataset for modelling.

```r
set.seed(42)

trainIndex <- sample(1:nrow(X), 0.8 * nrow(X))

# 80% of data is used for training
X.train <- X[trainIndex, ]
y.train <- y[trainIndex]

# 20% of data is used for testing
X.test <- X[-trainIndex, ]
y.test <- y[-trainIndex]
```

Before moving to modelling step, it is beneficial to check the dimensions and balance of the datasets.

```r
# Number of samples in train data
dim(X.train)
```

```
[1] 59598    26
```

```r
train.size <- dim(X.train)[1]

# Number of samples in test data
dim(X.test)
```

```
[1] 14900    26
```

```r
test.size <- dim(X.test)[1]

# Proportion of stayed employees for train data
prop.table(table(y.train))
```

```
y.train
        0         1 
0.4752341 0.5247659 
```

```r
# Proportion of stayed employees for test data
prop.table(table(y.test))
```

```
y.test
       0        1 
0.472953 0.527047 
```

We can observe that the train and test datasets are balanced within themselves. Also the train data is representative of test data.

Predictive Classification Models

Predictive classification models are a type of machine learning algorithm used to predict the category or class label of new, unseen instances based on historical data. These models are trained using a labelled dataset where the input features (independent variables) are associated with known class labels (dependent variable). The goal of the model is to learn the relationship between the features and the class labels so that it can accurately classify new data points into one of the predefined categories.

In this project we aim to find the risk of an employee leaving the company (class 0) and the factors affecting employee retention. So we will develop several classification models and examine their performances.

Logistic Regression

The logistic regression model estimates the odds of the dependent variable occurring and applies the logit (log-odds) transformation to express this relationship.

$$g(\pi_i) = \mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) \in (-\infty, +\infty)$$

Basic Logistic Classifier

```r
# First of all we check the model statistics with all the features
glm.FULL <- glm(y.train ~ ., data = X.train, family = binomial)

summary(glm.FULL)
```

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

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 -1.505e+00  9.372e-02 -16.057  < 2e-16 ***
Age                          5.657e-03  9.538e-04   5.931 3.00e-09 ***
GenderMale                   5.510e-01  1.967e-02  28.008  < 2e-16 ***
Years.at.Company             1.263e-02  1.117e-03  11.305  < 2e-16 ***
Job.RoleFinance              5.782e-02  4.596e-02   1.258  0.20839    
Job.RoleHealthcare           3.478e-02  4.017e-02   0.866  0.38663    
Job.RoleMedia                1.053e-01  3.433e-02   3.068  0.00216 ** 
Job.RoleTechnology           5.270e-02  4.601e-02   1.145  0.25202    
Monthly.Income               1.164e-05  7.834e-06   1.486  0.13740    
Work.Life.Balance           -1.814e-01  1.038e-02 -17.478  < 2e-16 ***
Job.Satisfaction            -1.235e-01  7.955e-03 -15.522  < 2e-16 ***
Performance.Rating          -9.274e-02  1.020e-02  -9.091  < 2e-16 ***
Number.of.Promotions         2.248e-01  9.933e-03  22.631  < 2e-16 ***
OvertimeYes                 -3.351e-01  2.079e-02 -16.123  < 2e-16 ***
Distance.from.Home          -8.487e-03  3.442e-04 -24.660  < 2e-16 ***
Education.Level              1.280e-01  8.077e-03  15.844  < 2e-16 ***
Marital.StatusMarried        2.625e-01  2.808e-02   9.348  < 2e-16 ***
Marital.StatusSingle        -1.400e+00  3.032e-02 -46.162  < 2e-16 ***
Number.of.Dependents         1.320e-01  6.315e-03  20.907  < 2e-16 ***
Job.Level                    1.143e+00  1.422e-02  80.401  < 2e-16 ***
Company.Size                -1.026e-01  1.392e-02  -7.369 1.72e-13 ***
Company.Tenure               8.568e-05  4.270e-04   0.201  0.84095    
Remote.WorkYes               1.612e+00  2.754e-02  58.548  < 2e-16 ***
Leadership.OpportunitiesYes  1.071e-01  4.508e-02   2.376  0.01751 *  
Innovation.OpportunitiesYes  1.204e-01  2.650e-02   4.545 5.49e-06 ***
Company.Reputation          -1.200e-01  1.118e-02 -10.731  < 2e-16 ***
```

```
Employee.Recognition        -3.518e-03  1.141e-02  -0.308  0.75783
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 82474  on 59597  degrees of freedom
Residual deviance: 63074  on 59571  degrees of freedom
AIC: 63128

Number of Fisher Scoring iterations: 4
```

The above model statistics indicate that p-values of Company.Tenure and Employee Recognition are above 0.5 indicating that these features are insignificant to the results. Additionally, some Job.Roles and Monthly.Income also have high p-values indicating that their effect on Attrition is less significant compared to other features. However, for now we would like to keep all the features in the model and apply feature selection later.

In order to understand how well the model fits the data we can make use of 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. The formula for R^2 is:

$$R^2 = 1 - \frac{RSS}{ESS}$$

Where:

- RSS is the sum of squares of the residuals (the differences between observed and predicted values), i.e. the deviance of the fitted model
- ESS is the total sum of squares due to regression (the differences between the observed values and the mean of the observed values)

```r
R2 <- 1 - (summary(glm.FULL)$deviance/summary(glm.FULL)$null.deviance)

R2
```

```
[1] 0.2352228
```

```r
##
```

With the full model the value of R^2 0.2352228 indicates that approximately 23.52% of the variance in the target can be explained by the features in the model. Since 23.52% is relatively low, it suggests that the model is not capturing much of the underlying pattern in the data.

Multicollinearity can be a reason for a low R^2 value, as it can make it difficult to determine the individual effect of each predictor on the target. Calculating the Variance Inflation Factor (VIF) can help to check for multicollinearity among the features.

```r
library(car)
vif(glm.FULL)
```

```
                         Age                   GenderMale
                    1.401734                     1.013865
            Years.at.Company               Job.RoleFinance
                    1.644621                     2.698555
          Job.RoleHealthcare                  Job.RoleMedia
                    3.012243                     1.678633
         Job.RoleTechnology                Monthly.Income
                    4.303772                     2.998647
           Work.Life.Balance              Job.Satisfaction
                    1.006181                     1.004924
          Performance.Rating          Number.of.Promotions
                    1.001971                     1.009719
                 OvertimeYes            Distance.from.Home
                    1.005366                     1.009721
             Education.Level         Marital.StatusMarried
                    1.004629                     2.081804
        Marital.StatusSingle          Number.of.Dependents
                    2.158749                     1.008324
                   Job.Level                  Company.Size
                    1.093709                     1.001481
              Company.Tenure                Remote.WorkYes
                    1.238531                     1.058169
Leadership.OpportunitiesYes Innovation.OpportunitiesYes
                    1.000765                     1.000480
          Company.Reputation          Employee.Recognition
                    1.002558                     1.000406
```

Logistic Regression with Backward Variable Selection

Logistic Regression with Shrinkage Method

ROC Curve & Comparison of Logistic Classifiers

Another Classification Model

Model Results

Performance Metrics and Confusion Matrix