MARYAM TAJ

MA335 FINAL PROJECT

MSC APPLIED DATA SCIENCE

REGISTRATION ID: 2211714

DATED: 20TH JUNE 2023

Abstract

In this project, we aimed to investigate the relationship between various characteristics of Alzheimer's disease and the diagnosis of Alzheimer's (Demented) or not (Nondemented). Using R programming language, the provided dataset (project data.csv) was analyzed through descriptive statistics, clustering algorithms, logistic regression modeling, and feature selection methods. The findings contribute to understanding Alzheimer's disease and have implications for early detection and management.

Contents:

Word Count:

The total word count for this report excluding cover page and appendix is 1900.

INTRODUCTION

This report presents the analysis of a dataset on Alzheimer's disease, aiming to investigate the relationship between various characteristics and the diagnosis of Alzheimer's (Demented) or non-Alzheimer's (Nondemented). By analyzing the dataset and applying various statistical techniques, this report aims to contribute to the understanding of the characteristics associated with Alzheimer's disease. The insights gained from this analysis can potentially inform clinical decision-making and further research in the field of Alzheimer's disease.

RESEARCH QUESTION

As the dataset mainly includes the characteristics of the Alzheimer's disease. The main objective of this report is to investigate the relationship between those characteristics and the diagnosis, i.e., Alzheimer (Demented) or not (Nondemented).

PRELIMINARY ANALYSIS OF THE DATA

1. Descriptive Statistics

   In this dataset, we have the x and y variables. Our dependent variable (y) is **Group** and the remaining variables which are the characteristics of the Alzheimer's disease are the x variables. These are named as:
   - Gender (M.F)
   - Age
   - Year of Education (EDUC)
   - Socioeconomic Status (SES)
   - Mini Mental State Examination (MMSE)
   - Clinical Dementia Rating (CDR)
   - Estimated Total Intracranial Volume (eTIV)
   - Normalize Whole Brain Volume (nWBV) and Atlas Scaling Factor (ASF)

> summary(data)

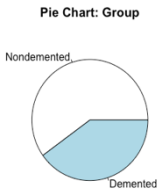
Table 01

Group	**M.F**	**Age**	**EDUC**	**SES**
Length:317	Min: 0.0000	Min: 60.00	Min: 6.00	Min: 1.000
Class:character	1st Qu: 0.0000	1st Qu: 71.00	1st Qu: 12.00	1st Qu: 2.000
Mode:character	Median: 0.0000	Median: 76.00	Median: 15.00	Median: 2.000
	Mean: 0.4322	Mean: 76.72	Mean: 14.62	Mean: 2.546
	3rd Qu: 1.0000	3rd Qu: 82.00	3rd Qu: 16.00	3rd Qu: 3.000
	Max: 1.0000	Max: 98.00	Max: 23.00	Max: 5.000
MMSE	**CDR**	**eTIV**	**nWBV**	**ASF**
Min: 4.00	Min: 0.0000	Min: 1106	Min: 0.6440	Min: 0.876
1st Qu: 27.00	1st Qu: 0.0000	1st Qu: 1358	1st Qu: 0.7000	1st Qu: 1.098
Median: 29.00	Median: 0.0000	Median: 1476	Median: 0.7320	Median: 1.189
Mean: 27.26	Mean: 0.2729	Mean: 1494	Mean: 0.7306	Mean: 1.192
3rd Qu: 30.00	3rd Qu: 0.5000	3rd Qu: 1599	3rd Qu: 0.7570	3rd Qu: 1.293
Max: 30.00	Max: 2.0000	Max: 2004	Max: 0.8370	Max: 1.587

The summary table provides an overview of the dataset, including the count, minimum, maximum, and quartiles for each variable. The "Group" variable represents the diagnosis of individuals as either "Nondemented," "Demented," or "Other." The "M.F" variable indicates the gender, with 2 representing females and 1 representing males.
These descriptive statistics provide initial insights into the dataset. For example, the mean age of the individuals is approximately 76.72 years, with a range of 60 to 98 years. The mean education level is approximately 14.62 years, with a range of 6 to 23 years. The mean CDR score is approximately 0.2729, indicating a relatively low severity of dementia on average. These statistics help us understand the distribution and range of values for each variable, which will be further analyzed in the subsequent sections of the report.

