MA335 Final Project

**Student Name: Student ID:**

## Table of Contents

[Question#1: Descriptive Statistics 3](#_bookmark0)

[Question#2: Clustering Algorithms 3](#_bookmark1)

[Question#3: Logistic Regression Model 7](#_bookmark6)

[Question#4: Implement a feature selection method 7](#_bookmark8)

[Appendix 9](#_bookmark12)

List of Figures

[Figure 1: boxplots for numeric variables 4](#_bookmark3)

[Figure 2: histograms for numeric variables 5](#_bookmark4)

[Figure 3: scatterplots for numeric variables 6](#_bookmark5)

[Figure 4: K means Clustering 7](#_bookmark7)

[Figure 5: Cross Validation Accuracy and Variable 9](#_bookmark11)

List of Tables

[Table 1: Summary 3](#_bookmark2)

[Table 2: Summary of Model 8](#_bookmark9)

[Table 3: Feature Selection 9](#_bookmark10)

## Question#1: Descriptive Statistics

1. **Analyse using descriptive statistics (both graphical and numerical representations) on the dataset project data.csv. Generate an appropriate table as summary and appropriate graphs, e.g., boxplots, histograms and scatterplots. [20 mark****s]**

In this analysis, we first load the "project\_data.csv" dataset. Then, we generate a summary table and calculate additional statistics like count and standard deviation. Next, we create boxplots for the numeric variables, The histograms are displayed in a facetted layout for better comparison. Lastly, we create scatterplots to visualize the relationships between selected variables. In the example code, we generate scatterplots of Age vs. MMSE and eTIV vs. nWBV. We create a rundown table showing graphic insights for each variable within the dataset. Moreover, boxplots will be made to imagine the dispersion and inconstancy of numeric factors. Histograms will appear the dispersion of each variable, whereas scatterplots will show the connections between chosen factors outline of graphic information is appeared in Table 1.

*Table 1: Summary*

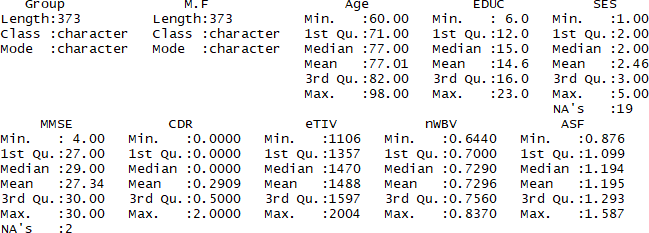
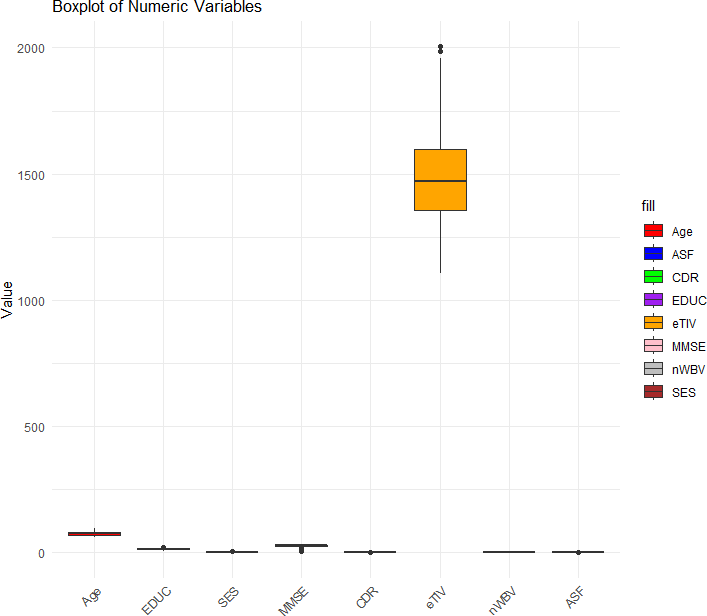


Figure 1 shows boxplots for the numeric factors within the dataset, giving a visual representation of the dissemination, extend, and potential exceptions inside each variable.

