ALZHEIMER'S ANALYSIS

INFX 502

Modupe T Olayinka

UL LAFAYETTE | C00293417

Table of Contents

DATA

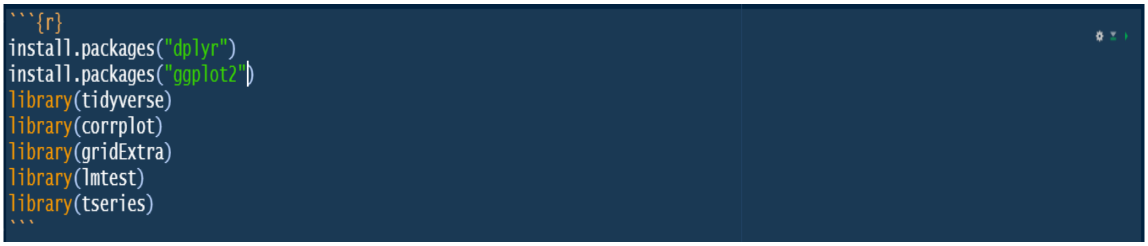
Dataset Description

The participants in this long-term research at OASIS range in age from 60 to 96. Each patient underwent 373 imaging sessions throughout at least one visit, each spaced by at least a year. Three or four different T1-weighted MRI scans taken during one scan session are presented to everyone. Both right-handed men and women are included in the topics. The following link contains the dataset submitted from an open data source.

OASIS Brains - Open Access Series of Imaging Studies (oasis-brains.org)

Loading Libraries

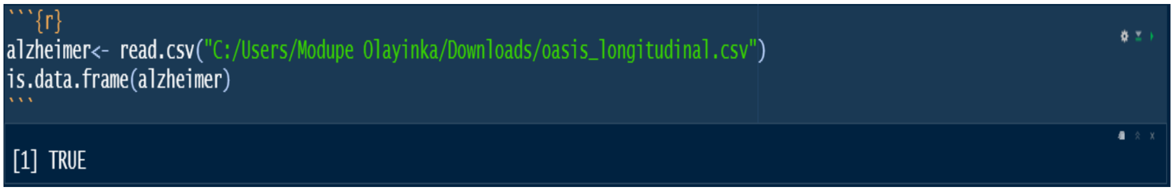
A few packages were installed, and a few libraries were loaded to complete my task. The code below was used to accomplish this.



~~~
```{r}
install.packages("dplyr")
install.packages("ggplot2")
library(tidyverse)
library(corrplot)
library(gridExtra)
library(lmtest)
library(tseries)
```
~~~

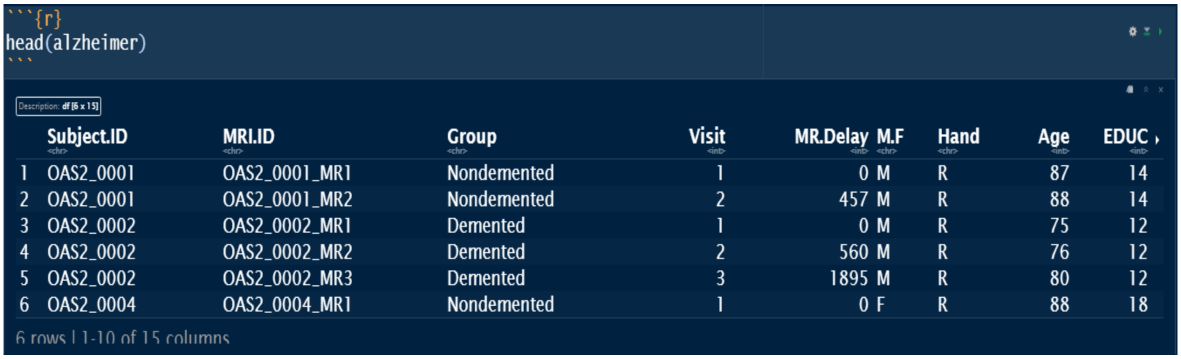
Loading the Dataset

The dataset was loaded using the code below from an Excel CSV file, and a data frame is used to verify that the dataset was loaded correctly.



~~~
```{r}
alzheimer<- read.csv("C:/Users/Modupe Olayinka/Downloads/oasis_longitudinal.csv")
is.data.frame(alzheimer)
```

[1] TRUE
~~~

The head () function displays the dataset's first six rows.



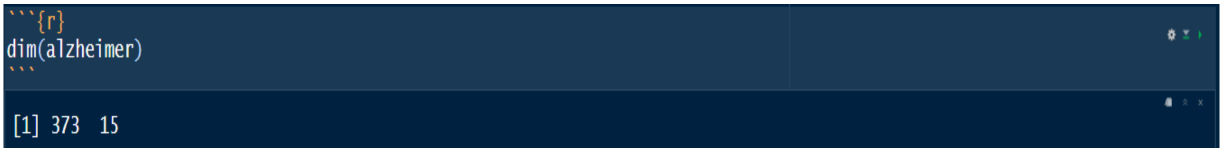
~~~
```{r}
head(alzheimer)
```
~~~

Description: df [6 x 15]

	Subject.ID <chr>	MRI.ID <chr>	Group <chr>	Visit <int>	MR.Delay <int>	M.F <chr>	Hand <chr>	Age <int>	EDUC <int>
1	OAS2_0001	OAS2_0001_MR1	Nondemented	1	0	M	R	87	14
2	OAS2_0001	OAS2_0001_MR2	Nondemented	2	457	M	R	88	14
3	OAS2_0002	OAS2_0002_MR1	Demented	1	0	M	R	75	12
4	OAS2_0002	OAS2_0002_MR2	Demented	2	560	M	R	76	12
5	OAS2_0002	OAS2_0002_MR3	Demented	3	1895	M	R	80	12
6	OAS2_0004	OAS2_0004_MR1	Nondemented	1	0	F	R	88	18

6 rows | 1-10 of 15 columns

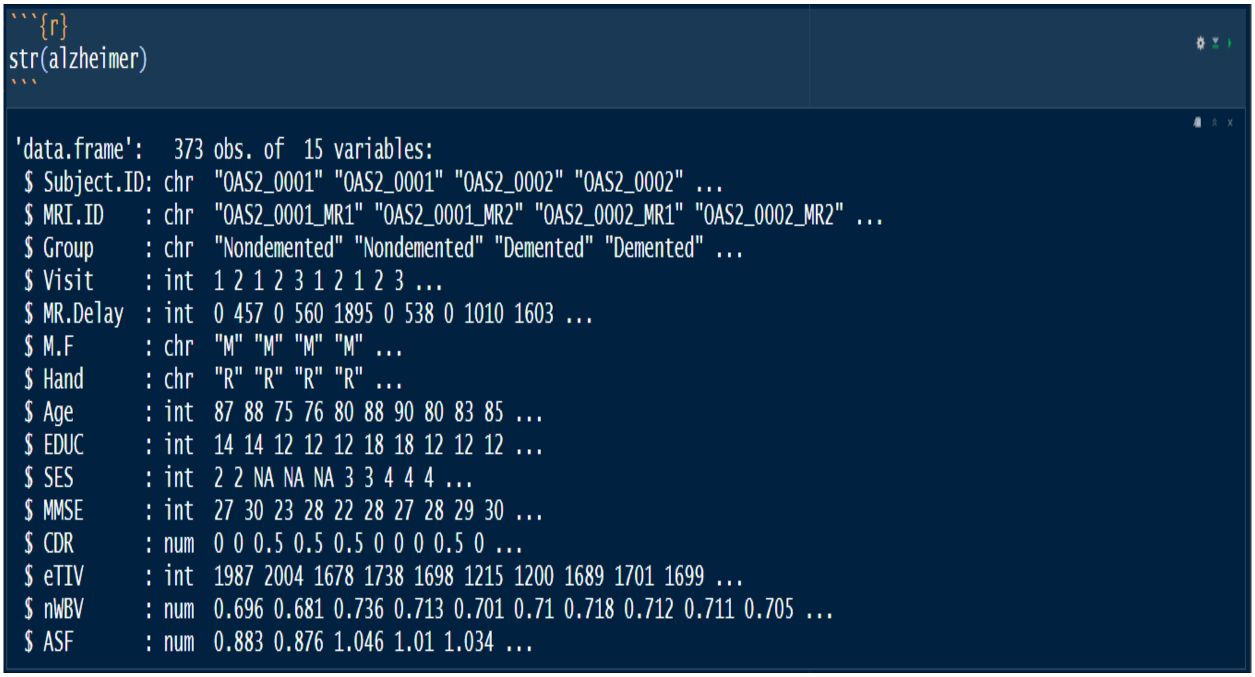
The dim () function was employed to determine the dataset's number of columns and rows. There are 15 columns and 373 rows in this dataset.



~~~
```{r}
dim(alzheimer)
```

[1] 373  15
~~~

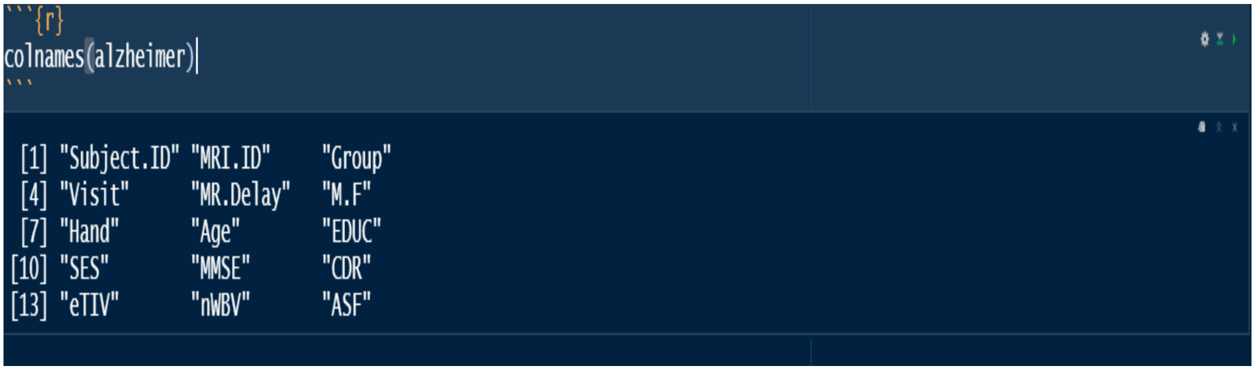
The original structures

The str () function was used to determine the modes of the variables in the dataset, and the output shown below reveals that the original dataset had three modes: num, int, and Chr. Before the study began, the dataset only needed to be lightly cleaned.



~~~
```{r}
str(alzheimer)
```

'data.frame':   373 obs. of  15 variables:
 $ Subject.ID: chr  "OAS2_0001" "OAS2_0001" "OAS2_0002" "OAS2_0002" ...
 $ MRI.ID    : chr  "OAS2_0001_MR1" "OAS2_0001_MR2" "OAS2_0002_MR1" "OAS2_0002_MR2" ...
 $ Group     : chr  "Nondemented" "Nondemented" "Demented" "Demented" ...
 $ Visit     : int  1 2 1 2 3 1 2 1 2 3 ...
 $ MR.Delay  : int  0 457 0 560 1895 0 538 0 1010 1603 ...
 $ M.F       : chr  "M" "M" "M" "M" ...
 $ Hand      : chr  "R" "R" "R" "R" ...
 $ Age       : int  87 88 75 76 80 88 90 80 83 85 ...
 $ EDUC      : int  14 14 12 12 12 18 18 12 12 12 ...
 $ SES       : int  2 2 NA NA NA 3 3 4 4 4 ...
 $ MMSE      : int  27 30 23 28 22 28 27 28 29 30 ...
 $ CDR       : num  0 0 0.5 0.5 0.5 0 0 0 0.5 0 ...
 $ eTIV      : int  1987 2004 1678 1738 1698 1215 1200 1689 1701 1699 ...
 $ nWBV      : num  0.696 0.681 0.736 0.713 0.701 0.71 0.718 0.712 0.711 0.705 ...
 $ ASF       : num  0.883 0.876 1.046 1.01 1.034 ...
~~~

Using the colnames() method, the dataset's list of columns was shown, as can be seen below.

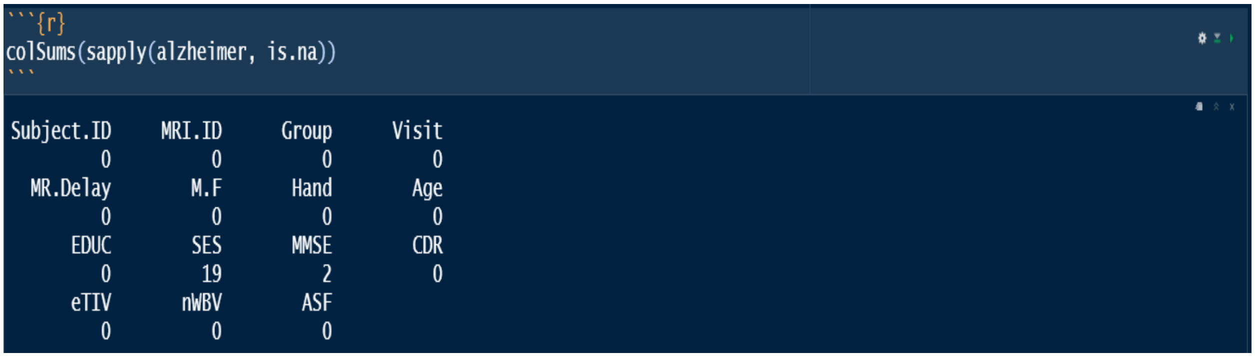


~~~
```{r}
colnames(alzheimer)
```
~~~

```
 [1] "Subject.ID" "MRI.ID"     "Group"
 [4] "Visit"      "MR.Delay"   "M.F"
 [7] "Hand"       "Age"        "EDUC"
[10] "SES"        "MMSE"       "CDR"
[13] "eTIV"       "nWBV"       "ASF"
```

Data Cleaning

I began by examining the dataset to see if there were equal numbers of values in each column. The code listed below was used to look for any specific missing data.

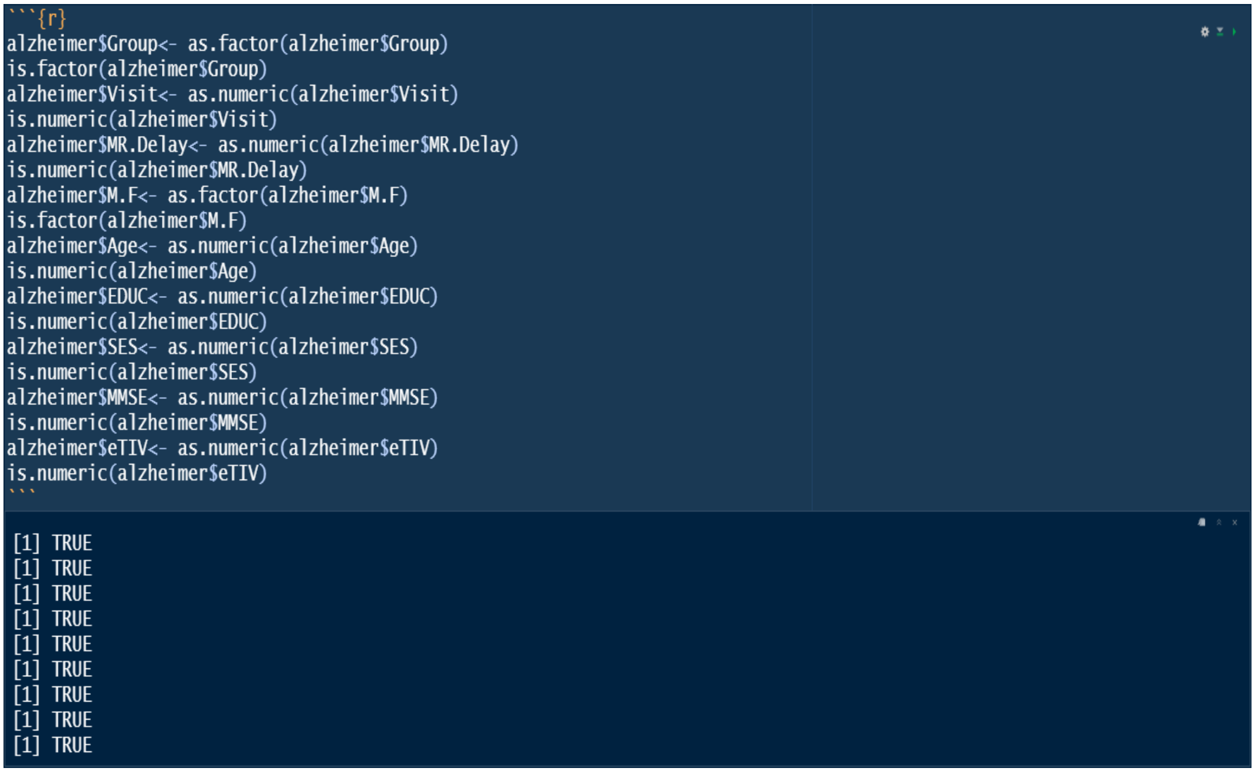


~~~
```{r}
colSums(sapply(alzheimer, is.na))
```
~~~

```
Subject.ID     MRI.ID      Group      Visit
         0          0          0          0
  MR.Delay        M.F       Hand        Age
         0          0          0          0
      EDUC        SES       MMSE        CDR
         0         19          2          0
      eTIV       nWBV        ASF
         0          0          0
```

The SES and MMSE features both lack some values. It is impossible just to fill in the blanks with random values because they are taken from a real OASIS study. The data will continue to be cleaned up.

