**Abstract**

This report comprises of the results of the descriptive statistical analysis of a dataset of patients with Alzheimer's disease. This dataset includes information on the patients' age, gender, education, socioeconomic status, Mini Mental State Examination (MMSE) score, Clinical Dementia Rating (CDR) score, estimated total intracranial volume (eTIV), normalized whole brain volume (nWBV), and Atlas scaling factor (ASF). The analysis is structured into four different tasks. Firstly, the descriptive summary of the dataset is generated and different graphical representations such as boxplots & histograms, to understand the dataset and give insight on the relationships between variables. Secondly, the k means clustering algorithm is used to identify patterns and the groupings within the dataset. Thirdly, a logistic regression model is fitted using the other variables to predict the dependent variable, "Group" and even the logistic model’s summary provides the important information about the model and its statistical significance. Lastly, the feature selection method is implemented to identify the significant features in understanding the Alzheimer's data set.

**Contents**

1. Introduction
2. Descriptive Statistics
3. Clustering
4. Logistic Regression
5. Feature Selection
6. References
7. Appendix

**1)Introduction**

Alzheimer's disease is a progressive neurological disorder which affects millions of people worldwide. As a data science consultant, I have been tasked with analyzing a dataset that which has different characteristics related to Alzheimer's. The primary objective of this report is to explore the relationship between these characteristics and the diagnosis of the Alzheimer's disease, distinguishing between individuals classified as "Demented" and "Nondemented." By examining the dataset and employing statistical analysis techniques, we can gain valuable insights into the potential risk factors and markers that may contribute to the development of the disease.

The dataset at hand provides a comprehensive set of variables that capture different aspects related to Alzheimer's disease. These variables include the group diagnosis (categorized as "Nondemented," "Demented," or "Other"), gender, age, years of education, socioeconomic status, mini-mental state examination score, clinical dementia rating, estimated total intracranial volume, normalized whole brain volume, and atlas scaling factor. Each of these variables may hold valuable information in understanding the relationship between the characteristics and the diagnosis.

To conduct the analysis, we will utilize R, a powerful programming language for statistical computing. R provides a wide range of tools and packages that enable us to perform descriptive statistics, visualization, clustering algorithms, logistic regression modeling, and feature selection methods. By leveraging these techniques, we can uncover patterns, identify important variables, and ultimately gain a deeper understanding of the factors influencing Alzheimer's disease.

This report aims to present a thorough analysis of the dataset, showcasing the insights and findings obtained through statistical analysis. The report will adhere to the guidelines provided, ensuring proper structure, documentation of the R code used, and a clear account of assumptions made during the analysis process.

By exploring the relationship between various characteristics and the diagnosis of Alzheimer's disease, we hope to contribute to the existing knowledge base surrounding the disease. The findings may have implications for early detection, risk assessment, and potentially lead to advancements in preventive measures and therapeutic interventions.Alzheimer's disease is a neurodegenerative disorder that affects millions of individuals worldwide, causing progressive cognitive decline and memory loss. Understanding the factors associated with the development and progression of Alzheimer's is of paramount importance for early detection, accurate diagnosis, and effective treatment strategies.

The dataset contains multiple variables, which explains specific aspects relevant to Alzheimer's disease. These variables include the Group, which represents the diagnosis (Nondemented, Demented, or Other), as well as demographic and clinical factors such as gender (M/F), age, years of education (EDUC), socioeconomic status (SES), mini-mental state examination (MMSE) scores, clinical dementia rating (CDR), estimated total intracranial volume (eTIV), normalized whole brain volume (nWBV), and atlas scaling factor (ASF).By thoroughly examining and analyzing these variables, we aim to uncover meaningful patterns, associations, and potential predictors of Alzheimer's disease. Descriptive statistics, graphical representations, and advanced statistical techniques will be employed to gain comprehensive insights into the dataset and establish a solid foundation for subsequent analysis.Through this investigation, we strive to contribute to the existing body of knowledge surrounding Alzheimer's disease, enhance our understanding of its underlying characteristics, and potentially identify valuable markers for early detection and intervention. The findings of this analysis have the potential to inform healthcare professionals, researchers, and policymakers, ultimately leading to improved patient care, personalized treatment approaches, and a better understanding of this devastating neurological condition.

