MA335 – FINAL PROJECT

Modelling experimental and observational data

Amin Nizar Ali - 2213409

Submitted on: 21st June 2023

Ali, Amin N
aa22824@essex.ac.uk

Word Count Total	3192
Word Count excluding Appendix, Cover page and Content Table (Page 3 to 8)	1502

Abstract

In this project, the aim is to analyze a dataset on Alzheimer's disease and investigate the relationship between various characteristics and the diagnosis of Alzheimer's. The tasks involve conducting descriptive statistics, including graphical and numerical representations, implementing clustering algorithms, fitting a logistic regression model, and performing feature selection. The analysis will provide insights into the dataset, identify important features, and contribute to understanding the relationship between the variables and the diagnosis of Alzheimer's disease.

Table of Contents

Contents

Introduction:

In this assignment I, as a data science consultant, was tasked with conducting an analysis on a comprehensive dataset that focuses on the various characteristics associated with Alzheimer's disease. The primary objective of this investigation was to explore the relationship between these characteristics and the diagnosis of Alzheimer's (referred to as "Demented") or the absence of the disease (referred to as "Nondemented"). The dataset encompasses a range of variables, including demographic factors such as gender and age, as well as cognitive measures like the Mini Mental State Examination (MMSE) and Clinical Dementia Rating (CDR). Additionally, it incorporates structural brain imaging metrics such as estimated total intracranial volume (eTIV), normalized whole brain volume (nWBV), and atlas scaling factor (ASF). By thoroughly analyzing these variables, I aimed to gain valuable insights into the potential associations between the identified characteristics and the diagnosis of Alzheimer's disease.

The details of the data provided is:

No.	NAME OF VARIABLE	DESCRIPTION
1	Group	Group of the diagnosis (Nondemented, Demented, Other)
2	M.F	Gender
3	Age	Age
4	EDUC	Year of education
5	SES	Socioeconomic Status (1-5, 1-low, 5-high)
6	MMSE	Mini mental state examination
7	CDR	Clinical dementia rating
8	eTIV	Estimated total intracranial volume
9	nWBV	Normalize whole brain volume
10	ASF	Atlas scaling factor

Preliminary Analysis

To begin the analysis, I first installed and imported the necessary libraries required for the entire study. These libraries included **corrplot** for correlation plots, **caret** for data modeling and machine learning, **MASS** for various statistical functions, **factoextra** for data visualization and clustering, **Boruta** for feature selection, **flexdashboard** for creating interactive dashboards, **ggplot2** for data visualization, **plotly** for interactive plots,

broom for tidy model outputs, **tidyverse** for data manipulation and visualization, and **dplyr** for data manipulation.

Next, I read the provided CSV file into a dataframe, which contained a total of 373 records. I conducted an initial analysis of the data by examining the file's contents. To gain a comprehensive understanding of the dataset, I utilized the **summary** function to obtain insights into various statistical measures, including central tendencies and distributions of the variables.

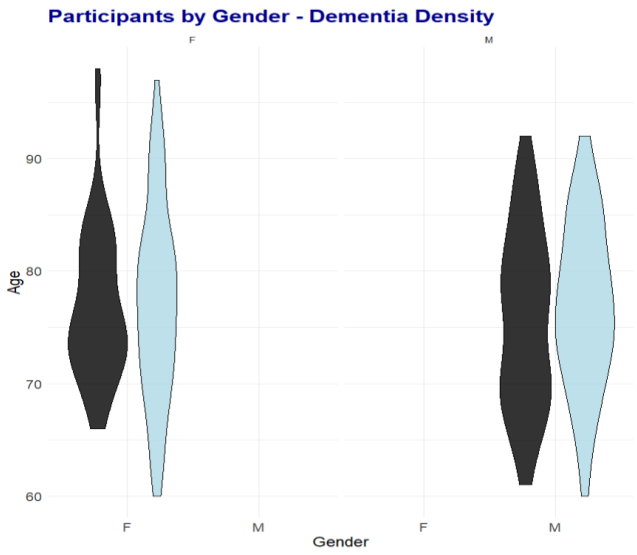
By leveraging these libraries and performing preliminary data exploration, I laid the foundation for further analysis and visualization of the dataset. This ensured that the subsequent steps in the project would be supported by the necessary tools and a solid understanding of the data's characteristics.

