MA335 Final Project Report

Sourav Ghosh Hansda

Date-20/06/2023

Abstract

The report explores the relationship between different characteristics related to Alzheimer's disease and the diagnosis of individuals as "Demented" or "Nondemented." Through descriptive statistics and their analysis based on a battery of techniques, namely, cluster analysis, logistic regression and forward selection stepwise method, valuable insights are gained into potential factors contributing to the development of the disease. The dataset can be segmented into 2 or 3 clusters with some overlaps. Men, individuals with few years of education, and those with high socioeconomic status appear more vulnerable to dementia.

Contents

1) Introduction

Alzheimer is a progressive disease that destroys memory. Millions of people are afflicted with Alzheimer worldwide. As a data science consultant, I am tasked to analyse a dataset having different characteristics related to Alzheimer's. The report attempts to explore the relationship between these characteristics and the diagnosis on Alzheimer, distinguishing people as "Demented" or "Nondemented." By examining descriptive statistics and applying statistical techniques to the dataset, we can gain valuable insights into the potential factors that may contribute to the development of the disease.

The rest of the report is organized as under. Section 2 presents descriptive statistics and boxplots/histogram on the interrelationship of variables, based on the given 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). In Section 3, the k means clustering algorithm is used to identify patterns and the groupings within the dataset. In Section 4, a

logistic regression model is fitted while using the remaining observations to check the robustness of the model. In Section 5, a forward stepwise selection method is implemented to identify the significant features in understanding the Alzheimer's data set. Concluding observations are articulated in Section 6.

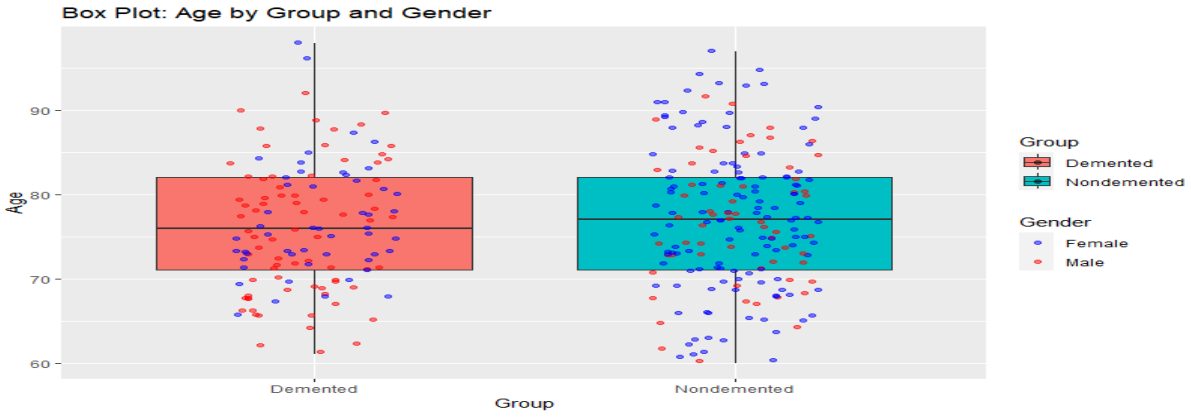
2) Descriptive Statistics

Summary Statistics of the Dataset

variables	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
M.F	317	0.43	0.50	0	0.416	0	0	1	1	0.27	-1.93	0.0279
Age	317	76.72	7.81	76	76.55	8.90	60	98	38	0.19	-0.43	0.44
EDUC	317	14.62	2.93	15	14.64	4.45	6	23	17	-0.091	0.0402	0.164
SES	317	2.55	1.12	2	2.53	1.48	1	5	4	0.153	-1.076	0.063
MMSE	317	27.26	3.86	29	28.07	1.48	4	30	26	-2.29	6.75	0.217
CDR	317	0.27	0.38	0	0.21	0	0	2	2	1.479	2.586	0.021
eTIV	317	1493.58	179.72	1476	1483.77	176.43	1106	2004	898	0.506	-0.179	10.094
nWBV	317	0.73	0.04	0.73	0.73	0.042	0.64	0.84	0.19	0.184	-0.502	0.00214
ASF	317	1.192	0.14	1.189	1.19	0.151	0.88	1.59	0.71	0.07	-0.260	0.00784

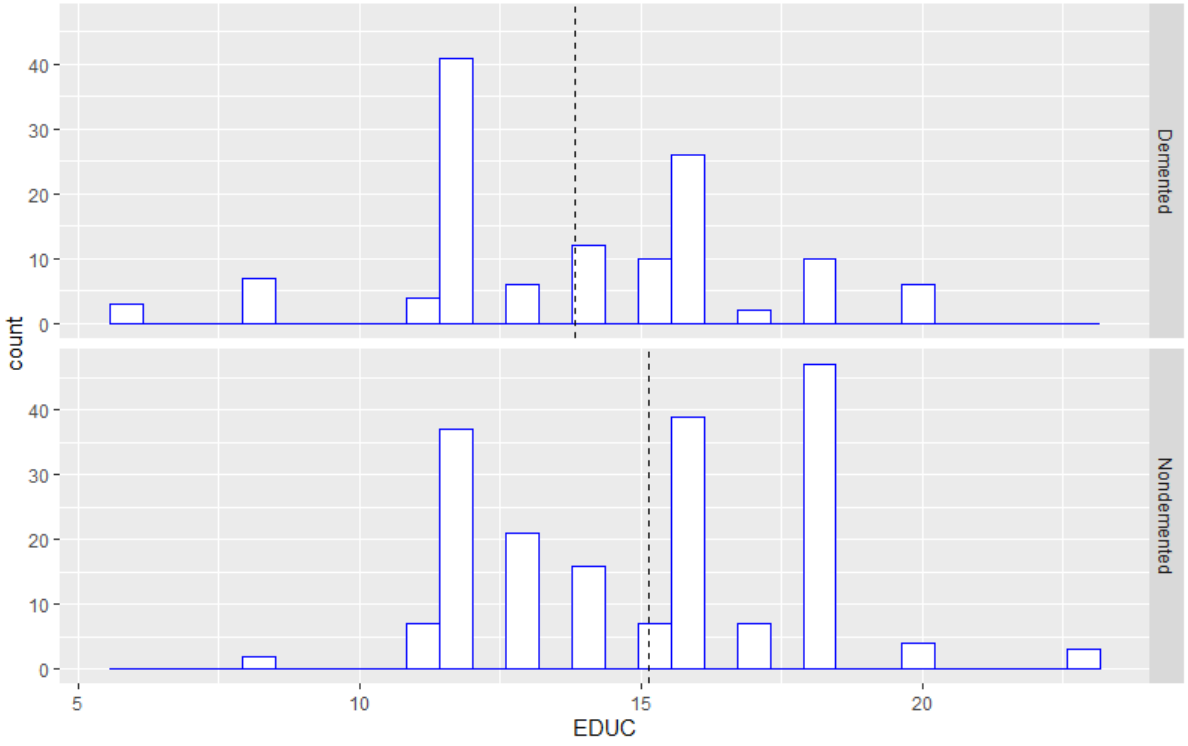
M.F: M=1; F=0.

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.

