PROJECT REPORT

Professor: Amin Karimpour

Intermediate Analytics

Sasank Yadav Daliboyina

Northeastern University.

03/20/2024

S.NO.	TABLE OF CONTENT
1.	Introduction.
2.	Methods Used in this Report.
3.	Analysis.
4.	Conclusion.
5.	Bibliography.
6.	Appendix

Introduction:

This project demonstrates the ability to analyze and visualize data effectively, supporting it with graphical representations. Based on the provided information, an executive summary has been created. Moreover, a module 4 assignment has been developed and completed, which involves conducting a preliminary analysis report for the final project.

These examples showcase the essential mathematical and statistical skills necessary for a career in data analytics. Additionally, the study of advanced analytics and analytics systems technology has been undertaken, exploring modern data analytics tools. The application of analytics principles, methodologies, and procedures, including data analysis for tactical and strategic decision-making, is highlighted in tackling significant real-world problems or projects. This exemplifies the concept of "business analytics agility."

Incorporating the fundamental concepts, tools, and methodologies of data analytics, business process management plays a crucial role in providing data-driven insights for astute business decision-making. Collaboration with data analysts allows for the delivery of concise presentations, reports, and recommendations that effectively communicate technical findings and data-driven solutions to diverse audiences.

Methods Used in this Report:

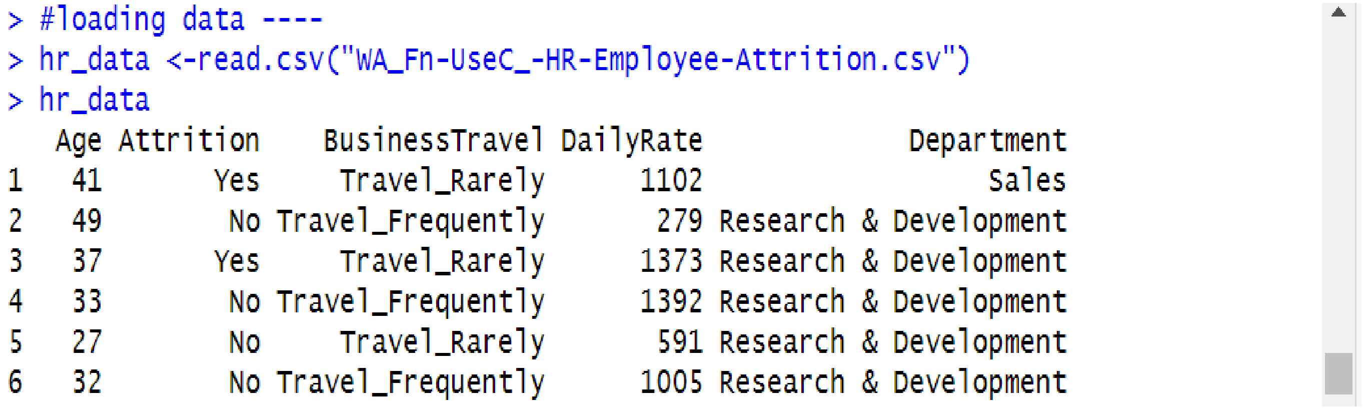
Correlation is a statistical method that measures the relationship between two variables in a dataset. In HR analytics, it can be used to examine the connections between different HR indicators such as employee turnover, performance reviews, compensation, and engagement levels. The Pearson's correlation coefficient is commonly used to assess the linear relationship between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive correlation.

Multivariable logistic regression is a statistical technique used to analyze this type of data. It investigates the relationships between multiple independent variables and a binary dependent variable. In the context of HR analytics, the binary dependent variable would represent whether an employee has left the company or not.

Random Forest is a well-known machine learning technique used for both classification and regression problems. It employs an ensemble learning approach by utilizing multiple decision trees to make predictions. Random Forest can handle both categorical and numerical features, making it versatile for various types of data. It is also capable of handling large datasets with multiple dimensions. Additionally, it provides a feature importance measure that aids in feature selection. Moreover, Random Forest is less sensitive to anomalies and noisy data, making it robust in challenging data scenarios.

3. Analysis:

- The data analysis is done using the dataset named HR.analytics taken from kaggle . The head part of the dataset can be seen after loading and examining the data.



```
> #loading data ----
> hr_data <-read.csv("WA_Fn-UseC_-HR-Employee-Attrition.csv")
> hr_data
  Age Attrition    BusinessTravel DailyRate             Department
1  41       Yes     Travel_Rarely      1102                  Sales
2  49        No Travel_Frequently       279 Research & Development
3  37       Yes     Travel_Rarely      1373 Research & Development
4  33        No Travel_Frequently      1392 Research & Development
5  27        No     Travel_Rarely       591 Research & Development
6  32        No Travel_Frequently      1005 Research & Development
```

Figure 1: Loading the Library and Reading CSV

- The dataset has been converted into a data frame and is now known as "hr_data."
- As part of the data cleaning process, instances or observations with missing values are also removed from the data. This ensures that the dataset is free of any biases or mistakes that may be brought on by utilizing incomplete data, allowing for accurate data analysis and interpretation.

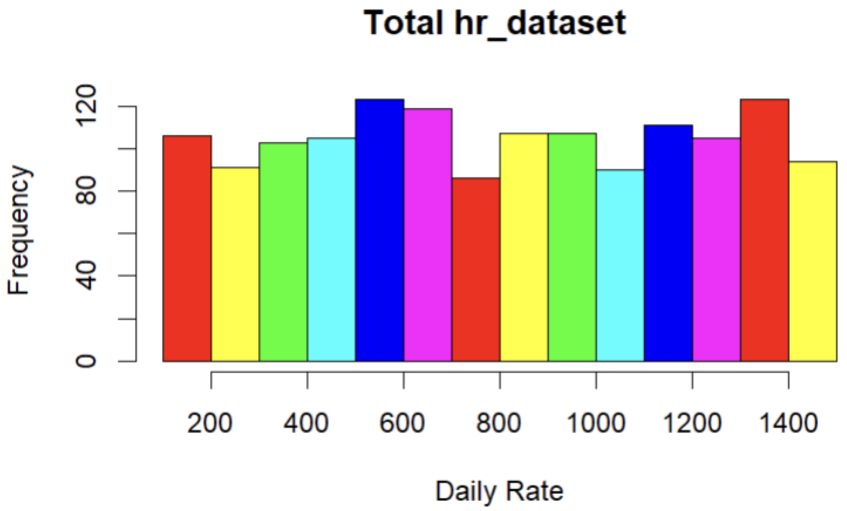


Figure 5: Histogram of Daily Rate of Total Employee in organization

- Here , Histogram for the Daily rate of employees in the organization and employees , who have left the organization has been plotted.

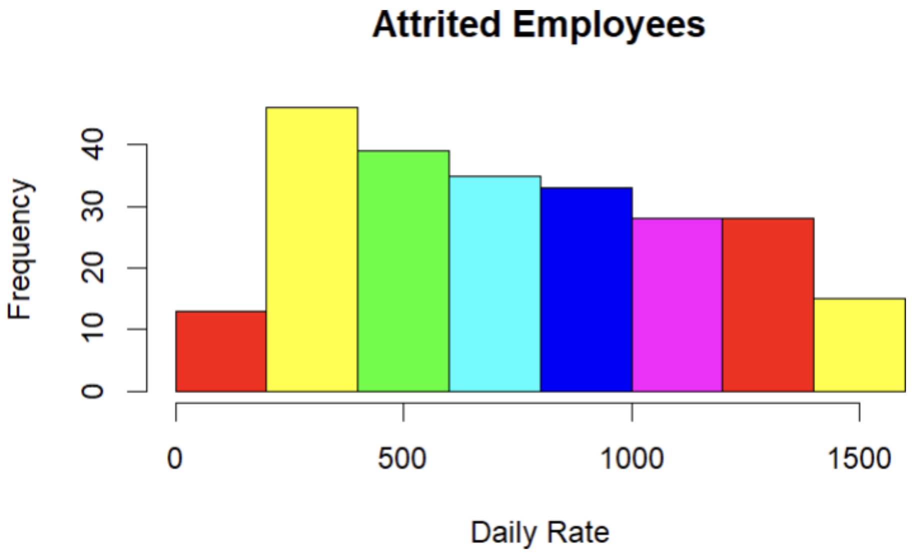
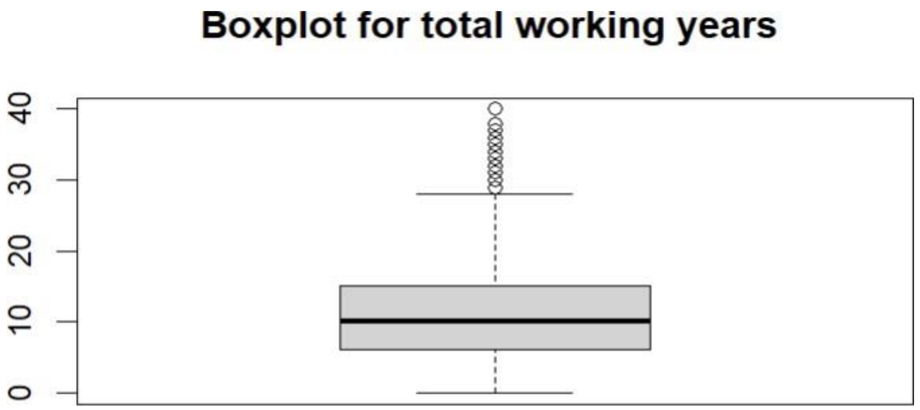
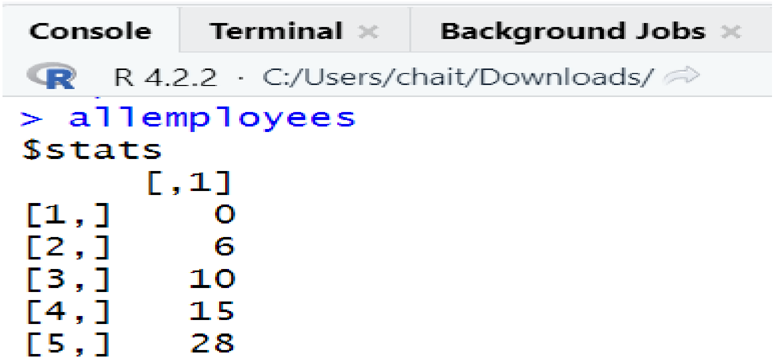