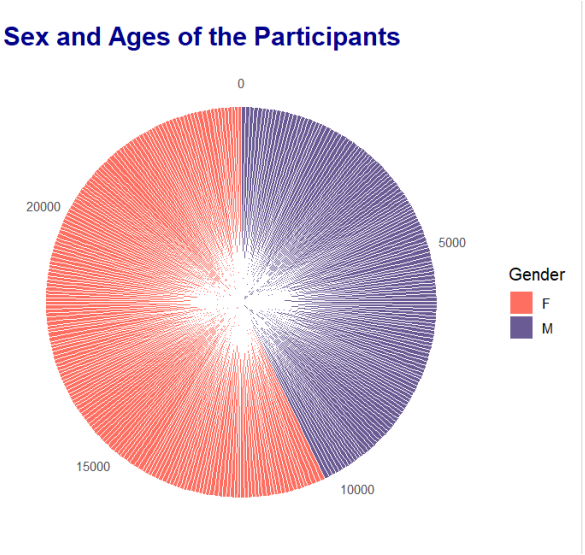
Out of the 10 variables in the dataset, 2 were initially character type. I converted them to factors for analysis. The remaining 8 variables were numeric. After identifying and removing 21 null values, the dataset was cleansed to ensure accurate analysis. Additional data cleaning and preprocessing steps were performed, including handling missing values, addressing outliers, and standardizing variables. These steps guarantee the quality and reliability of the analysis. The resulting sample dataset serves as a reference, containing the necessary variables and transformations for meaningful insights.

Analysis:

Visualizations:

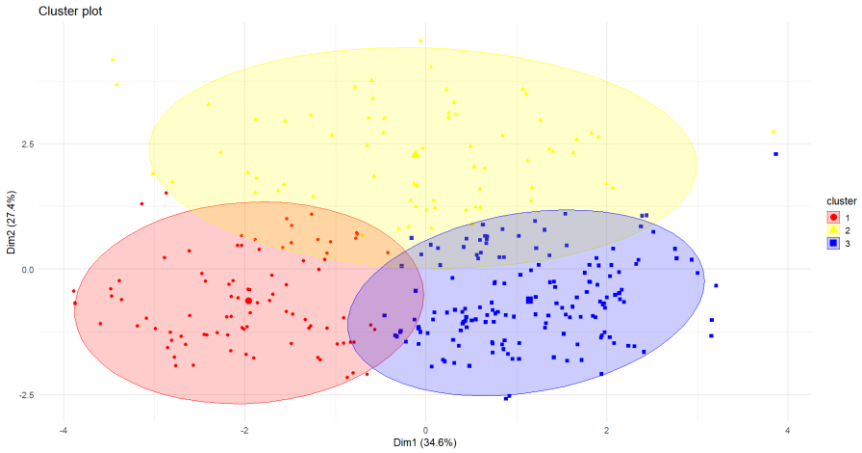
Upon further data analysis and visualization, several key findings were uncovered. Firstly, there were 180 female participants, with 51 of them being diagnosed with dementia, while there were 137 male participants, with 76 of them being diagnosed with dementia. This indicates a higher likelihood of males being affected by dementia. The average age for males was 77 years, and for females, it was 76 years, suggesting that age alone may not be a significant factor contributing to the observed differences in dementia rates between genders. When examining socio-economic status, the data revealed a wider spread among males compared to females, with females predominantly falling within the range of 2 to 4 on the socio-economic scale, while males spanned the range of 1 to 4. Furthermore, among females, there was a positive correlation between education and dementia, whereas this association was not observed among males, indicating a potentially different relationship between education and dementia based on gender. Two visualizations are shared below, rest are attached in the end in the appendix after the code.