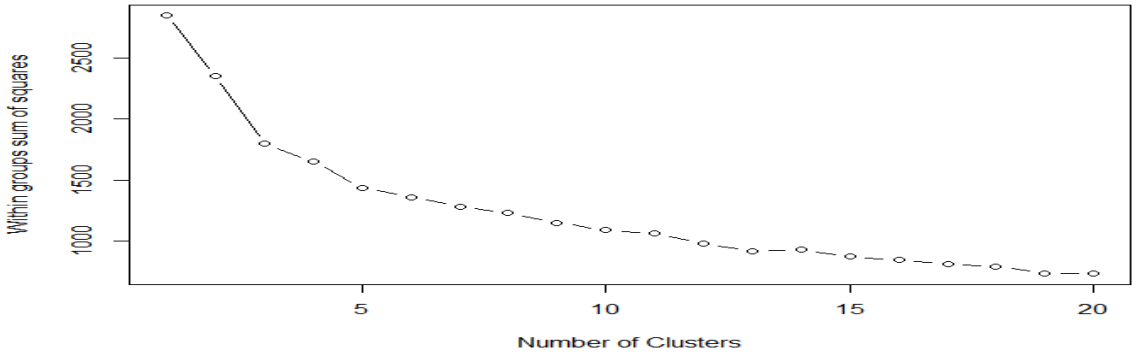


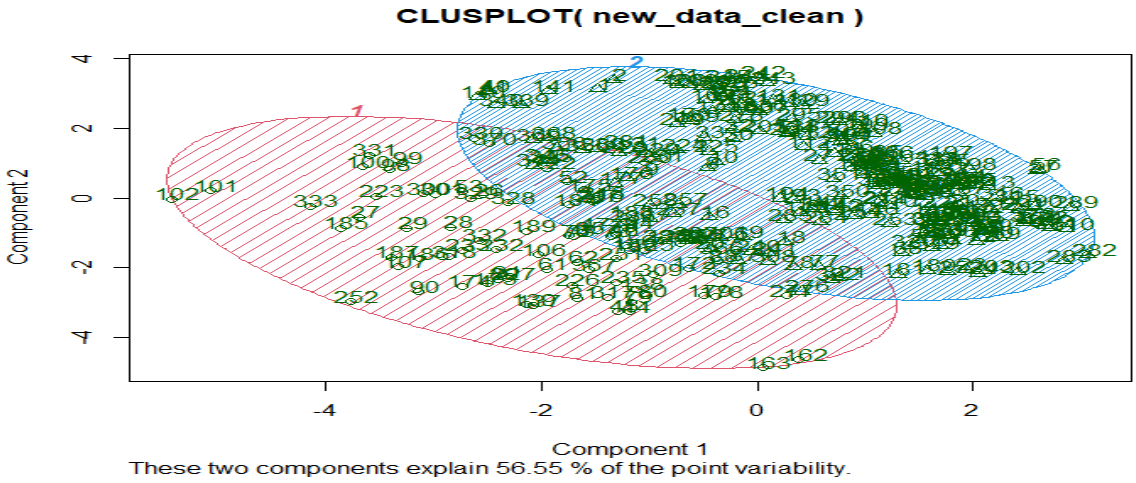
As per the above histogram, the average years of education at around 14 years for the demented group stand out lower than that at around 15 years for the nondemented group. Further, the number of demented persons peaks at around 12 years of education as against 18 years in the nondemented group.

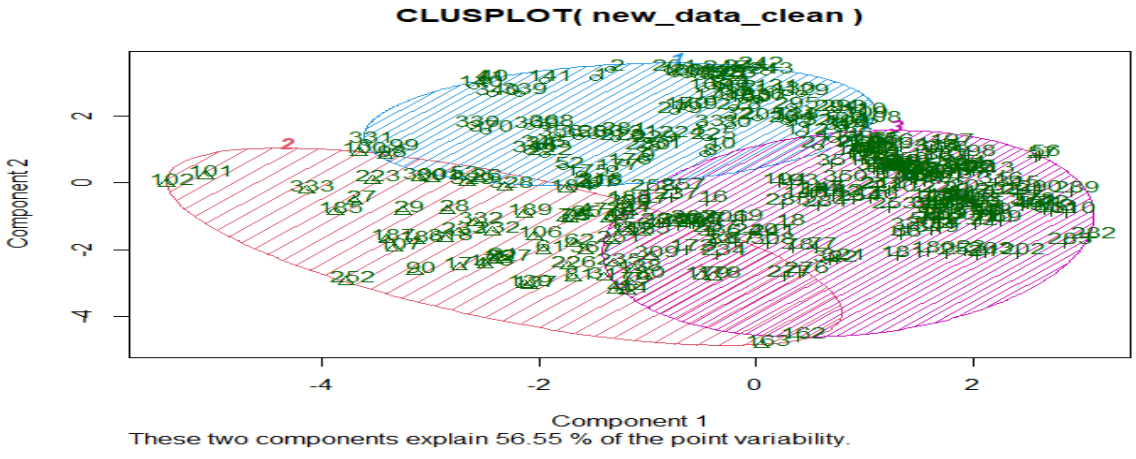
3) Clustering

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 male (on average), higher average of age, lower average education, higher average SES, lower average MMSE, higher average CDR, lower average eTIV, lower average nWBV, higher average ASF. Cluster 2 is associated with lower chance of being female (on average), while the other characteristics are likely to be

the opposite vis-a-vis those in Cluster 1. K mean clustering with 3 clusters contains 161, 67 and 89 observations in the clusters respectively. The magnitude of within cluster sum of squares is much smaller as compared with Cluster 2. The ratio between_SS cluster / total_SS cluster is 36.8%, higher than Cluster2.







As we increase the number of clusters, the within sum of squares reduce and the ratio between_SS cluster / total_SS cluster increases, eg, with 4 clusters, it is 43.6%. The gain in

ratio is not that high when we go from Cluster 3 to Cluster 4 as compared to that from Cluster 2 to Cluster 3. Therefore, Cluster 2 or Cluster 3 using k mean clustering may be appropriate.

4) Logistic Regression

We begin by taking log of the variables to scale down the values. Further we divide the data into training and testing data sets (80% training data and 20% testing data). In light of the interrelationship between demented/nondemented group and the independent variables as plotted /explained in the appendix, we construct the following model after several permutations and combinations: Group ~ Age + EDUC +MMSE+nWBV +SES + error

The coefficients of Age, EDUC, MMSE and nWBV are negative. Therefore, with an increase in any of these variables, a person is likely to be nondemented. The lowest SES (ie, represented by intercept) is statistically significant as compared to higher SES. Using confusion matrix, we find that the misclassification error in Training data set is 14.17% and it's 14.28% in testing dataset (Appendix).

	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

5) Feature Selection

We use a forward stepwise selection 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. So, we stopped the procedure. The final model turns out to be:

Group ~ 4.23 + (1.05*CDR) - (0.18*EDU) + (0.17*M.F) – (0.51*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

The Group variable moves with CDR and M.F in the same direction and with EDUC and eTIV in the opposite direction. The magnitude of impact from changes in CDR on the Group is the maximum.

6) Conclusion

It looks like there are more men having dementia in comparison to women. Individuals having low number of years of education or belonging to the highest socio-economic status are likely to be particularly vulnerable to dementia. The demented and nondemented groups are not very different in respective of characteristics like eTIV and ASF. The observations can at the most be segmented into 2 or 3 clusters albeit with some overlapping borders. A forward step-wise selection method produces a 4 variable predictor model for the group with the CDR having the maximum impact on the Group variable.

Appendix

Graph & Tables:

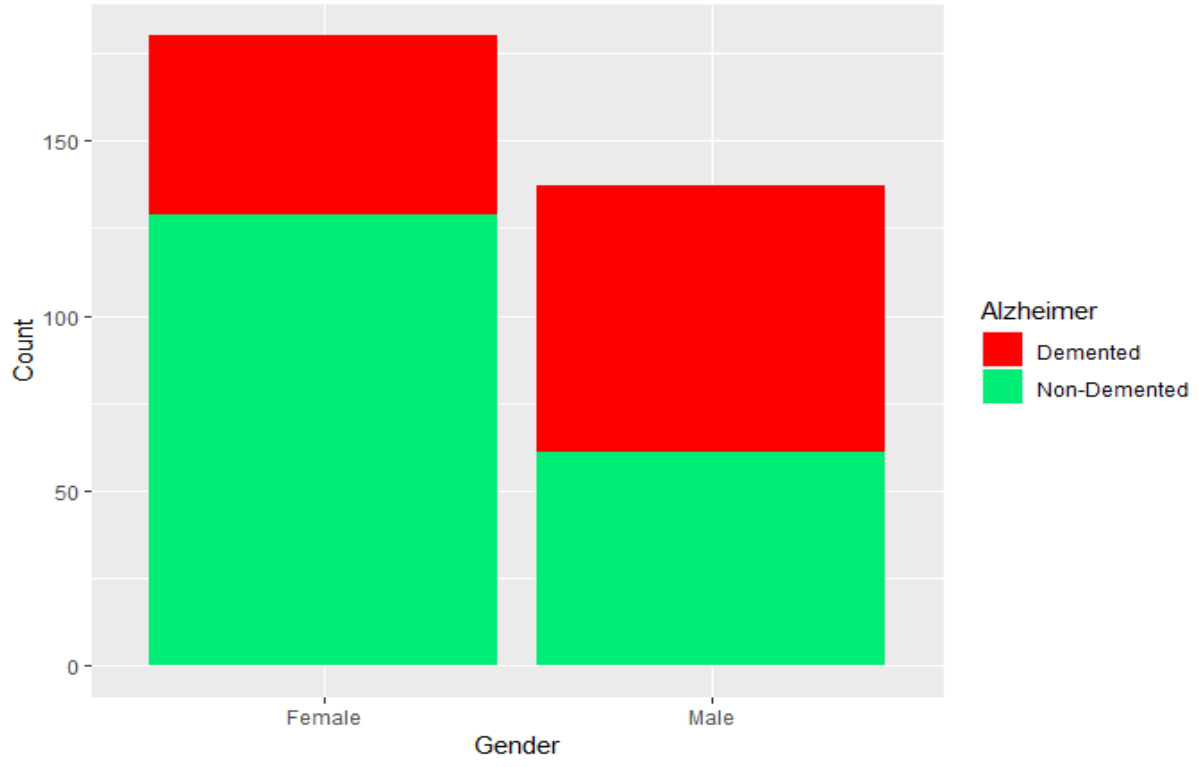
Summary Statistics of the Demented

variables	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
M.F	127	0.60	0.49	1	0.62	0	0	1	1	-0.40	-1.857	0.044
Age	127	76.20	7.35	76	75.95	8.90	61	98	37	0.35	-0.26	0.65
EDUC	127	13.83	3.03	14	13.90	2.97	6	20	14	-0.19	0.0582	0.27
SES	127	2.77	1.20	3	2.79	1.48	1	5	4	-0.14	-1.18	0.106
MMSE	127	24.32	4.66	26	24.83	4.45	4	30	26	-1.30	2.579	0.413
CDR	127	0.67	0.30	0.5	0.63	0	0.5	2	1.5	2.16	5.94	0.027
eTIV	127	1490.70	172.38	1477	1480.35	127.50	1143	1957	814	0.60	0.16	15.30
nWBV	127	0.72	0.033	0.711	0.71	0.03	0.646	0.806	0.16	0.37	-0.47	0.0030
ASF	127	1.19	0.13	1.188	1.19	0.11	0.897	1.535	0.638	0.030	-0.098	0.012