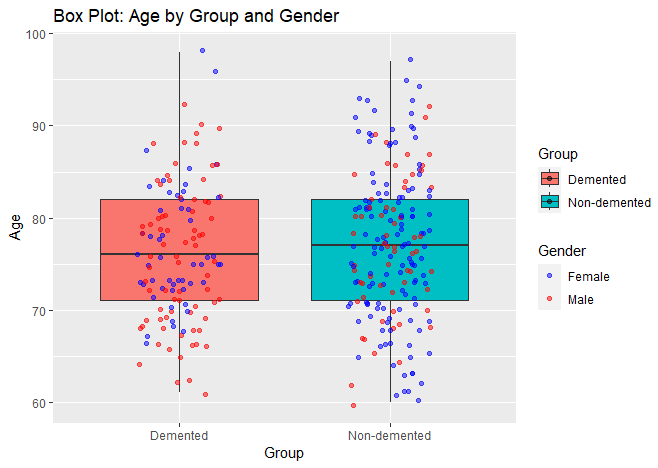
In the following sections of this report, we will present the methodology, results, and discussions pertaining to the analysis of this dataset, utilizing the powerful statistical programming language R. Through a systematic and rigorous approach, we aim to provide comprehensive insights into the relationship between the characteristics within the dataset and the diagnosis of Alzheimer's disease.

**2)Descriptive Statistics**

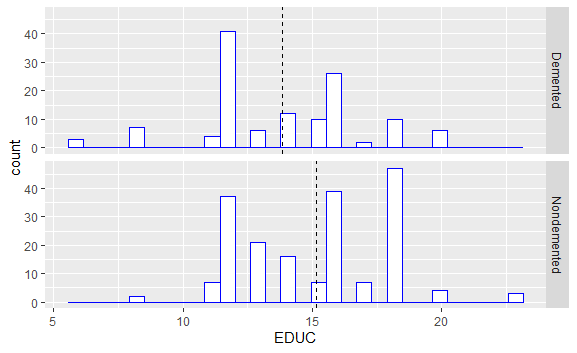
Summary Statistics of the Dataset



Group: Demented=1; Nondemented=0. M.F: M=1; F=0.

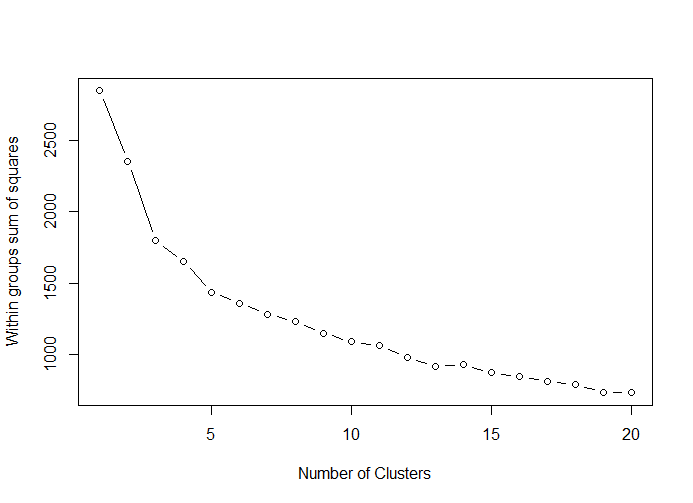
In the sample of 317 individuals, the group mean of 0.4 being closer to zero indicates larger proportion of nondemented individuals. Similarly, M.F mean of 0.43 implies larger proportion of females in the group. Individuals of 60 to 98 years are included in the group with an average age of 76 years. The number of years of education has been varying widely from 6 years to 23 years. While mean MMSE of 27.26 is below the mark of 30, it remains higher than 24, which is usually taken as the indication of possible cognitive impairment or dementia. This is also supported by the low mean CDR at 0.27. The low skewness of 0.07 for ASF suggests a roughly symmetric distribution while the kurtosis value of -0.26 implies a relatively flat distribution. 

The minimum and maximum ages of the demented group is higher than those of the nondemented group. However, the median age of the demented group is lower than the nondemented group. There are more males in the demented group than females in contrast to the nondemented group. As the number of males are less than the number of females in the dataset, this means Alzheimer seems more prevalent in gents than in ladies.