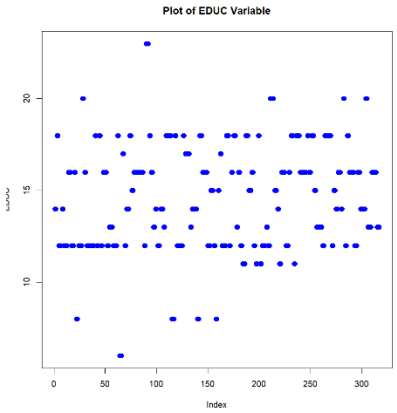
Clustering Algorithms

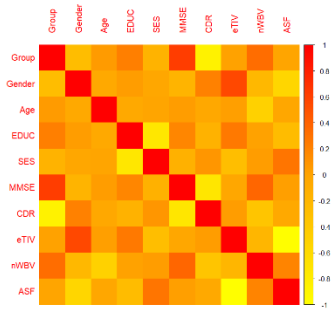
Using clustering algorithms, I grouped similar items based on their features, simplifying the data to 8 variables and adjusting their scaling. By measuring distances between data points and their centers, I formed meaningful clusters and identified correlations between variables. The correlation plot revealed weak to strong relationships, with the strongest positive correlation found between nWBV and MMSE. Age showed the weakest relationship, with the most significant correlation observed with nWBV. Euclidean distances between data points further confirmed these findings, providing insights into the data's underlying patterns.



Logistic Regression Model

Employing logistic regression, I predicted the "Group" variable and explored the relationship between predictors and outcomes. Using the Boruta method, I assessed variable relevance by creating shadow attributes. Examining correlations, I found non-colinearity between SES and CDR, indicating their independent contribution, while nWBV exhibited negative colinearity. Slight variations in Age, MMSE, eTIV, and nWBV suggested potential associations.

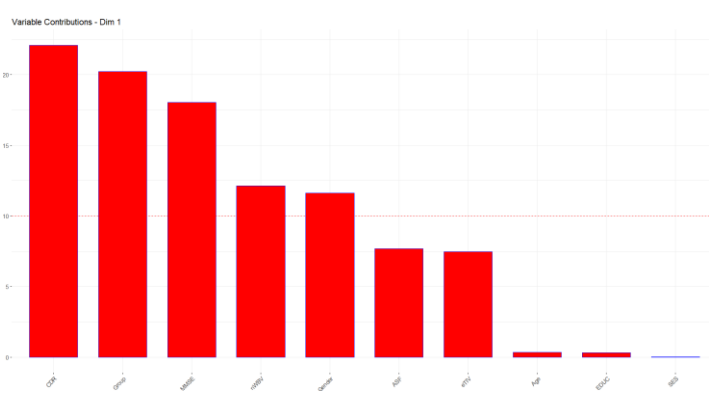


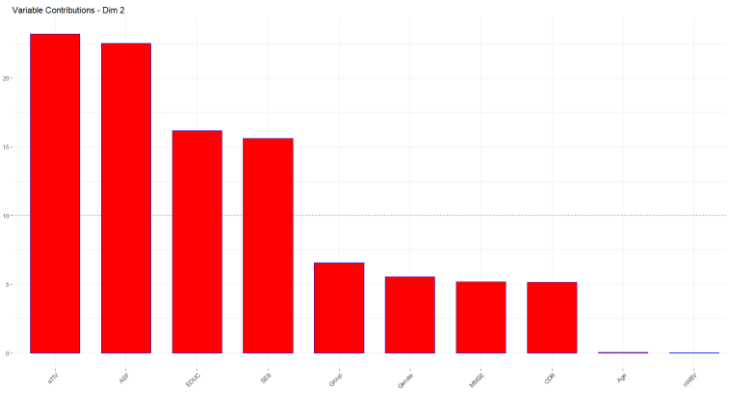


Logistic regression proved effective for binary classification, providing insights into the factors influencing group classification. By training on 80% of the data, I achieved robust model learning and evaluation, shedding light on the patterns and dynamics of Alzheimer's disease prediction.

Feature Selection Method

Implementing Principal Component Analysis (PCA), I optimized my analysis by reducing data dimensionality while retaining essential information. By rotating the data, I identified the most significant components and obtained eigenvalues. Creating a shadow dataset, I ranked variables based on prediction accuracy. PC1 accounted for 30.8% of variance, highlighting its critical role. PC2 contributed 26.8%, capturing important patterns. PCA provided insights into important components and their relevance in optimizing predictions.