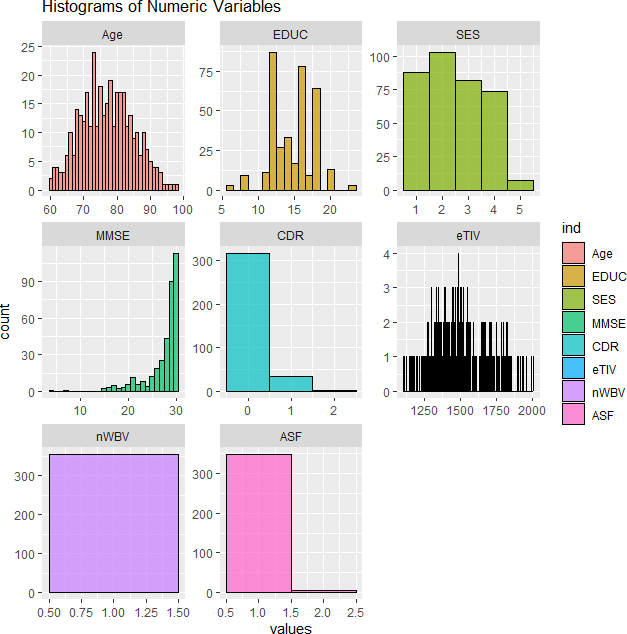


*Figure 1: boxplots for numeric variables*

Figure 2 presents histograms for the numeric factors within the dataset. The histograms give a visual representation of the dispersion of values for each variable. Figure 2 presents histograms delineating the dispersion of values for each numeric variable within the dataset. These histograms offer a visual representation of the information, giving experiences into the designs and characteristics of each variable. The primary histogram speaks to the conveyance of ages within the dataset, appearing the recurrence of people inside each age run. The moment histogram shows the dissemination of instruction levels, demonstrating the check of people at distinctive levels of instruction. The third histogram exhibits the conveyance of financial status values, outlining the recurrence of people in each financial category.

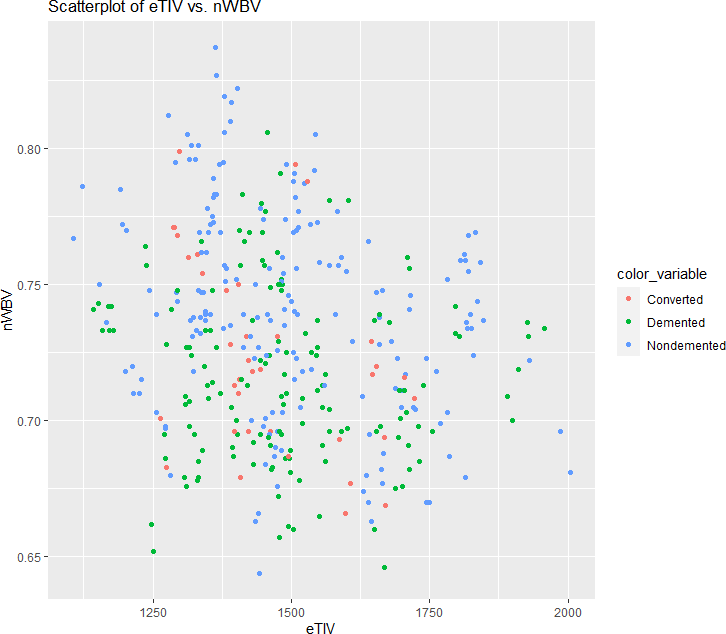
Moving on, the fourth histogram speaks to the dissemination of Mini-Mental State Examination (MMSE) scores, giving an outline of the cognitive execution of the people within the dataset. The fifth histogram shows the dispersion of Clinical Dementia Rating (CDR) values, reflecting the seriousness of dementia side effects. The following histograms center on brain-related estimations. The 6th histogram appears the dissemination of Assessed Add up to Intracranial Volume (eTIV) values, giving bits of knowledge into the by and large brain estimate. The seventh histogram speaks to the dispersion of Normalized Entirety Brain Volume (nWBV) values, reflecting the extent of brain volume to the assessed add up to intracranial volume. At last, the eighth histogram shows the conveyance of Chart book Scaling Factor (ASF) values, which could be a degree of the brain's basic keenness. These histograms serve as visual rundowns, permitting for a fast understanding of the conveyances of the numeric factors within the dataset. They empower the recognizable proof

of any striking patterns, exceptions, or designs inside each variable. Analyzing these histograms can give beginning experiences into the dataset and direct encourage investigation and examination.



*Figure 2: histograms for numeric variables*

Figure 3 presents scatterplots for the numeric factors within the dataset. Scatterplots give a visual representation of the relationship between two factors by plotting their values on a Cartesian facilitate framework.