As per the above histogram, the average years of education at around 14 years for the demented group stand out lower than that at around 15 years for the nondemented group. Further, the number of demented persons peaks at around 12 years of education as against 18 years in the nondemented group.

**3) Clustering**

Towards implementing clustering algorithm, we first normalize the data by subtracting from the mean and dividing by standard deviation. We then calculate the Euclidean distance and use the following scree plot.

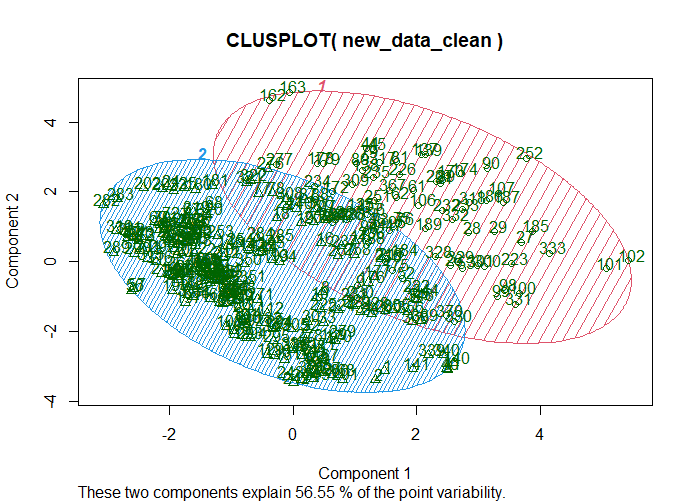
The highest drop in the within groups sum of squares happens at number of clusters: 2 and 3.

We have applied the technique of K mean clustering with 2 clusters, resulting in 90 and 227 observations in the clusters respectively.

The within cluster sum of squares is 723.6541 and 1629.7171, respectively. Ideally, it should be as small as possible. The ratio between\_SS cluster / total\_SS cluster is 17.3%. Ideally, this ratio should be as large as possible.

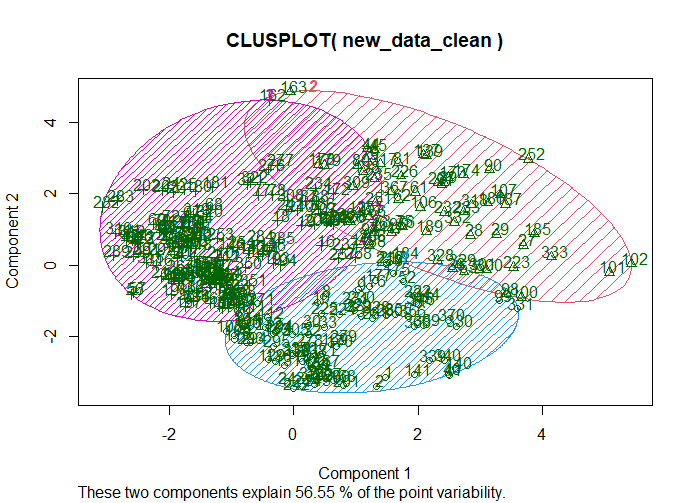
Cluster 1 is associated with higher chance of being female (on average), higher average of age, lower average education, higher average socio economic status, lower average mental state examination, higher average clinical dementia rating, lower average total intracranial volume, lower average normalize whole brain volume, higher average atlas scaling factor.

Cluster 2 is associated with lower chance of being female (on average), lower average of age, higher average education, lower average socio economic status, higher average mental state examination, lower average clinical dementia rating, higher average total intracranial volume, higher average normalize whole brain volume, lower average atlas scaling factor.



K mean clustering with 3 clusters contain 161,67 and 89 observations in each cluster.

The within cluster sum of squares is 807.1084 506.5956 and 484.9509. Ideally, the within cluster sum of squares should be as small as possible. These numbers are much smaller as compared with cluster 2. The ratio between\_SS cluster / total\_SS cluster is 36.8%. This is higher than cluster 2.



Further as we go up more clusters the within sum of square will reduce and the ratio between\_SS cluster / total\_SS cluster will increase (With cluster 4 it is 43.6%)The gain is not that high when we go from Cluster 2 to cluster 3 (It is 19%) as compared to cluster 3 to cluster 4 (It is 8%).