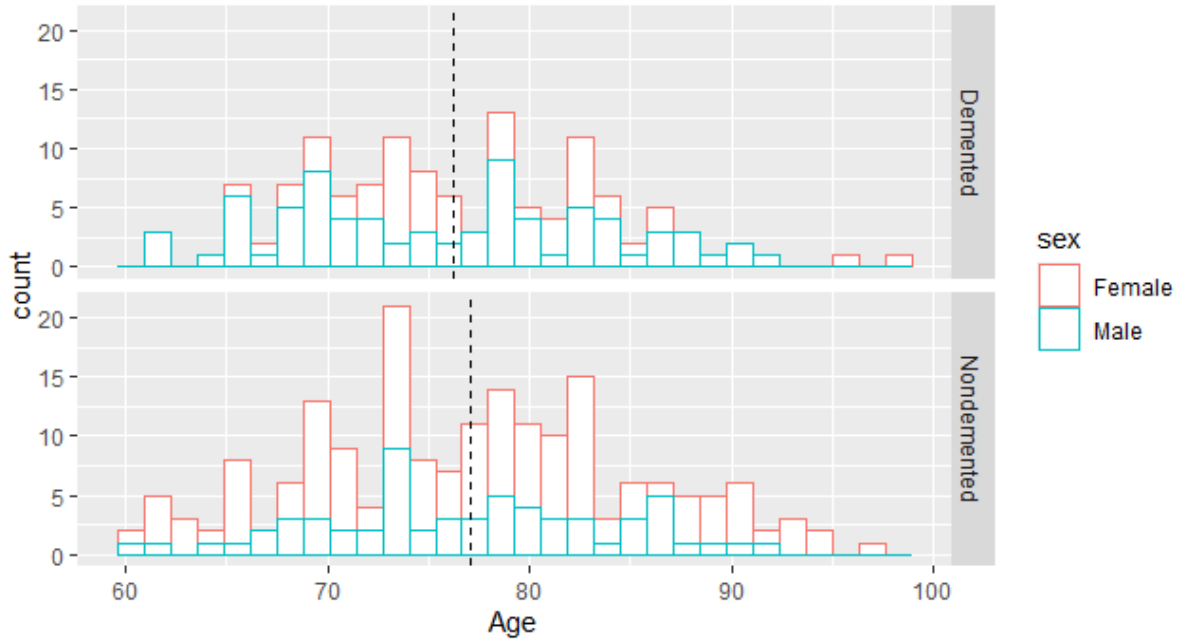
Summary statistics of the Non-Demented

variables	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
M.F	190	0.32	0.468	0	0.276	0	0	1	1	0.761	-1.43	0.034
Age	190	77.058	8.096	77	76.99	8.896	60	97	37	0.089	-0.529	0.587
EDUC	190	15.14	2.74	16	15.13	2.965	8	23	15	0.122	-0.371	0.1989
SES	190	2.39	1.048	2	2.36	1.483	1	5	4	0.319	-0.900	0.0760
MMSE	190	29.23	0.883	29	29.36	1.483	26	30	4	-1.050	0.601	0.0640
CDR	190	0.0053	0.051	0	0	0	0	0.5	0.5	9.517	89.03	0.00371
eTIV	190	1495.5	184.89	1474.5	1486.87	185.33	1106	2004	898	0.448	-0.39	13.41
nWBV	190	0.74	0.038	0.739	0.741	0.0400	0.644	0.837	0.193	-0.0081	-0.39	0.00274
ASF	190	1.19	0.144	1.19	1.189	0.160	0.876	1.587	0.711	0.096	-0.386	0.0104

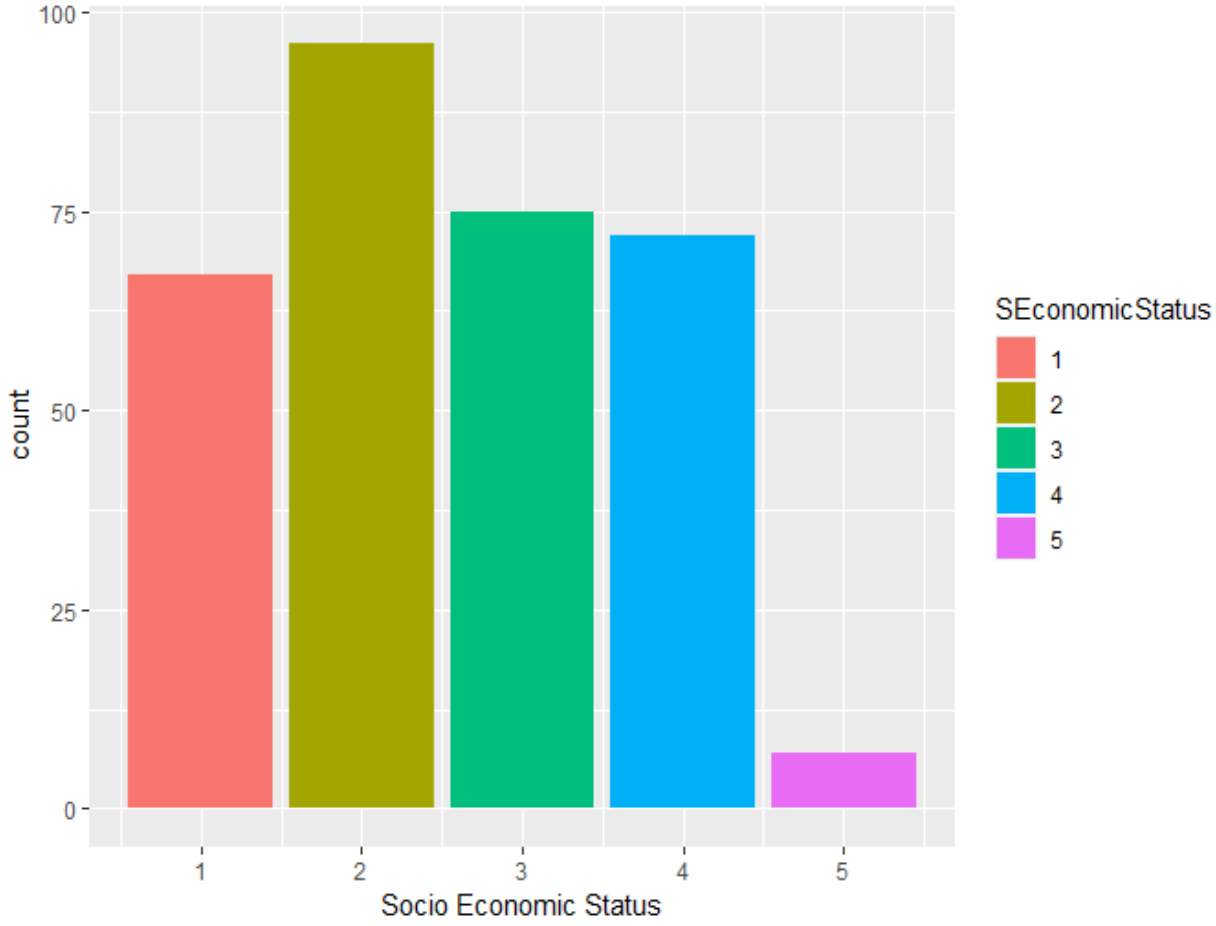
Gender and Alzheimer (Demented or Non-Demented)