2. Graphical Representation of data:

In this section, graphical representations were used to visualize the data. We saw the overall representation of the data through the help of pairs plot which are mentioned in appendix (Figure 01). Those plots show the correlation, density, and scatter plots between the variables of the dataset. Through density plots, we can see the distribution of the variables.
In the pie chart, the nondemented are more in our data than demented. The counts for Non demented are 190 and demented are 127.
In the next graph, the distribution of gender by group can be seen. We can see that **the proportion of nondemented female is more that demented men**. It has been proved from graph that males are more demented to Alzheimer than females.

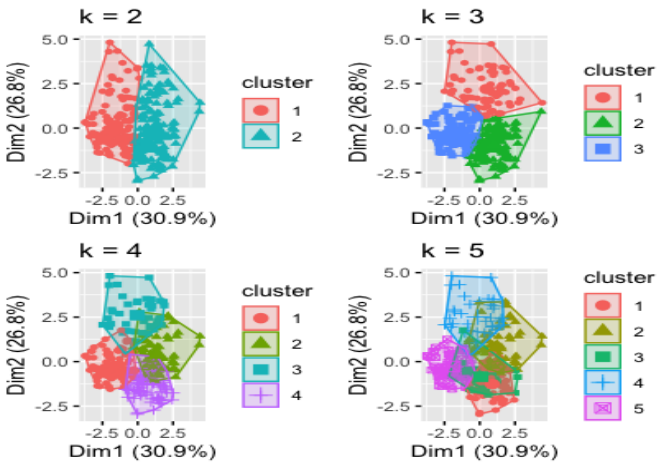


The next analysis was with the help of boxplot in which the variables SES, eTIV, ASF, CDR, nWBV were visualized across different groups (Demented or non demented). All the graphs are included in the appendix (Figure 02). In SES, the outlier is present in non demented, and its median is much lower than demented. Demented group has higher SES on average and spread is more for it. We can say that individuals with Alzheimer's have higher SES than others and greater diversity in socioeconomic backgrounds. In eTIV, the median for both groups were same but spread is different which means the difference in brain size and volume in non demented people. But due to outliers, there are some exceptional cases in demented group. The variable CDR shows Individuals in demented group have high CDR rating is not informative or varied enough to distinguish between individuals without dementia.

ANALYSIS:
In this section, we delve deeper into the dataset, applying various analytical techniques to gain insights and draw meaningful conclusions. The analysis includes the following components:

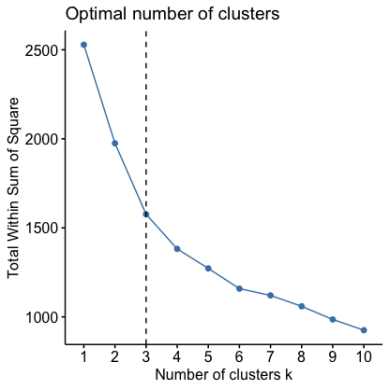
1) **Clustering Algorithm:**
   K means, and hierarchical cluster analysis can be used in our study on Alzheimer's disease to find distinct subgroups or trends among the patient population. In the **K means**, the Euclidean distance matrix was calculated and visualzed it through the heatmap can be seen in appendix (Figure 03). Darker colors indicate higher dissimilarity or distance, while lighter colors indicate lower dissimilarity. In the diagram, these scatterplots shows that each data point is represented as a point and the color indicates its assigned cluster. The following points have been observed:

- K=2: The data points are divided into two distinct clusters, each with its own centroid. The clusters appear well-separated, indicating a clear division in the data.
- For k = 3: The data points are now grouped into three clusters. We can see that there is less overlapping.
- For k = 4: With an additional cluster, we observe a further subdivision of the data, resulting in more refined groupings. **Overlapping** is observed in it.
- For k = 5: We observe a further subdivision of the data. More overlapping can be seen in it.