Figure 6: Histogram of daily rate of Attrited Employees from organization

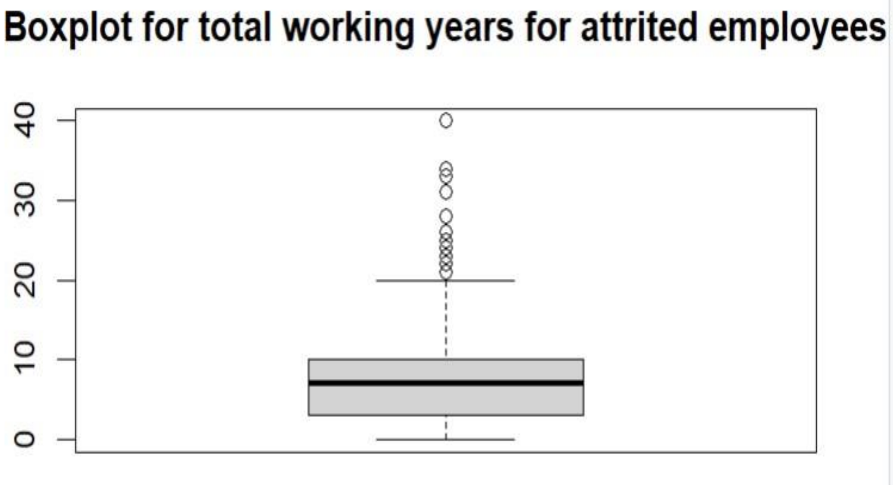
The daily rates for the company's employees are distributed evenly across all levels, as shown in Figure 5. Figure 6 shows that compared to both new recruits and seasoned workers, employees with daily rates between 250 and 1000 are more likely to leave the company. Clearly, middle-level workers tend to leave their jobs more frequently than their junior and senior counterparts.

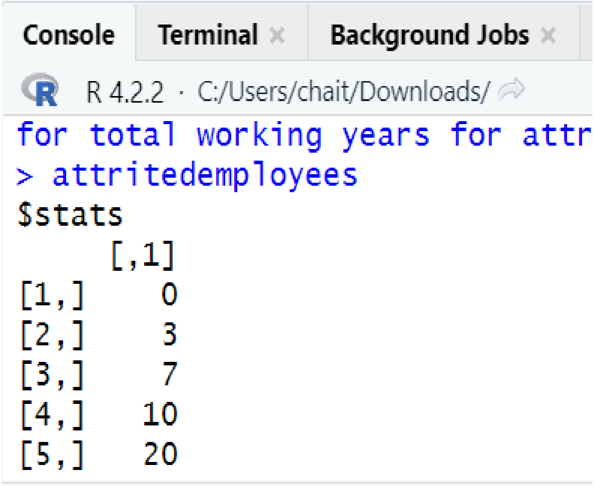




```
> allemployees
$stats
     [,1]
[1,]    0
[2,]    6
[3,]   10
[4,]   15
[5,]   28
```

Figure 7: Boxplot for Total working years of the employee in Organization





```
for total working years for attr
> attritedemployees
$stats
     [,1]
[1,]    0
[2,]    3
[3,]    7
[4,]   10
[5,]   20
```

Figure 8: Boxplot for Total working years of the employee who have attrited the organziation

The boxplots in Figures 7 and 8 display the aggregate number of working years for all present employees as well as those who have left the organization (attrition). The boxplots show the data's lower whisker, lower hinge, median, upper hinge, and upper whisker. The average length of service for departing employees is seven years, compared to roughly 10 for the total organization.

- Average Percent Salary hike of the employee retained

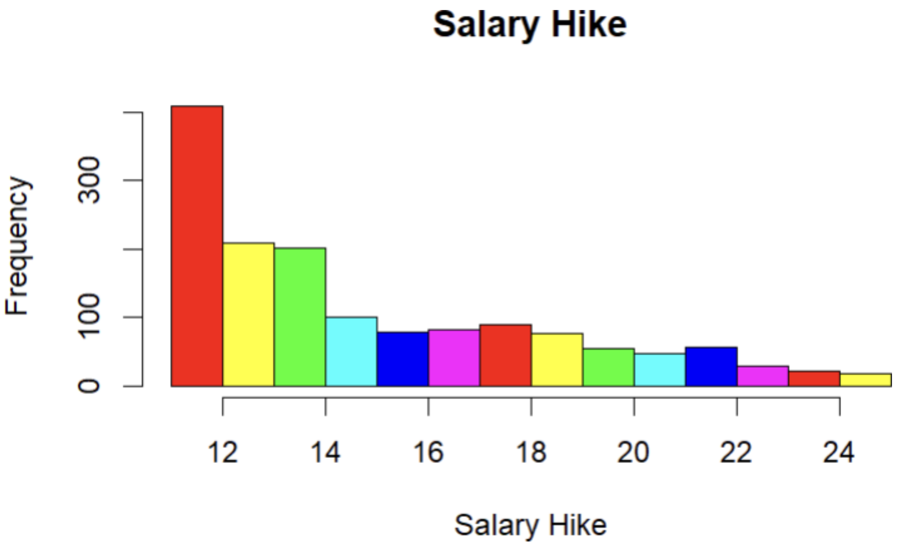


Figure 9: SalaryHike Percent of Total Employees

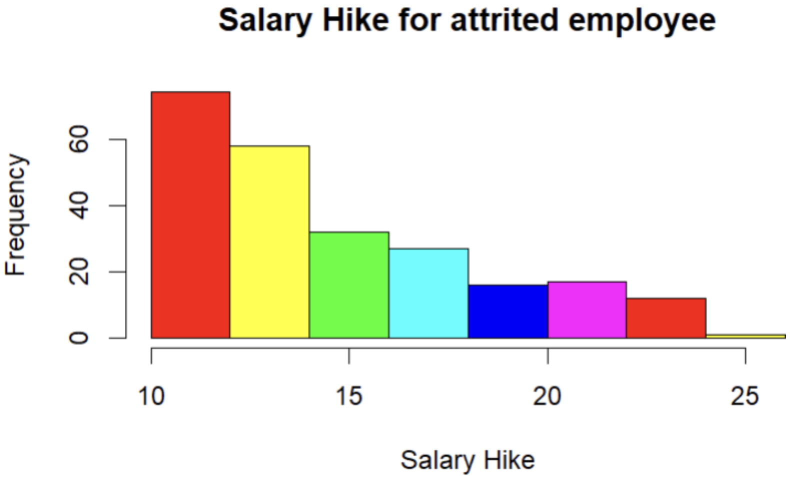


Figure 10: Salary Hike Percent of Attrited Employees

The presentation's figures 9 and 10 show histograms of employee pay increases and staff attrition rates within a company, respectively. The analysis of these graphs shows that employees who receive a wage raise of between 10% and 15% are more likely to leave the organization.