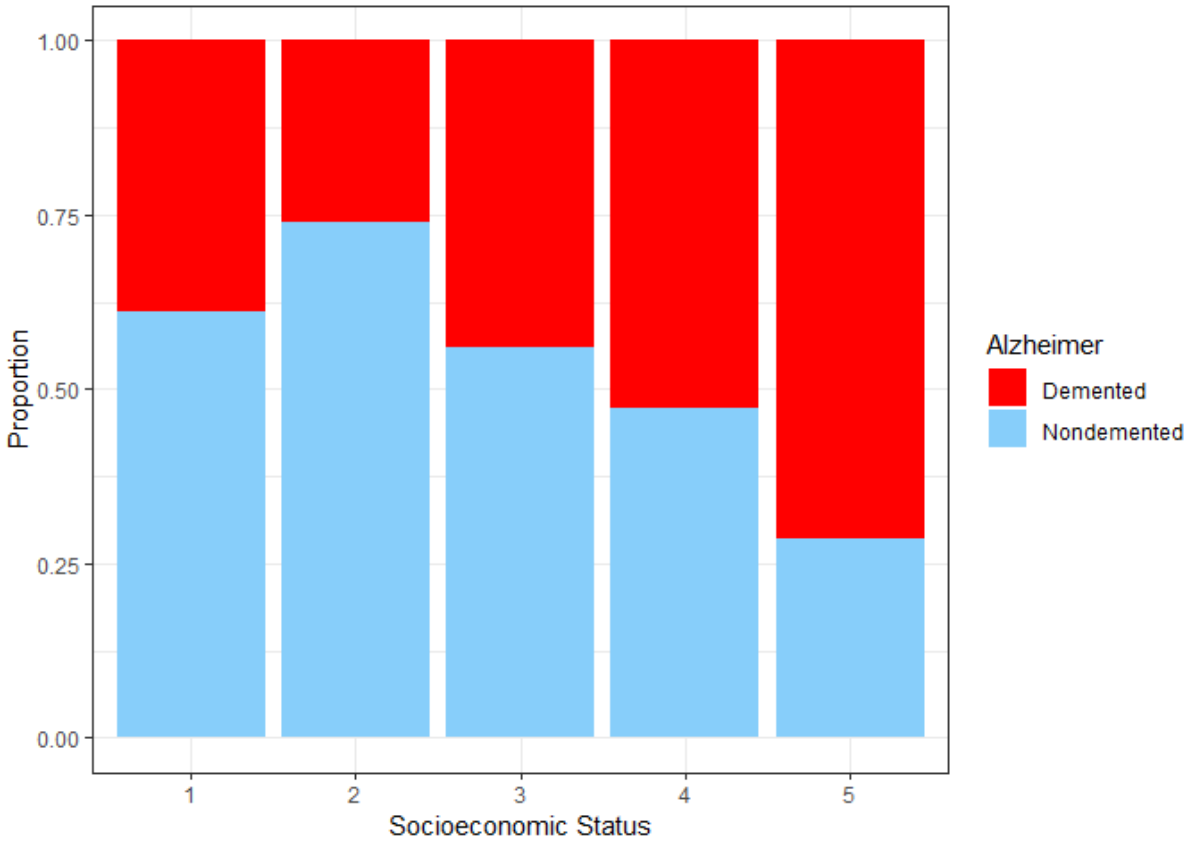


Age and Alzheimer (Demented or Non-Demented)

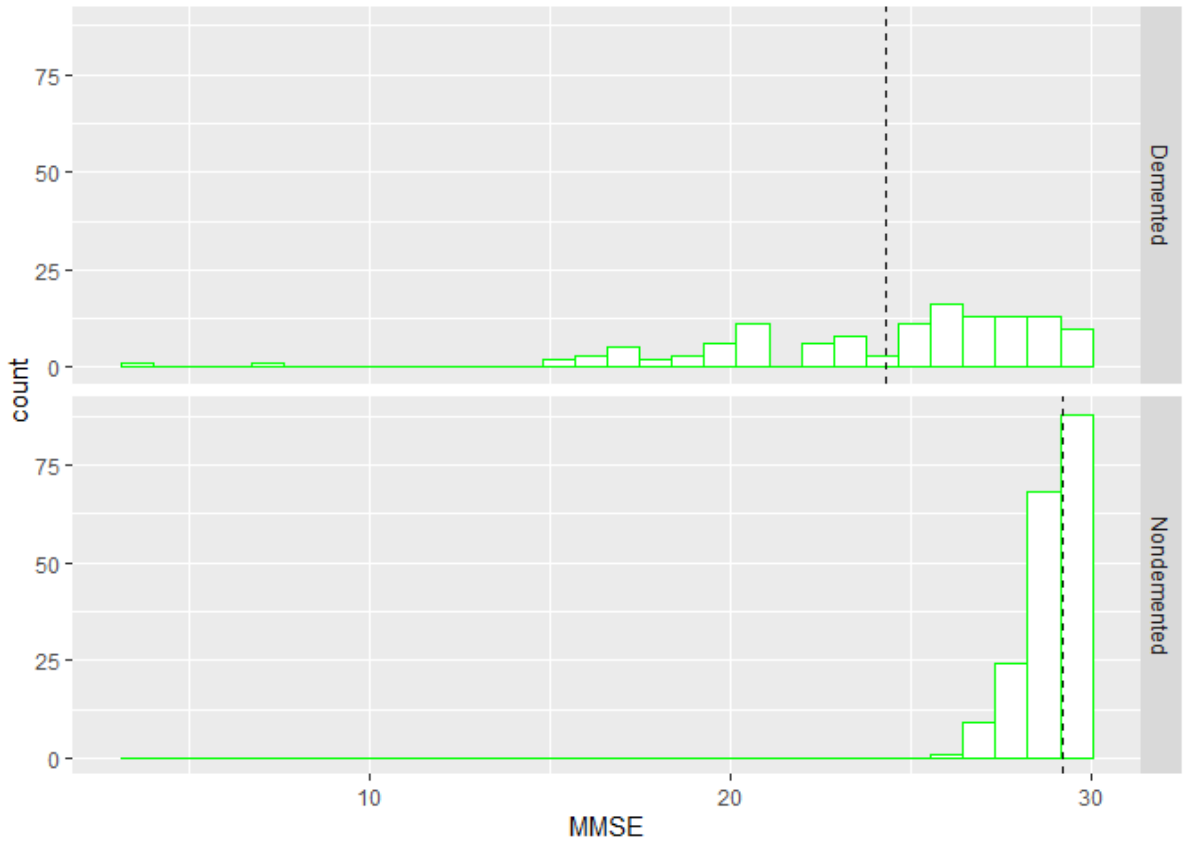


Socioeconomic Status and Alzheimer (Demented or Non-Demented)