To Determine the optimal number of clusters, we use the two methods:

The one is **within cluster sum of squares (WCSS)** method for different values of K.
According to the elbow method, it can be seen in the diagram that adding more clusters beyond the K=3 does not significantly improve model's fit. So, the optimal number of clusters are three. The other method is **Silhouette** and in figure 04 in appendix it can be seen the peak of the points is K=3 and the value of the silhouette width is above 0.20 which proves that there is somehow overlapping in the clusters but not much.



In **Hierarchical Clustering,** four linkage methods were performed on the dataset e.g., single, complete, average, and centroid. The purpose was to identify clusters or groups of observations based on their similarities or dissimilarities. For each method, we plotted the resulting dendrogram, which visually displays the hierarchical structure of the clusters. The dendrogram in appendix (figure 05) shows how observations are merged or divided at different levels of similarity. We chose to cut the dendrogram into four clusters and highlighted them with red borders to aid in interpretation.

2) Logistic Regression Analysis

Logistic regression is performed to predict the binary variable "Group" based on several predictor variables including "M.F", "Age", "EDUC", "SES", "MMSE", "CDR", and "nWBV". The dummy variables have been made for the group variable, non demented=0 and demented=1. First, we checked the multicollinearity between the predictor variables and check their vif values. The vif values for the eTIV and ASF were very high, and it was also proved from the corr plot in appendix (figure 06).

Interpretation of the model:

Coefficients	Estimate	Std. Error	Z value	P value
M.F	157.7023	7825.56	0.020	0.98
Age	-18.16	8568.87	-0.002	0.998
EDUC	-0.7371	0.4474	-1.647	0.0995
SES	0.9961	0.9349	1.065	0.2867
MMSE	-2.1511	1.6216	-1.327	0.1847
CDR	104.32	23208.0337	0.004	0.9964
nWBV	-118.337	75.71	-1.563	0.1181
Null Deviance: 342.68 on 252 degrees of freedom				
Residual deviance: 8.74				
AIC: 24.74				

Deviance Residuals: They measure the goodness of fit of the model. In this model, they suggest that model has some discrepancies with the observed values as they range from -1.8 to 1.4.
Coefficients: The negative coefficient of age suggests that as age increases, the odds of having Alzheimer's disease decrease. High coefficient of education also leads to increase odds of having Alzheimer. The CDR variable indicates the highest coefficient value.
The significance of the coefficients is assessed based on the p-values. The intercept term is not significant (p = 0.9839), suggesting that the predicted log-odds of being in the "Demented" group does not differ significantly from zero when all other predictor variables are zero. None of the predictor variables are statistically significant at the conventional significance level of 0.05, except for Age which shows a borderline significance (p = 0.0995).
Null Deviance and Residual Deviance: The lower the value, the better the model can predict the value of the response variable. In this model, the null deviance is 342.6858, and the residual deviance is 8.7465, indicating an improvement in model fit.
AIC Value: It relatively shows the quality of the model, and the lowers AIC proves that our model is representing well.

3) Feature Selection Method

In this section, we performed three methods for feature selection in our dataset:

i. Forward Selection:

According to this method, the variables which are more important are analyzed with stepwise regression analysis using the "**forward**" method. The analysis starts with a null model (Group ~ 1) and then proceeds to add predictors one by one based on their **AIC (Akaike Information Criterion)** values. The final selected model is:

Group ~ CDR + M.F + eTIV + EDUC

Interpretation of this model:

Coefficient	Estimate	St. Error	T value	P value
Intercept	$7.257e^{-1}$	1.340	5.417	$1.21e^{-07}$

CDR	1.047	3.657	28.62	$<2e^{-16}$
M.F	1.732	3.375	5.130	$5.10e^{-7}$
eTIV	-3.397	9.464	-3.589	0.000385
EDUC	-1.219	4.870	-2.504	0.012
Null Deviance: 76.12 on 316 degrees of freedom				
Residual Deviance: 17.881 on 312 degrees of freedom				
AIC: 0.17949				

The significance codes show the **p-values** for each coefficient, indicating the level of significance. In this case, all predictors except eTIV are statistically significant at the conventional significance level ($p < 0.05$).
The **AIC** of 0.17949 indicates a good fit of the model to the data.