Discussion

Based on visualizations and analyses, several insights emerge. Firstly, there is a higher prevalence of dementia among males, despite a smaller sample size, suggesting that gender may play a significant role in the disease's development. Secondly, the similar average ages of males and females with dementia indicate that age alone may not be the sole determinant of dementia occurrence. Other factors beyond age, such as genetics, lifestyle, or environment, could contribute to the disease. Additionally, the wider spread of socio-economic classes among males suggests that socio-economic status may have a more diverse impact on dementia risk for males compared to females. Furthermore, the positive correlation between education and dementia observed in females indicates that higher education levels may act as a protective factor against dementia in women. These insights underscore the complex nature of Alzheimer's disease and highlight the importance of considering multiple factors, including gender, age, socio-economic status, and education, in understanding and addressing the disease effectively.

Moving on to the clustering results, they reveal meaningful patterns and relationships within the data. The correlations between variables provide insights into the interdependencies among different factors. Notably, the strong positive correlation between normalized whole brain volume (nWBV) and Mini Mental State Examination (MMSE) suggests that as the brain volume increases, cognitive performance tends to improve. Additionally, the negative correlation between the atlas scaling factor (ASF) and the estimated total intracranial volume (eTIV) highlights an inverse relationship. These findings shed light on potential connections between brain structure, cognitive abilities, and age. Overall, clustering has allowed us to uncover valuable insights and understand the underlying patterns in the data.

Moving forward, the logistic regression analysis and Boruta method provide valuable insights into the relevance of predictors for classifying individuals into "Demented" or "Non-demented" groups. Variables such as socio-economic status (SES) and Clinical Dementia Rating (CDR) demonstrate independent contributions, while nWBV exhibits negative correlations. Logistic regression proves to be an effective approach for this binary prediction task, offering a comprehensive understanding of Alzheimer's disease classification and revealing important patterns and dynamics.

Lastly, insights from the Principal Component Analysis (PCA) reveal that PC1 and PC2 hold significant information about the underlying patterns in the data, contributing 30.8% and 26.8% to the total variance, respectively. These principal components provide valuable insights into the variables that strongly influence the diagnosis of Alzheimer's disease. Leveraging this knowledge enhances our predictive models and deepens our understanding of the factors at play in Alzheimer's disease diagnosis.

Conclusion

In conclusion, the insights gained from visualizations, clustering analysis, logistic regression, and Principal Component Analysis (PCA) provide a comprehensive understanding of the complexities surrounding Alzheimer's disease. Gender is a significant factor, with higher dementia prevalence among males, emphasizing the need for gender-specific research. Age alone does not determine dementia occurrence, necessitating consideration of genetics, lifestyle, and environment. Socio-economic status exhibits diverse impacts on dementia risk between genders, requiring tailored approaches. Clustering analysis reveals meaningful patterns, with strong correlations between brain structure and cognitive performance. Logistic regression and Boruta identify variables contributing to dementia classification, while PCA highlights influential components. These insights guide the development of accurate predictive models and deepen our understanding of the disease.

By integrating these findings into future research and interventions, we can make strides in early detection, prevention, and treatment of Alzheimer's. It is crucial to consider gender, age, socio-economic status, education, brain structure, and cognitive abilities to improve outcomes for individuals affected by this debilitating condition. Together, these insights offer a promising pathway for advancements in Alzheimer's research, ultimately enhancing the lives of those impacted by the disease.

Appendix

Code:

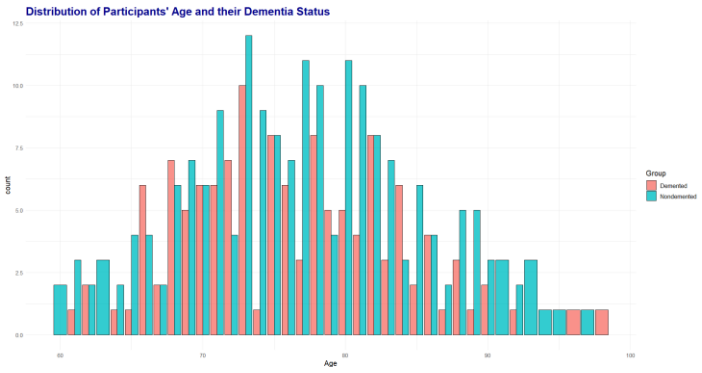
```
# Load necessary libraries
library(corrplot)
library(caret)
library(MASS)
library(factoextra)
library(Boruta)
library(flexdashboard)
library(ggplot2)
library(plotly)
library(broom)
library(tidyverse)
library(dplyr)
library(viridis)
# Read the CSV data into a dataframe
dataframe <- read.csv("C:/Personal Data/Essex/Term 2/Modelling/Final Project/project data.csv", header = TRUE)
# Display summary statistics of the dataframe
summary(dataframe)
# Display the column names of the dataframe
dataframe_names <- names(dataframe)
dataframe_names
# Display the structure of the dataframe
dataframe_structure <- str(dataframe)
dataframe_structure
# Count the number of NULLS/NAs values in every column
missing_values <- sum(is.na(dataframe))
missing_values
# Create a new dataframe by removing rows with missing values
df_new <- na.omit(dataframe)
# Get unique values of the "Group" column
unique_group <- unique(df_new$Group)
unique_group
# Subset the dataframe by removing rows with "Converted" in the "Group" column
DATA_AD <- subset(df_new, Group != "Converted")
# Convert the dataframe to a data frame and convert the "Gender" column to factor
DATA_AD <- as.data.frame(unclass(DATA_AD), stringsAsFactors = TRUE)
names(DATA_AD)[2] <- "Gender"
DATA_AD <- DATA_AD %>% mutate(Gender = factor(Gender, levels = c("1", "0"), labels = c("F", "M")))
# Display summary statistics of the DATA_AD dataframe
summary_data_AD <- summary(DATA_AD)
summary_data_AD
# Set up the layout for plotting multiple plots in one row
par(mfrow = c(1, 3))
# Attach the DATA_AD dataframe for easy access to column names
attach(DATA_AD)
# Plotting and analysis code goes here...
# Load necessary libraries
# Bar plot showing the distribution of participants' age and their dementia status
age_distribution <- ggplot(data = DATA_AD, aes(x = Age, colour = Group, fill = Group)) +
  geom_bar(position = 'dodge', color = "black", alpha = 0.8) +
  ggtitle(label = "Distribution of Participants' Age and their Dementia Status") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 18, face = "bold", color = "darkblue"),
    plot.subtitle = element_text(size = 14, face = "italic"),
    axis.title = element_text(size = 12),
    legend.title = element_text(size = 12),
    legend.text = element_text(size = 10)
  )
# Polar plot showing the sex and ages of the participants in the groups
sex_and_age <- ggplot(DATA_AD, aes(x = "", y = Age, fill = Gender)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y", start = 0) +
  ggtitle("Sex and Ages of the Participants") +
  xlab("") +
```