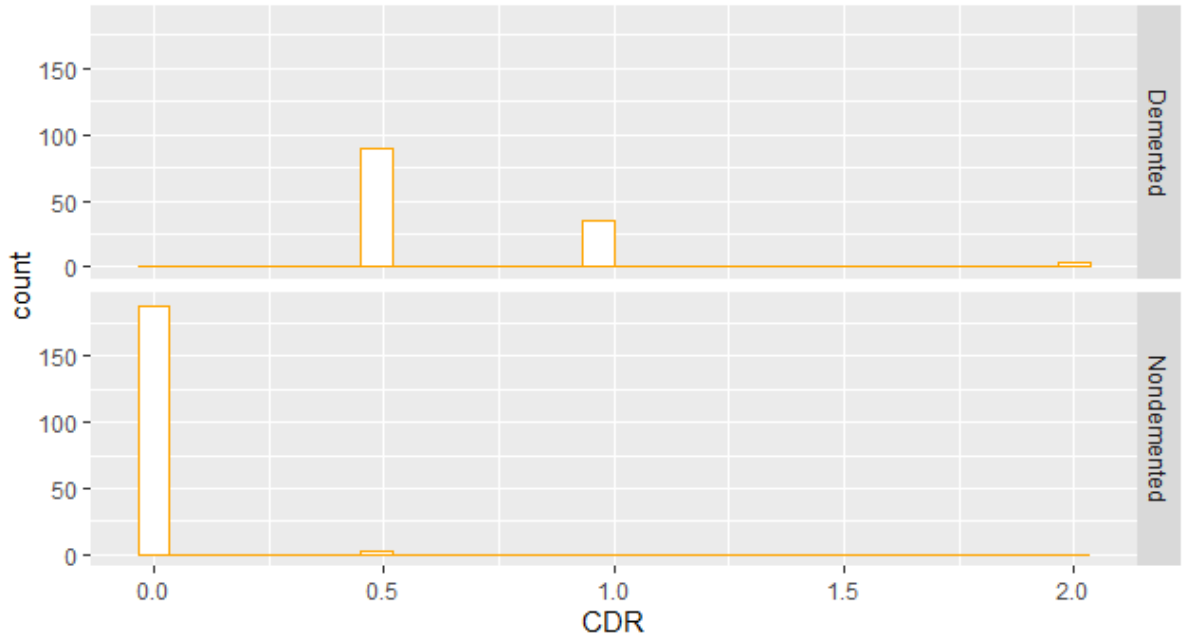




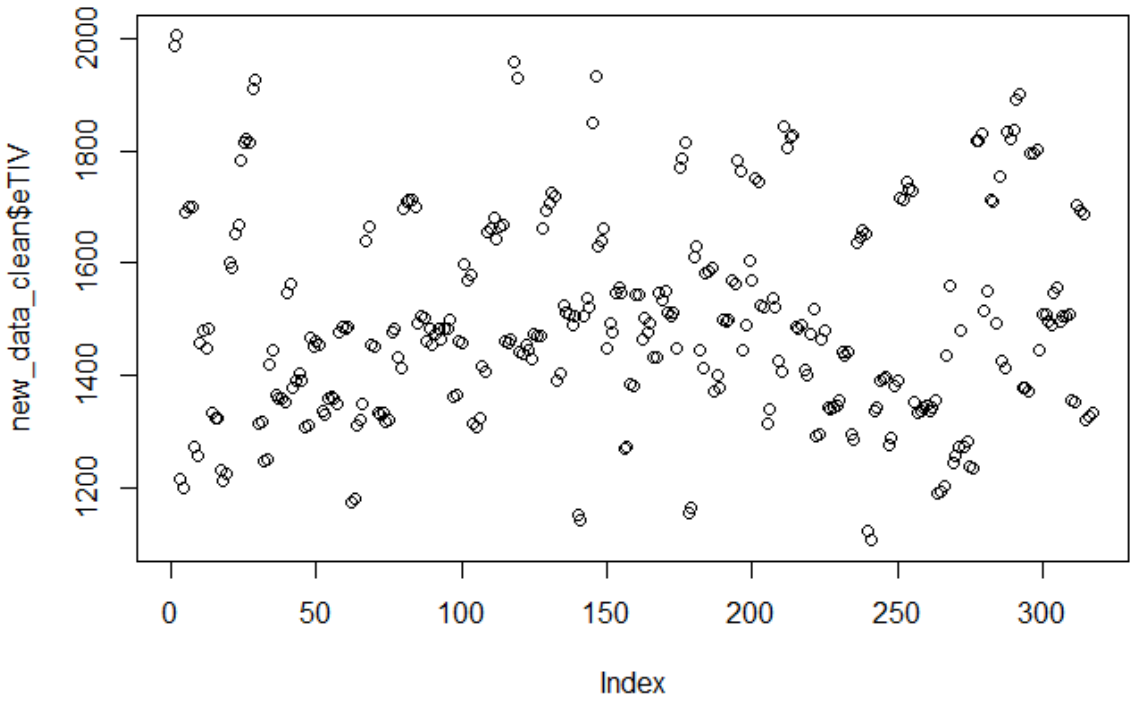
Mini Mental State Examination and Alzheimer (Demented or Non-Demented)

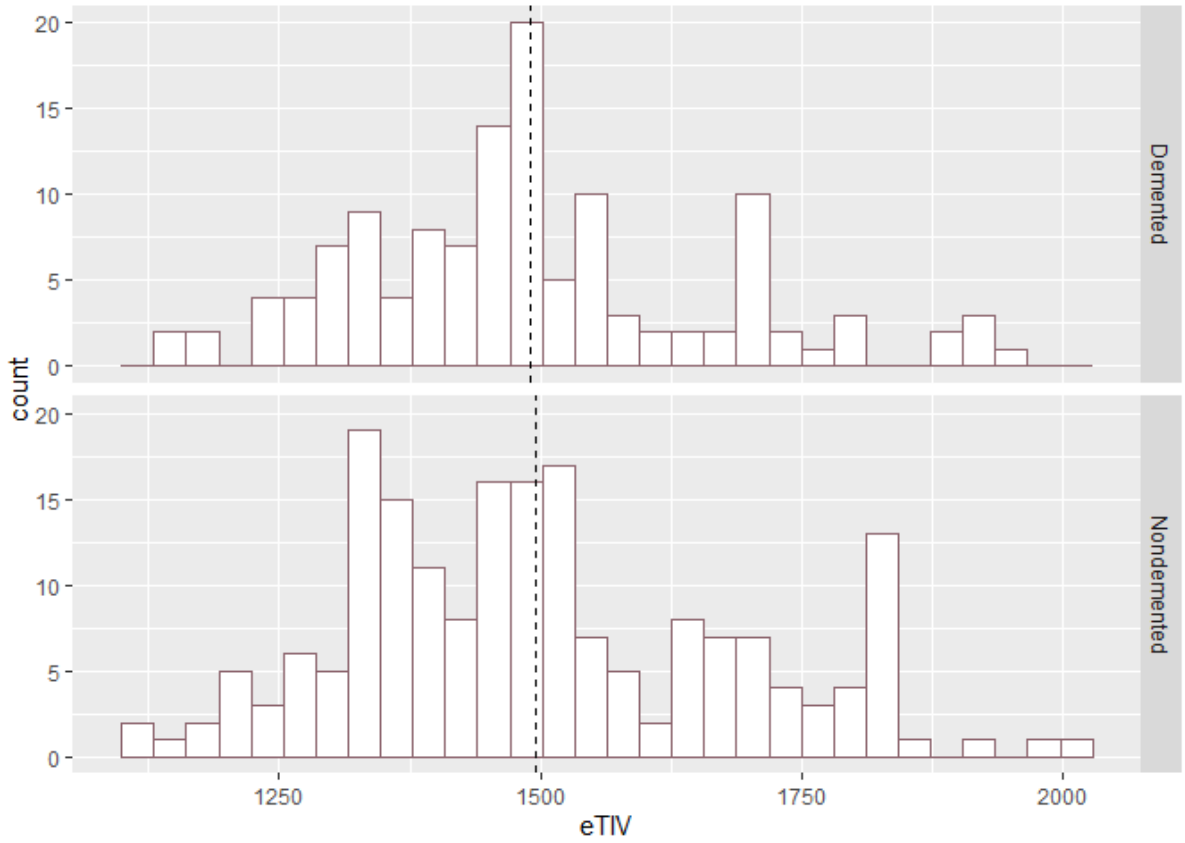


Clinical Dementia Rating and Alzheimer (Demented or Non-Demented)