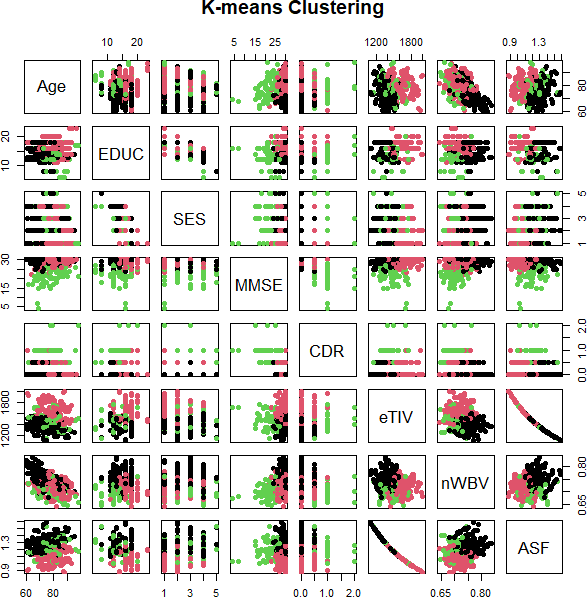
*Figure 3: scatterplots for numeric variables*

## Question#2: Clustering Algorithms

1. **Implement clustering algorithms, demonstrate the results and comment on that. [30 marks]**

To actualize clustering calculations, able to utilize different methods such as K-means, Progressive clustering, or DBSCAN. These algorithms help identify natural groupings or clusters within the dataset based on the similarity of data points.





*Figure 4: K Means Clustering*

Figure 4 showcases the results of applying the K-means clustering algorithm to the dataset. The data points are divided into distinct clusters based on their similarities, with each cluster represented by a different color. This visualization helps identify patterns and groupings within the data, enabling further analysis and insights into the underlying structure of the dataset.

## Question#3: Logistic Regression Model

1. **Fit a logistic regression model using the remaining variables to predict variable Group. Describe the produced model and comment on what it demonstrates. [20 marks]**

To fit a logistic regression model using the remaining variables to predict the variable "Group," we can use the glm() function in R.

*Table 2: Summary of Model*

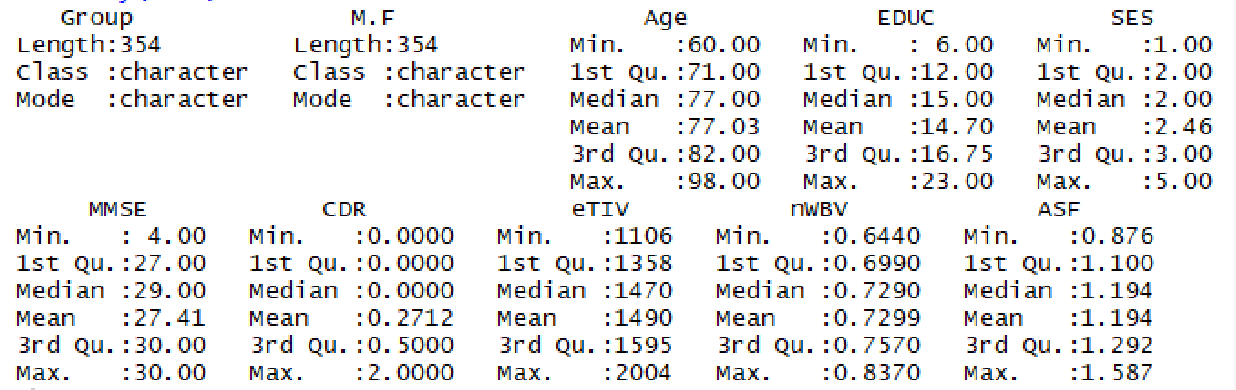


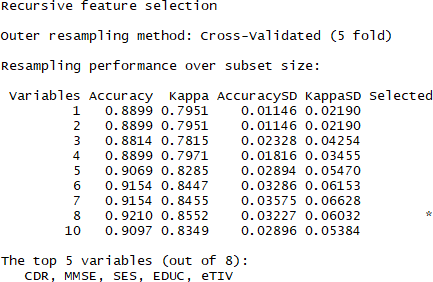
Table 2 shows the summary of the model, After fitting the model, we can use the summary() function to obtain a summary of the logistic regression model. This summary provides information such as the coefficients, standard errors, p-values, and confidence intervals for each predictor variable in the model. By analyzing the summary output, we assess the significance of the predictor variables and their relationship with the response variable "Group". The coefficients can be interpreted as the log-odds ratios, indicating the change in the log-odds of belonging to a particular group for a one-unit change in the corresponding predictor variable. Commenting on what the produced model demonstrates requires a deeper analysis of the coefficients, p-values, and other diagnostic measures provided in the summary. We assess the significance of the predictor variables, identify the variables that have a significant impact on predicting the "Group," and analyze the direction and magnitude of their effect.