ii. Backward Selection:

After applying the backward selection method for the selection of the important features, the variables come different from those in the forward process. The final selected model is:

Group ~ M.F + Age + EDUC + CDR + nWBV + ASF

Here's the interpretation of the coefficients:

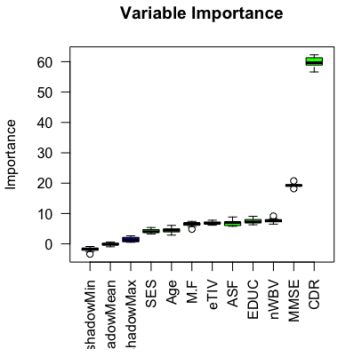
Coefficients	**Estimate**	**Std Error**	**T value**	**P value**
Intercept	0.5464	0.484	1.129	0.2599
M.F	0.1620	0.0336	4.822	$2.24e^{-06}$
Age	-0.0031	0.002	-1.524	0.128
EDUC	-0.013	0.004	-2.696	0.0074
CDR	1.020	0.039	25.60	$<2e^{-16}$
nWBV	-0.8135	0.463	-1.755	0.080
ASF	0.44872	0.120	3.722	0.00023
Null Deviance: 76.12 on 316 degrees of freedom				
Residual Deviance: 17.67 on 310 degrees of freedom				
AIC: 0.4143				

The significance codes show the p-values for each coefficient, indicating the level of significance. In this case, M.F, EDUC, CDR, and ASF are statistically significant predictors ($p < 0.05$).
The AIC of 0.41434 indicates a good fit of the model to the data.

iii. Boruta Package for Feature Selection:

The Boruta package was used for feature selection in this analysis. It iteratively identifies important attributes by comparing their importance with that of randomized

shadow attributes. After 10 iterations, the Boruta algorithm confirmed 9 attributes as important: M.F, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, and ASF. In this plot, all the variables are of the green color which proves that all variables seem to be important.



DISCUSSION AND CONCLUSION:

In this report, our main objective was to find the relationship between the predictor variables which leads to the development of the Alzheimer's disease.

First, we analyze the data in a descriptive way (numerically and graphically). We get to know the structure and variability in our dataset.

Secondly, clustering analysis is performed using the k-means algorithm to identify natural groupings or clusters within the dataset. Hierarchical clustering is also employed to assess similarities and groupings within the dataset. The resulting dendrograms help visualize the hierarchical relationships between observations.

Thirdly, a logistic regression model was built to predict the presence or absence of the condition based on the given predictors. Our model provides the accuracy of 98% which indicates that the variables which are most important in developing Alzheimer's are M.F, Age, EDUC, SES, MMSE, CDR, and nWBV.

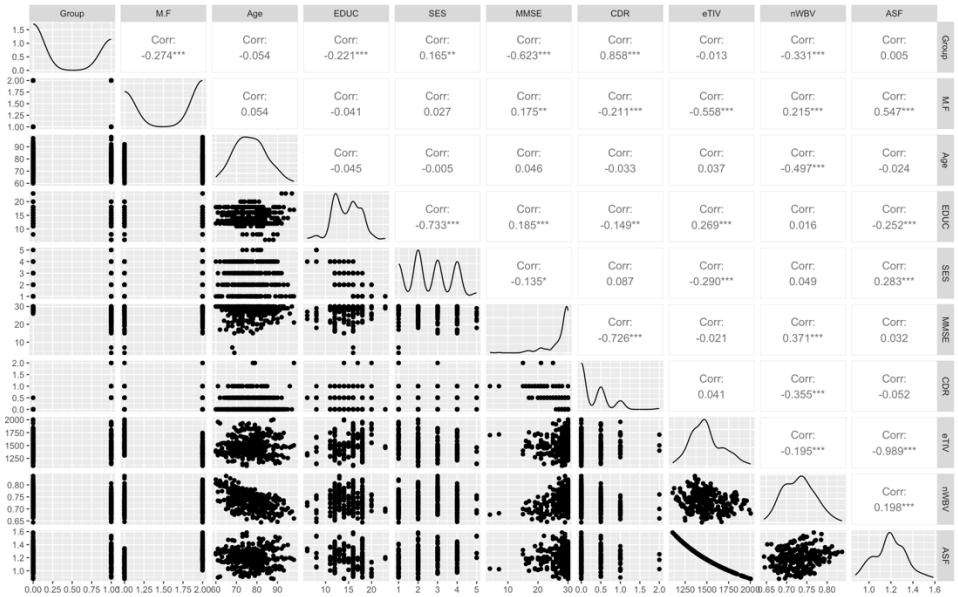
In addition to all this analysis, feature selection methods are also used to see the important variables e.g., forward, backward and Boruta methods. All these methods showed different variables which leads to Alzheimer's.