As integers were used to record the following variables, "Visit," "MR Delay," "Age," "EDUC," "SES," "MMSE," and “eTIV,” the mode was changed from "int" to "num" using the as. numeric () function. The variables "Group" and "M. F” were saved as characters to make easy changes during the investigation possible. It has therefore been treated as a string value. We will use the as. factor () function to change the variable from qualitative to quantitative. Two of those columns are ID which will not be used in the dataset analysis and will be left as chr. The command that was executed to modify the dataset is shown below.

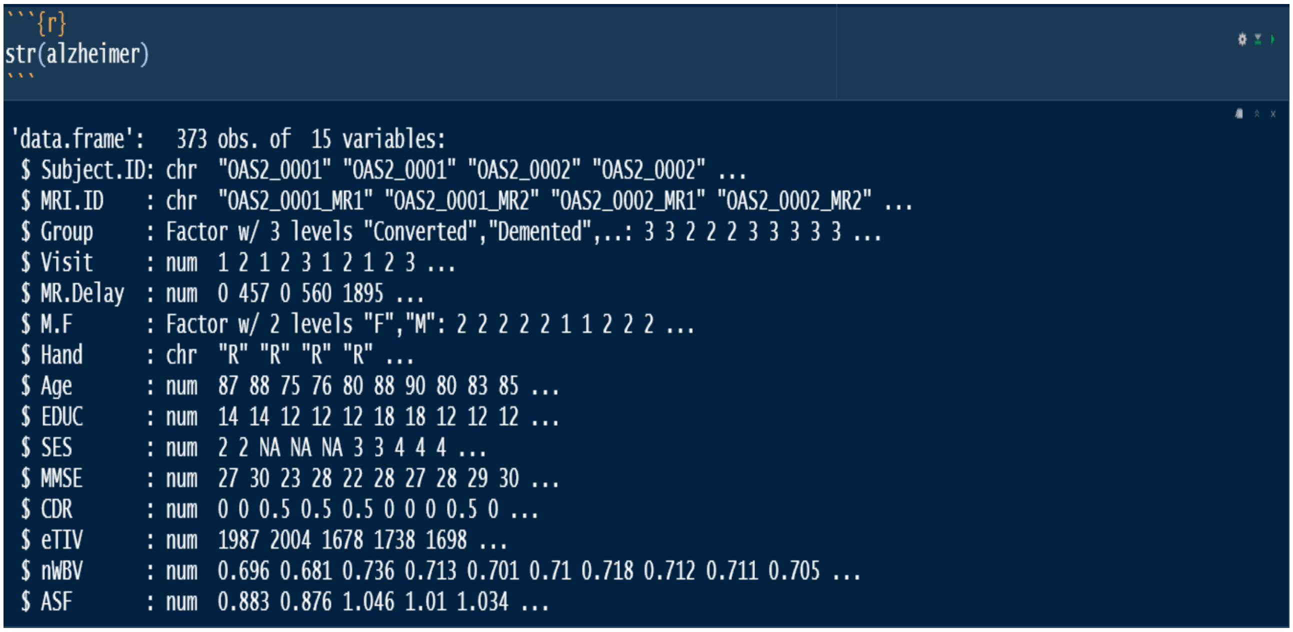


~~~
```{r}
alzheimer$Group<- as.factor(alzheimer$Group)
is.factor(alzheimer$Group)
alzheimer$Visit<- as.numeric(alzheimer$Visit)
is.numeric(alzheimer$Visit)
alzheimer$MR.Delay<- as.numeric(alzheimer$MR.Delay)
is.numeric(alzheimer$MR.Delay)
alzheimer$M.F<- as.factor(alzheimer$M.F)
is.factor(alzheimer$M.F)
alzheimer$Age<- as.numeric(alzheimer$Age)
is.numeric(alzheimer$Age)
alzheimer$EDUC<- as.numeric(alzheimer$EDUC)
is.numeric(alzheimer$EDUC)
alzheimer$SES<- as.numeric(alzheimer$SES)
is.numeric(alzheimer$SES)
alzheimer$MMSE<- as.numeric(alzheimer$MMSE)
is.numeric(alzheimer$MMSE)
alzheimer$eTIV<- as.numeric(alzheimer$eTIV)
is.numeric(alzheimer$eTIV)
```
~~~

```
[1] TRUE
[1] TRUE
[1] TRUE
[1] TRUE
[1] TRUE
[1] TRUE
[1] TRUE
[1] TRUE
[1] TRUE
```

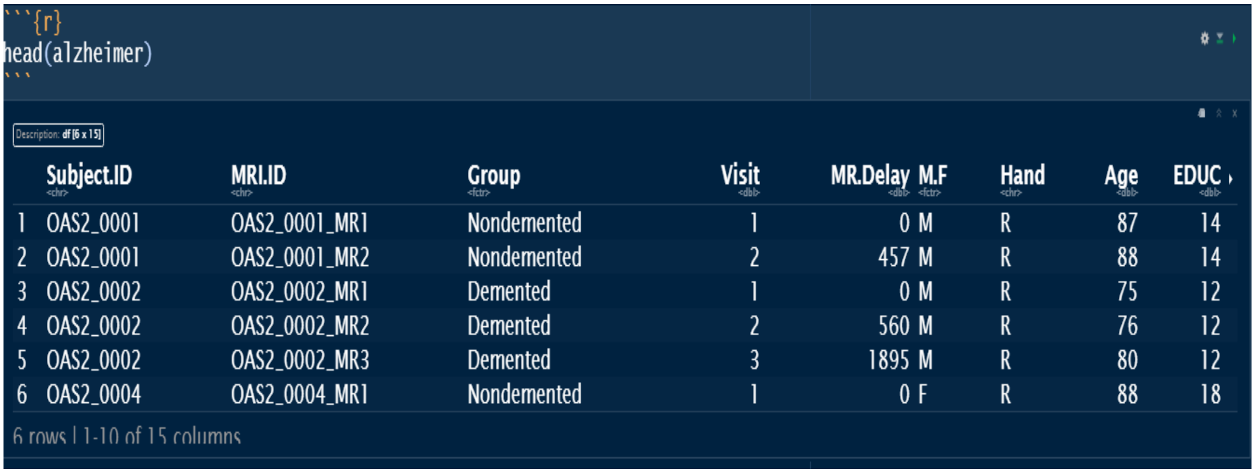
Cleaned Dataset

The dataset's new data structure was provided using the str () function. It was also possible to view the dataset's first and last six observations using the head () and tail () functions.

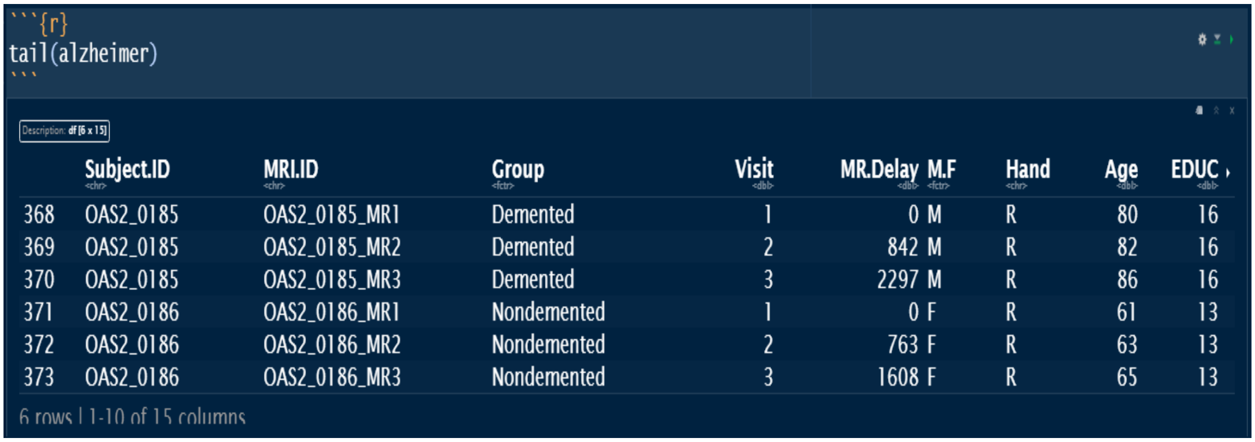


~~~
```{r}
str(alzheimer)
```
~~~

```
'data.frame':	373 obs. of  15 variables:
 $ Subject.ID: chr  "OAS2_0001" "OAS2_0001" "OAS2_0002" "OAS2_0002" ...
 $ MRI.ID    : chr  "OAS2_0001_MR1" "OAS2_0001_MR2" "OAS2_0002_MR1" "OAS2_0002_MR2" ...
 $ Group     : Factor w/ 3 levels "Converted","Demented",..: 3 3 2 2 2 3 3 3 3 3 ...
 $ Visit     : num  1 2 1 2 3 1 2 1 2 3 ...
 $ MR.Delay  : num  0 457 0 560 1895 ...
 $ M.F       : Factor w/ 2 levels "F","M": 2 2 2 2 2 1 1 2 2 2 ...
 $ Hand      : chr  "R" "R" "R" "R" ...
 $ Age       : num  87 88 75 76 80 88 90 80 83 85 ...
 $ EDUC      : num  14 14 12 12 12 18 18 12 12 12 ...
 $ SES       : num  2 2 NA NA NA 3 3 4 4 4 ...
 $ MMSE      : num  27 30 23 28 22 28 27 28 29 30 ...
 $ CDR       : num  0 0 0.5 0.5 0.5 0 0 0 0.5 0 ...
 $ eTIV      : num  1987 2004 1678 1738 1698 ...
 $ nWBV      : num  0.696 0.681 0.736 0.713 0.701 0.71 0.718 0.712 0.711 0.705 ...
 $ ASF       : num  0.883 0.876 1.046 1.01 1.034 ...
```



~~~
```{r}
head(alzheimer)
```

Description: df [6 x 15]

  Subject.ID  MRI.ID         Group        Visit  MR.Delay M.F  Hand  Age  EDUC ▸
  <chr>       <chr>          <fctr>       <dbl>  <dbl>    <fctr> <chr> <dbl> <dbl>
1 OAS2_0001   OAS2_0001_MR1  Nondemented      1         0 M    R      87   14
2 OAS2_0001   OAS2_0001_MR2  Nondemented      2       457 M    R      88   14
3 OAS2_0002   OAS2_0002_MR1  Demented         1         0 M    R      75   12
4 OAS2_0002   OAS2_0002_MR2  Demented         2       560 M    R      76   12
5 OAS2_0002   OAS2_0002_MR3  Demented         3      1895 M    R      80   12
6 OAS2_0004   OAS2_0004_MR1  Nondemented      1         0 F    R      88   18
6 rows | 1-10 of 15 columns
~~~



~~~
```{r}
tail(alzheimer)
```

Description: df [6 x 15]

    Subject.ID  MRI.ID         Group        Visit  MR.Delay M.F  Hand  Age  EDUC ▸
    <chr>       <chr>          <fctr>       <dbl>  <dbl>    <fctr> <chr> <dbl> <dbl>
368 OAS2_0185   OAS2_0185_MR1  Demented         1         0 M    R      80   16
369 OAS2_0185   OAS2_0185_MR2  Demented         2       842 M    R      82   16
370 OAS2_0185   OAS2_0185_MR3  Demented         3      2297 M    R      86   16
371 OAS2_0186   OAS2_0186_MR1  Nondemented      1         0 F    R      61   13
372 OAS2_0186   OAS2_0186_MR2  Nondemented      2       763 F    R      63   13
373 OAS2_0186   OAS2_0186_MR3  Nondemented      3      1608 F    R      65   13
6 rows | 1-10 of 15 columns
~~~

Dataset Description

The longitudinal study ranged in age from 60 to 96 at the OASIS laboratory. The dataset is broken down into many variables in the table below.

Variable Names	Variable Description
ID	Identification
M/F	Gender (M for Male, F for female
Hand	Handedness (L for left, R for right)
Age	Age in years
EDUC	Years of Education
SES	Socioeconomic Status

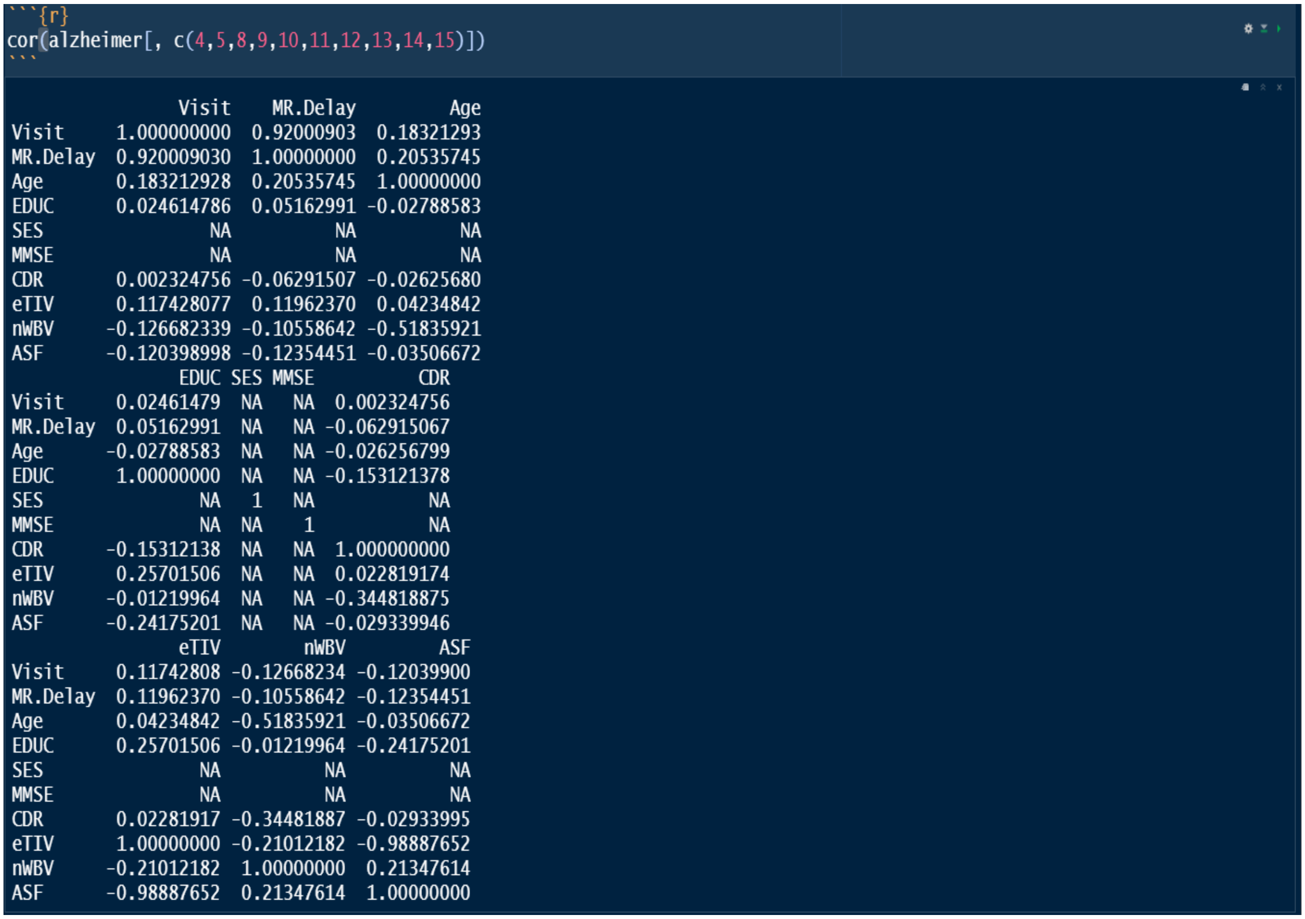
MMSE	Mini-Mental State Examination
CDR	Clinical Dementia Rating
eTIV	Estimated Total Intracranial Volume
nWBV	Normalize Whole Brain Volume
ASF	Atlas Scaling Factor
MR Delay	Delay

Expectations

This data analysis intends to provide a trustworthy and accurate model to estimate the primary cause of Alzheimer's/Dementia in persons between the ages of 60 and 96, which will be of paramount relevance to everyone inside and outside of the US. People are curious about the main component of Alzheimer's and how some of these characteristics affect Alzheimer's. To determine whether some of the causes of Alzheimer's are thought to exist. Health professionals can also help patients by advising them on what to look out for in the early identification of Alzheimer's by being aware of the leading causes of Alzheimer's cases. My domain knowledge leads me to believe there should be a negative and positive relationship between the CDR, the response variable, and the independent factors or predictor variable. I think that the patient's left or right-handedness shouldn't be related to their diagnosis of Alzheimer's. The factors in which I am most interested are age, education, M.F. (gender), ASF, Visit, and nWBV because I want to know if they have an impact on a patient's Alzheimer’s status.