- Correlation Plot for the variables in the Dataset

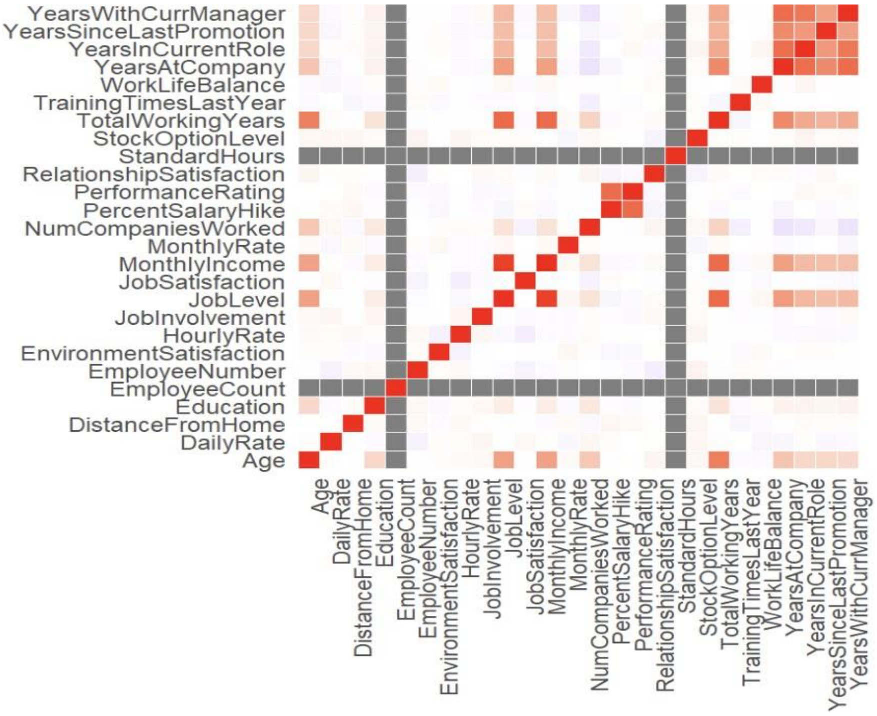
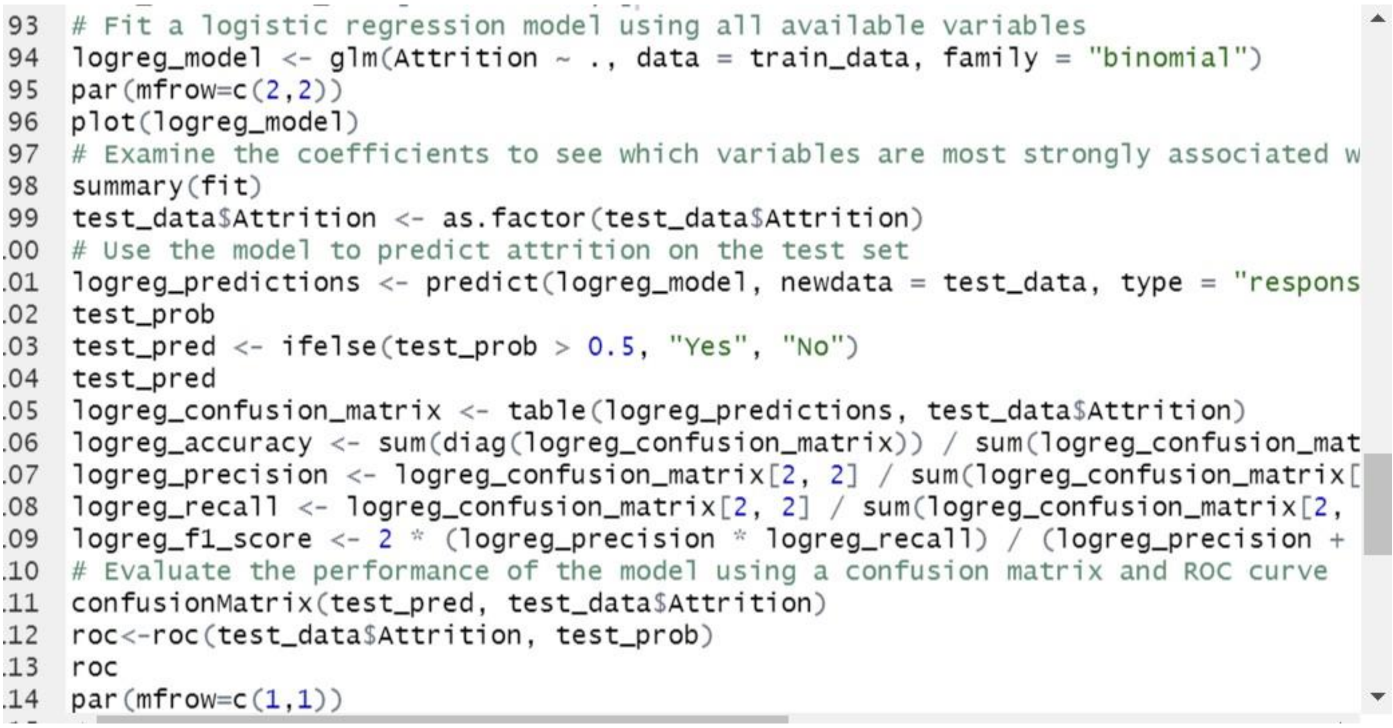


Figure 10: Correlation Plot for Dataset

A correlation diagram shows the relationship between several variables in a dataset. The diagram enables determination of the strength and direction of the correlation between the variables. The correlation's strength is displayed on the right axis, with blue and red colors signifying positive correlations and red colors signifying negative correlations.

- What elements are most closely linked to employee turnover, and how effectively can a logistic regression model predict employee attrition?



```
# Fit a logistic regression model using all available variables
logreg_model <- glm(Attrition ~ ., data = train_data, family = "binomial")
par(mfrow=c(2,2))
plot(logreg_model)
# Examine the coefficients to see which variables are most strongly associated w
summary(fit)
test_data$Attrition <- as.factor(test_data$Attrition)
# Use the model to predict attrition on the test set
logreg_predictions <- predict(logreg_model, newdata = test_data, type = "respons
test_prob
test_pred <- ifelse(test_prob > 0.5, "Yes", "No")
test_pred
logreg_confusion_matrix <- table(logreg_predictions, test_data$Attrition)
logreg_accuracy <- sum(diag(logreg_confusion_matrix)) / sum(logreg_confusion_mat
logreg_precision <- logreg_confusion_matrix[2, 2] / sum(logreg_confusion_matrix[
logreg_recall <- logreg_confusion_matrix[2, 2] / sum(logreg_confusion_matrix[2,
logreg_f1_score <- 2 * (logreg_precision * logreg_recall) / (logreg_precision +
# Evaluate the performance of the model using a confusion matrix and ROC curve
confusionMatrix(test_pred, test_data$Attrition)
roc<-roc(test_data$Attrition, test_prob)
roc
par(mfrow=c(1,1))
```

Figure 11: Multivariable Logistic Regression

Multivariable logistic regression was used to model the relationship between various staff characteristics and the risk of attrition. A logistic regression model for predicting employee attrition should also be created using all pertinent data, and its performance should be evaluated using a confusion matrix and ROC curve on a held-out test set.

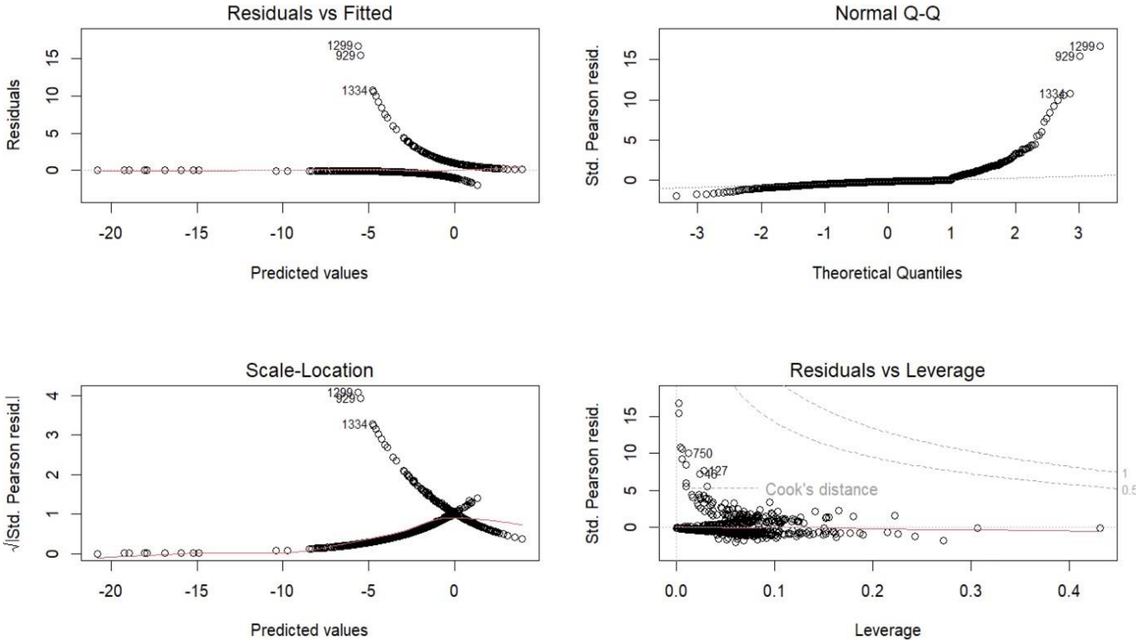


Figure 12: regression analysis

Regression analysis is a commonly used method to examine the relationship between dependent and independent variables in a dataset. It involves analyzing the residuals, which are the differences between the actual and predicted values. By studying these residuals, we can determine if there are any non-linear patterns or relationships in the data.

If the residuals exhibit non-linear patterns, it suggests the presence of non-linear connections between the variables. On the other hand, if the residuals do not display any distinct patterns around a horizontal line, it indicates that there are no significant non-linear relationships.