In **conclusion,** this study demonstrated the ability to accurately predict the presence or absence of the medical condition based on a set of significant predictors and provide valuable insights for clinical decision-making and risk assessment.

Appendix

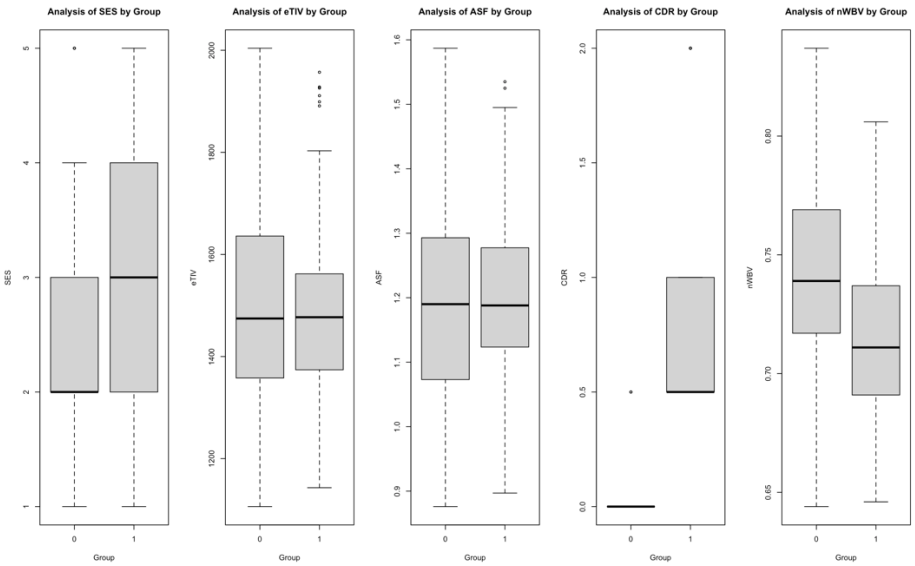
```
#Load the necessary libraries
library(dplyr)
library(magrittr)
library(ggplot2)
library(GGally)
library(reshape2)
#Set the working directory
setwd("/Users/zeeshan 1/Desktop/MA335MOD")
#Load the dataset
data <- read.csv("project data.csv")
# Convert M/F to numeric values
data$M.F <- ifelse(data$M.F == "M", 1, 2)
# Remove rows with Group = "Converted" and missing values
data <- data %>% filter(Group != "Converted" & !is.na(Group))
data <- na.omit(data)  # Remove rows with missing values
# Convert Group to binary format (0: Nondemented, 1: Demented)
data$Group <- ifelse(data$Group == "Nondemented", 0, 1)
#Structure of the dataset
str(data)
# Create an appropriate table as summary and appropriate graphs
summary_table<-summary(data)
summary_table
#Analysis using the ggpairs
ggpairs(data)
```