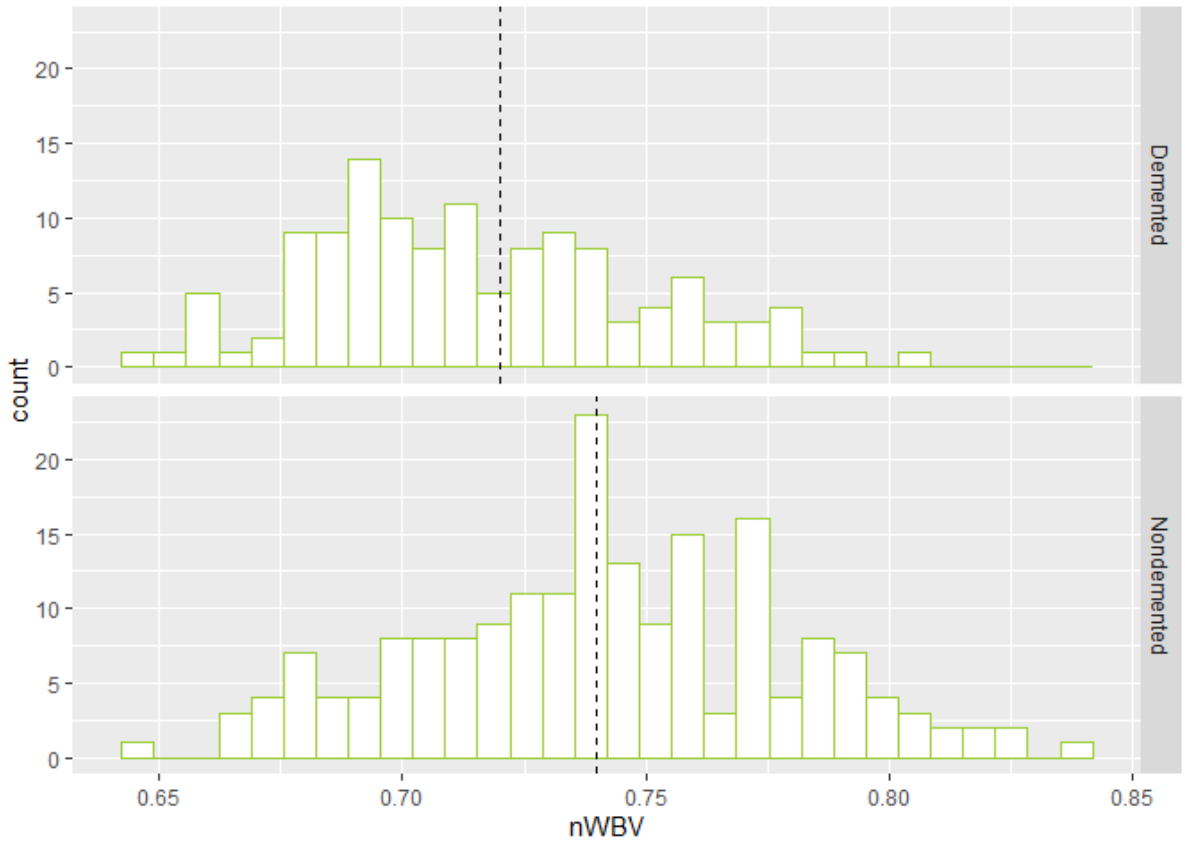


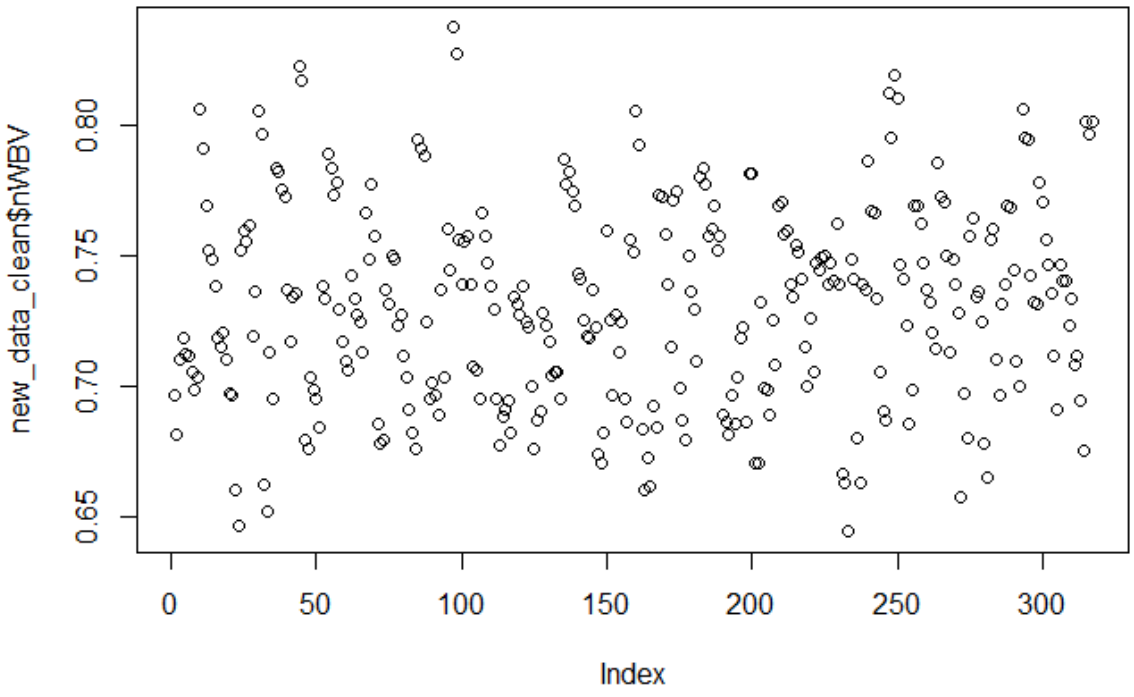
Estimated total intracranial volume and Alzheimer (Demented or Non-Demented)



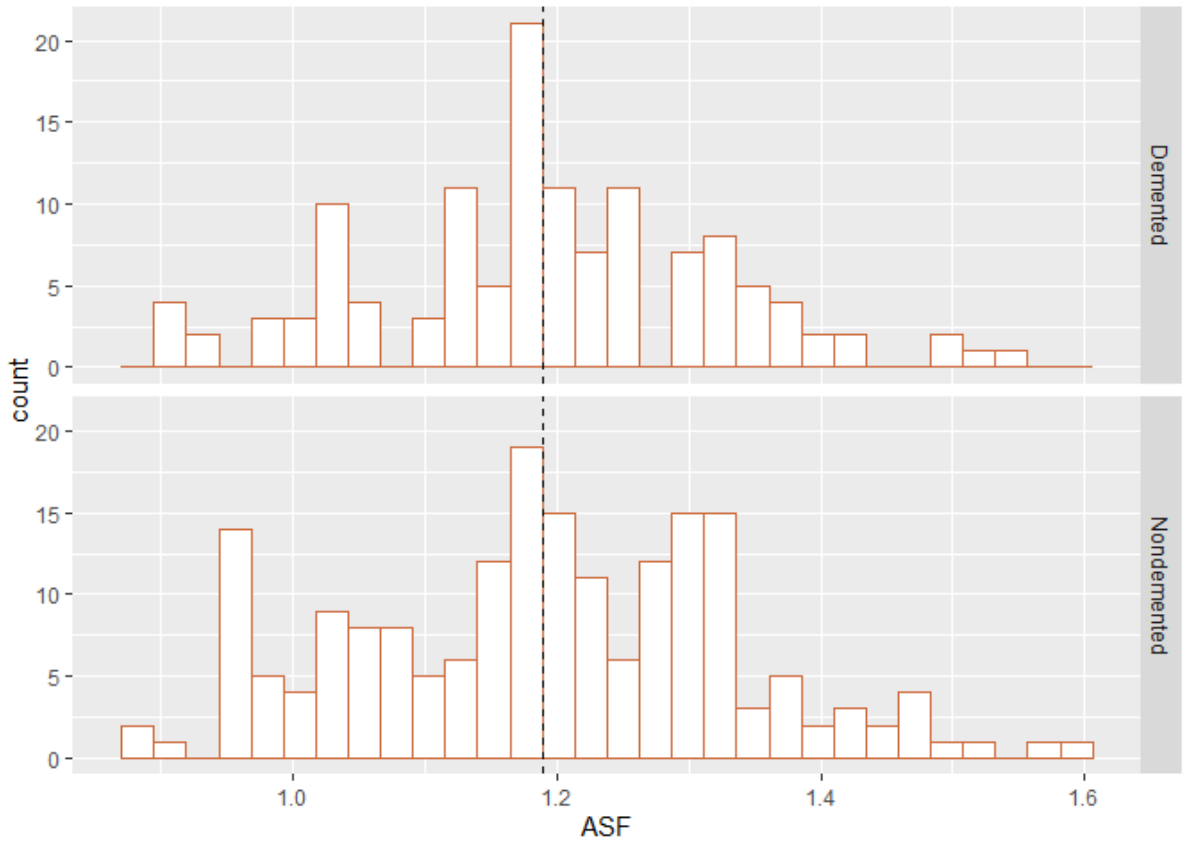


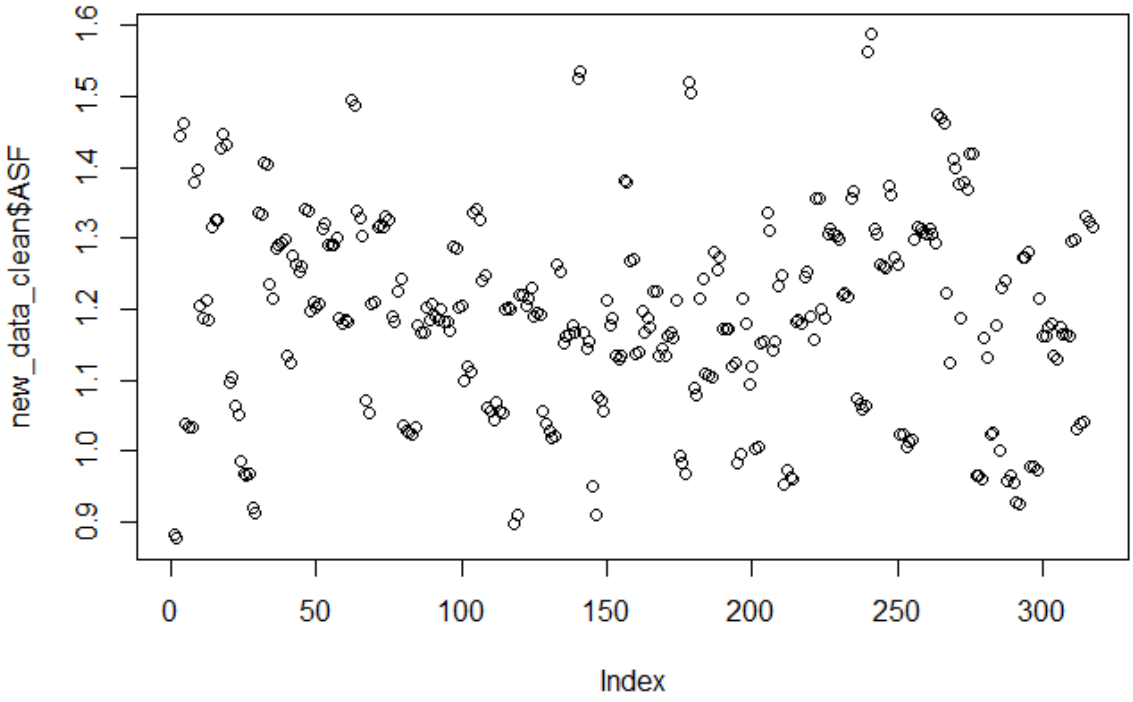
Normalize whole brain volume and Alzheimer (Demented or Non-Demented)