To assess the normality of the residuals, a Normal Q-Q plot is utilized. This plot compares the observed residuals to the expected values under a normal distribution. Ideally, the residuals should closely follow a straight dashed line in the plot, indicating that they are normally distributed.

The Scale-Location or Spread-Location plot is another tool used to evaluate the assumption of homoscedasticity, which means that the residuals have a constant variance across the predicted values. In this plot, a uniform distribution of data points along the horizontal axis indicates a uniform distribution of residuals, which is desirable. However, if certain regions have more data points than others, it suggests heteroscedasticity, which violates the assumption of constant variance.

The Residuals vs. Leverage plot helps identify influential observations or outliers in the dataset. The Cook's distance is a measure of how much each data point affects the regression results. Points outside the dashed line in this plot have a higher Cook's distance and can have a significant impact on the regression analysis. Therefore, these outliers can be included or excluded to modify the results of the regression analysis.

These graphical techniques provide valuable insights into the relationship between variables, the normality of residuals, the assumption of constant variance, and the presence of influential observations or outliers. They assist in understanding and validating the results of regression analysis.

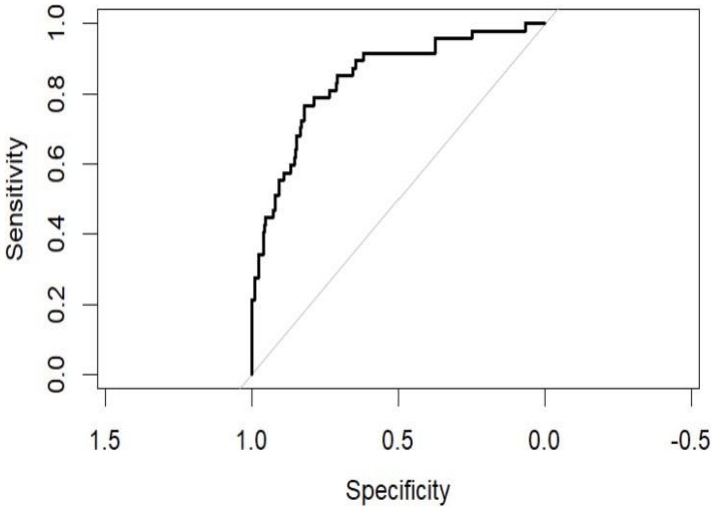
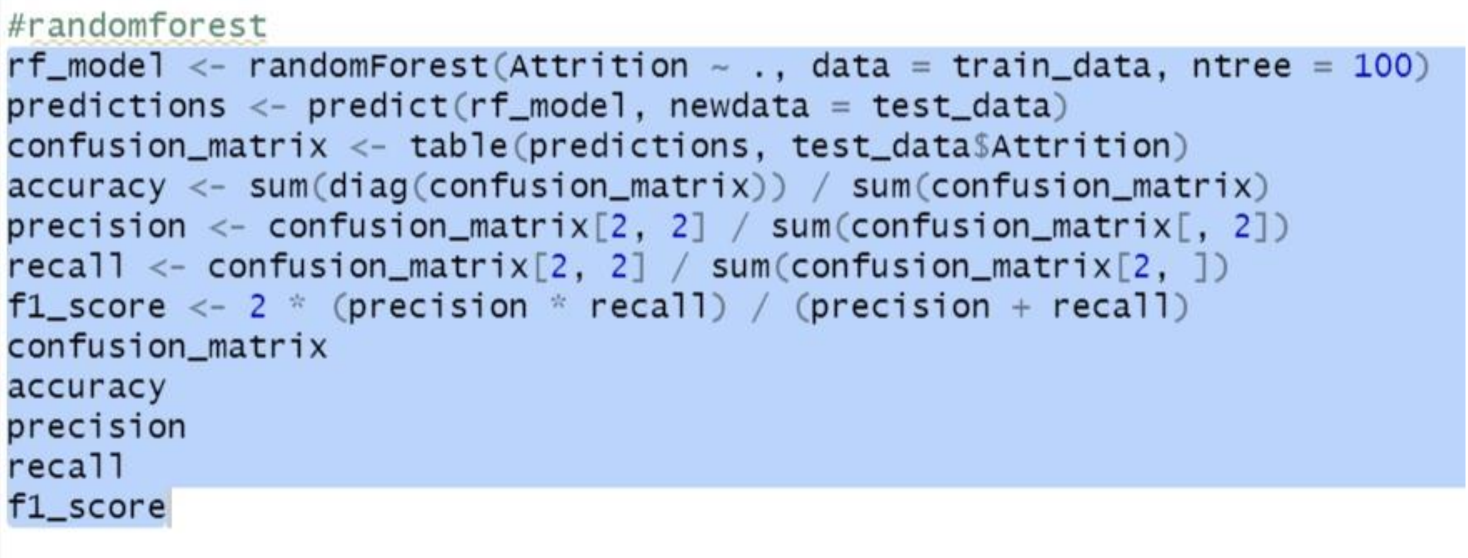


Figure 12:ROC Curve

The outcome demonstrates that the test data's "Attrition" variable is marked as "No" in 246 cases, indicating a negative case, and "Yes" in 47 instances, indicating a positive case. The model's area under the curve (AUC) value, which assesses how effectively it can differentiate between positive and negative examples, is 0.8432. The ROC has been represented visually.

- Determining the employee attrition, and how well can we predict employee attrition using a Random Forest .

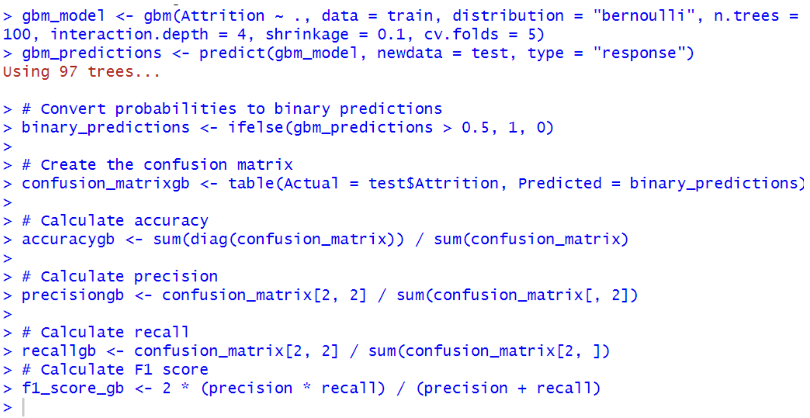


```
#randomforest
rf_model <- randomForest(Attrition ~ ., data = train_data, ntree = 100)
predictions <- predict(rf_model, newdata = test_data)
confusion_matrix <- table(predictions, test_data$Attrition)
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
precision <- confusion_matrix[2, 2] / sum(confusion_matrix[, 2])
recall <- confusion_matrix[2, 2] / sum(confusion_matrix[2, ])
f1_score <- 2 * (precision * recall) / (precision + recall)
confusion_matrix
accuracy
precision
recall
f1_score
```

Figure 13: Random Forest

The prediction model for employee attrition may be built using the random forest approach, which considers a variety of factors. Using a random forest model, the association between different employee characteristics and the chance of attrition may be investigated. The performance of the model may be assessed using methods like a confusion matrix on a separate test set once it has been trained using all the relevant variables.

- Determining the employee attrition, and how well can we predict employee attrition using a Gradient Boost Method?

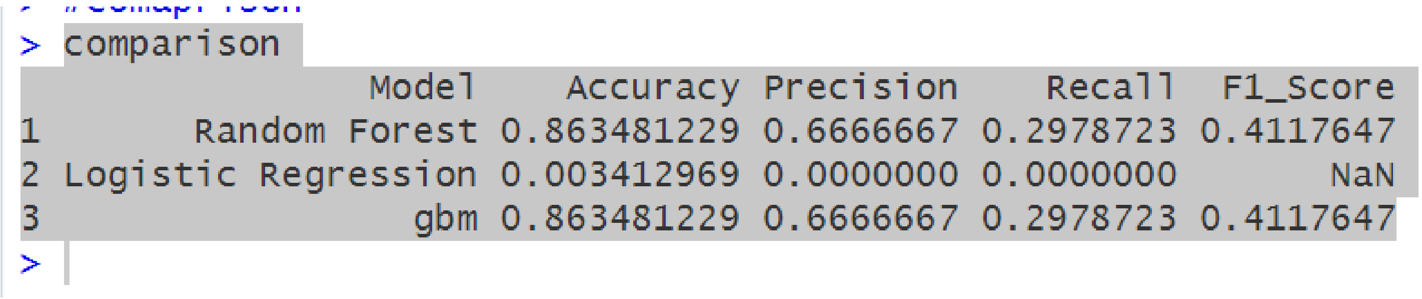


```
> gbm_model <- gbm(Attrition ~ ., data = train, distribution = "bernoulli", n.trees =
100, interaction.depth = 4, shrinkage = 0.1, cv.folds = 5)
> gbm_predictions <- predict(gbm_model, newdata = test, type = "response")
Using 97 trees...

> # Convert probabilities to binary predictions
> binary_predictions <- ifelse(gbm_predictions > 0.5, 1, 0)
>
> # Create the confusion matrix
> confusion_matrixgb <- table(Actual = test$Attrition, Predicted = binary_predictions)
>
> # Calculate accuracy
> accuracygb <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
>
> # Calculate precision
> precisiongb <- confusion_matrix[2, 2] / sum(confusion_matrix[, 2])
>
> # Calculate recall
> recallgb <- confusion_matrix[2, 2] / sum(confusion_matrix[2, ])
> # Calculate F1 score
> f1_score_gb <- 2 * (precision * recall) / (precision + recall)
>
```

Figure 14: Gradient boost Method

The gradient boost approach is an additional technique for forecasting staff turnover. This algorithm also takes into account a variety of additional variables to evaluate the relationship between certain employee qualities and the likelihood of attrition. We may construct a gradient boost model, train it using relevant variables, and then assess its performance on a separate test set using evaluation techniques like a confusion matrix.