We should go for cluster 2 or cluster 3 using k mean clustering.

**4)Logistic Regression**

We take log of several variables to scale down the values. Further we divide the data into training and testing data set (80% training data and 20% testing data).

After several permutations and combinations, we construct the following model,

Group ~ Age + Year of education + Mini mental state examination + Normalize whole brain volume + Socio Economic Status + error

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | z value | Pr(>|z|) |
| (Intercept) | 206.17 | 32.55 | 6.33 | 0.00\*\*\* |
| Age | -13.06 | 3.03 | -4.32 | 0.00\*\*\* |
| EDUC | -3.20 | 1.62 | -1.97 | 0.05\* |
| MMSE | -35.78 | 5.77 | -6.20 | 0.00\*\*\* |
| nWBV | -29.64 | 7.53 | -3.94 | 0.00\*\*\* |
| SES2 | -1.18 | 0.63 | -1.88 | 0.06. |
| SES3 | 0.08 | 0.68 | 0.12 | 0.91 |
| SES4 | -0.59 | 0.78 | -0.75 | 0.45 |
| SES5 | -5.82 | 4.86 | -1.20 | 0.23 |

Age, EDUC, MMSE and nWBV are negative. This show that an increase in any of these variables, the log of odds of getting dementia goes down (A person likely to be nondemented). These results confirm with the graphs and summary statistics. Lower Socio Economic status is significant as compared to higher socio economic status.

The misclassification error in Training data set is 14.17%. The train data set correctly classifies 85.82%.

The misclassification error in Training data set is 14.28%. The test data set correctly classifies 85.71%.

**5)Feature Selection**

We use a forward stepwise section method to find the most important features.

First, we fit the intercept-only model. This model had an AIC of -450.23.

Next, we fit every possible one-predictor model. The model that produced the lowest AIC and also had a statistically significant reduction in AIC compared to the intercept-only model used the predictor CDR. This model had an AIC of -870.39.

Next, we fit every possible two-predictor model. The model that produced the lowest AIC and also had a statistically significant reduction in AIC compared to the single-predictor model added the predictor EDU. This model had an AIC of -880.95.

Next, we fit every possible three-predictor model. The model that produced the lowest AIC and also had a statistically significant reduction in AIC compared to the two-predictor model added the predictor M/F. This model had an AIC of -892.35.

Next, we fit every possible four-predictor model. It turned out that none of these models produced a significant reduction in AIC, thus we stopped the procedure.

The final model turns out to be:

Group ~ 4.23 + (1.05\*CDR) - (0.18\*EDU) + (0.17\*M/F) – (0.50\*eTIV )

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Step | Df | Deviance | Resid. Df | Resid. Dev | AIC |
|  | NA | NA | 316 | 76.119 | -450.22 |
| + CDR | -1 | 56.022 | 315 | 20.097 | -870.38 |
| +EDUC | -1 | 0.780 | 314 | 19.316 | -880.950 |
| +M.F | -1 | 0.799 | 313 | 18.516 | -892.35 |
| +eTIV | -1 | 0.725 | 312 | 17.791 | -903.02 |

6)References

7) Appendix

# Load the necessary libraries

library(ggplot2)

library(psych)

library(plyr)

library(ggcorrplot)

library("cluster")

myproject <- read.csv("C:/Users/Admin/OneDrive - University of Essex/MA335/Final project-20230509 2023-05-09 09\_39\_38/project data.csv", header = TRUE, sep = ",")

# Converted Male/Female into numerical value

myproject$M.F <- ifelse(myproject$M.F == "F", 0, 1)

# Remove rows with Group="Converted"

myproject <- subset(myproject, !(Group == "Converted"))

# Remove missing data

new\_data\_clean <- na.omit(myproject)

# Descriptive Statistics

# Boxplot

ggplot(new\_data\_clean, aes(x = factor(Group), y = Age, fill = Group)) +

geom\_boxplot() +

geom\_jitter(aes(color = factor(M.F)), width = 0.2, alpha = 0.5) +

labs(x = "Group", y = "Age", fill = "Group", color = "Gender") +

scale\_color\_manual(values = c("blue", "red"), labels = c("Female", "Male")) +

ggtitle("Box Plot: Age by Group and Gender")