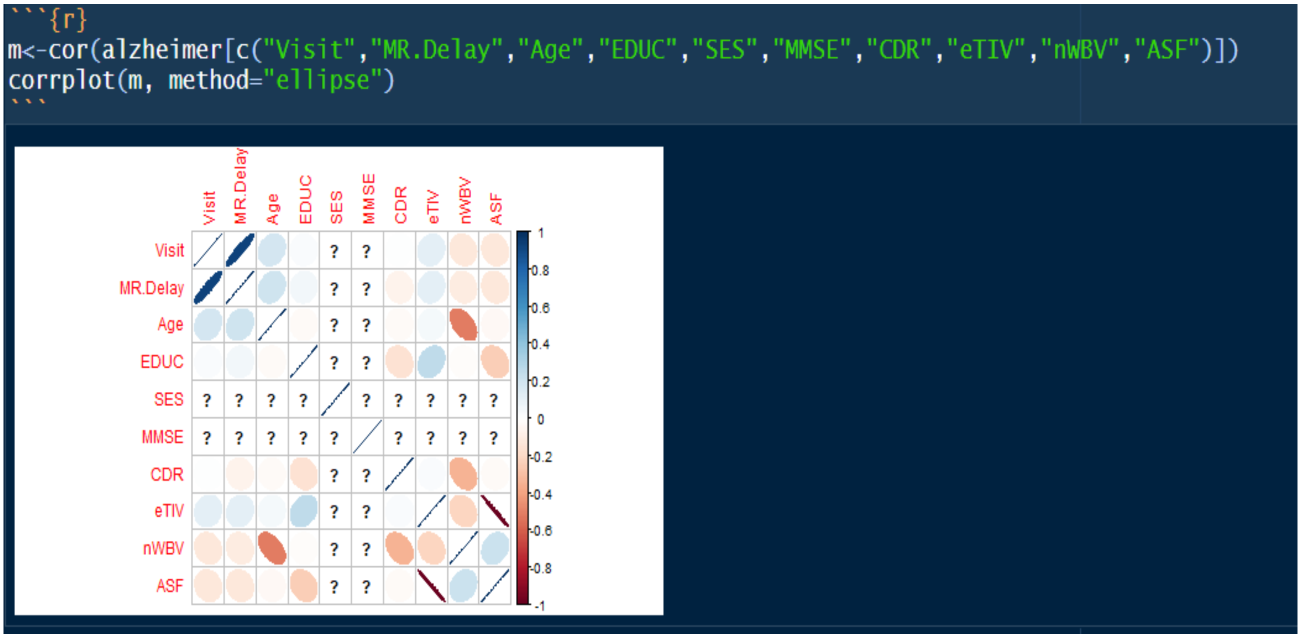
Data Analysis

Plotting Continuous variable for the Dataset



~~~
```{r}
cor(alzheimer[, c(4,5,8,9,10,11,12,13,14,15)])
```
~~~

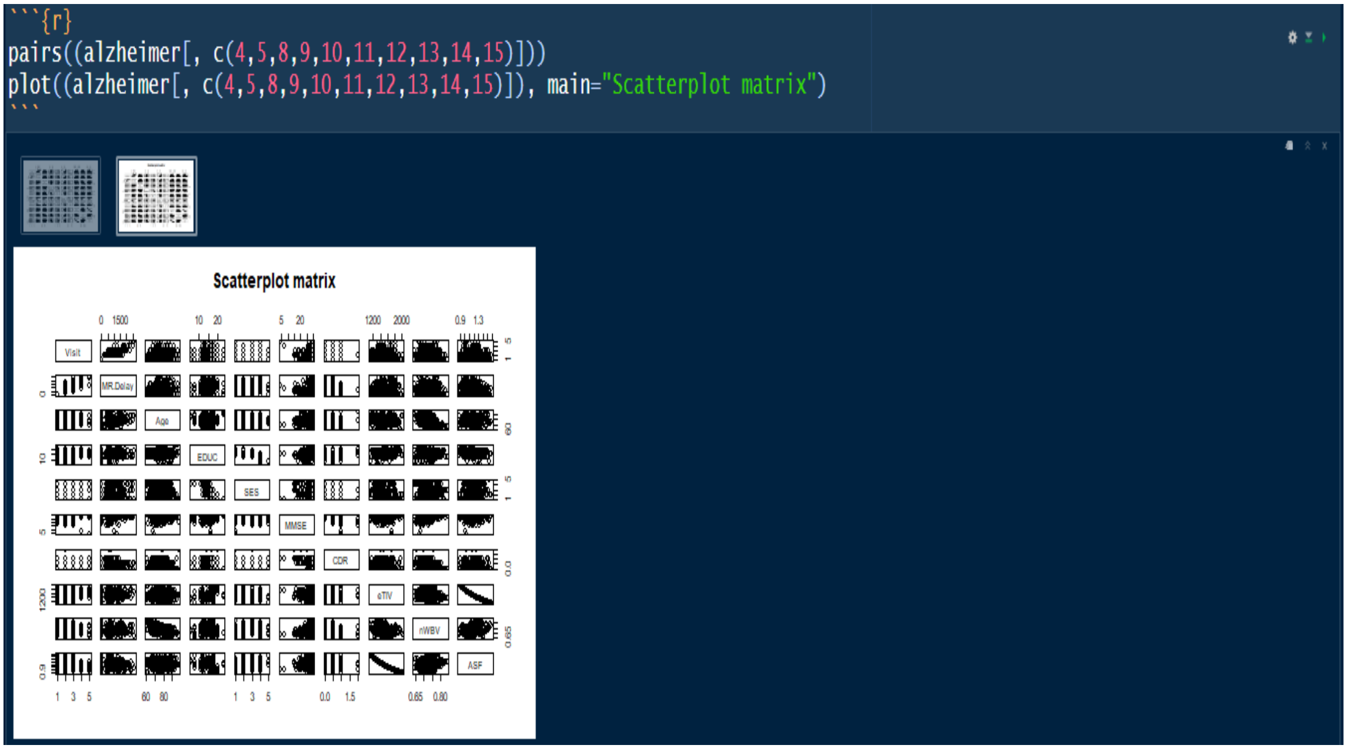
```
                Visit    MR.Delay         Age
Visit     1.000000000  0.92000903  0.18321293
MR.Delay  0.920009030  1.00000000  0.20535745
Age       0.183212928  0.20535745  1.00000000
EDUC      0.024614786  0.05162991 -0.02788583
SES                NA          NA          NA
MMSE               NA          NA          NA
CDR       0.002324756 -0.06291507 -0.02625680
eTIV      0.117428077  0.11962370  0.04234842
nWBV     -0.126682339 -0.10558642 -0.51835921
ASF      -0.120398998 -0.12354451 -0.03506672
                EDUC SES MMSE          CDR
Visit     0.02461479  NA   NA  0.002324756
MR.Delay  0.05162991  NA   NA -0.062915067
Age      -0.02788583  NA   NA -0.026256799
EDUC      1.00000000  NA   NA -0.153121378
SES               NA   1   NA           NA
MMSE              NA  NA    1           NA
CDR      -0.15312138  NA   NA  1.000000000
eTIV      0.25701506  NA   NA  0.022819174
nWBV     -0.01219964  NA   NA -0.344818875
ASF      -0.24175201  NA   NA -0.029339946
                eTIV        nWBV         ASF
Visit     0.11742808 -0.12668234 -0.12039900
MR.Delay  0.11962370 -0.10558642 -0.12354451
Age       0.04234842 -0.51835921 -0.03506672
EDUC      0.25701506 -0.01219964 -0.24175201
SES               NA          NA          NA
MMSE              NA          NA          NA
CDR       0.02281917 -0.34481887 -0.02933995
eTIV      1.00000000 -0.21012182 -0.98887652
nWBV     -0.21012182  1.00000000  0.21347614
ASF      -0.98887652  0.21347614  1.00000000
```



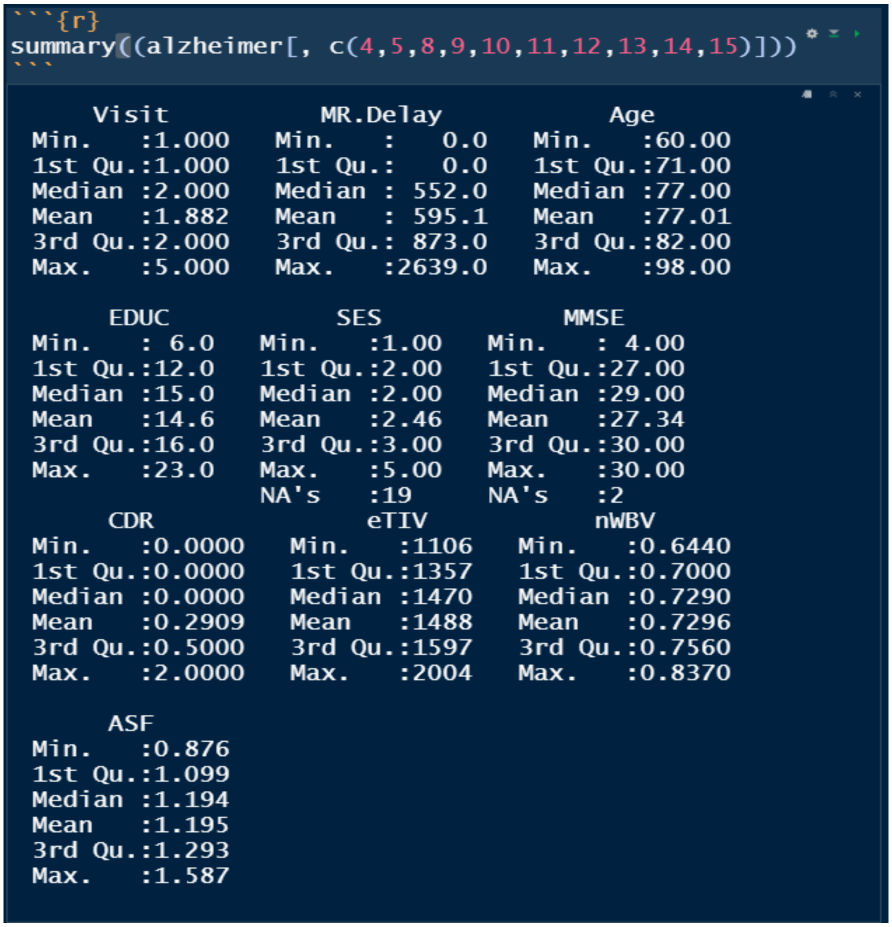
~~~
```{r}
m<-cor(alzheimer[c("Visit","MR.Delay","Age","EDUC","SES","MMSE","CDR","eTIV","nWBV","ASF")])
corrplot(m, method="ellipse")
```
~~~

The correlation matrix, which the correlation test generated, showed some weak negative correlations between the dataset's numerical variables. Age and CDR have a weakly negative association. The CDR and EDUC likewise have a slender negative association. Additionally, there is a weak positive association between the eTIV variable and EDUC. Numerous other factors, such as the MMSE and eTIV variables, are not correlated, meaning there is no discernible connection between the variables. Last, there is a strong negative association between the Age and nWBV variables.

Creating a scatter plot matrix figure for the numerical using the pairs () function.

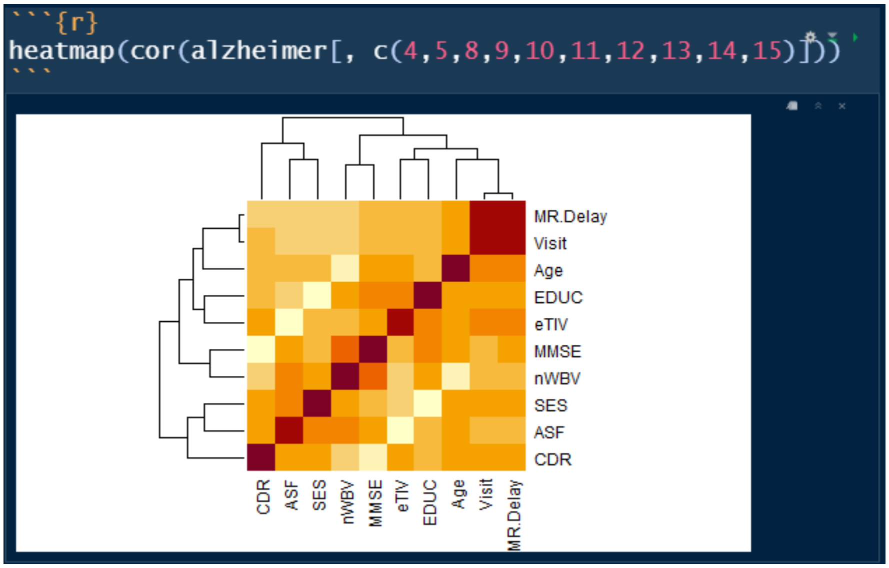


```{r}
pairs((alzheimer[, c(4,5,8,9,10,11,12,13,14,15)]))
plot((alzheimer[, c(4,5,8,9,10,11,12,13,14,15)]), main="Scatterplot matrix")
```



~~~
```{r}
summary((alzheimer[, c(4,5,8,9,10,11,12,13,14,15)]))
```

     Visit          MR.Delay            Age       
 Min.   :1.000   Min.   :   0.0   Min.   :60.00  
 1st Qu.:1.000   1st Qu.:   0.0   1st Qu.:71.00  
 Median :2.000   Median : 552.0   Median :77.00  
 Mean   :1.882   Mean   : 595.1   Mean   :77.01  
 3rd Qu.:2.000   3rd Qu.: 873.0   3rd Qu.:82.00  
 Max.   :5.000   Max.   :2639.0   Max.   :98.00  
                                                 
      EDUC           SES             MMSE      
 Min.   : 6.0   Min.   :1.00   Min.   : 4.00  
 1st Qu.:12.0   1st Qu.:2.00   1st Qu.:27.00  
 Median :15.0   Median :2.00   Median :29.00  
 Mean   :14.6   Mean   :2.46   Mean   :27.34  
 3rd Qu.:16.0   3rd Qu.:3.00   3rd Qu.:30.00  
 Max.   :23.0   Max.   :5.00   Max.   :30.00  
                NA's   :19     NA's   :2      
      CDR              eTIV           nWBV       
 Min.   :0.0000   Min.   :1106   Min.   :0.6440  
 1st Qu.:0.0000   1st Qu.:1357   1st Qu.:0.7000  
 Median :0.0000   Median :1470   Median :0.7290  
 Mean   :0.2909   Mean   :1488   Mean   :0.7296  
 3rd Qu.:0.5000   3rd Qu.:1597   3rd Qu.:0.7560  
 Max.   :2.0000   Max.   :2004   Max.   :0.8370  
                                                 
      ASF       
 Min.   :0.876  
 1st Qu.:1.099  
 Median :1.194  
 Mean   :1.195  
 3rd Qu.:1.293  
 Max.   :1.587  
~~~

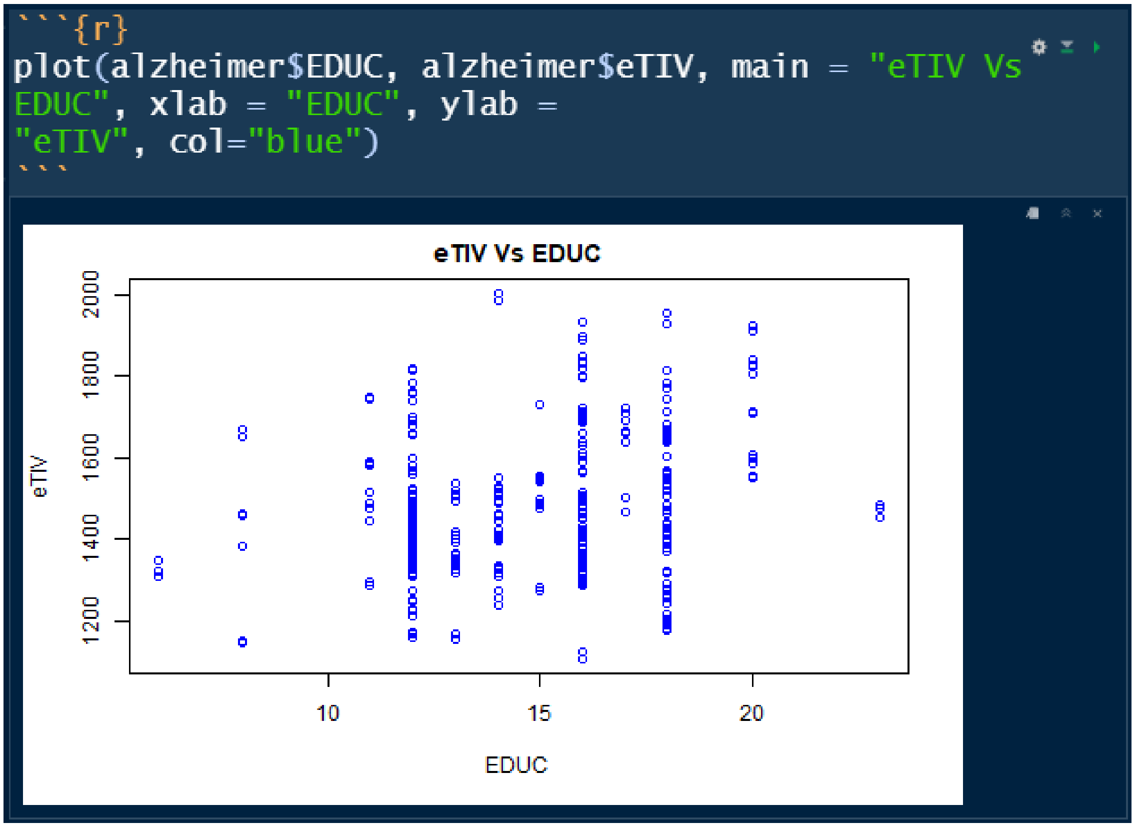
Using the heatmap () function, I created a heat map showing how the numerical variables correlated.



~~~
```{r}
heatmap(cor(alzheimer[, c(4,5,8,9,10,11,12,13,14,15)]))
```
~~~

I conducted extra analysis by making individual plots for some variables that looked related, according to the pairs () approach used above for more Visualization.