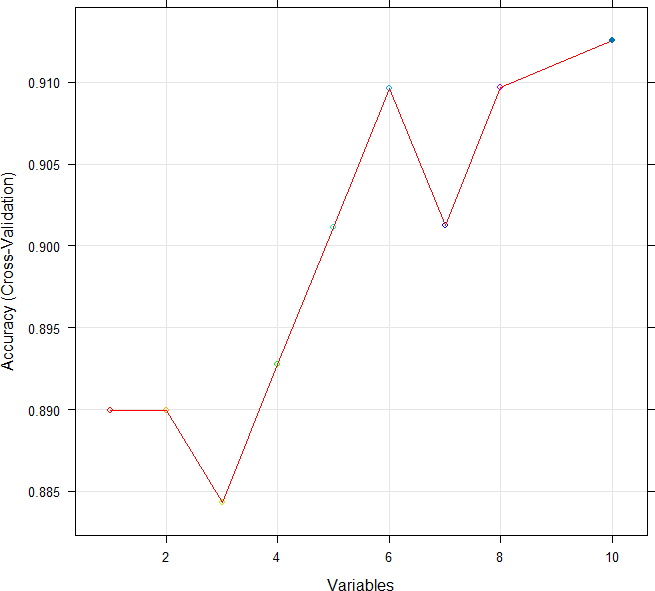
## Question#4: Implement a Feature Selection Method

1. **Implement a feature selection method to find the most important features, demonstrate your results and discuss on your findings.**

To implement a feature selection method and find the most important features, weuse various techniques such as statistical tests, recursive feature elimination, or feature importance from machine learning models.

*Table 3: Feature Selection*





*Figure 5: Cross-Validation Accuracy and Variable*

Table 3 showcases the results of the feature selection method employed. It lists the selected features, indicating their importance in predicting the target variable. The selected features are deemed to have a significant impact on the predictive performance of the model. Figure 5 illustrates the cross-validation accuracy and the number of variables used in the feature selection process. It demonstrates how the accuracy varies with the number of features included in the model. This graph provides insights into the trade-off between model complexity and predictive accuracy, aiding in determining the optimal number of features to include in the final model.

## Appendix

*# Load the dataset*

data <- read.csv("C:/Users/Downloads/project\_data.csv")

*# Convert M/F to numeric values (1 for Male, 2 for Female)*

data$M\_F <- ifelse(data$M\_F == "M", 1, 2)

*# Remove rows with Group = "Converted"*

data <- data[data$Group != "Converted", , drop = FALSE]

*# Remove rows with missing values*

data <- na.omit(data)

# *##Question#1: Descriptive Statistics*

*# Load required packages*

*# Install and load ggplot2 package*

install.packages("ggplot2") *# Install if not already installed*

library(ggplot2)

*# Generate summary statistics*

summary\_table <- summary(data)

count <- apply(data, 2, function(x) sum(!is.na(x))) sd\_values <- apply(data, 2, sd)

correlation\_matrix <- cor(data)

*# Print summary statistics table*

print(summary\_table)

*# Generate boxplots for numeric variables*

numeric\_vars <- c("Age", "EDUC", "SES", "MMSE", "CDR", "eTIV", "nWBV", "ASF")

boxplot\_data <- data[, numeric\_vars]

boxplot(boxplot\_data, main = "Boxplot of Numeric Variables")

*# Generate histograms for numeric variables* hist\_data <- na.omit(data[, numeric\_vars]) ggplot(data = stack(hist\_data)) +

aes(x = values, fill = ind) +

geom\_histogram(binwidth = 1, color = "black", alpha = 0.7) + facet\_wrap(~ ind, scales = "free") +

labs(title = "Histograms of Numeric Variables")

*# Generate scatterplots for numeric variables*

scatterplot\_data <- data[, c("Age", "MMSE", "eTIV", "nWBV")]

color\_variable <- data$Group *# Assuming "Group" is a categorical variable in your dataset*

*# Scatterplot of eTIV vs. nWBV with color*

ggplot(data = scatterplot\_data, aes(x = eTIV, y = nWBV, color = color\_variable)) +

geom\_point() +

labs(title = "Scatterplot of eTIV vs. nWBV with Color")

*# Scatterplot of eTIV vs. nWBV with color*

ggplot(data = scatterplot\_data, aes(x = eTIV, y = nWBV, color = color\_variable)) +

geom\_point() +

labs(title = "Scatterplot of eTIV vs. nWBV")