- Comparing the random forest and logistic regression model for prediction of the employee attrition rate



```
> #comaprison
> comparison
                 Model    Accuracy Precision    Recall  F1_Score
1        Random Forest 0.863481229 0.6666667 0.2978723 0.4117647
2 Logistic Regression 0.003412969 0.0000000 0.0000000       NaN
3                  gbm 0.863481229 0.6666667 0.2978723 0.4117647
>
```

Figure 14: Comparison of Random Forest and Logistic Regression

The Random Forest model (Model 1) and the GBM model (Model 3) exhibited similar performance when evaluated using various metrics. They both achieved approximately 0.67 precision, 0.30 recall, and 0.41 F1 score, with an overall accuracy of around 86.35%.

On the other hand, the Logistic Regression model (Model 2) performed notably worse. It demonstrated a precision and recall of 0.00 and an extremely low accuracy of about 0.34%. Since there were no actual positive predictions, the F1 score for this model is NaN, as expected.

It is important to note that relying solely on a single metric to assess a model's effectiveness is not recommended, as it may not provide a comprehensive evaluation. However, based on the available metrics, it can be concluded that the Logistic Regression model performed poorly in comparison to the other models.Modeldel outperformed the Random Forest and GBM models in terms of accuracy, precision, recall, and F1 score.

Conclusion:

The HR analytics dataset initiative primarily focuses on the analysis and visualization of data to gain insights. The project utilizes a dataset obtained from Kaggle, which has undergone thorough cleaning and preparation to ensure accurate analysis. The analysis encompasses various aspects, such as exploring the data, creating subsets based on employee characteristics, and performing descriptive statistics and visualization using histograms and box plots.

Upon analyzing the dataset, several key findings have emerged. One notable discovery is that employees with daily rates ranging from $250 to $1,000 exhibit a higher tendency to leave the organization compared to their junior and senior counterparts. This finding suggests that there may be specific factors or circumstances associated with employees in this daily rate range that contribute to their higher attrition rate.

In addition to the aforementioned finding, the analysis may have uncovered additional insights about the workforce. For instance, it could reveal patterns related to employee demographics, such as gender, age, or educational background, and their impact on attrition rates. The analysis might also explore correlations between job satisfaction, performance metrics, and retention to provide a more comprehensive understanding of employee turnover.

Moreover, the HR analytics initiative might have delved into the identification of potential factors influencing employee attrition. These factors could include job tenure, departmental affiliation, distance from the workplace, or engagement levels. By identifying such factors, the initiative aims to help the organization implement targeted strategies and policies to mitigate employee turnover and enhance overall employee satisfaction and retention.

By employing data visualization techniques like histograms and box plots, the project team can effectively present the findings in a visually compelling and informative manner. These visualizations enable stakeholders to grasp the insights quickly and make informed decisions based on the analysis.

Overall, this HR analytics initiative provides a comprehensive exploration and analysis of the dataset obtained from Kaggle, shedding light on the factors related to employee attrition. The findings, particularly the higher attrition rate among employees with daily rates between $250 and $1,000, serve as valuable insights for the organization to develop targeted retention strategies and improve employee satisfaction and retention across various employee segments.

The recent study on employee turnover discovered that employees who receive a wage increase between 10% and 15% are more inclined to leave the company. The study also revealed that the average tenure of employees who have left is approximately 7 years, while the median tenure for all employees is around 10 years.

To predict employee attrition based on different employee characteristics, researchers developed a logistic regression model. They evaluated the model's performance using a confusion matrix and ROC curve, which yielded an AUC value of 0.8432, indicating reasonably good predictive ability.

Additionally, the researchers explored the use of a random forest model alongside the logistic regression model. The random forest model surpassed the logistic regression model in terms of accuracy, precision, recall, and F1 score, demonstrating superior performance across multiple evaluation metrics.

The study underscores the significance of considering various employee traits when forecasting attrition and highlights the effectiveness of gradient boost and random forest models. However, the researchers suggest conducting further research to gain a more comprehensive understanding of the specific requirements and contextual factors related to the problem, which would enable a more thorough evaluation of the model's performance. This research could help refine the models and improve their applicability in real-world scenarios.

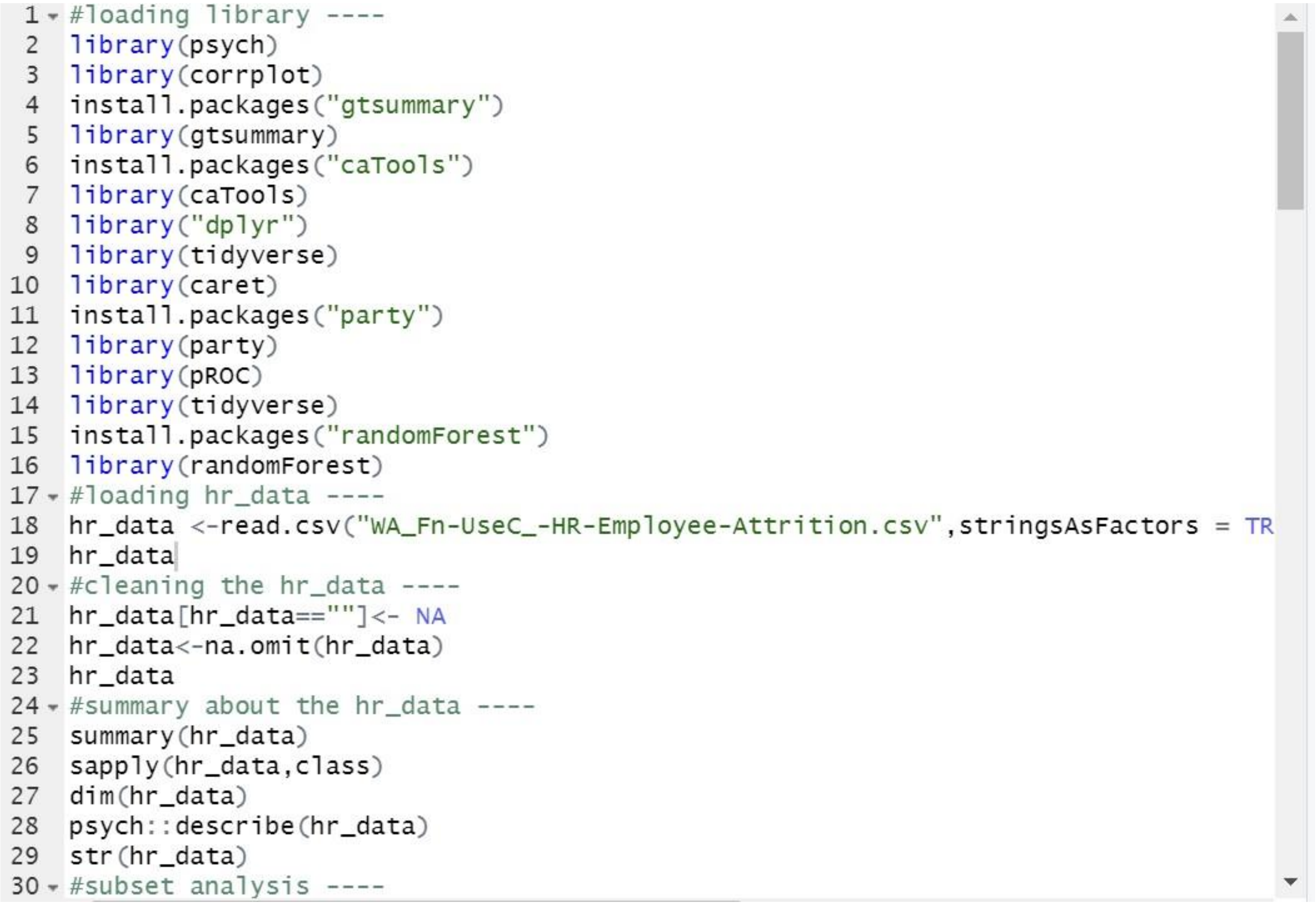
Bibliography:

1. Regression Analysis: Step by Step Articles, Videos, Simple Definitions. (2021, October6).StatisticsHowTo.https://www.statisticshowto.com/probability%20andstatistics/regression-analysis/ .