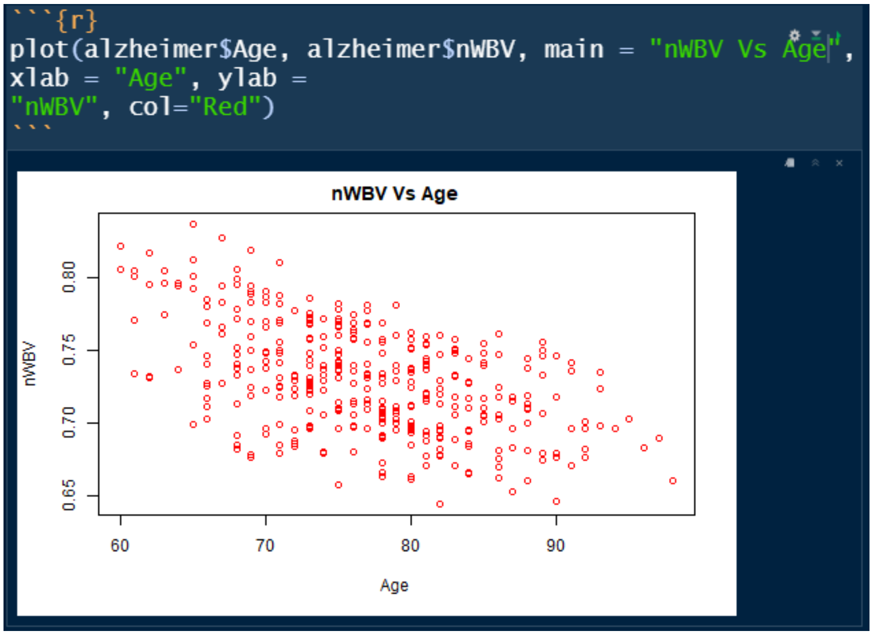
eTIV vs EDUC



~~~
```{r}
plot(alzheimer$EDUC, alzheimer$eTIV, main = "eTIV Vs
EDUC", xlab = "EDUC", ylab =
"eTIV", col="blue")
```
~~~

This graph demonstrates a weak positive correlation between eTIV and the individuals' educational attainment (EDUC). The figure shows that the correlation between the two variables is positive, with a value of 0.25701506 in the correlation matrix.

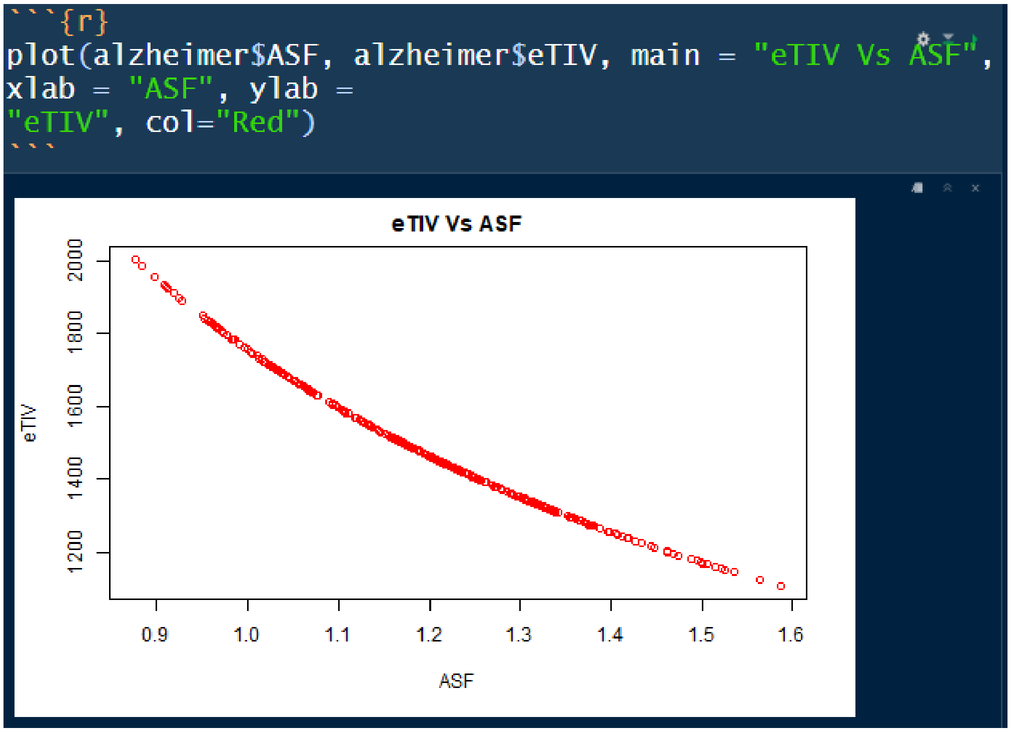
nWBV vs. Age



~~~
```{r}
plot(alzheimer$Age, alzheimer$nWBV, main = "nWBV Vs Age",
xlab = "Age", ylab =
"nWBV", col="Red")
```
~~~

Age and nWBV are pitted against one another in the graph above. The association appears to be moderately negative in the above chart. The correlation coefficient is a moderately negative association of -0.51835921.

eTIV vs. ASF

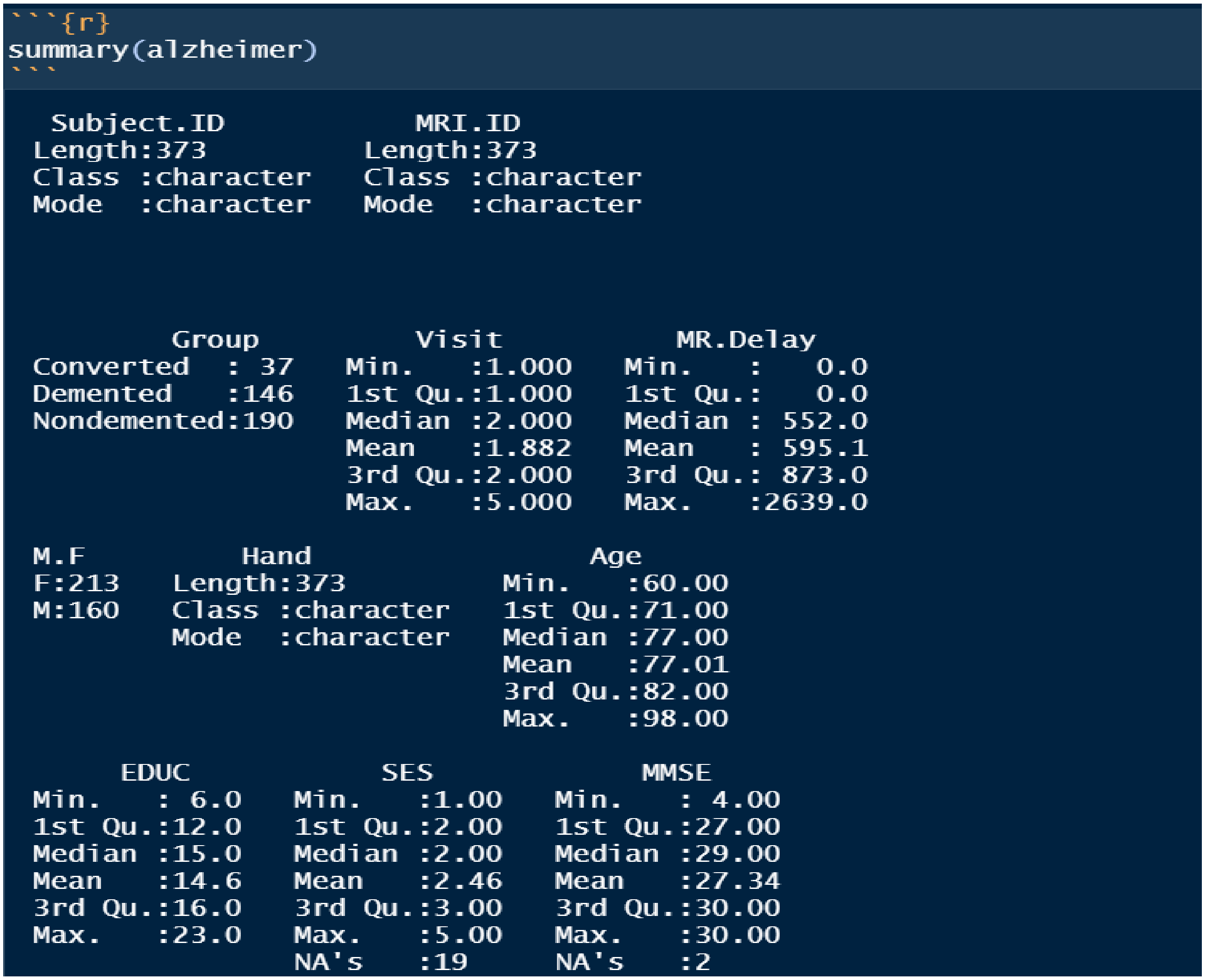


~~~
```{r}
plot(alzheimer$ASF, alzheimer$eTIV, main = "eTIV Vs ASF",
xlab = "ASF", ylab =
"eTIV", col="Red")
```
~~~

This explains a strong negative correlation between the computed scaling factor that converts the native-space brain and skull (ASF) to the atlas target and the estimated total intracranial volume of the patient's brain (eTIV). According to the correlation matrix above, the correlation coefficient for both variables is -0.98887652.

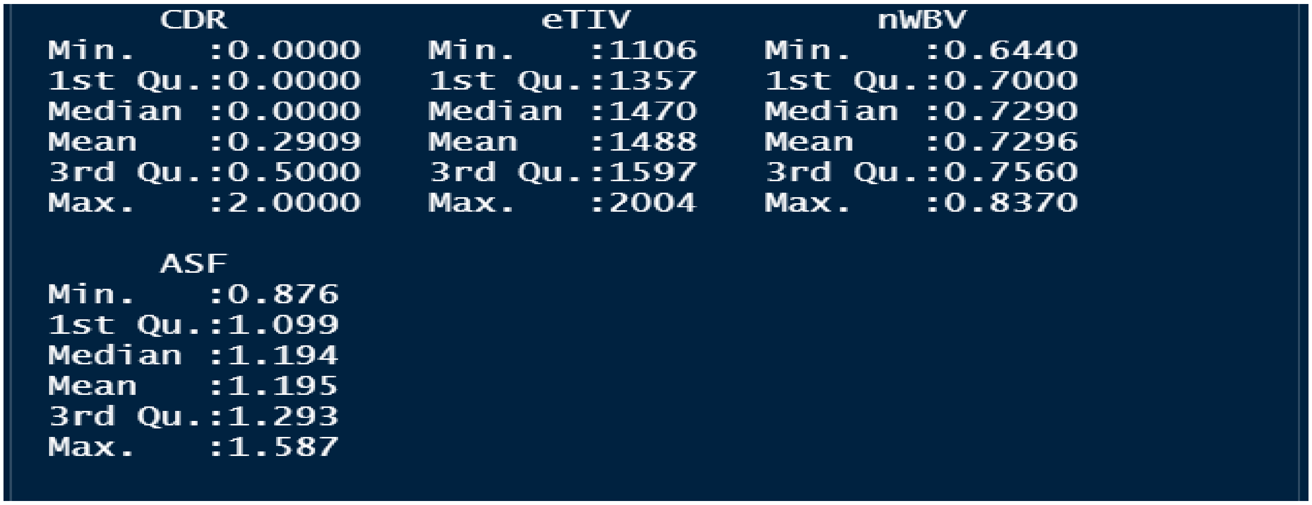
Summary Statistics of the variables and visualizations

To describe the statistical properties and distributions of the variables. The summary () tool was used to view the distribution of the dataset at a high level.



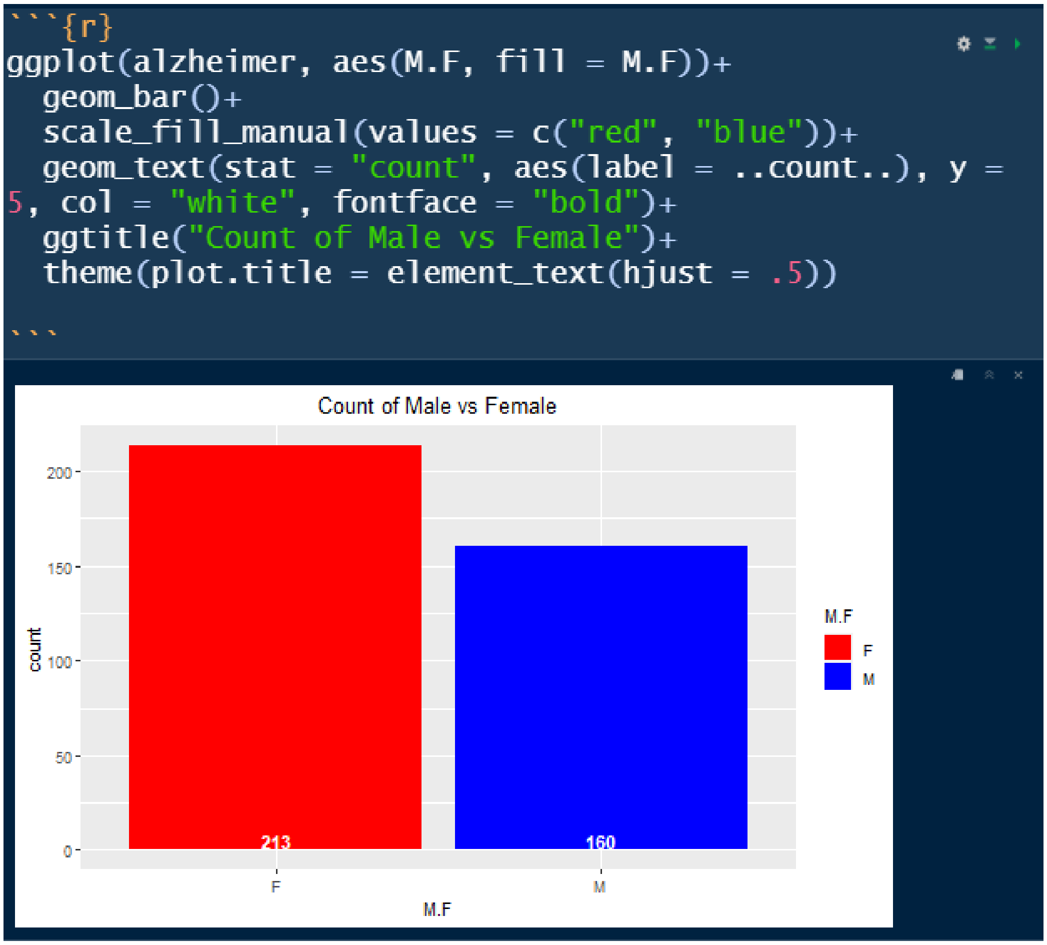
~~~
```{r}
summary(alzheimer)
```
~~~

```
  Subject.ID          MRI.ID
 Length:373         Length:373
 Class :character   Class :character
 Mode  :character   Mode  :character



         Group        Visit         MR.Delay
 Converted  : 37   Min.   :1.000   Min.   :   0.0
 Demented   :146   1st Qu.:1.000   1st Qu.:   0.0
 Nondemented:190   Median :2.000   Median : 552.0
                   Mean   :1.882   Mean   : 595.1
                   3rd Qu.:2.000   3rd Qu.: 873.0
                   Max.   :5.000   Max.   :2639.0

 M.F         Hand                 Age
 F:213   Length:373          Min.   :60.00
 M:160   Class :character   1st Qu.:71.00
         Mode  :character   Median :77.00
                            Mean   :77.01
                            3rd Qu.:82.00
                            Max.   :98.00

      EDUC           SES            MMSE
 Min.   : 6.0   Min.   :1.00   Min.   : 4.00
 1st Qu.:12.0   1st Qu.:2.00   1st Qu.:27.00
 Median :15.0   Median :2.00   Median :29.00
 Mean   :14.6   Mean   :2.46   Mean   :27.34
 3rd Qu.:16.0   3rd Qu.:3.00   3rd Qu.:30.00
 Max.   :23.0   Max.   :5.00   Max.   :30.00
                NA's   :19     NA's   :2
```



```
      CDR                  eTIV               nWBV       
 Min.   :0.0000   Min.   :1106   Min.   :0.6440  
 1st Qu.:0.0000   1st Qu.:1357   1st Qu.:0.7000  
 Median :0.0000   Median :1470   Median :0.7290  
 Mean   :0.2909   Mean   :1488   Mean   :0.7296  
 3rd Qu.:0.5000   3rd Qu.:1597   3rd Qu.:0.7560  
 Max.   :2.0000   Max.   :2004   Max.   :0.8370  

      ASF       
 Min.   :0.876  
 1st Qu.:1.099  
 Median :1.194  
 Mean   :1.195  
 3rd Qu.:1.293  
 Max.   :1.587  
```