Figure 1



```
# Pie chart for Group variable
group_counts <- table(data$Group)
group_labels <- c("Nondemented", "Demented")
pie(group_counts, labels = group_labels, main = "Pie Chart: Group")
print(group_counts)
#Boxplot
boxplot(SES ~ Group, data = data, main = "Analysis of SES by Group", xlab = "Group", ylab = "SES")
boxplot(eTIV ~ Group, data = data, main = "Analysis of eTIV by Group", xlab = "Group", ylab = "SES")
boxplot(ASF ~ Group, data = data, main = "Analysis of ASF by Group", xlab = "Group", ylab = "SES")
boxplot(CDR ~ Group, data = data, main = "Analysis of CDR by Group", xlab = "Group", ylab = "SES")
boxplot(nWBV ~ Group, data = data, main = "Analysis of nWBV by Group", xlab = "Group", ylab = "SES")
# Select the variables to include in the boxplots
variables <- c("SES", "eTIV", "ASF","CDR","nWBV")
# Create a new figure
par(mfrow = c(1, length(variables)))
# Generate boxplots for each variable
for (variable in variables) {
  boxplot(data[[variable]] ~ data$Group, data = data,
          main = paste("Analysis of", variable, "by Group"),
          xlab = "Group", ylab = variable)
}
```

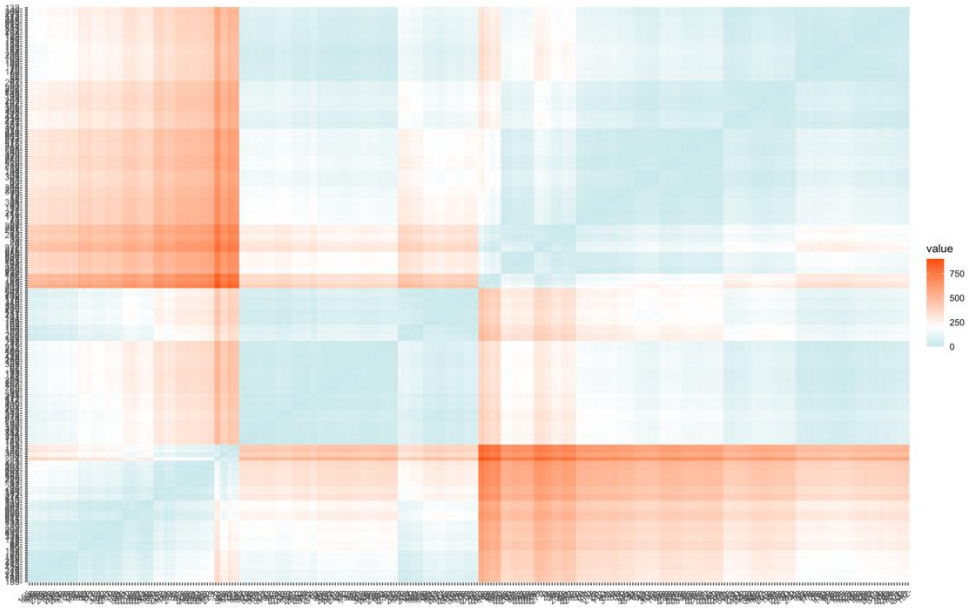
Figure 02



```
# Histogram
hist(data$EDUC, main = "Distribution of Year of Education", xlab = "Year of Education")
hist(data$MMSE, main = "Distribution of MMSE Scores", xlab = "MMSE Score")
#Gender distribution by groups
ggplot(data, aes(x = factor(Group), fill = factor(M.F))) +
  geom_bar(position = "fill") +
  labs(title = "Gender Distribution by Group", x = "Group", y = "Proportion") +
  scale_fill_manual(values = c("blue", "pink"),
                    labels = c("Male", "Female"),
                    breaks = c("1", "2"),
                    name = "Gender") +
  guides(fill = guide_legend(title = "Group")) +
  theme(legend.position = "bottom")

#Implementing clustering algorithms
library(factoextra)
#Calculating the Euclidean distance matrix
distance.Euclidean <- get_dist(data)
#Visualize the matrix using color scheme
fviz_dist(distance.Euclidean, gradient = list(low = "#00AFBB", mid = "white", high =
"#FC4E07"))
```