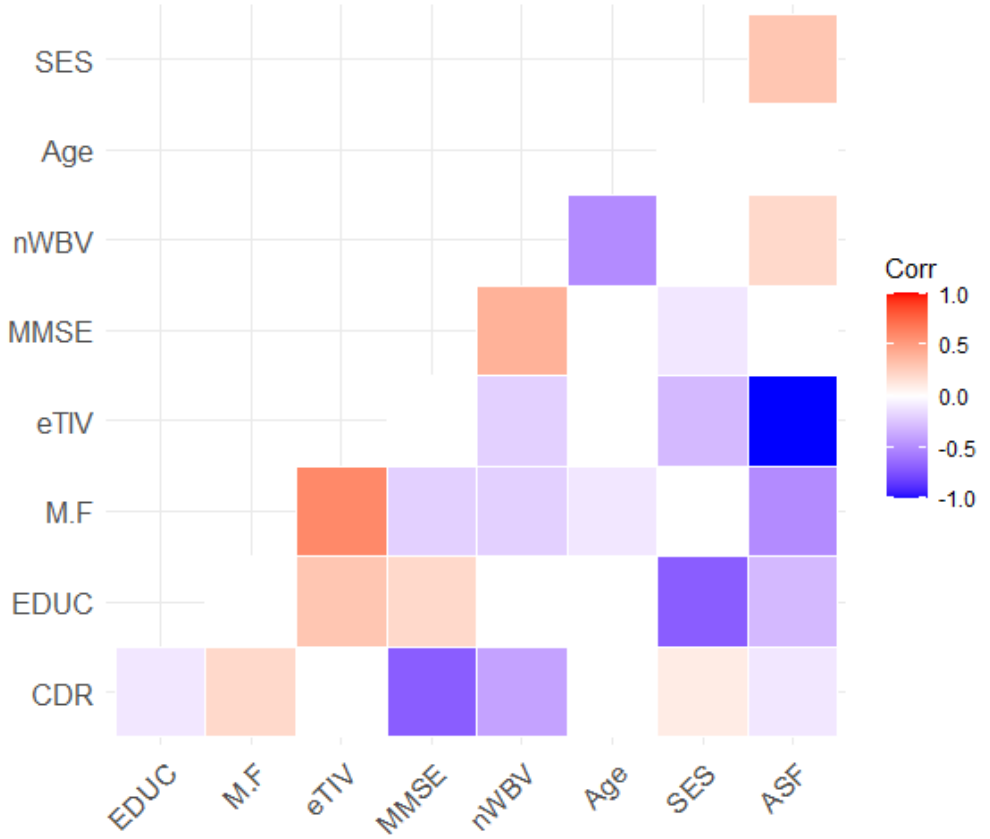


Atlas scaling factor and Alzheimer (Demented or Non-Demented)