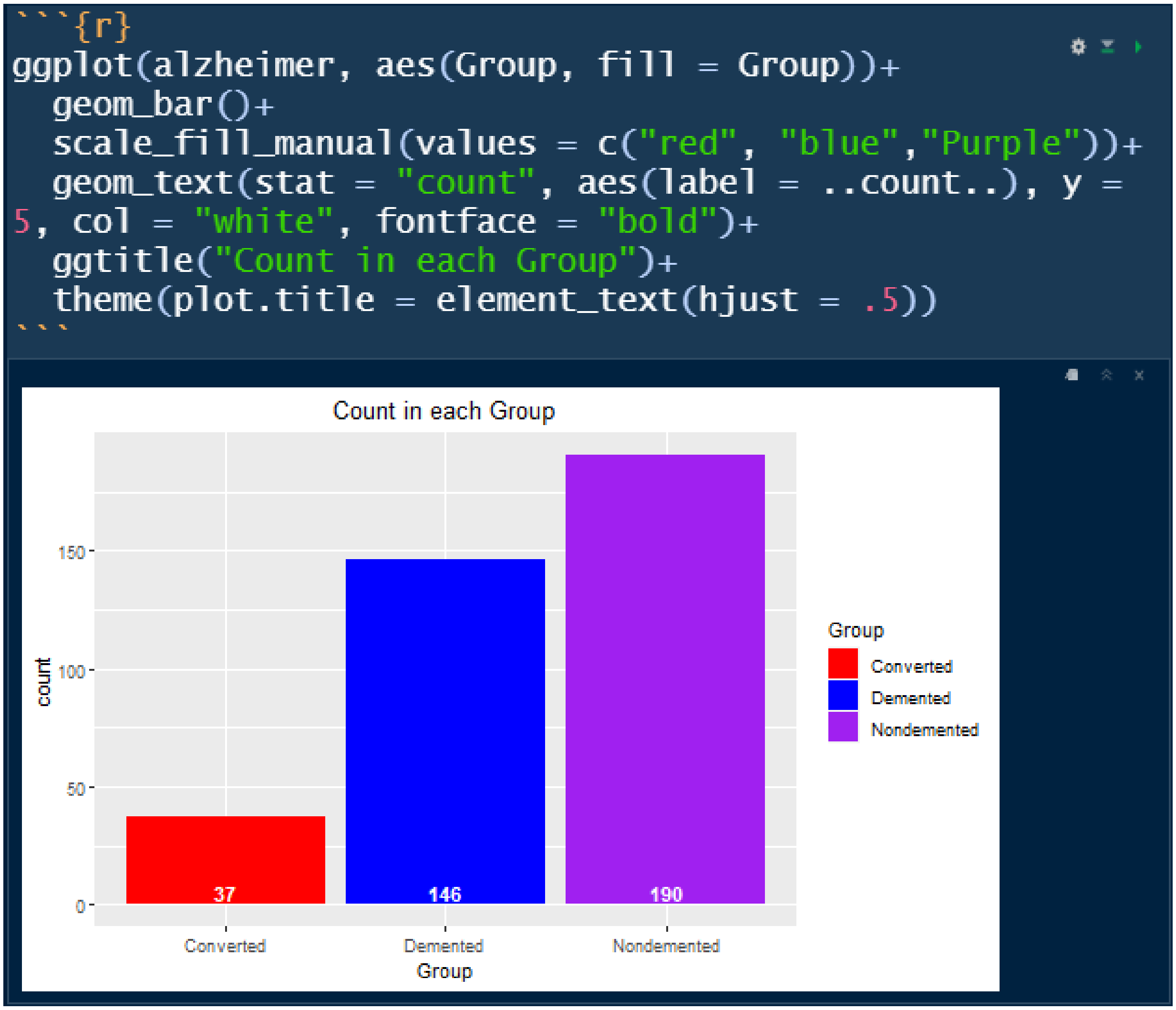
Count of Observation for Males and Females



~~~
```{r}
ggplot(alzheimer, aes(M.F, fill = M.F))+
  geom_bar()+
  scale_fill_manual(values = c("red", "blue"))+
  geom_text(stat = "count", aes(label = ..count..), y =
5, col = "white", fontface = "bold")+
  ggtitle("Count of Male vs Female")+
  theme(plot.title = element_text(hjust = .5))

```
~~~

There are 373 total observations, as may be seen in the barplot below. There are 213 observations for females and 160 for males. The information is well visualized in the graphic below.

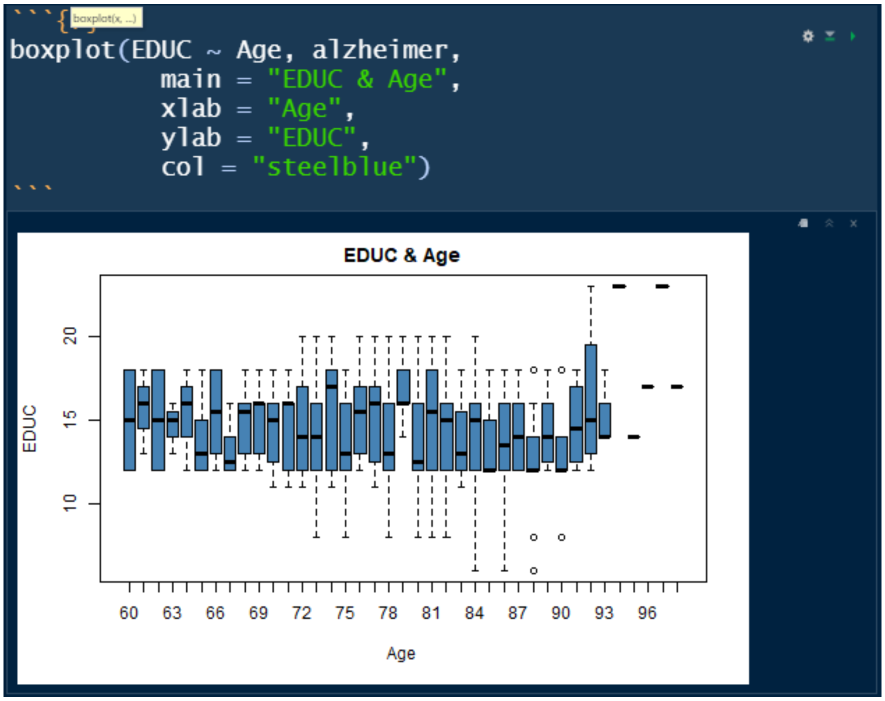
Number of cases in the Group



~~~
```{r}
ggplot(alzheimer, aes(Group, fill = Group))+
  geom_bar()+
  scale_fill_manual(values = c("red", "blue","Purple"))+
  geom_text(stat = "count", aes(label = ..count..), y =
5, col = "white", fontface = "bold")+
  ggtitle("Count in each Group")+
  theme(plot.title = element_text(hjust = .5))
```
~~~

This barplot displays the number of observations made by each group. According to the graph below, out of the 373 total observations, 37 are converted, 146 are demented, and 190 are not demented. A clear depiction of this information is shown in the figure below.

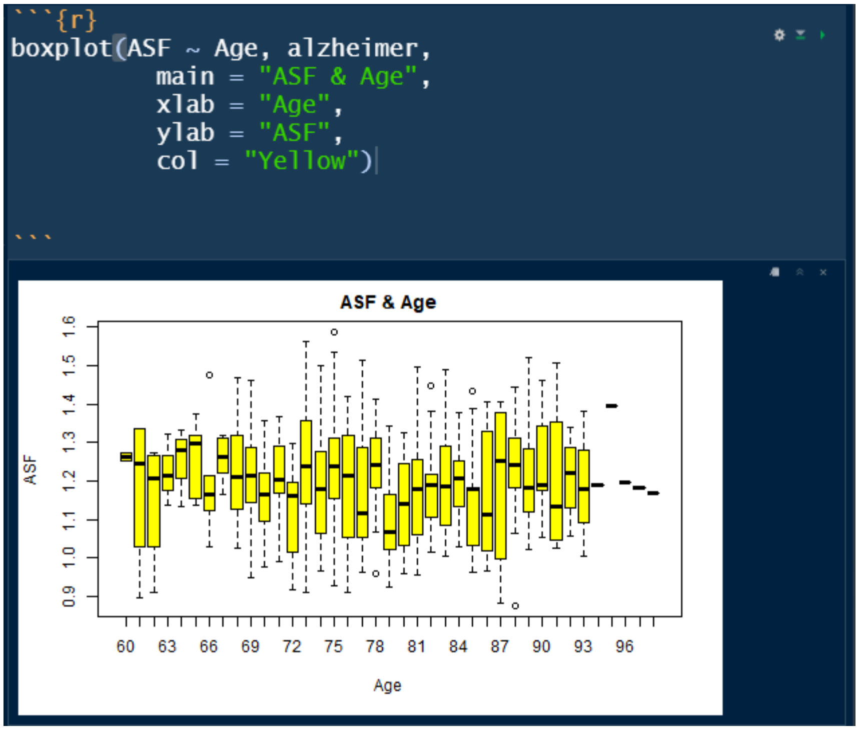
EDUC vs Age



~~~
```{
boxplot(EDUC ~ Age, alzheimer,
          main = "EDUC & Age",
          xlab = "Age",
          ylab = "EDUC",
          col = "steelblue")
```
~~~

The association between the EDUC and Age variables is depicted in the boxplot. According to the graph below, patients who are 94 years old or older have the greatest average education level.

ASF vs Age



~~~
```{r}
boxplot(ASF ~ Age, alzheimer,
         main = "ASF & Age",
         xlab = "Age",
         ylab = "ASF",
         col = "Yellow")


```
~~~

ASF and Age variables' relationships are displayed in a boxplot. The figure below shows that for the computed scaling factor, which converts the native-space brain and skull to the atlas target, the age range of 95 has the highest average result (ASF).

Analysis of the Continuous Variables with Categorical Variables

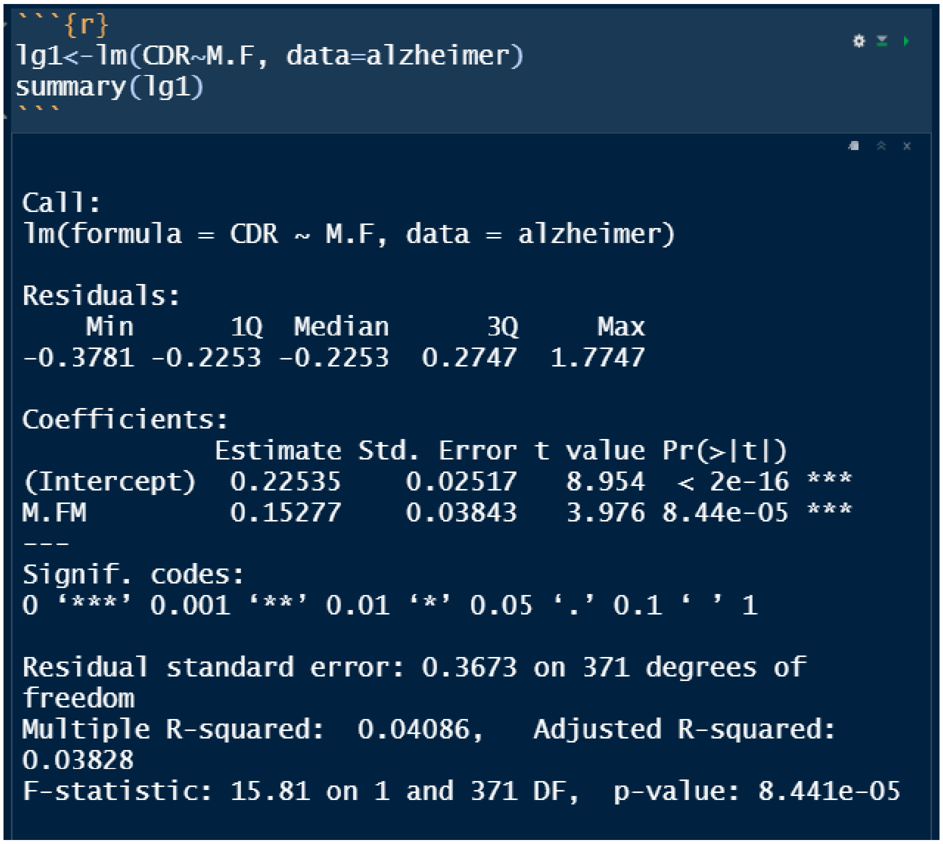
Comparing the categorical and numerical variables, I decided to compare the two categorical variables in the dataset with the dependent variable of the CDR. The categorical variables that are present in the data set are Group, and M.F. Plots against CDR (response variable) were made for each of these factors (used as predictor variables).

For M.F variable:

Based on the clinical dementia rate (CDR), I predict that the prevalence of Alzheimer's will be higher in men than in women. I employed simple linear regression and testing to verify this hypothesis.