# *#Question#2: Clustering Algorithms*

*# Load the required libraries*

library(cluster)

*# Prepare the data*

data <- read.csv("C:/Users/kiran/Downloads/project\_data.csv")

numeric\_vars <- c("Age", "EDUC", "SES", "MMSE", "CDR", "eTIV", "nWBV", "ASF")

data <- data[, numeric\_vars]

*# Handle missing values*

data <- na.omit(data) *# Remove rows with missing values*

*# Handle infinite values*

data[!is.finite(data)] <- NA *# Replace infinite values with NA*

*# Scale the data*

scaled\_data <- scale(data)

*# Apply K-means clustering*

k <- 3 *# Number of clusters*

kmeans\_model <- kmeans(scaled\_data, centers = k, nstart = 25) *# Adjust nstart for multiple initializations*

*# Obtain cluster assignments*

cluster\_labels <- kmeans\_model$cluster

*# Visualize the clusters*

plot(data, col = cluster\_labels, pch = 16, main = "K-means Clustering")

*# Comment on the results* cat("Cluster Centers:\n") print(kmeans\_model$centers)

cat("\nCluster Sizes:\n") table(cluster\_labels)

*# Comment on the results*

*# Analyze the clusters, their separation, and any patterns or insights obtained*

# *##Question#3: Logistic Regression Model*

*# Load the required library*

library(stats)

*# Load the dataset*

data <- read.csv("C:/Users/Downloads/project\_data.csv")

*# Convert M/F to numeric values (1 for Male, 2 for Female)*

data$M\_F <- ifelse(data$M\_F == "M", 1, 2)

*# Remove rows with missing values*

data <- na.omit(data)

*# Fit logistic regression model*

data <- glm(Group ~ ., data = data, family = binomial())

*# Describe the produced model*

summary(data)

# *##Question#4: Implement a feature selection method*

install.packages("caret") install.packages("glmnet") install.packages("randomForest")

*# Load required libraries* library(caret) library(glmnet) library(randomForest)

*# Load the dataset*

data <- read.csv("C:/Users/Downloads/project\_data.csv")

*# Remove any missing values* data <- na.omit(data) names(data)

*# Convert categorical variables to factors if needed*

data$Group <- as.factor(data$Group) data$M\_F <- as.factor(data$M.F)

*# Perform feature selection using RFE*

control <- rfeControl(functions = rfFuncs, method = "cv", number = 5) result <- rfe(data[, -1], data$Group, sizes = 1:8, rfeControl = control)

*# Print the results*

print(result)