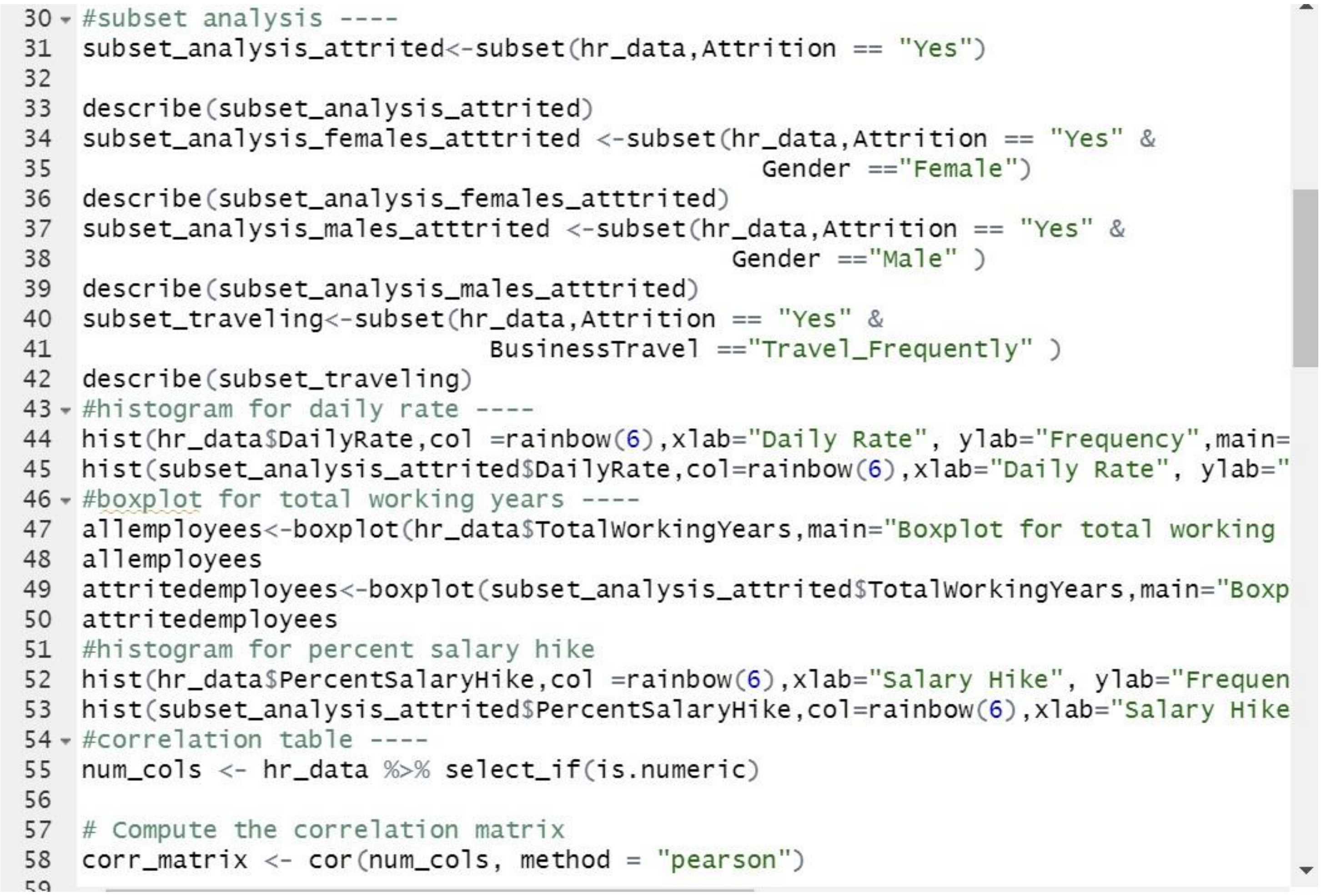
2. Quick-R: Descriptives. (2022). Dcd.https://www.statmethods.net/stats/descriptives.html .

3.1.9 Subgroup analyses: finding means and standard deviations for subgroups. (n.d.). Unknown. https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R-Manual/RManual7.h

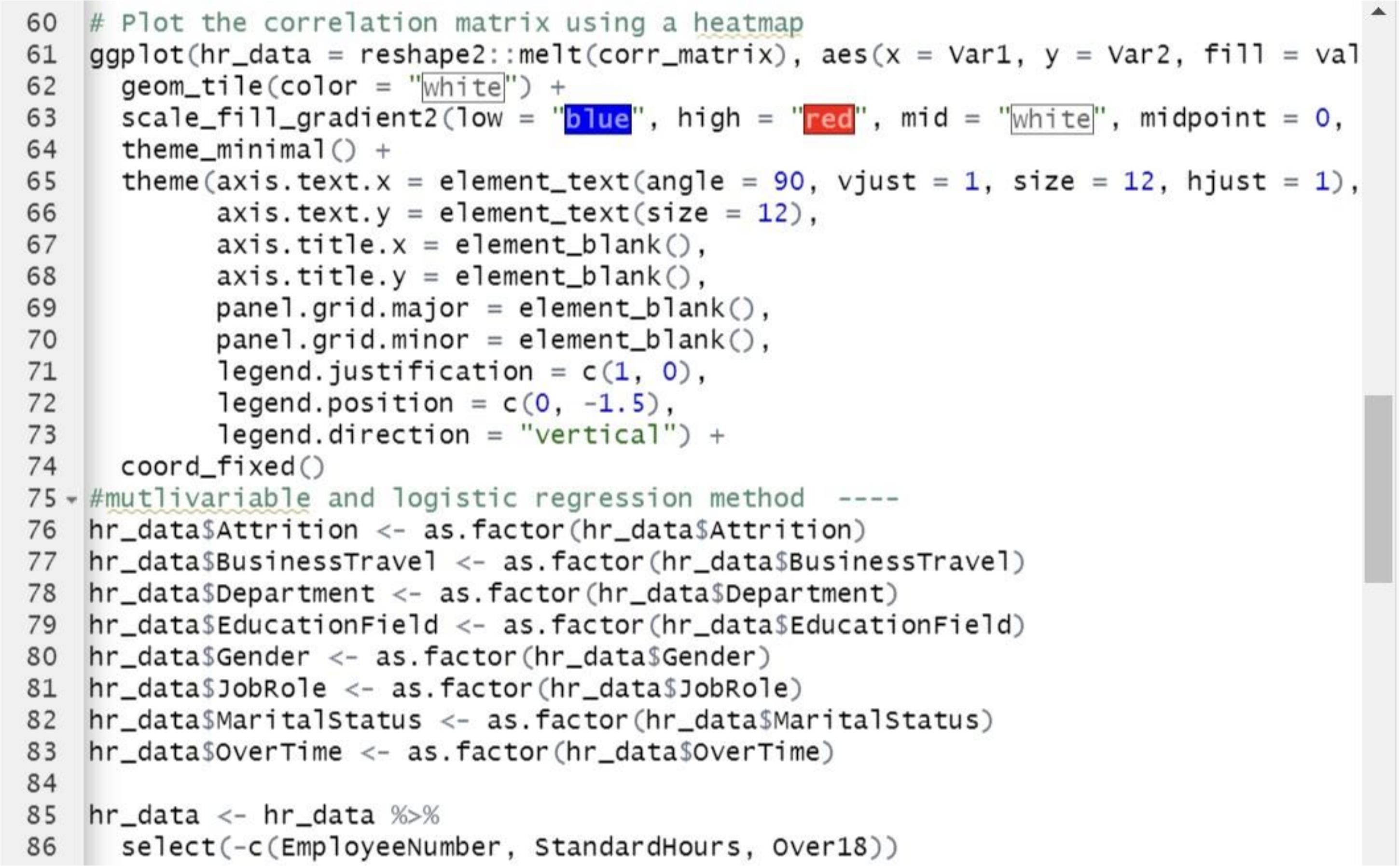
Appendix:



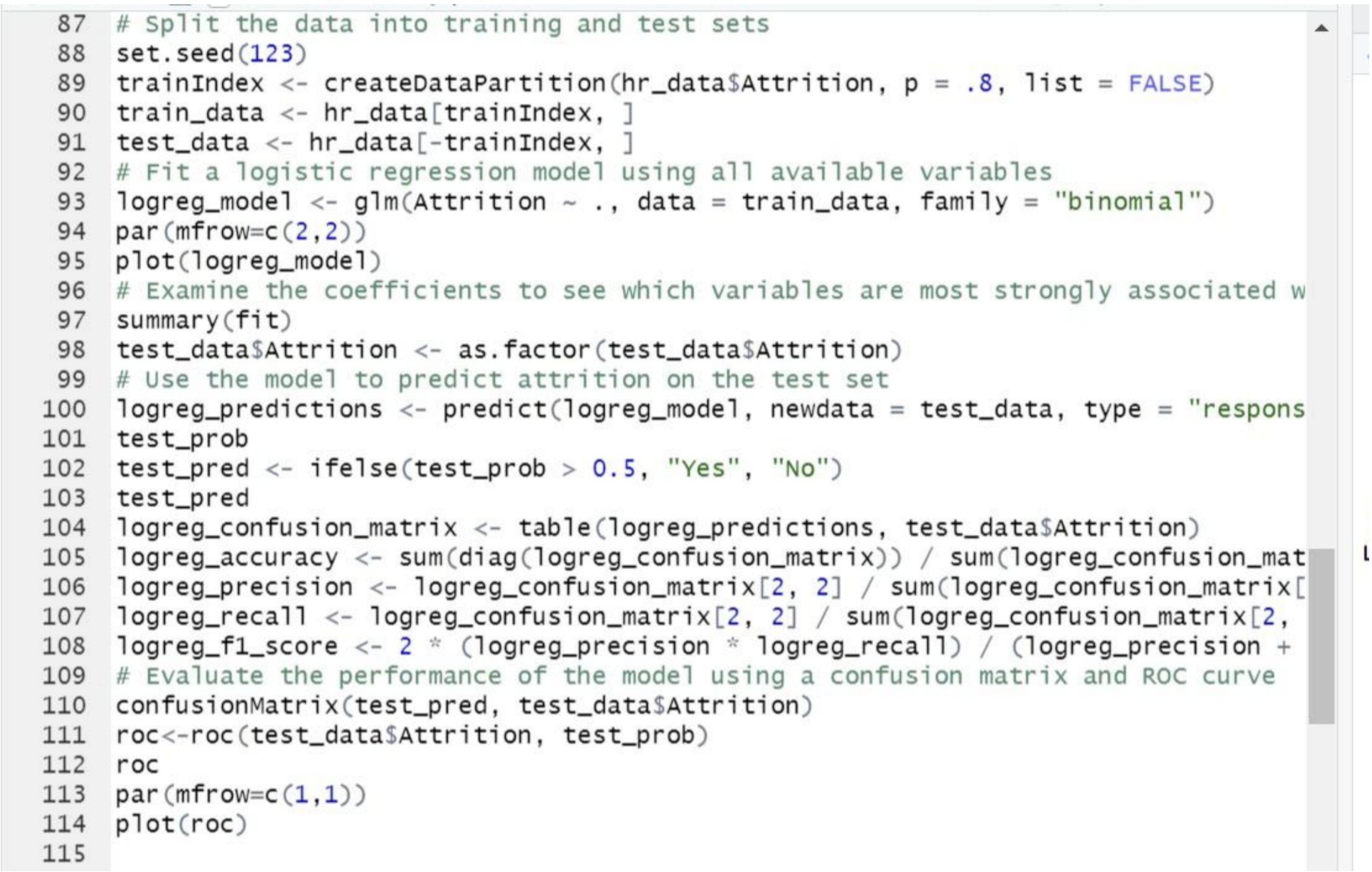
```
#loading library ----
library(psych)
library(corrplot)
install.packages("gtsummary")
library(gtsummary)
install.packages("caTools")
library(caTools)
library("dplyr")
library(tidyverse)
library(caret)
install.packages("party")
library(party)
library(pROC)
library(tidyverse)
install.packages("randomForest")
library(randomForest)
#loading hr_data ----
hr_data <-read.csv("WA_Fn-UseC_-HR-Employee-Attrition.csv",stringsAsFactors = TR
hr_data
#cleaning the hr_data ----
hr_data[hr_data==""]<- NA
hr_data<-na.omit(hr_data)
hr_data
#summary about the hr_data ----
summary(hr_data)
sapply(hr_data,class)
dim(hr_data)
psych::describe(hr_data)
str(hr_data)
#subset analysis ----
```



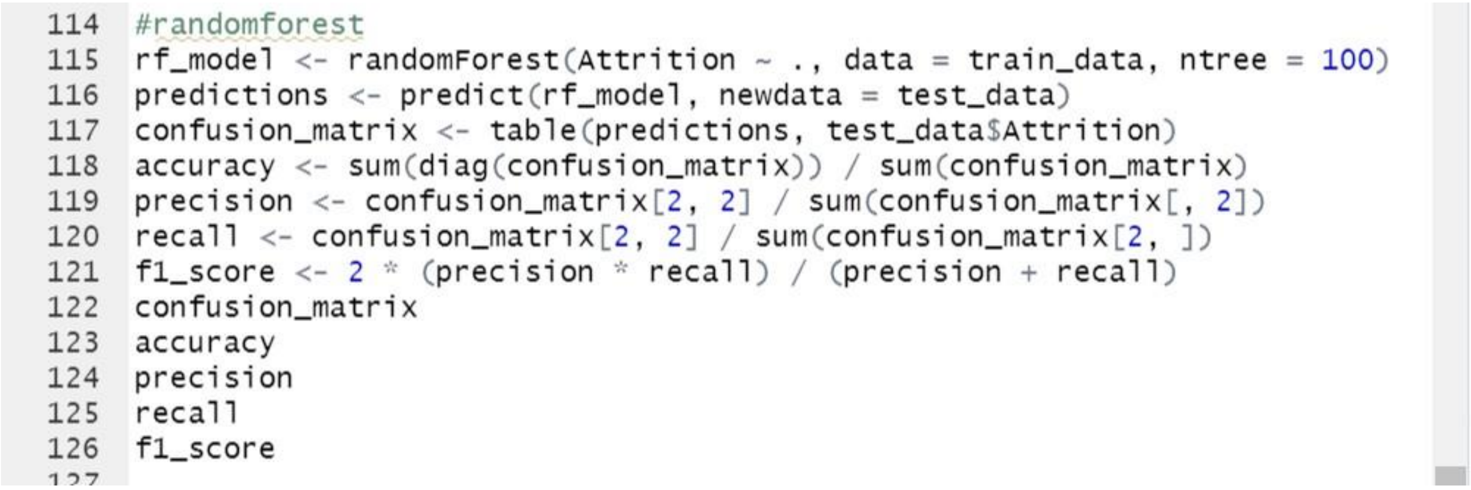
```
#subset analysis ----
subset_analysis_attrited<-subset(hr_data,Attrition == "Yes")

describe(subset_analysis_attrited)
subset_analysis_females_atttrited <-subset(hr_data,Attrition == "Yes" &
                                             Gender =="Female")
describe(subset_analysis_females_atttrited)
subset_analysis_males_atttrited <-subset(hr_data,Attrition == "Yes" &
                                           Gender =="Male" )
describe(subset_analysis_males_atttrited)
subset_traveling<-subset(hr_data,Attrition == "Yes" &
                           BusinessTravel =="Travel_Frequently" )
describe(subset_traveling)
#histogram for daily rate ----
hist(hr_data$DailyRate,col =rainbow(6),xlab="Daily Rate", ylab="Frequency",main=
hist(subset_analysis_attrited$DailyRate,col=rainbow(6),xlab="Daily Rate", ylab="
#boxplot for total working years ----
allemployees<-boxplot(hr_data$TotalWorkingYears,main="Boxplot for total working
allemployees
attritedemployees<-boxplot(subset_analysis_attrited$TotalWorkingYears,main="Boxp
attritedemployees
#histogram for percent salary hike
hist(hr_data$PercentSalaryHike,col =rainbow(6),xlab="Salary Hike", ylab="Frequen
hist(subset_analysis_attrited$PercentSalaryHike,col=rainbow(6),xlab="Salary Hike
#correlation table ----
num_cols <- hr_data %>% select_if(is.numeric)

# Compute the correlation matrix
corr_matrix <- cor(num_cols, method = "pearson")
```