H0: β1 = 0

H1: β1 ≠ 0



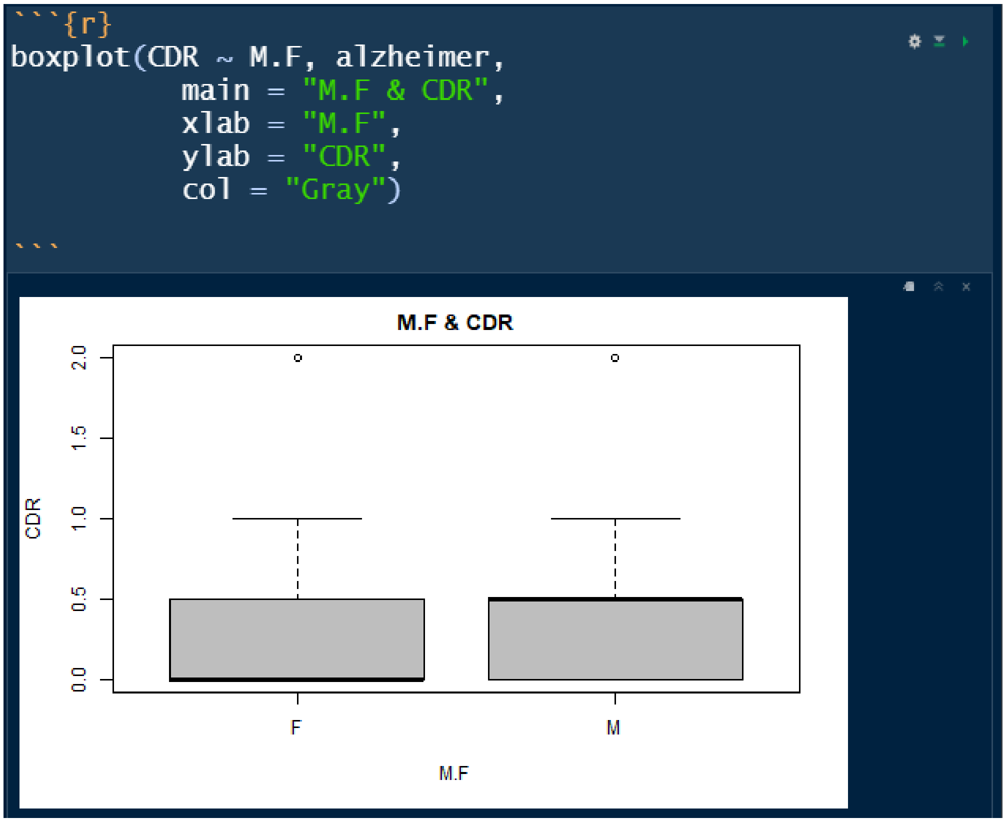
~~~
```{r}
lg1<-lm(CDR~M.F, data=alzheimer)
summary(lg1)
```

Call:
lm(formula = CDR ~ M.F, data = alzheimer)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.3781 -0.2253 -0.2253  0.2747  1.7747 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.22535    0.02517   8.954  < 2e-16 ***
M.FM         0.15277    0.03843   3.976 8.44e-05 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3673 on 371 degrees of
freedom
Multiple R-squared:  0.04086,	Adjusted R-squared:
0.03828 
F-statistic: 15.81 on 1 and 371 DF,  p-value: 8.441e-05
~~~

From the linear regression results, it can be inferred that the p-value associated with the M.F. is less than 0.05, and the t-value is low. As a result, we can rule out the null hypothesis and conclude that there is a connection between the patient's M.F. (gender) and the clinical dementia rate (CDR). It is possible to interpret the intercept β0 based on CDR as evidence of Alzheimer's disease in females and the intercept β0 + β1 as evidence of Alzheimer's disease in males with CDR.

The equation becomes:

CDR = 0.22535 + 0.15277 M.F + €

For the link between the Gender and CDR variables, I created a boxplot. In this study, the average CDR for men is higher than the average CDR for women.



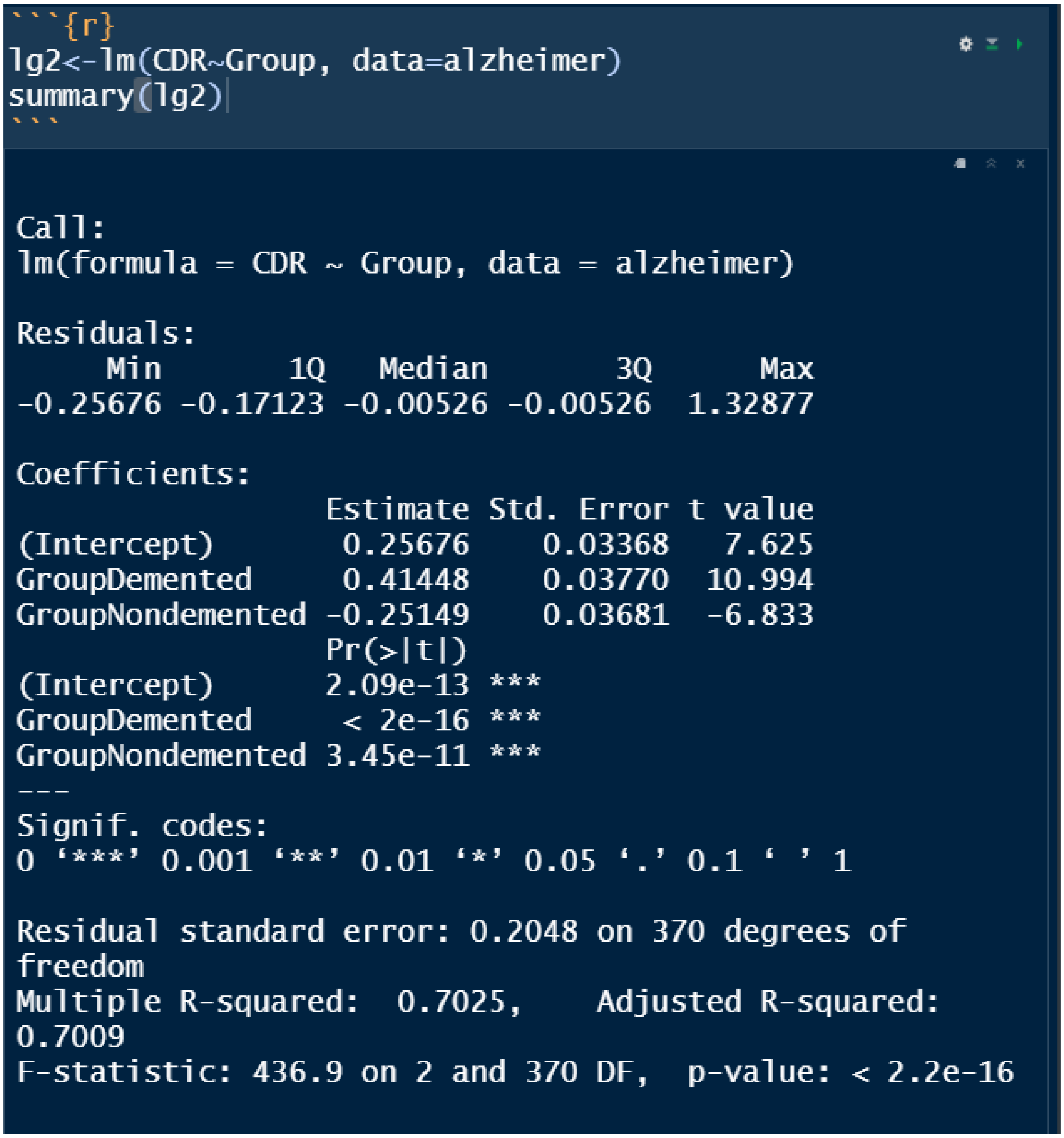
~~~
```{r}
boxplot(CDR ~ M.F, alzheimer,
          main = "M.F & CDR",
          xlab = "M.F",
          ylab = "CDR",
          col = "Gray")

```
~~~

For Group Variable:

I anticipated that people with dementia would have higher CDR values. I employed simple linear regression and testing to verify this hypothesis.

H0: β1 = β2 = 0

H1: β1 ≠ β2 ≠ 0



~~~
```{r}
lg2<-lm(CDR~Group, data=alzheimer)
summary(lg2)
```
~~~

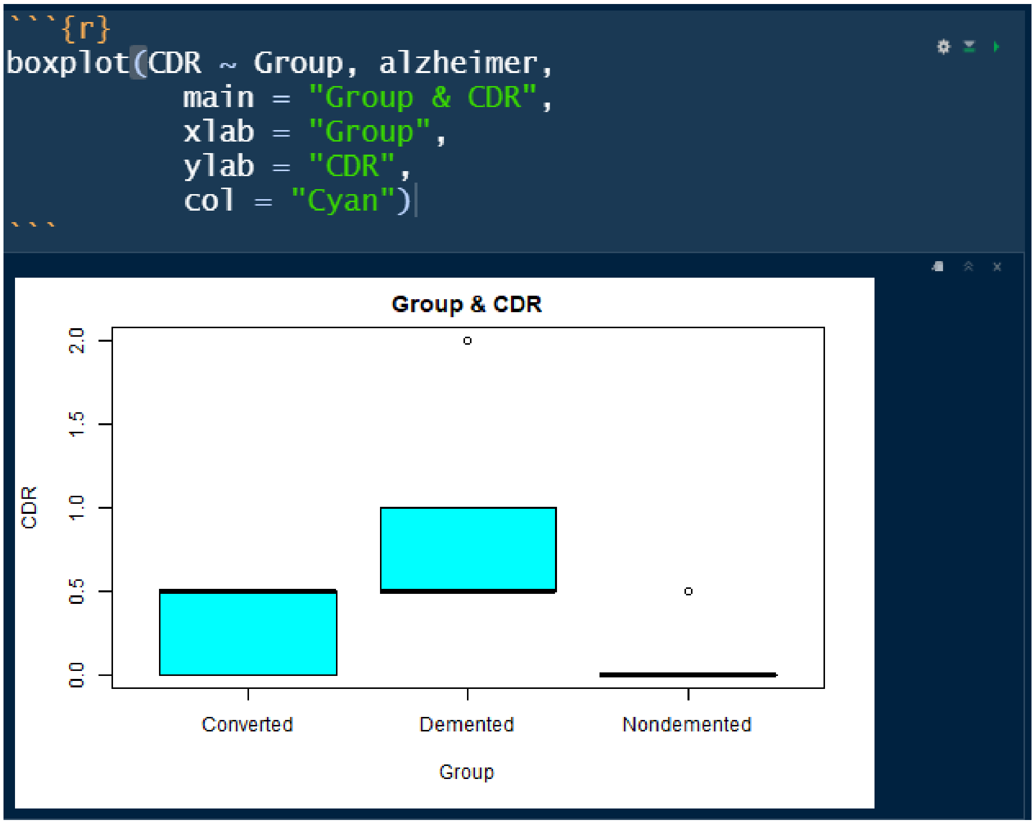
```
Call:
lm(formula = CDR ~ Group, data = alzheimer)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.25676 -0.17123 -0.00526 -0.00526  1.32877 

Coefficients:
                 Estimate Std. Error t value
(Intercept)       0.25676    0.03368   7.625
GroupDemented     0.41448    0.03770  10.994
GroupNondemented -0.25149    0.03681  -6.833
                 Pr(>|t|)    
(Intercept)      2.09e-13 ***
GroupDemented     < 2e-16 ***
GroupNondemented 3.45e-11 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2048 on 370 degrees of
freedom
Multiple R-squared:  0.7025,	Adjusted R-squared:
0.7009 
F-statistic: 436.9 on 2 and 370 DF,  p-value: < 2.2e-16
```

It is clear from the results of the linear regression that the Group variable has a significant p-value of less than 0.05 for the F-statistic. Thus, we conclude that the clinical dementia rate (CDR) and Alzheimer's status (Converted, Demented, and Nondemented) are significantly correlated.

The equation becomes:

CDR = 0.25676 + 0.41448 GroupDemented – 0.25149 GroupNondemented + €

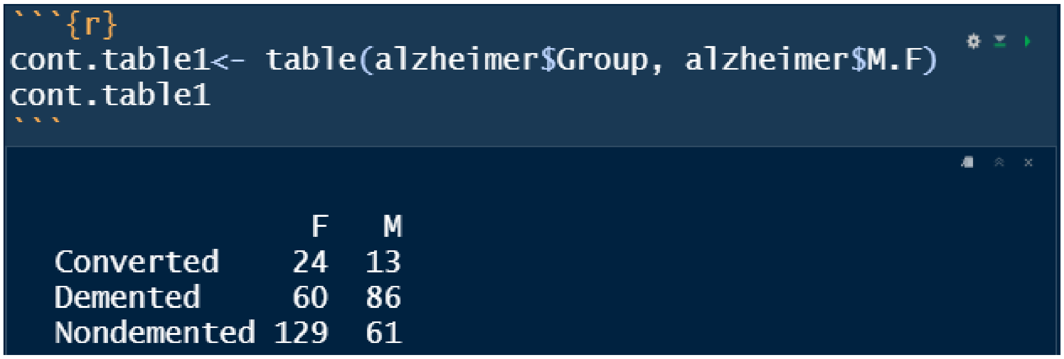
I plotted a boxplot for the relationship between the Group and CDR variables. The average CDR for the demented is the highest, showing that dementia increases with an increase in the value of CDR.



~~~
```{r}
boxplot(CDR ~ Group, alzheimer,
          main = "Group & CDR",
          xlab = "Group",
          ylab = "CDR",
          col = "Cyan")
```
~~~

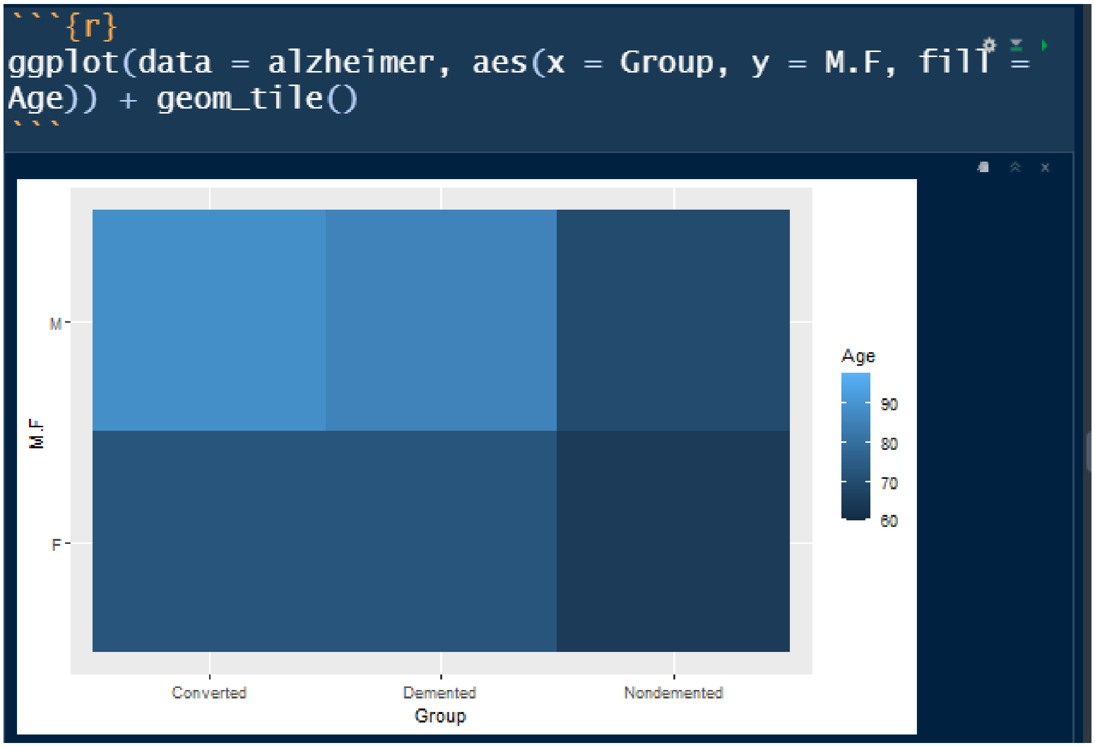
Contingency tables for Categorical Variables.

The process examines the association between two categorical variables using the contingency table.

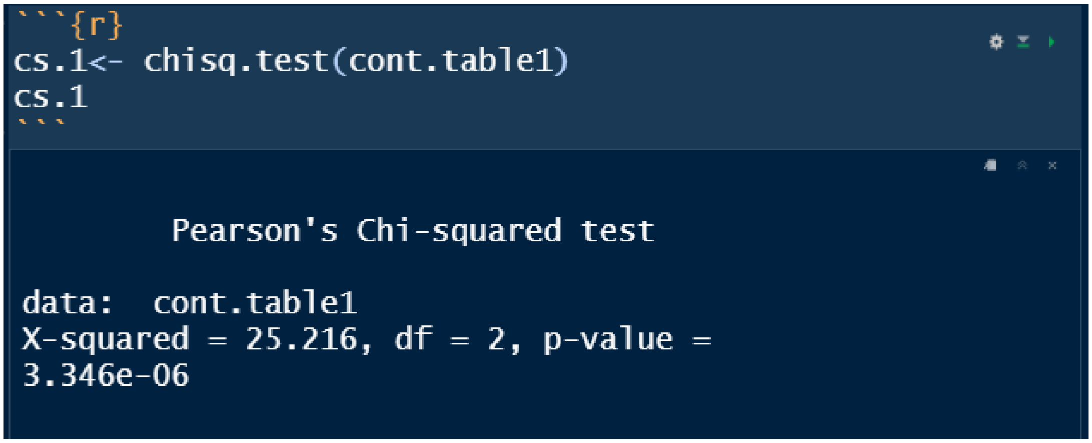


~~~
```{r}
cont.table1<- table(alzheimer$Group, alzheimer$M.F)
cont.table1
```
~~~

```
              F   M
  Converted    24  13
  Demented     60  86
  Nondemented 129  61
```



~~~
```{r}
ggplot(data = alzheimer, aes(x = Group, y = M.F, fill =
Age)) + geom_tile()
```
~~~



~~~
```{r}
cs.1<- chisq.test(cont.table1)
cs.1
```

        Pearson's Chi-squared test

data:  cont.table1
X-squared = 25.216, df = 2, p-value =
3.346e-06
~~~

Multiple Linear regression

I will employ a mixed selection method to choose the ideal model for my dataset.