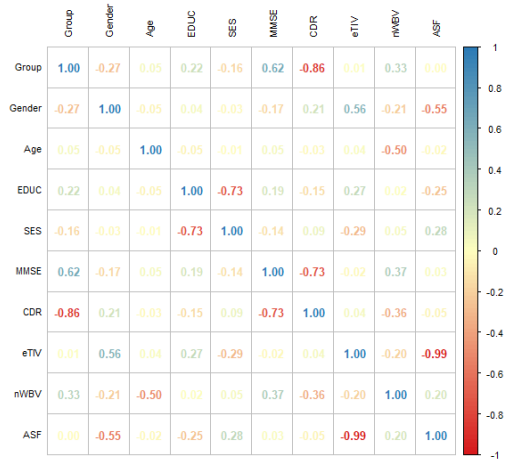
```
  ylab("") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 18, face = "bold", color = "darkblue"),
    plot.subtitle = element_text(size = 14, face = "italic"),
    axis.title = element_blank(),
    legend.title = element_text(size = 12),
    legend.text = element_text(size = 10),
    panel.grid = element_blank()  # Remove grid lines for a cleaner look
  ) +
  scale_fill_manual(values = c("#FF6F61", "#6B5B95"))  # Specify custom colors
# Boxplot showing the socio-economic status of demented participants by gender
ad_demented <- DATA_AD %>% filter(Group == "Demented") %>% group_by(Age)
demented_by_gender <- ggplot(ad_demented , aes(x = Gender, y = SES, fill = Gender)) +
  geom_boxplot(color = "black", fill = "red", outlier.color = "yellow") +
  ggtitle(label = "Demented Participants by Gender") +
  xlab("Gender") +
  ylab("Socio Economic Status") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 18, face = "bold", color = "darkblue"),
    plot.subtitle = element_text(size = 14, face = "italic"),
    axis.title = element_text(size = 12),
    legend.title = element_blank(),
    legend.text = element_blank()
  )
#############################################################################################
# Violin plot showing the distribution of participants' age by gender and dementia status
age_distribution_violin <- ggplot(data = DATA_AD, aes(x = Gender, y = Age, fill = Group)) +
  geom_violin(color = "black", alpha = 0.8) +
  scale_fill_manual(values = c("black", "lightblue")) +
  facet_wrap(~ Gender) +
  ggtitle(label = "Participants by Gender - Dementia Density") +
  xlab("Gender") +
  ylab("Age") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 18, face = "bold", color = "darkblue"),
    axis.title = element_text(size = 14),
    axis.text = element_text(size = 12),
    legend.position = "none"
  )
age_distribution_violin
#############################################################################
# Convert the first two columns of DATA_AD to numeric
Data_AD1 <- DATA_AD %>% mutate_at(c(1:2), as.numeric)
# Display summary statistics of Data_AD1
summary(Data_AD1)
# Compute the correlation matrix of Data_AD1
cor_AD <- cor(Data_AD1)
# Create a correlation plot using corrplot
corrplot(cor_AD, method = "number",
         tl.col = "black",
         tl.cex = 0.8,
         col = colorRampPalette(c("#D7191C", "#FFFFBF", "#2C7BB6"))(100))
# Compute the pairwise distance matrix of Data_AD1 using Pearson correlation
Dist_AD <- get_dist(Data_AD1, stand = TRUE, method = "pearson")
# Visualize the distance matrix using a heatmap
fviz_dist(Dist_AD, gradient = list(low = "blue", mid = "lightblue", high = "white"))
#########################################################################################
# Select the predictor variables
x <- Data_AD1[, 2:10]
# Select the target variable
y <- Data_AD1$Group
# Perform Boruta feature selection
AD_Boruta <- Boruta(y ~ ., data = Data_AD1, doTrace = 1)
# Get the final decision of the Boruta algorithm
decisions <- AD_Boruta$finalDecision
# Select the significant features
```