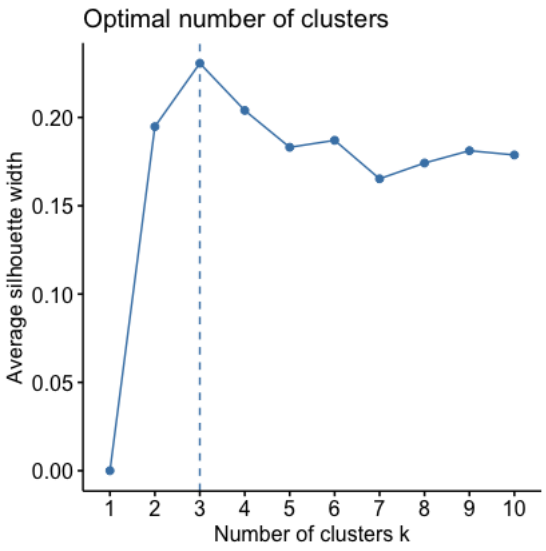
Figure 03




```
# Select variables for clustering
clustering_data <- data[, c("Age", "EDUC", "SES", "MMSE", "CDR", "eTIV", "nWBV",
"ASF")]
# Standardize the variables
clustering_data <- scale(clustering_data)
# Remove rows with missing values
```

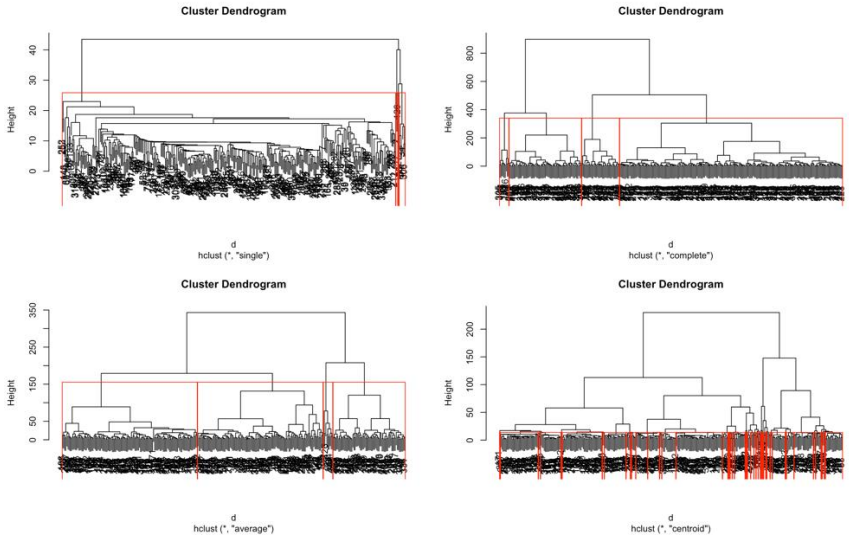
```
clean_data <- na.omit(clustering_data)
set.seed(123)
#Perform k-means clustering with k=2, 3, 4, and 5
kmeans2 <- kmeans(clean_data, centers = 2, nstart = 20)
kmeans3 <- kmeans(clean_data, centers = 3, nstart = 20)
kmeans4 <- kmeans(clean_data, centers = 4, nstart = 20)
kmeans5 <- kmeans(clean_data, centers = 5, nstart = 20)
kmeans2
str(kmeans2)
#Individual plots for each cluster
fviz_cluster(kmeans2, data = clean_data)
fviz_cluster(kmeans3, data = clean_data)
fviz_cluster(kmeans4, data = clean_data)
fviz_cluster(kmeans5, data = clean_data)
f1 <- fviz_cluster(kmeans2, geom = "point", data = clean_data) + ggtitle("k = 2")
f2 <- fviz_cluster(kmeans3, geom = "point", data = clean_data) + ggtitle("k = 3")
f3 <- fviz_cluster(kmeans4, geom = "point", data = clean_data) + ggtitle("k = 4")
f4 <- fviz_cluster(kmeans5, geom = "point", data = clean_data) + ggtitle("k = 5")
library(gridExtra)
#All the plots in grid layout
grid.arrange(f1, f2, f3,f4, nrow = 2)
#Determining the optimal number of clusters using within-cluster sum of squares (wss)
fviz_nbclust(clean_data, kmeans, method = "wss")+
  geom_vline(xintercept = 3, linetype = 2)
library(cluster)
fviz_nbclust(clean_data, kmeans, method = "silhouette")#optimal number of clusters by
silhouette coefficient
```