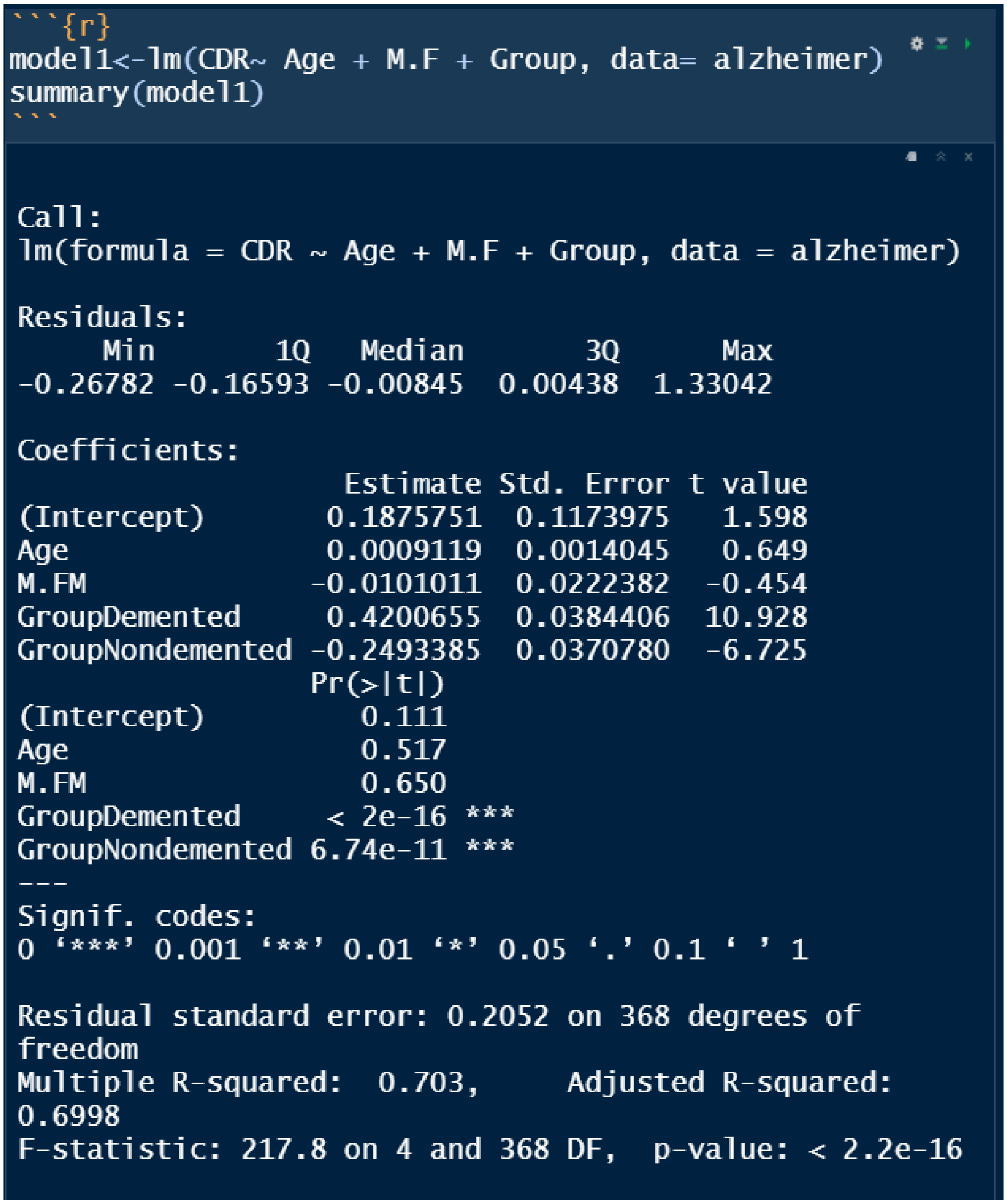
Model 1

$H0: \beta1 = \beta2 = \beta3 = \beta3 = \beta4 = 0$

$H1: \beta1 \neq \beta2 \neq \beta3 \neq \beta3 \neq \beta4 \neq 0$

The equation becomes:

$Y= \beta0 + \beta1Age + \beta2M.F + \beta3GroupDemented + \beta4GroupNondemented + \text{€}$



~~~
```{r}
model1<-lm(CDR~ Age + M.F + Group, data= alzheimer)
summary(model1)
```
~~~

```
Call:
lm(formula = CDR ~ Age + M.F + Group, data = alzheimer)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.26782 -0.16593 -0.00845  0.00438  1.33042 

Coefficients:
                   Estimate Std. Error t value
(Intercept)       0.1875751  0.1173975   1.598
Age               0.0009119  0.0014045   0.649
M.FM             -0.0101011  0.0222382  -0.454
GroupDemented     0.4200655  0.0384406  10.928
GroupNondemented -0.2493385  0.0370780  -6.725
                 Pr(>|t|)    
(Intercept)         0.111    
Age                 0.517    
M.FM                0.650    
GroupDemented     < 2e-16 ***
GroupNondemented 6.74e-11 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2052 on 368 degrees of 
freedom
Multiple R-squared:  0.703,	Adjusted R-squared:  
0.6998 
F-statistic: 217.8 on 4 and 368 DF,  p-value: < 2.2e-16
```

This multiple linear regression results indicated that the p-value and the variables' F-statistic are related. Given that the t-statistics have a low p-value, I reject the null hypothesis. Since Group is a significant variable and the t-statistics have a high p-value, the link between CDR and age and gender is unaffected.

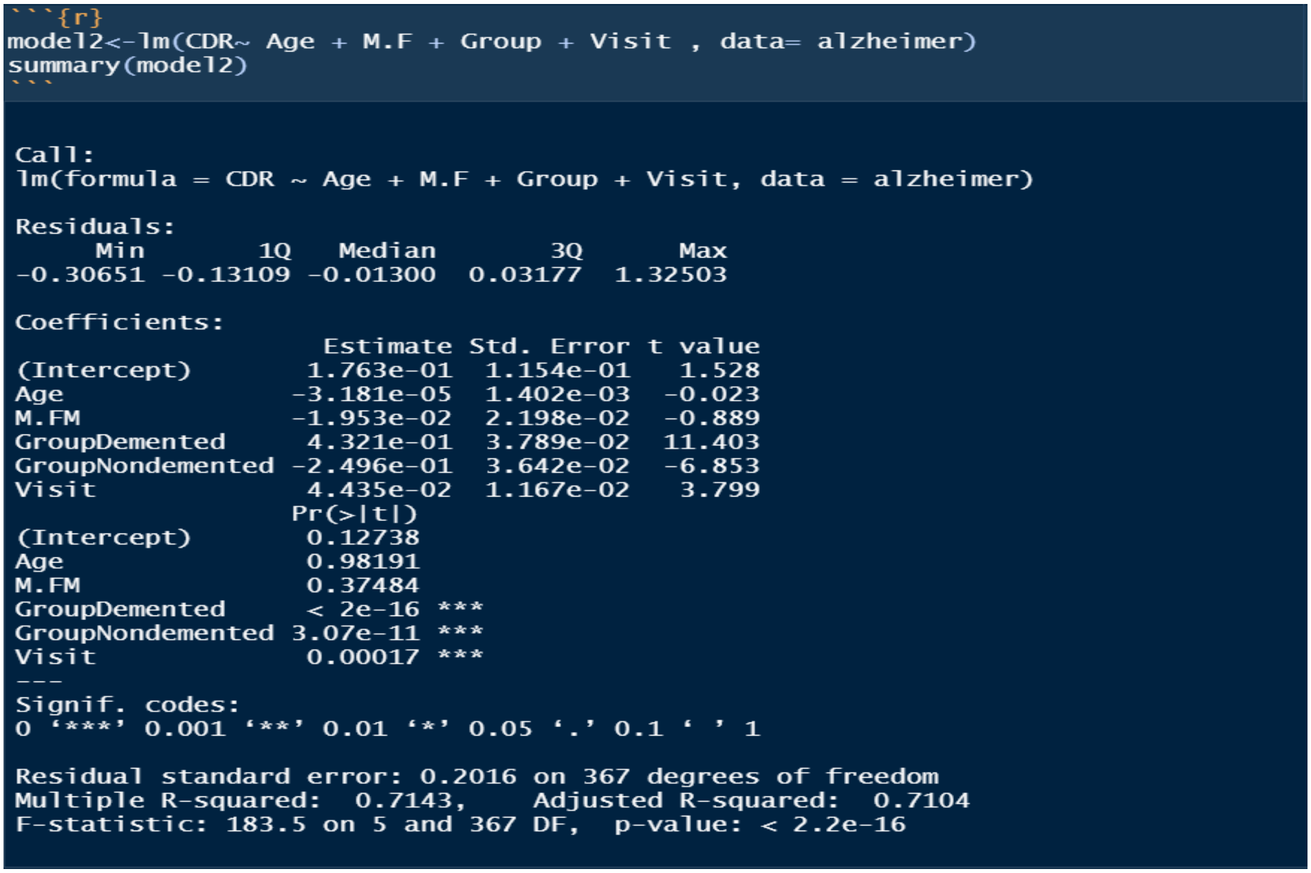
Model 2

H0: β1 = β2 = β3= β4= β5 = 0

H1: β1 ≠ β2 ≠ β3 ≠ β4 ≠ β5≠ 0

The equation becomes:

$Y= \beta_0 + \beta_1 Age + \beta_2 M.F + \beta_3 GroupDemented + \beta_4 GroupNondemented + \beta_5 Visit + \text{€}$



~~~
```{r}
model2<-lm(CDR~ Age + M.F + Group + Visit , data= alzheimer)
summary(model2)
```

 Call:
 lm(formula = CDR ~ Age + M.F + Group + Visit, data = alzheimer)

 Residuals:
      Min       1Q   Median       3Q      Max 
 -0.30651 -0.13109 -0.01300  0.03177  1.32503 

 Coefficients:
                     Estimate Std. Error t value
 (Intercept)        1.763e-01  1.154e-01   1.528
 Age               -3.181e-05  1.402e-03  -0.023
 M.FM              -1.953e-02  2.198e-02  -0.889
 GroupDemented      4.321e-01  3.789e-02  11.403
 GroupNondemented  -2.496e-01  3.642e-02  -6.853
 Visit              4.435e-02  1.167e-02   3.799
                   Pr(>|t|)    
 (Intercept)        0.12738    
 Age                0.98191    
 M.FM               0.37484    
 GroupDemented      < 2e-16 ***
 GroupNondemented 3.07e-11 ***
 Visit              0.00017 ***
 ---
 Signif. codes:  
 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Residual standard error: 0.2016 on 367 degrees of freedom
 Multiple R-squared:  0.7143,	Adjusted R-squared:  0.7104 
 F-statistic: 183.5 on 5 and 367 DF,  p-value: < 2.2e-16
~~~

This multiple linear regression results indicated that the p-value and the variables' F-statistic are related. Given the low p-value of the t-statistics, I reject the null hypothesis. The association between CDR and Gender and Age is unaffected since Group and Visit are significant factors and the t-statistics have a high p-value.

Model 3

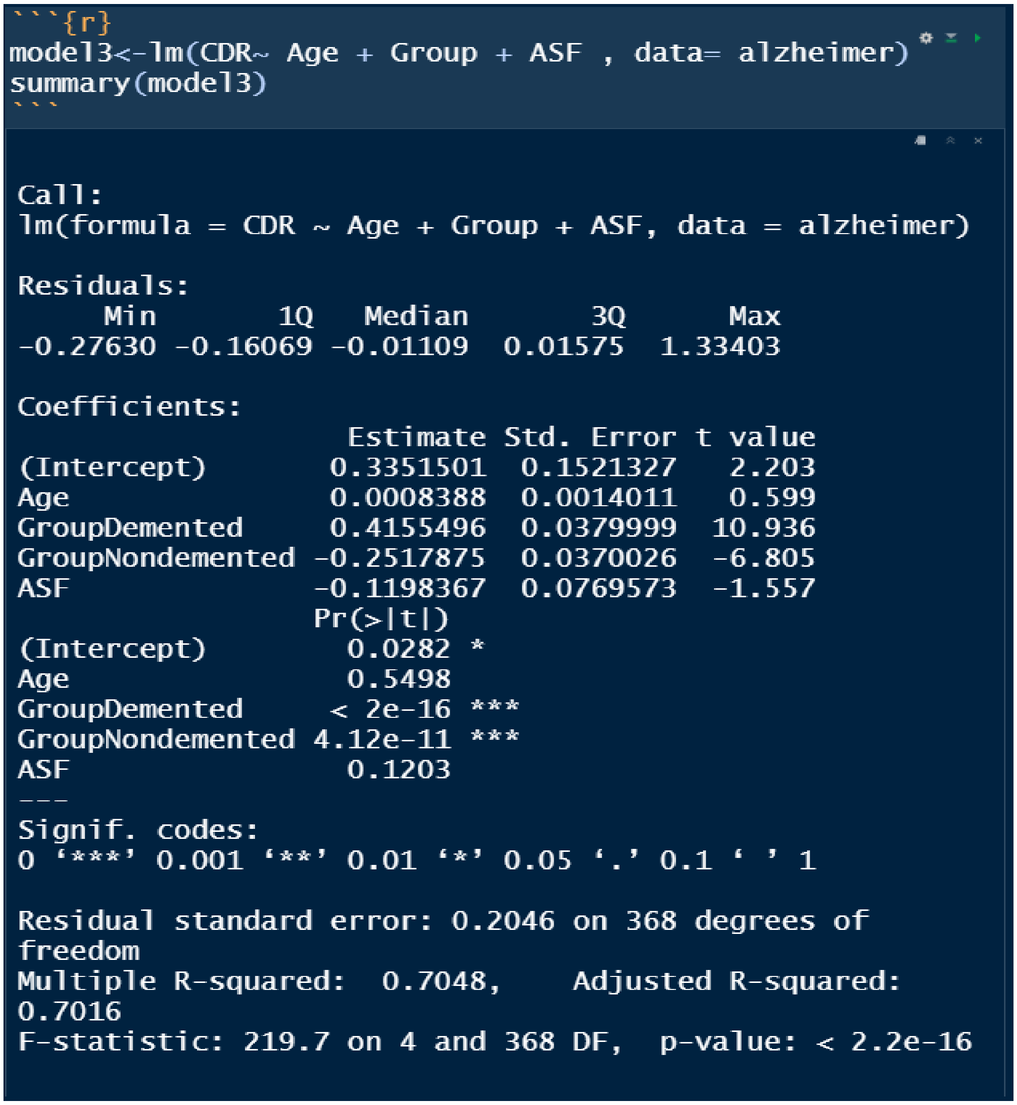
I discovered that from model 2, the M.F variable is not related to the response variable; I removed both the Visit and M.F variables from the model and replaced them with the ASF variable to see the output.

$H0: \beta1 = \beta2 = \beta3= \beta4= 0$

$H1: \beta1 \neq \beta2 \neq \beta3 \neq \beta4 \neq 0$

The equation becomes:

$Y= \beta0 + \beta1Age + \beta2GroupDemented + \beta3GroupNondemented + \beta4ASF + \text{€}$



~~~
```{r}
model3<-lm(CDR~ Age + Group + ASF , data= alzheimer)
summary(model3)
```
~~~

```
Call:
lm(formula = CDR ~ Age + Group + ASF, data = alzheimer)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.27630 -0.16069 -0.01109  0.01575  1.33403 

Coefficients:
                   Estimate Std. Error t value
(Intercept)       0.3351501  0.1521327   2.203
Age               0.0008388  0.0014011   0.599
GroupDemented     0.4155496  0.0379999  10.936
GroupNondemented -0.2517875  0.0370026  -6.805
ASF              -0.1198367  0.0769573  -1.557
                 Pr(>|t|)    
(Intercept)        0.0282 *  
Age                0.5498    
GroupDemented     < 2e-16 ***
GroupNondemented 4.12e-11 ***
ASF                0.1203    
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2046 on 368 degrees of
freedom
Multiple R-squared:  0.7048,	Adjusted R-squared:  
0.7016 
F-statistic: 219.7 on 4 and 368 DF,  p-value: < 2.2e-16
```

This multiple linear regression results indicated that the p-value and the variables' F-statistic are related. Given that the t-statistics have a low p-value, I reject the null hypothesis. Given that Group is a significant variable, and the t-statistics have a high p-value, the link between CDR, Age, and ASF is unaffected. Indicating that the variance of this model 3's mistakes is around 70% smaller than the variance of the response variable, the Adjusted R2 value for this model 3 is 0.70.