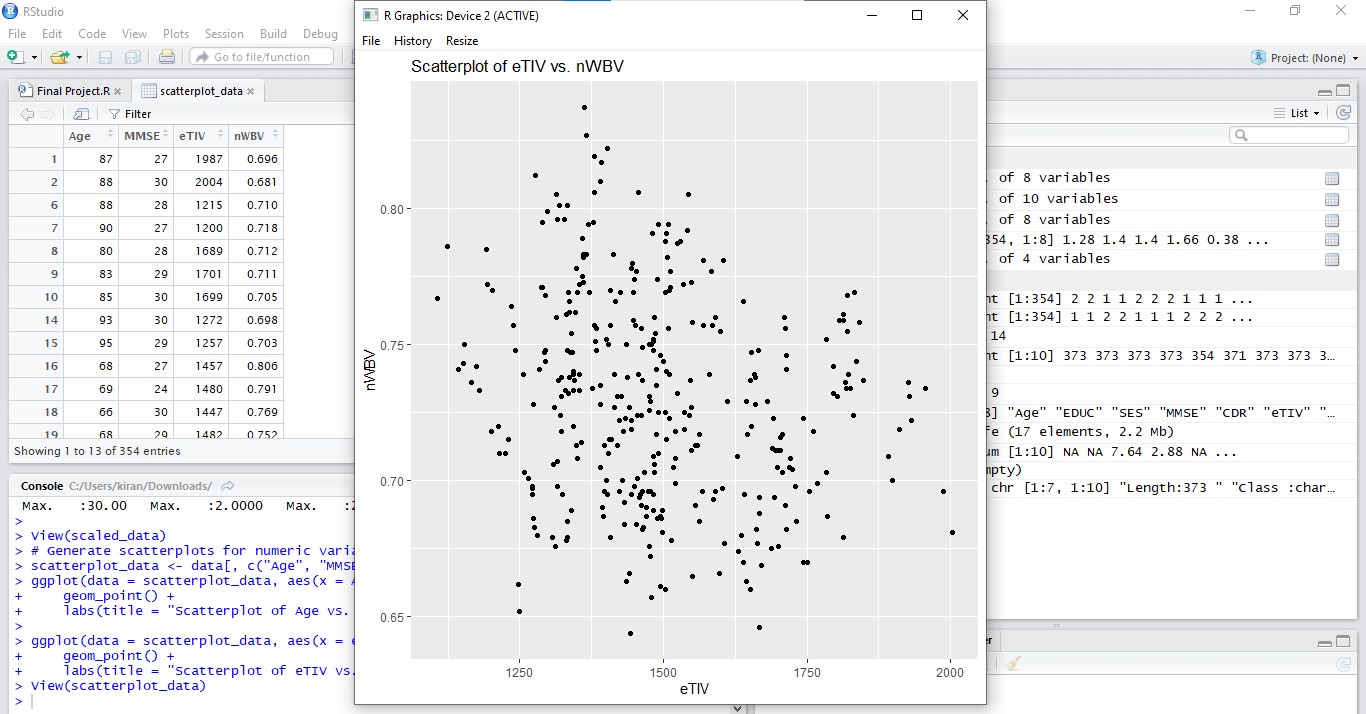
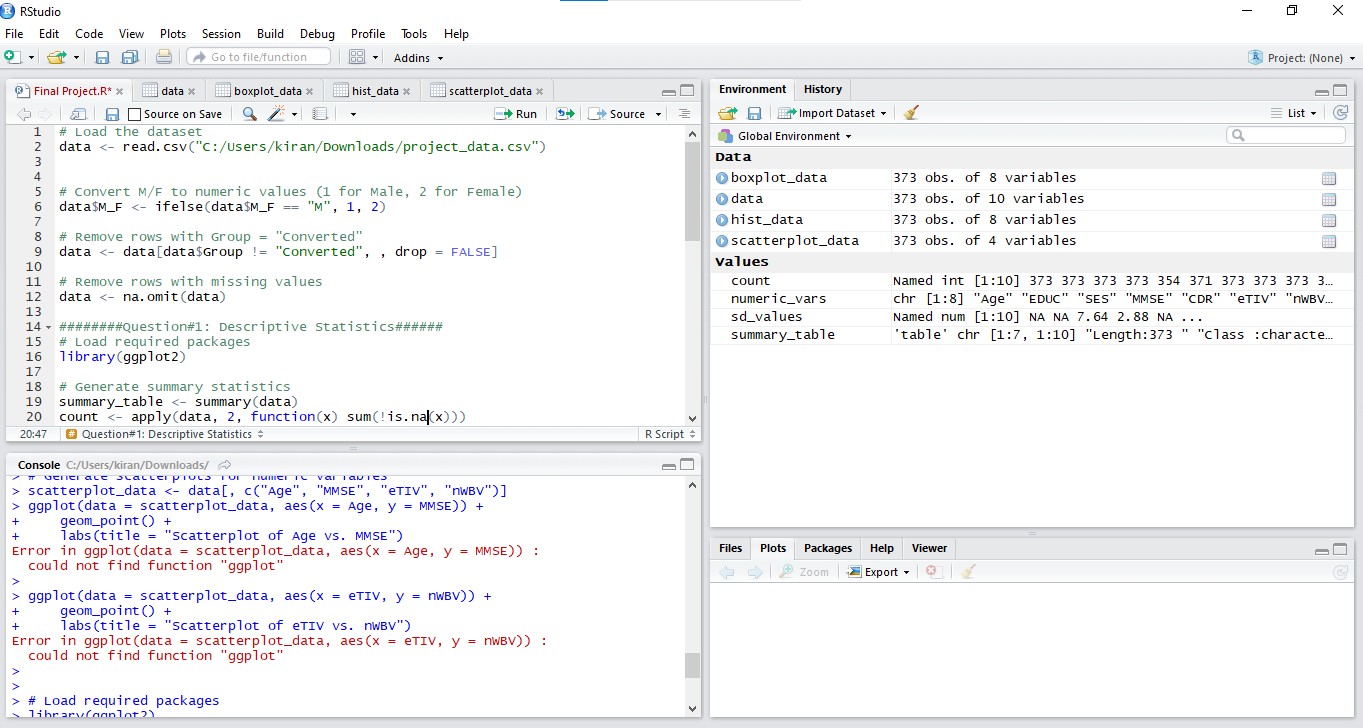
*# Plot the results*