Figure 04



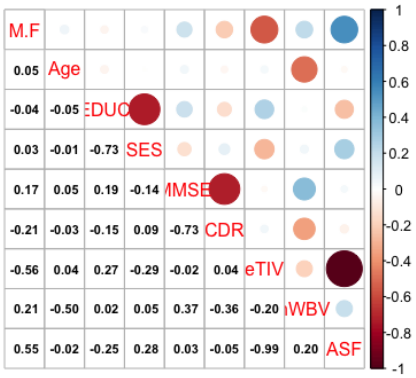
```
#Hierarchial Clustering
#Start calculating the distance matrix
d <- dist(data, method = "euclidean")
#Apply hierarchical clustering for different linkage methods
fit.single <- hclust(d, method="single")
fit.complete <- hclust(d, method="complete")
fit.average <- hclust(d, method="average")
fit.centroid <-hclust(d, method="centroid")
par(mfrow=c(2,2))
plot(fit.single) # print the dendrogram
groups.fit.single <- cutree(fit.single, k=4) # cut tree into k=4 clusters
# draw dendrogram with red borders around the 4 clusters
rect.hclust(fit.single, k=4, border="red")
#Checking how many observations are in each cluster
table(groups.fit.single)
#Dendogram for complete method
plot(fit.complete)
groups.fit.complete <- cutree(fit.complete, k=4)
rect.hclust(fit.complete, k=4, border="red")
#Dendogram for average linkage method
table(groups.fit.complete)
plot(fit.average)
groups.fit.average <- cutree(fit.average, k=4)
rect.hclust(fit.average, k=4, border="red")
#Dendogram for centroid linkage method
table(groups.fit.average)
plot(fit.centroid)
groups.fit.centroid <- cutree(fit.centroid, k=4)
rect.hclust(fit.centroid, k=46, border="red")
table(groups.fit.centroid)
#Calculating the mean of each variable by clusters for centroid:
aggregate(data, by=list(cluster=groups.fit.centroid), mean)
```

Figure 05



#Logistic Regreesion Analysis
library(faraway)
library(corrplot)
#Checking the multicollinearity
X<-data[,2:10]
cor_matrix <- cor(data[, 2:10])
print(cor_matrix)
corrplot.mixed(cor_matrix, lower.col = "black", number.cex = .7)

Figure 06



#VIF values checking
vif(X)
# Create a training and test set
set.seed(123)
index <- sample(nrow(data), nrow(data) * 0.8)
train <- data[index,]
test <- data[-index,]
# Fit the model
model <- glm(Group ~ M.F + Age + EDUC + SES + MMSE + CDR + nWBV, data = train, family = binomial)
# Assess the model
summary(model)
# Check the accuracy of the model
predictions <- predict(model, test, type = "response")
actual <- test$Group
# Calculate the accuracy
accuracy <- mean(predictions >= 0.5 & actual == 1) + mean(predictions < 0.5 & actual == 0)
# Print the accuracy
print(accuracy)
# Predicting probabilities using the logistic regression model
glm.probs <- predict(model, type = "response") # Pr(Y=1|X)

```
# Converting probabilities to class labels
glm.predicted <- rep("Nondemented", length(data$Group))
glm.predicted[glm.probs > 0.5] <- "Demented"
# Creating a confusion matrix
confusion_matrix <- table(glm.predicted, data$Group)
# Displaying the confusion matrix and accuracy
confusion_matrix

#Feature Selection Methods
#Forward Selection
x<-data[,2:10]
y<-data[,1]
model1<-glm(Group~ 1, data= data)
step1<-step(model1, scope= formula(model2),method='forward')
summary(step1)
#Using Backward method
model2<-glm(y~.,data=X)
summary(model2)
step2<-step(model2,method="backward")
summary(step2)
#Feature selection through using boruta package
library(Boruta)
boruta1 <- Boruta(Group ~., data=data, doTrace=1)
decision<-boruta1$finalDecision
signif <- decision[boruta1$finalDecision %in% c("Confirmed")]
print(signif)
plot(boruta1, xlab="", main="Variable Importance", las=2)
#Attribute statistics from the Boruta object
attStats(boruta1)
print(boruta1)
```