Correlation Plot of all Independent Variables



Misclassification error in Training Data using Confusion Matrix

	Actual	
Predicted	0	1
0	146	27
1	10	78

Misclassification error in Testing Data using Confusion Matrix

	Actual	
Predicted	0	1
0	31	5
1	3	17

Code:

```
#libraries loaded

library(ggplot2)

library(psych)

library("cluster")

library(ggcorrplot)

library(plyr)

# Read the csv file

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=",")

#print column names

colnames(myproject)

# print first few rows of the dataset

head(myproject)

# print last few rows of the dataset

tail(myproject)
```

```
# number of rows in the dataset
nrow(myproject)
#count the frequencies of each unique value in the Group column
table(myproject$Group)
#count the frequencies of each unique value in the M.F column
table(myproject$M.F)
#Convert Male/Female into numerical value
myproject$M.F <- ifelse( myproject$M.F == "F", 0, 1)
myproject$M.F
#count missing values in dataset
num_na_values <- sum(is.na(myproject))
num_na_values
# Remove missing values
data_clean <- na.omit(myproject)
nrow(data_clean)
#Remove rows with Group="Converted"
new_data_clean <- subset(data_clean, !(Group == "Converted"))
# number of rows in the clean dataset
nrow(new_data_clean) #317
#Summary of cleaned dataset
summary(new_data_clean)
# count the frequencies of each unique value in the group column
table(new_data_clean$Group)
#getting unique values of the group column
```

```
unique(new_data_clean$Group)
############################Summary Statistics############
#summary statistics of cleaned dataset
head(new_data_clean)
nrow(new_data_clean) #317
colnames(new_data_clean)
ncol(new_data_clean)
describe(new_data_clean[,2:10])
# Subset the cleaned dataset to include only rows with 'Group' equal to "Demented"
new_data_clean_D <- subset(new_data_clean, (Group == "Demented"))
head(new_data_clean_D)
nrow(new_data_clean_D) #127
colnames(new_data_clean_D)
ncol(new_data_clean_D)
describe(new_data_clean_D[,2:10])
# Subset the cleaned dataset to include only rows with 'Group' equal to "Nondemented"
new_data_clean_ND <- subset(new_data_clean, (Group == "Nondemented"))
head(new_data_clean_ND)
nrow(new_data_clean_ND) #190
colnames(new_data_clean_ND)
ncol(new_data_clean_ND)
describe(new_data_clean_ND[,2:10])
# Calculate the total number of rows in the 'Demented' and 'Nondemented' groups
nrow(new_data_clean_D)+nrow(new_data_clean_ND)#317
```

```
# Create a box plot to visualize the relationship between 'Group' and 'Age' with color-coded 'Gender'
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")
############Each independent variable relation with dependent variable###############
################Gender vs Alzeimer#########################
# Subset the data for males
M<-subset(new_data_clean, M.F == 1)
nrow(M) #137
# Subset the data for females
F<-subset(new_data_clean, M.F == 0)
nrow(F) #180
nrow(M)+nrow(F)#317
# Subset the data for demented males
MD<- subset(new_data_clean, M.F == 1 & Group == 'Demented')
nrow(MD) #76
# Subset the data for demented females
FD<- subset(new_data_clean, M.F == 0 & Group == 'Demented')
nrow(FD) #51
# Subset the data for nondemented males
MND<- subset(new_data_clean, M.F == 1 & Group == 'Nondemented')
```

```
nrow(MND) #61
# Subset the data for nondemented females
FND<- subset(new_data_clean, M.F == 0 & Group == 'Nondemented')
nrow(FND) #129
nrow(MD)+nrow(FD)+nrow(MND)+nrow(FND)#317
# Calculate the proportion of males with dementia
proportionMales<-round(76/137,2) #55% Males
proportionMales
# Calculate the proportion of females with dementia
proportionfemales<-round(51/180,2) #28% Females
proportionfemales
# Create a data frame for gender and demented count
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)
# Create a stacked bar plot to visualize the gender vs. count with fill representing dementia status
ggplot(df, aes(x = gender, y = count, fill = demented)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Gender", y = "Count", fill = "Alzheimer") +
  scale_fill_manual(values = c("red", "springgreen2"))
############################Age vs Alzheimer#########################
# Assign "Male" or "Female" based on M.F variable
sex <- ifelse(new_data_clean$M.F == 1, "Male", "Female")
```

```
# Calculate the mean age for demented and non-demented groups
mu <- ddply(new_data_clean, "Group", summarise, grp.mean=round(mean(Age),2))
mu #Demented Mean Age is 76.20
  #NonDemented Mean Age is 77.06
# Create a histogram of Age with color representing sex and facet by Group
p<-ggplot(new_data_clean, aes(x=Age, color=sex))+
 geom_histogram(fill="white")+
 facet_grid(Group ~ .)
p
#  Add mean lines for demented and non-demented groups
p+geom_vline(data=mu, aes(xintercept=grp.mean),linetype="dashed")
##########################Education vs Alzemier##################
# Calculate the mean education level for demented and non-demented groups
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
# Create a histogram of EDUC with color representing Group
q<-ggplot(new_data_clean, aes(x=EDUC))+
 geom_histogram(color="blue", fill="white")+
 facet_grid(Group ~ .)
q
# Add mean lines for demented and non-demented groups
q+geom_vline(data=nu, aes(xintercept=grp.mean),linetype="dashed")
##########################Socio Economic Status vs Alzeimer################
```

```
# Create a bar plot to visualize the distribution of Socio Economic Status
SEconomicStatus<-as.factor(new_data_clean$SES)
ggplot(new_data_clean) +
  geom_bar(mapping = aes(x = SES,fill=SEconomicStatus)) +
  labs(x = "Socio Economic Status")
# Create a bar plot to visualize the proportion of demented vs. non-demented based on Socioeconomic Status
ggplot(new_data_clean, aes(x = factor(SES), fill = factor(Group))) +
  geom_bar(position = "fill") +
  scale_fill_manual(values = c("red", "lightskyblue")) +
  labs(x = "Socioeconomic Status", y = "Proportion", fill = "Alzheimer") +
  theme_bw()
########################Mini Mental State Examination vs Alzeimer#################
new_data_clean$MMSE
count(new_data_clean$MMSE)
# Calculate the mean MMSE score for demented and non-demented groups
wu <- ddply(new_data_clean, "Group", summarise, grp.mean=round(mean(MMSE),2))
wu #Demented Mean MMSE is 24.32
#NonDemented Mean MMSE is 29.23
# Create a histogram of MMSE with color representing Group
g<-ggplot(new_data_clean, aes(x=MMSE))+
  geom_histogram(color="green", fill="white")+
  facet_grid(Group ~ .)
g
# Add mean lines for demented and non-demented groups
```

```
g+geom_vline(data=wu, aes(xintercept=grp.mean),linetype="dashed")
######################Clinical Dimentia Rating vs Alzeimer################
new_data_clean$CDR
# Calculate the mean CDR score for demented and non-demented groups
yu <- ddply(new_data_clean, "Group", summarise, grp.mean=round(mean(CDR),2))
yu #Demented Mean CDR is 0.67
#NonDemented Mean CDR is 0.01
count(new_data_clean$CDR)
# Create a histogram of CDR with color representing Group
f<-ggplot(new_data_clean, aes(x=CDR))+
  geom_histogram(color="orange", fill="white")+
  facet_grid(Group ~ .)
f
####################Estimated total intracranial volume vs Alzeimer#################
new_data_clean$eTIV
plot(new_data_clean$eTIV)
# Calculate the mean eTIV for demented and non-demented groups
ru <- ddply(new_data_clean, "Group", summarise, grp.mean=round(mean(eTIV),2))
ru #Demented Mean eTIV is 1490.7
#NonDemented Mean eTIV is 1495.5
count(new_data_clean$CDR)
# Create a histogram of eTIV with color representing Group
s<-ggplot(new_data_clean, aes(x=eTIV))+
  geom_histogram(color="pink4", fill="white")+
```

```
  facet_grid(Group ~ .)
s
# Add mean lines for demented and non-demented groups
s+geom_vline(data=ru, aes(xintercept=grp.mean),linetype="dashed")
################Normalize whole brain volume vs Alzeimer########
new_data_clean$nWBV
plot(new_data_clean$nWBV)
# Calculate the mean nWBV for demented and non-demented groups
pu <- ddply(new_data_clean, "Group", summarise, grp.mean=round(mean(nWBV),2))
pu #Demented Mean nWBV is 0.72
#NonDemented Mean nWBV is 0.74
# Create a histogram of nWBV with color representing Group
r<-ggplot(new_data_clean, aes(x=nWBV))+
  geom_histogram(color="yellowgreen", fill="white")+
  facet_grid(Group ~ .)
r
# Add mean lines for demented and non-demented groups
r+geom_vline(data=pu, aes(xintercept=grp.mean),linetype="dashed")
################################Atlas scaling factor VS Alzeimer#################
new_data_clean$ASF
plot(new_data_clean$ASF)
# Calculate the mean ASF for demented and non-demented groups
du <- ddply(new_data_clean, "Group", summarise, grp.mean=round(mean(ASF),2))
du #Demented Mean ASF is 1.19
```

```
#NonDemented Mean ASF is 1.19
# Create a histogram of ASF with color representing Group
t<-ggplot(new_data_clean, aes(x=ASF))+
  geom_histogram(color="sienna3", fill="white")+
  facet_grid(Group ~ .)
t
# Add mean lines for demented and non-demented groups
t+geom_vline(data=du, aes(xintercept=grp.mean),linetype="dashed")
############Correlation Matrix######################
# Calculate the correlation matrix for the variables
corr <- round(cor(new_data_clean[,2:10]), 1)
head(corr)
# Visualize the correlation matrix using a heatmap
ggcorrplot(corr, hc.order = TRUE, type = "lower",
           outline.col = "white")
####################Clustering Algorithm##################
##################K mean Clustering##################
z <-new_data_clean[,-c(1,1)]
head(z)
# Calculate the mean and standard deviation for each variable
m<-apply(z,2,mean)
m
s<-apply(z,2,sd)
s
```

```
# Normalize the data by subtracting from the mean and dividing by the standard deviation
nor<-scale(z,center=m,scale =s)
head(nor)
# Calculate distance matrix
distance <- dist(nor)
print(distance, digits =2)
# Scree Plot to determine the optimal number of clusters
wss <- (nrow(nor)-1)*sum(apply(nor,2,var))
for (i in 2:20) wss[i] <- sum(kmeans(nor, centers=i)$withinss)
plot(1:20, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares")
# K-means clustering
set.seed(1234)
# Perform K-means clustering with k = 2
kc1<-kmeans(nor,2)
kc1
# Perform K-means clustering with k = 3
kc2<-kmeans(nor,3)
kc2
# Perform K-means clustering with k = 4
kc3<-kmeans(nor,4)
kc3
#Visualizing the clusters using clusplot
clusplot(new_data_clean,kc1$cluster,
        color=T,shade= T,
```

```
        labels=2, lines=0)
clusplot(new_data_clean,kc2$cluster,
        color=T,shade= T,
        labels=2, lines=0)
#########################Logistic Regression#####################
#Check the structure of the dataset
str(new_data_clean)
head(new_data_clean)
head(new_data_clean$Group)
tail(new_data_clean$Group)
#Convert the "Group" variable to binary values
group <- ifelse(new_data_clean$Group== "Nondemented", 0, 1)
head(group)
tail(group)
# Create a new data frame 'logisticdata' with the 'group' variable and remaining columns from 'new_data_clean'
logisticdata<-data.frame(group,new_data_clean[,-1])
head(logisticdata)
# Convert 'SES' column to a factor variable
logisticdata$SES<-as.factor(logisticdata$SES)
head(logisticdata$SES)
# Apply logarithmic transformation to 'Age' column
logisticdata$Age<-log(logisticdata$Age)
head(logisticdata$Age)
# Apply logarithmic transformation to 'EDUC' column
```

```
logisticdata$EDUC<-log(logisticdata$EDUC)
head(logisticdata$EDUC)
# Apply logarithmic transformation to 'MMSE' column
logisticdata$MMSE<-log(logisticdata$MMSE)
head(logisticdata$MMSE)
# Apply logarithmic transformation to 'eTIV' column
logisticdata$eTIV<-log(logisticdata$eTIV)
head(logisticdata$eTIV)
# Convert 'M.F' column to a factor variable
logisticdata$M.F<-as.factor(logisticdata$M.F)
head(logisticdata$M.F)
head(logisticdata)
str(logisticdata)
# Partition data - train (80%) & test (20%)
set.seed(1234)
ind <- sample(2, nrow(logisticdata), replace = T, prob = c(0.8, 0.2))
head(ind)
train <- logisticdata[ind==1,]
head(train)
test <- logisticdata[ind==2,]
head(test)
# Logistic regression model using the train data
mymodel <- glm(group ~ Age + EDUC + MMSE + nWBV + SES , data = train, family =
"binomial")
#summary of the logistic regression model
```

```
summary(mymodel)

# Prediction

p1 <- predict(mymodel, train, type = 'response')

head(p1)

tail(p1)

head(train)

tail(train)

# Misclassification error - train data

pred1 <- ifelse(p1>0.5, 1, 0)

tab1 <- table(Predicted = pred1, Actual = train$group)

tab1

1 - sum(diag(tab1))/sum(tab1)

# Misclassification error - test data

p2 <- predict(mymodel, test, type = 'response')

pred2 <- ifelse(p2>0.5, 1, 0)

tab2 <- table(Predicted = pred2, Actual = test$group)

tab2

1 - sum(diag(tab2))/sum(tab2)

################################Stepwise Forward Selection#################

#define intercept-only model

intercept_only <- lm(group ~ 1, data=logisticdata)

#define model with all predictors

all <- lm(group ~ ., data=logisticdata)

#perform forward stepwise regression
```

```
forward <- step(intercept_only, direction='forward', scope=formula(all), trace=1)

forward

#view results of forward stepwise regression

forward$anova

#view final model

round(forward$coefficients,2)
```