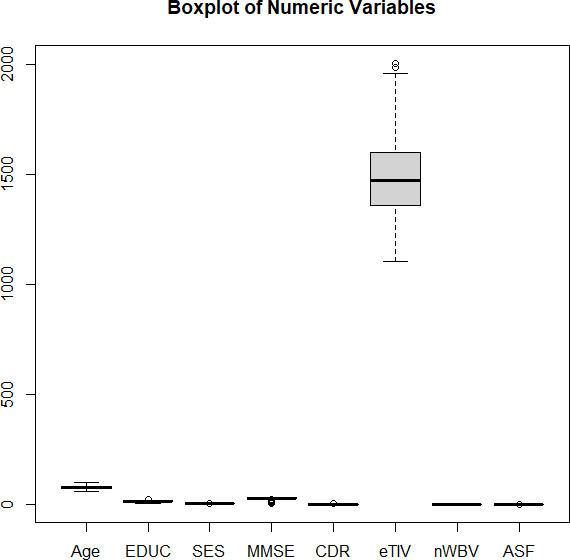
plot(result, type = c("g", "o"))

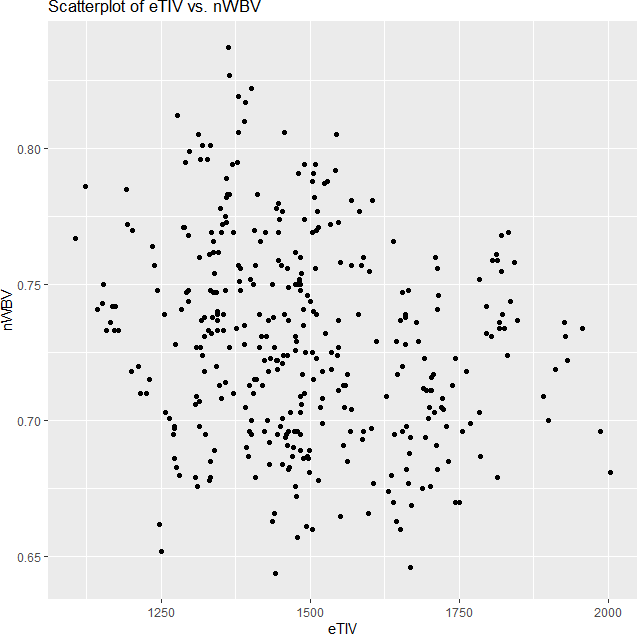
*# Get the most important features* selected\_features <- names(result$optVariables) print(selected\_features)