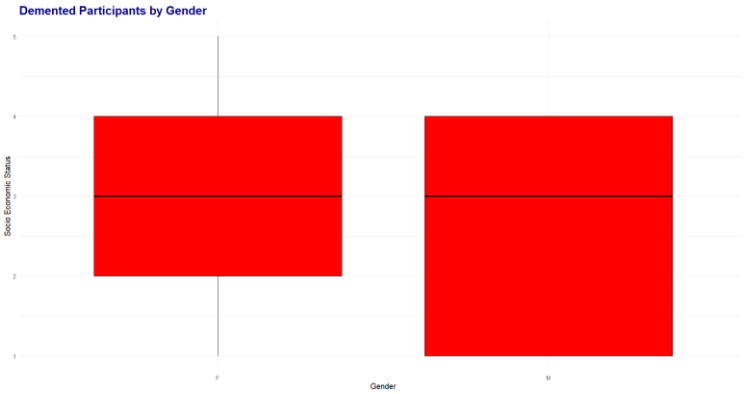
```
signif <- decisions[AD_Boruta$finalDecision %in% c("Confirmed")]
# Print the significant features
print(signif)
# Plot variable importance from Boruta
plot(AD_Boruta, xlab = " ", main = "Variable Importance")
# Print attribute statistics from Boruta
attStats(AD_Boruta)
# Set up the plotting layout
par(mfrow = c(1, 2))
# Select variables for clustering
Select_AD <- Data_AD1[, c(2, 4, 5, 6, 7, 8, 9, 10)]
# Scale the selected variables
Scale_AD <- scale(Select_AD)
# Calculate the distance matrix
Dist_AD <- get_dist(Scale_AD)
# Plot the silhouette method for determining the number of clusters
fviz_nbclust(Scale_AD, kmeans, method = "silhouette") +
  labs(subtitle = "Elbow Plot for Cluster Determination") +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 18, face = "bold"),
    plot.subtitle = element_text(size = 14, face = "italic"),
    axis.title = element_text(size = 12),
    axis.text = element_text(size = 10),
    legend.position = "none"
  ) +
  scale_fill_manual(values = c("lightblue", "lightgreen", "pink", "orange", "purple"))
#####################################################################
# Perform k-means clustering on the scaled data
AD_Km <- kmeans(Scale_AD, center = 3, nstart = 35)
# Print the sizes of each cluster
AD_Km$size
# Get the cluster assignments for each observation
Clusters_AD <- AD_Km$cluster
# Assign row names to Scale_AD using the combination of Group and index
rownames(Scale_AD) <- paste(DATA_AD$Group, 1:dim(Select_AD)[1], sep = "-")

# Visualize the clustering results
fviz_cluster(
  AD_Km, data = Select_AD, palette = c("red", "yellow", "blue"),
  ellipse.type = "t", geom = "point", pointsize = 2
) +
  theme_minimal(base_size = 14)
# Create contingency tables to examine cluster-group and cluster-gender relationships
table(Clusters_AD, DATA_AD$Group)
table(Clusters_AD, DATA_AD$Gender)
# Print the dimensions of the Select_AD data
dim(Select_AD)
# Create a pairs plot to visualize pairwise relationships between variables
pairs(Select_AD)
# Compute the correlation matrix of the Select_AD data
cor(Select_AD)
# Create a scatter plot of the EDUC variable
plot(
  Select_AD$EDUC, col = "blue", pch = 16, cex = 1.2,
  xlab = "Index", ylab = "EDUC",
  main = "Plot of EDUC Variable"
) +
  abline(1, 0, col = "red", lwd = 0.3)
#############################################################################
# Set the random seed for reproducibility
set.seed(123)
# Set the levels of Gender variable
levels(Gender) <- c(0, 1)
# Set the levels of Group variable
levels(Group) <- c(0, 1)
# Create a training set using 80% of the data
Training_AD <- Data_AD1$Group %>% createDataPartition(p = 0.8, list = FALSE)
# Split the data into training and test sets
```