# No. of Males

M<-subset(new\_data\_clean, M.F == 1)

nrow(M) #137

#No. of Females

F<-subset(new\_data\_clean, M.F == 0)

nrow(F) #180

nrow(M)+nrow(F)#317

#No. of Demented Males

MD<- subset(new\_data\_clean, M.F == 1 & Group == 'Demented')

nrow(MD) #76

#No. of Demented Females

FD<- subset(new\_data\_clean, M.F == 0 & Group == 'Demented')

nrow(FD) #51

#No. of NonDemented Males

MND<- subset(new\_data\_clean, M.F == 1 & Group == 'Nondemented')

nrow(MND) #61

#No. of NonDemented Females

FND<- subset(new\_data\_clean, M.F == 0 & Group == 'Nondemented')

nrow(FND) #129

nrow(MD)+nrow(FD)+nrow(MND)+nrow(FND)#317

proportionMales<-round(76/137,2) #55% Males

proportionMales

proportionfemales<-round(51/180,2) #28% Females

proportionfemales

gender <- c("Male", "Female", "Male", "Female")

demented <- c("Non-Demented", "Non-Demented", "Demented", "Demented")

count <- c(61,129,76,51)

df <- data.frame(gender, demented, count)

ggplot(df, aes(x = gender, y = count, fill = demented)) +

geom\_histogram(stat = "identity", position = "stack") +

labs(x = "Gender", y = "Count", fill = "Alzheimer") +

scale\_fill\_manual(values = c("red", "springgreen2"))

#Education vs Alzemier

nu <- ddply(new\_data\_clean, "Group", summarise, grp.mean=round(mean(EDUC),2))

nu #Demented Mean EDUC is 13.83

#NonDemented Mean EDUC is 15.14

q<-ggplot(new\_data\_clean, aes(x=EDUC))+

geom\_histogram(color="blue", fill="white")+

facet\_grid(Group ~ .)

q

# Add mean lines

q+geom\_vline(data=nu, aes(xintercept=grp.mean),linetype="dashed")

#Socio Economic Status vs Alzeimer

SEconomicStatus<-as.factor(new\_data\_clean$SES)

ggplot(new\_data\_clean) +

geom\_bar(mapping = aes(x = SES,fill=SEconomicStatus)) +

labs(x = "Socio Economic Status")

ggplot(new\_data\_clean, aes(x = factor(SES), fill = factor(Group))) +

geom\_bar(position = "fill") +

scale\_fill\_manual(values = c("red", "lightskyblue")) +

labs(x = "Socioeconomic Status", y = "Proportion", fill = "Alzheimer") +

theme\_bw()

#correlation

new\_data\_clean$Group <- ifelse(new\_data\_clean$Group == "Demented", 1, 0)

corr <- round(cor(new\_data\_clean[,1:10]), 1)

head(corr)

ggcorrplot(corr, hc.order = TRUE, type = "lower",

outline.col = "white")

describe(new\_data\_clean)

#Clustering

#Normalizing the data

z <-new\_data\_clean[,-c(1,1)]

head(z)

m<-apply(z,2,mean)

m

s<-apply(z,2,sd)

s

nor<-scale(z,center=m,scale =s)

head(nor)

# Set smaller margins

par(mar = c(5, 4, 4, 2) + 0.1, oma = c(0, 0, 0, 0))

# Scree Plot

wss <- (nrow(nor) - 1) \* sum(apply(nor, 2, var))

for (i in 2:20) wss[i] <- sum(kmeans(nor, centers = i)$withinss)

plot(1:20, wss, type = "b", xlab = "Number of Clusters", ylab = "Within groups sum of squares")

# Adjust outer margins

par(oma = c(0, 0, 2, 0))

# title

title(main = "Scree Plot")

# K-means clustering

set.seed(1234)

kc1<-kmeans(nor,2)

kc1

kc2<-kmeans(nor,3)

kc2

kc3<-kmeans(nor,4)

kc3

clusplot(new\_data\_clean,kc1$cluster,

color=T,shade= T,

labels=2, lines=0)

clusplot(new\_data\_clean,kc2$cluster,

color=T,shade= T,

labels=2, lines=0)