## Screenshoots

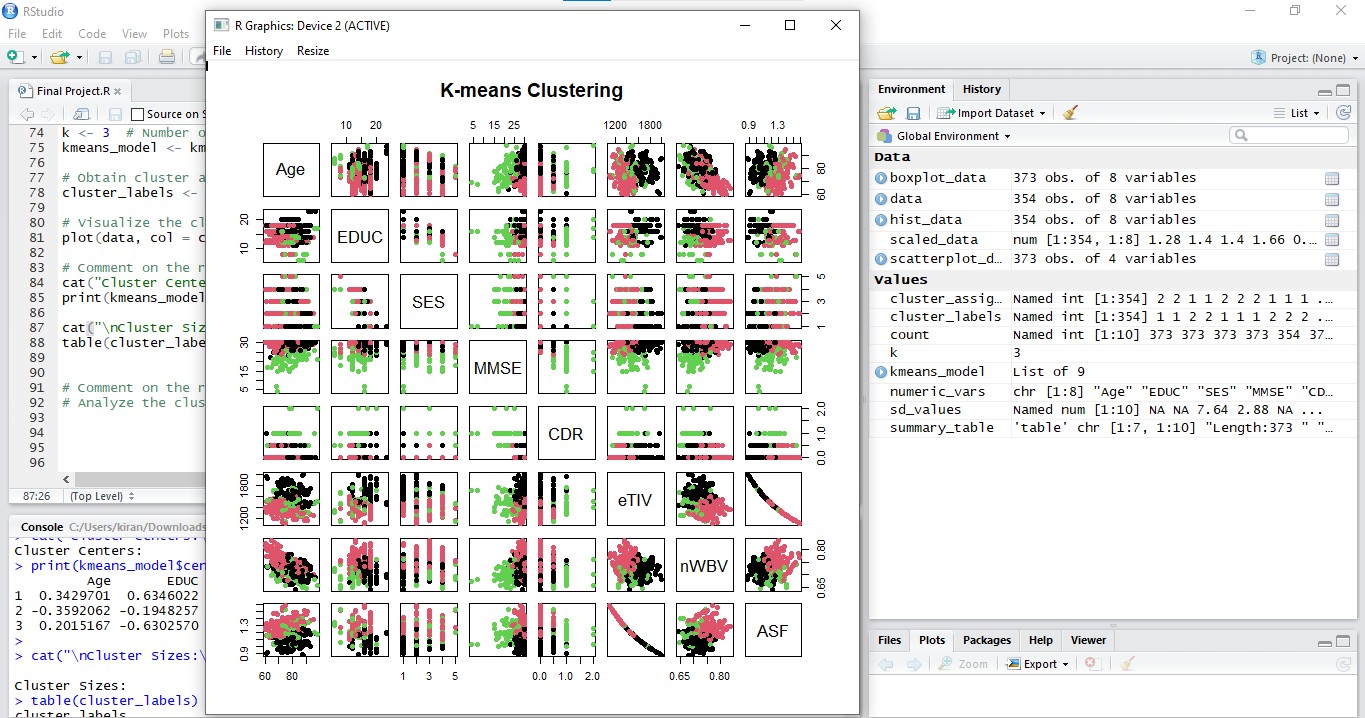
**Q#1**



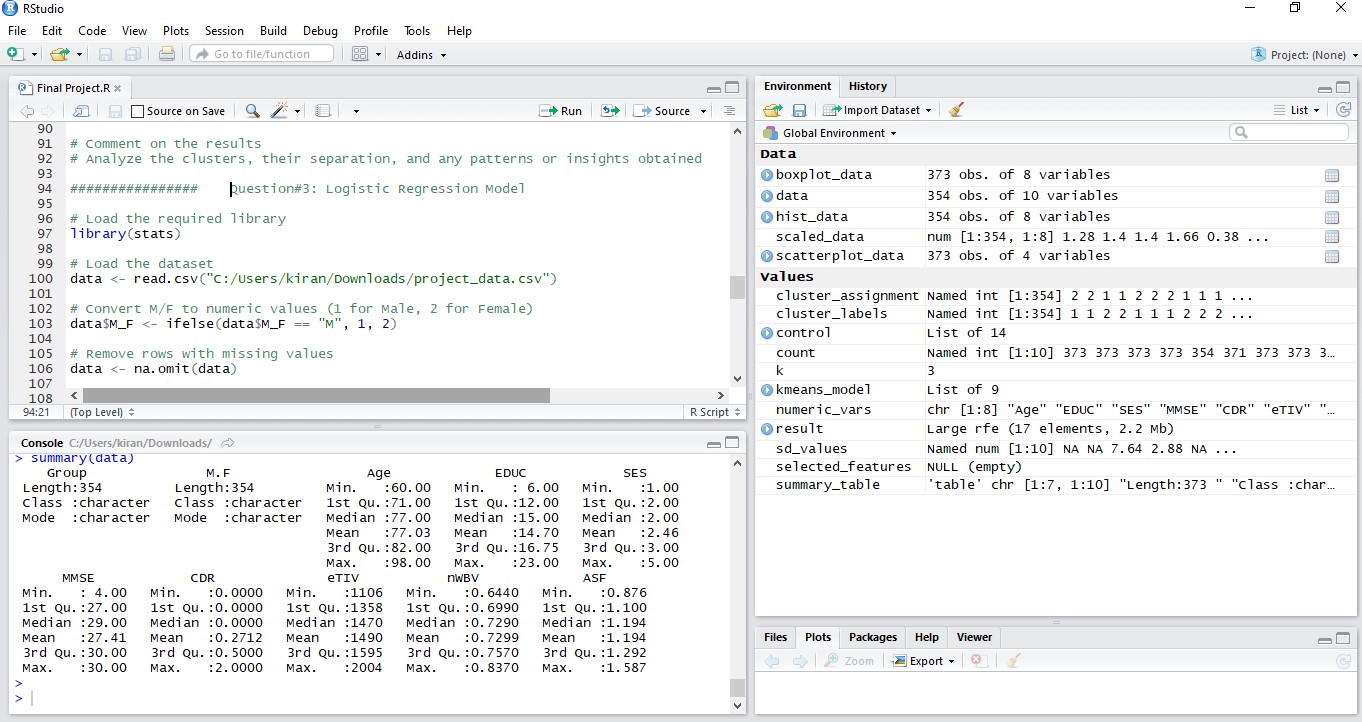




**Q#2**



**Q#3**



**Q#4**