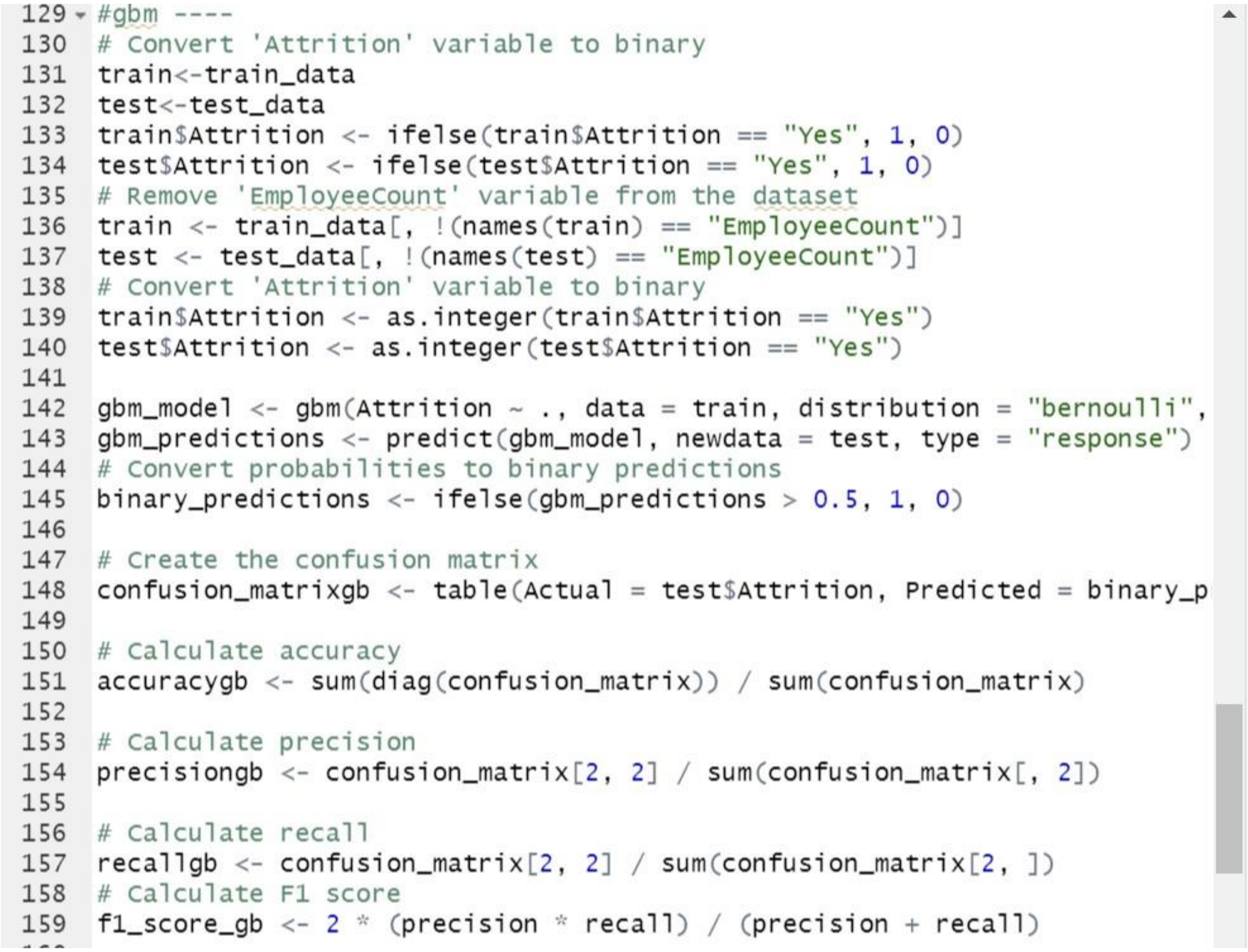
```
# Plot the correlation matrix using a heatmap
ggplot(hr_data = reshape2::melt(corr_matrix), aes(x = Var1, y = Var2, fill = val
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0,
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, size = 12, hjust = 1),
        axis.text.y = element_text(size = 12),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.justification = c(1, 0),
        legend.position = c(0, -1.5),
        legend.direction = "vertical") +
  coord_fixed()
#mutlivariable and logistic regression method  ----
hr_data$Attrition <- as.factor(hr_data$Attrition)
hr_data$BusinessTravel <- as.factor(hr_data$BusinessTravel)
hr_data$Department <- as.factor(hr_data$Department)
hr_data$EducationField <- as.factor(hr_data$EducationField)
hr_data$Gender <- as.factor(hr_data$Gender)
hr_data$JobRole <- as.factor(hr_data$JobRole)
hr_data$MaritalStatus <- as.factor(hr_data$MaritalStatus)
hr_data$OverTime <- as.factor(hr_data$OverTime)

hr_data <- hr_data %>%
  select(-c(EmployeeNumber, StandardHours, Over18))
```



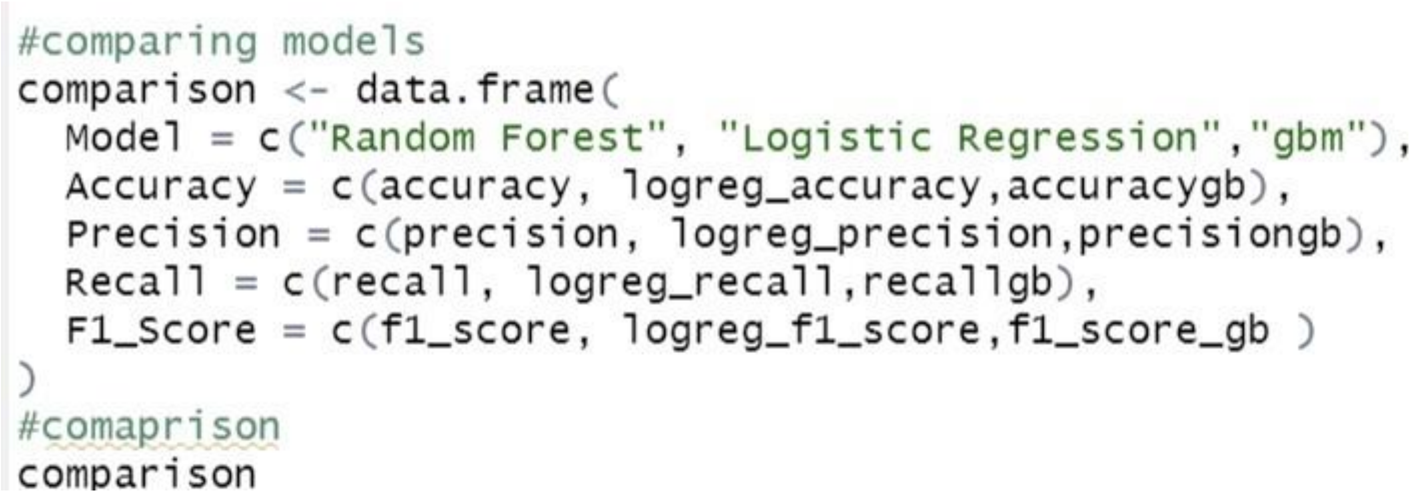
```r
# Split the data into training and test sets
set.seed(123)
trainIndex <- createDataPartition(hr_data$Attrition, p = .8, list = FALSE)
train_data <- hr_data[trainIndex, ]
test_data <- hr_data[-trainIndex, ]
# Fit a logistic regression model using all available variables
logreg_model <- glm(Attrition ~ ., data = train_data, family = "binomial")
par(mfrow=c(2,2))
plot(logreg_model)
# Examine the coefficients to see which variables are most strongly associated w
summary(fit)
test_data$Attrition <- as.factor(test_data$Attrition)
# Use the model to predict attrition on the test set
logreg_predictions <- predict(logreg_model, newdata = test_data, type = "respons
test_prob
test_pred <- ifelse(test_prob > 0.5, "Yes", "No")
test_pred
logreg_confusion_matrix <- table(logreg_predictions, test_data$Attrition)
logreg_accuracy <- sum(diag(logreg_confusion_matrix)) / sum(logreg_confusion_mat
logreg_precision <- logreg_confusion_matrix[2, 2] / sum(logreg_confusion_matrix[
logreg_recall <- logreg_confusion_matrix[2, 2] / sum(logreg_confusion_matrix[2,
logreg_f1_score <- 2 * (logreg_precision * logreg_recall) / (logreg_precision +
# Evaluate the performance of the model using a confusion matrix and ROC curve
confusionMatrix(test_pred, test_data$Attrition)
roc<-roc(test_data$Attrition, test_prob)
roc
par(mfrow=c(1,1))
plot(roc)

```



```r
#randomforest
rf_model <- randomForest(Attrition ~ ., data = train_data, ntree = 100)
predictions <- predict(rf_model, newdata = test_data)
confusion_matrix <- table(predictions, test_data$Attrition)
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
precision <- confusion_matrix[2, 2] / sum(confusion_matrix[, 2])
recall <- confusion_matrix[2, 2] / sum(confusion_matrix[2, ])
f1_score <- 2 * (precision * recall) / (precision + recall)
confusion_matrix
accuracy
precision
recall
f1_score
```



```
#gbm ----
# Convert 'Attrition' variable to binary
train<-train_data
test<-test_data
train$Attrition <- ifelse(train$Attrition == "Yes", 1, 0)
test$Attrition <- ifelse(test$Attrition == "Yes", 1, 0)
# Remove 'EmployeeCount' variable from the dataset
train <- train_data[, !(names(train) == "EmployeeCount")]
test <- test_data[, !(names(test) == "EmployeeCount")]
# Convert 'Attrition' variable to binary
train$Attrition <- as.integer(train$Attrition == "Yes")
test$Attrition <- as.integer(test$Attrition == "Yes")

gbm_model <- gbm(Attrition ~ ., data = train, distribution = "bernoulli",
gbm_predictions <- predict(gbm_model, newdata = test, type = "response")
# Convert probabilities to binary predictions
binary_predictions <- ifelse(gbm_predictions > 0.5, 1, 0)

# Create the confusion matrix
confusion_matrixgb <- table(Actual = test$Attrition, Predicted = binary_p

# Calculate accuracy
accuracygb <- sum(diag(confusion_matrix)) / sum(confusion_matrix)

# Calculate precision
precisiongb <- confusion_matrix[2, 2] / sum(confusion_matrix[, 2])

# Calculate recall
recallgb <- confusion_matrix[2, 2] / sum(confusion_matrix[2, ])
# Calculate F1 score
f1_score_gb <- 2 * (precision * recall) / (precision + recall)
```



```
#comparing models
comparison <- data.frame(
  Model = c("Random Forest", "Logistic Regression","gbm"),
  Accuracy = c(accuracy, logreg_accuracy,accuracygb),
  Precision = c(precision, logreg_precision,precisiongb),
  Recall = c(recall, logreg_recall,recallgb),
  F1_Score = c(f1_score, logreg_f1_score,f1_score_gb )
)
#comaprison
comparison
```