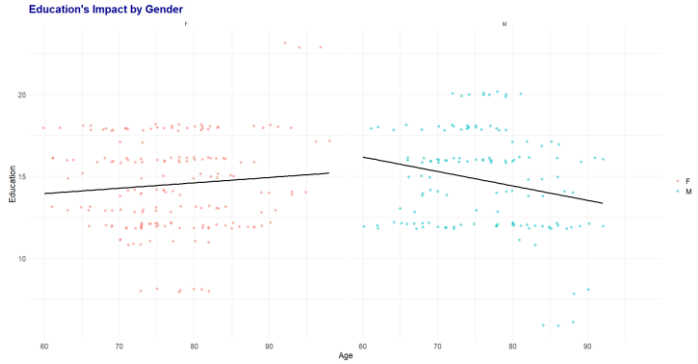
```
AD.TRAIN <- DATA_AD[Training_AD, ]
AD.TEST <- DATA_AD[-Training_AD, ]
# Build a logistic regression model using stepwise AIC
Model_AD <- glm(Group ~ ., data = AD.TRAIN, family = "binomial") %>% stepAIC(trace = FALSE)
# Display the summary of the logistic regression model
summary(Model_AD)
# Get the coefficients of the logistic regression model
coef(Model_AD)
# Predict probabilities using the logistic regression model on the training set
Probs_TRAIN <- Model_AD %>% predict(AD.TRAIN, type = "response")
# Predict probabilities using the logistic regression model on the test set
Probs_TEST <- Model_AD %>% predict(AD.TEST, type = "response")
# Assign the predicted group based on the probability threshold of 0.5
Grp_Predicted <- ifelse(Probs_TEST > 0.5, "up", "down")
Grp_Predicted
#########################################################################
# Calculate the column means of Data_AD1
apply(Data_AD1, 2, mean)
# Calculate the column variances of Data_AD1
apply(Data_AD1, 2, var)
# Visualize the correlation matrix using the corrplot function
corrplot(cor(Data_AD1), method = "color", col = colorRampPalette(c("yellow", "orange", "red"))(200))
# Perform Principal Component Analysis (PCA) on Data_AD1
PCA_1 <- prcomp(Data_AD1, scale = TRUE)
# Display the summary of the PCA results
summary(PCA_1)
# Print the PCA results with 2 decimal places
print(PCA_1, digit = 2)
# Get the eigen vectors (rotation) from the PCA results
eigen_vectors <- PCA_1$rotation
# Get the PCA scores
PCA_1.scores <- PCA_1$x
# Get the names of the components in PCA_1
names(PCA_1)
# Calculate the eigenvalues (variances) from the PCA results
Values_of_Eigen <- PCA_1$sdev^2
# Calculate the proportion of variance explained by each component
Prop_Var_Expl <- Values_of_Eigen / sum(Values_of_Eigen)
Prop_Var_Expl
# Calculate the cumulative proportion of variance explained
cumsum(Prop_Var_Expl)
# Plot the PCA results
plot(PCA_1, type = "l", col = "red", lwd = 4, cex.axis = 1.2, cex.lab = 1.5)
#####################################################################
# Plot the scree plot of the PCA results
plot(PCA_1, type = "l", col = "red", lwd = 4, cex.axis = 1.2, cex.lab = 1.5, main = "Data Screeplot of Alzheimer")
abline(1, 0, col = 'black', lwd = 2)
# Plot the percentage of variance explained by each principal component
plot(Prop_Var_Expl, xlab = "Principal Components", ylab = "Percent of Variance Explained", ylim = c(0, 1), type = "b", col = "green", pch = 16, cex.axis = 1.2, cex.lab = 1.5)
# Plot the cumulative percentage of variance explained
plot(cumsum(Prop_Var_Expl), xlab = "Principal Component", ylab = "Variance - Cumulative Percentage", ylim = c(0, 1), type = "b", col = "blue", pch = 16, cex.axis = 1.2, cex.lab = 1.5)
# Visualize the variable contributions to the first dimension
fviz_contrib(PCA_1, choice = "var", axes = 1, color = "blue", fill = "red", title = "Variable Contributions - Dim 1", title.size = 20, pointsize = 2)
# Visualize the variable contributions to the second dimension
fviz_contrib(PCA_1, choice = "var", axes = 2, color = "blue", fill = "red", title = "Variable Contributions - Dim 2", title.size = 20, pointsize = 2)
# Calculate the eigenvalues (variances) from the PCA results
Values_of_Eigen <- PCA_1$sdev^2
# Calculate the proportion of variance explained by each component
Prop_Var_Expl <- Values_of_Eigen / sum(Values_of_Eigen)
Prop_Var_Expl
# Calculate the cumulative proportion of variance explained
cumsum(Prop_Var_Expl)
# Plot the PCA results
plot(PCA_1, type = "l", col = "red", lwd = 4, cex.axis = 1.2, cex.lab = 1.5)
```

Plots:


Distribution of Participants' Age and their Dementia Status
count
Age
Group
Demented
Nondemented


Group Gender Age EDUC SES MMSE CDR eTIV nWBV ASF
Group 1.00 -0.27 0.05 0.22 -0.16 0.62 -0.86 0.01 0.33 0.00
Gender -0.27 1.00 -0.05 0.04 -0.03 -0.17 0.21 0.56 -0.21 -0.55
Age 0.05 -0.05 1.00 -0.05 -0.01 0.05 -0.03 0.04 -0.50 0.02
EDUC 0.22 0.04 -0.05 1.00 -0.73 0.19 -0.15 0.27 0.02 -0.25
SES -0.16 -0.03 -0.01 -0.73 1.00 -0.14 0.09 -0.29 0.05 0.28
MMSE 0.62 -0.17 0.05 0.19 -0.14 1.00 -0.73 0.02 0.37 0.03
CDR -0.86 0.21 -0.03 -0.15 0.09 -0.73 1.00 0.04 -0.36 -0.05
eTIV 0.01 0.56 0.04 0.27 -0.29 -0.02 0.04 1.00 -0.20 -0.99
nWBV 0.33 -0.21 -0.50 0.02 0.05 0.37 -0.36 -0.20 1.00 0.20
ASF 0.00 -0.55 0.02 -0.25 0.28 0.03 -0.05 -0.99 0.20 1.00


Demented Participants by Gender
Socio Economic Status
Gender


Education's Impact by Gender
Education
Age
F
M