Model 4

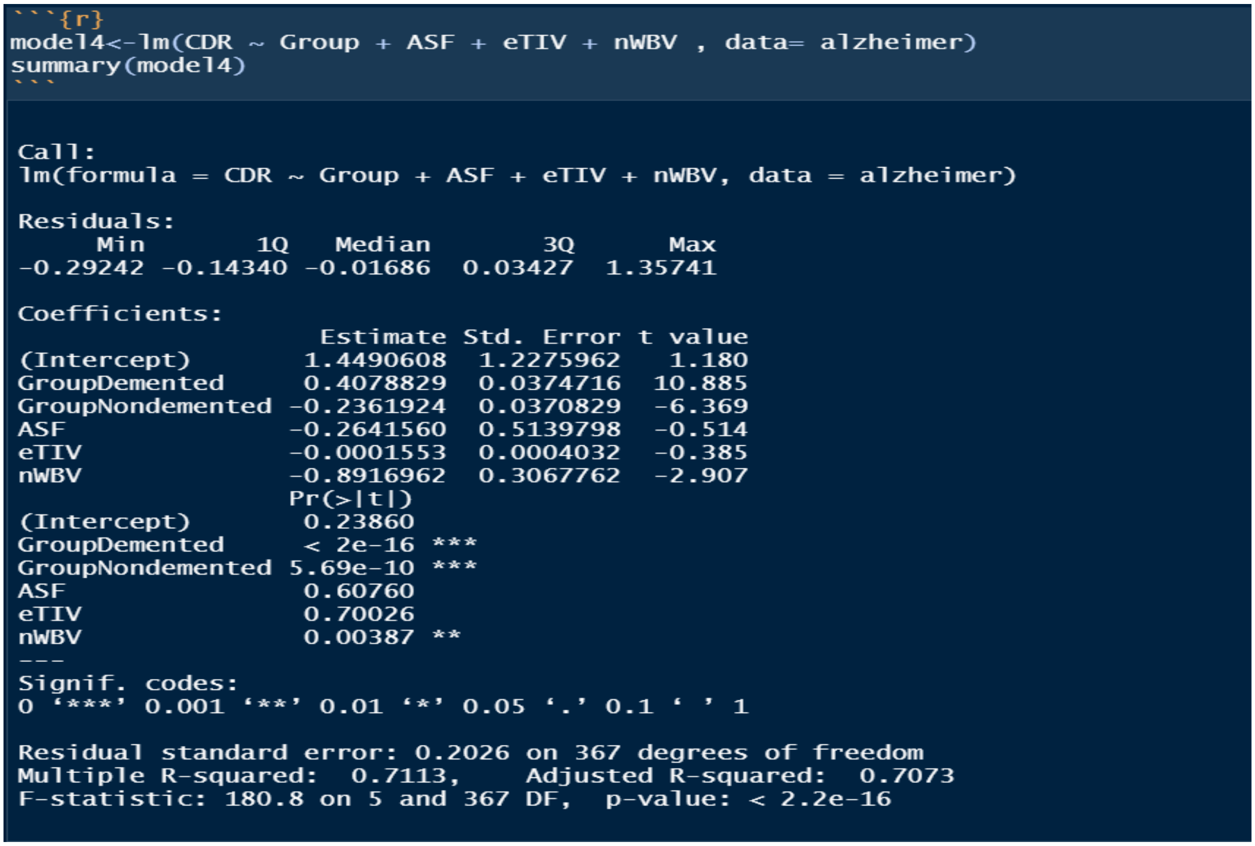
I performed another multiple linear regression, omitting age, which has no link with the response variable, and substituting eTIV and nWBV for it.

$H0 : \beta1 = \beta2 = \cdot \cdot \cdot = \beta p = 0$

$H1 : \beta j\ 5= 0$ for at least one j, $j = 1, . . . , p$

The equation becomes:

$Y= \beta0 + \beta1GroupDemented + \beta2GroupNondemented + \beta3ASF + \beta4eTIV + \beta5nWBV + €$



~~~
```{r}
model4<-lm(CDR ~ Group + ASF + eTIV + nWBV , data= alzheimer)
summary(model4)
```

Call:
lm(formula = CDR ~ Group + ASF + eTIV + nWBV, data = alzheimer)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.29242 -0.14340 -0.01686  0.03427  1.35741 

Coefficients:
                   Estimate Std. Error t value
(Intercept)       1.4490608  1.2275962   1.180
GroupDemented     0.4078829  0.0374716  10.885
GroupNondemented -0.2361924  0.0370829  -6.369
ASF              -0.2641560  0.5139798  -0.514
eTIV             -0.0001553  0.0004032  -0.385
nWBV             -0.8916962  0.3067762  -2.907
                 Pr(>|t|)    
(Intercept)       0.23860    
GroupDemented     < 2e-16 ***
GroupNondemented 5.69e-10 ***
ASF               0.60760    
eTIV              0.70026    
nWBV              0.00387 ** 
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2026 on 367 degrees of freedom
Multiple R-squared:  0.7113,	Adjusted R-squared:  0.7073 
F-statistic: 180.8 on 5 and 367 DF,  p-value: < 2.2e-16
~~~

ASF and eTIV are the factors in this model that are not statistically significant, according to Model 4's output. Given that this is a multiple regression, we examine the F-Statistic and associated p-value. The null hypothesis that all estimated betas are equal to zero can be rejected if the p-value corresponding to the F-statistics is less than 0.05. It can be said that some predictors are related to responses in some way. This model 4's adjusted R2 score is 0.71, which means that the variance of its mistakes is approximately 71% less than the variance of the response variable.

Model 5

Going through the correlation matrix, I discovered that there is a very weak negative association between the response variable and the EDUC variable; I decided to include the EDUC variable in the model to see the outcome.

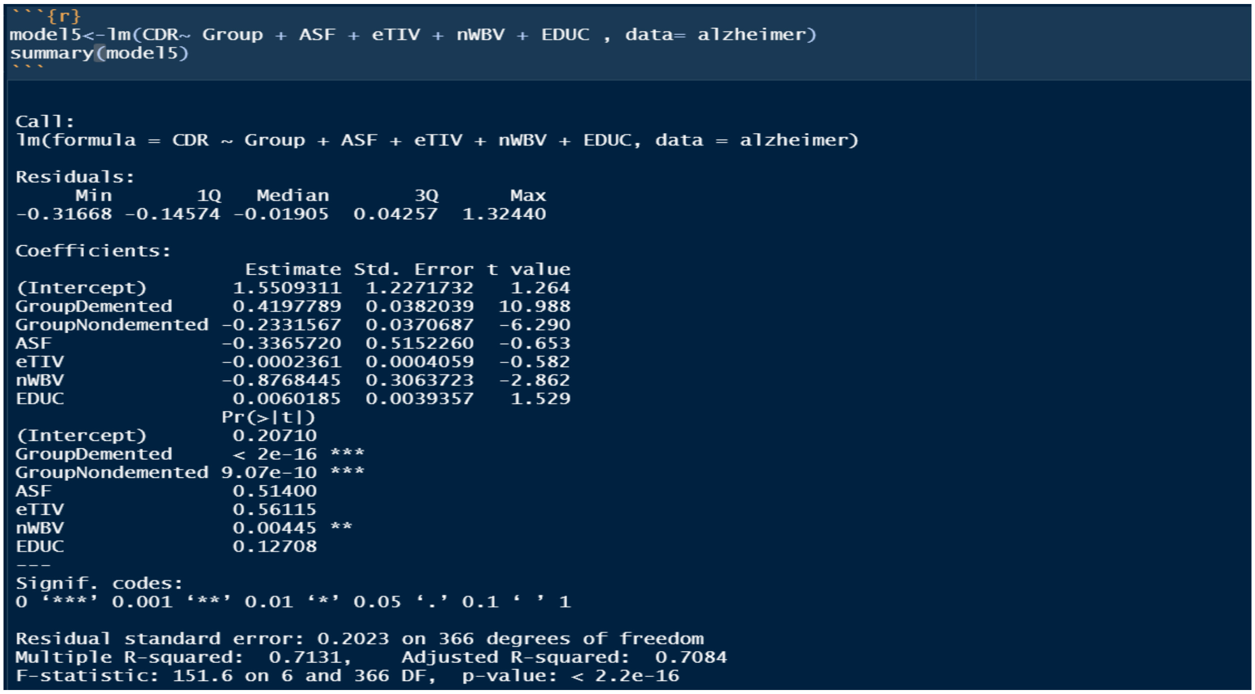
Equation:

$H0 : \beta 1 = \beta 2 = \cdot \cdot \cdot = \beta p = 0$

$H1 : \beta j\ 6 = 0$ for at least one j, j = 1, . . . , p

Therefore,

$Y = \beta 0 + \beta 1 GroupDemented + \beta 2 GroupNondemented + \beta 3 ASF + \beta 4 eTIV + \beta 5 nWBV + \beta 6 EDUC + €$



~~~
```{r}
model5<-lm(CDR~ Group + ASF + eTIV + nWBV + EDUC , data= alzheimer)
summary(model5)
```


Call:
lm(formula = CDR ~ Group + ASF + eTIV + nWBV + EDUC, data = alzheimer)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.31668 -0.14574 -0.01905  0.04257  1.32440 

Coefficients:
                   Estimate Std. Error t value
(Intercept)       1.5509311  1.2271732   1.264
GroupDemented     0.4197789  0.0382039  10.988
GroupNondemented -0.2331567  0.0370687  -6.290
ASF              -0.3365720  0.5152260  -0.653
eTIV             -0.0002361  0.0004059  -0.582
nWBV             -0.8768445  0.3063723  -2.862
EDUC              0.0060185  0.0039357   1.529
                 Pr(>|t|)    
(Intercept)       0.20710    
GroupDemented     < 2e-16 ***
GroupNondemented 9.07e-10 ***
ASF               0.51400    
eTIV              0.56115    
nWBV              0.00445 ** 
EDUC              0.12708    
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2023 on 366 degrees of freedom
Multiple R-squared:  0.7131,	Adjusted R-squared:  0.7084 
F-statistic: 151.6 on 6 and 366 DF,  p-value: < 2.2e-16
~~~

The model's output demonstrates that the two variables (Group and nWBV) are those that are statistically significant in this model. Given that this is a multiple regression, we examine the F-Statistic and associated p-value. The null hypothesis that all estimated betas are equal to zero can be rejected if the p-value corresponding to the F-statistics is less than 0.05. Indicating that the variance of this model V's mistakes is around 70% smaller than the variance of the response variable, the Adjusted R2 value for this model V is 0.70.

Model 6

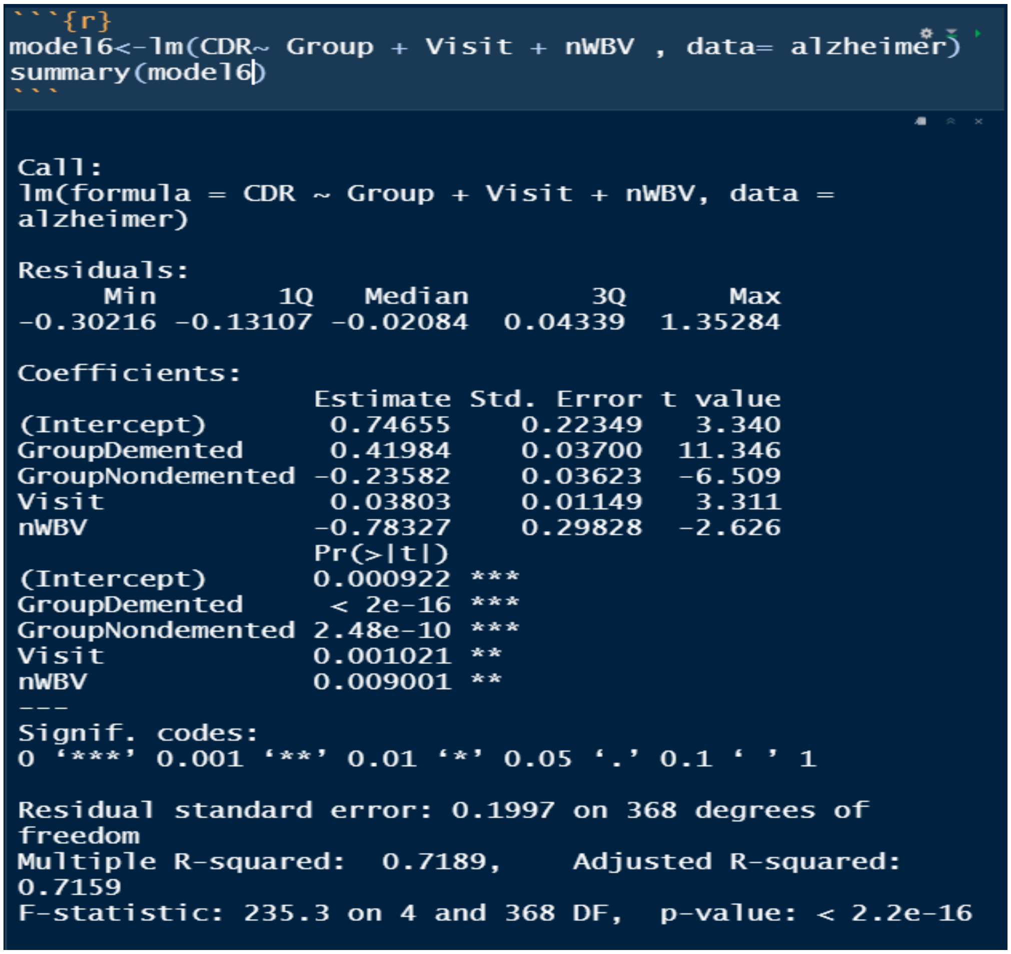
I eliminated all the irrelevant variables (EDUC, eTIV, ASF) from model 5 to obtain a very good model. I added the Visit variable due to its connection to the response variable from model 2.

$H0: \beta 1 = \beta 2 = \beta 3 = \beta 4 = 0$

$H1: \beta 1 \neq \beta 2 \neq \beta 3 \neq \beta 4 \neq 0$

The equation becomes:

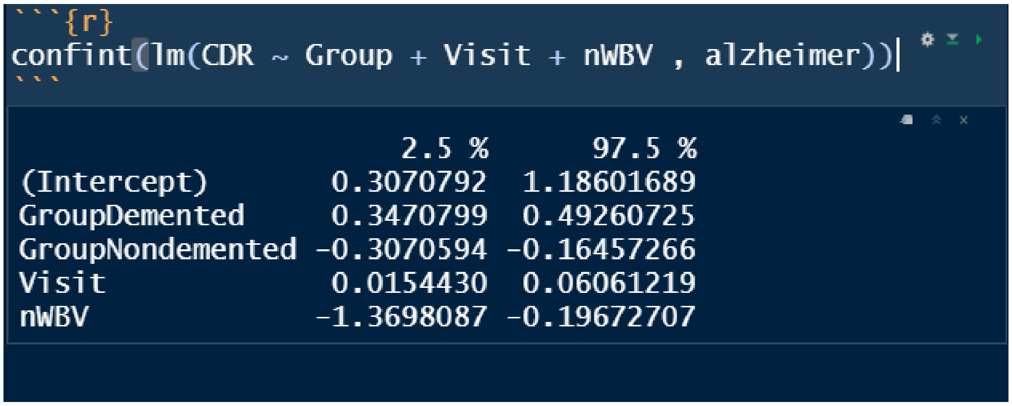
$Y= \beta 0 + \beta 1\text{GroupDemented} + \beta 2\text{GroupNondemented} + \beta 3\text{Visit} + \beta 4\text{nWBV} + \text{€}$



~~~
```{r}
model6<-lm(CDR~ Group + Visit + nWBV , data= alzheimer)
summary(model6)
```
~~~

```
Call:
lm(formula = CDR ~ Group + Visit + nWBV, data =
alzheimer)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30216 -0.13107 -0.02084  0.04339  1.35284

Coefficients:
                  Estimate Std. Error t value
(Intercept)        0.74655    0.22349   3.340
GroupDemented      0.41984    0.03700  11.346
GroupNondemented  -0.23582    0.03623  -6.509
Visit              0.03803    0.01149   3.311
nWBV              -0.78327    0.29828  -2.626
                  Pr(>|t|)
(Intercept)       0.000922 ***
GroupDemented      < 2e-16 ***
GroupNondemented  2.48e-10 ***
Visit             0.001021 **
nWBV              0.009001 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1997 on 368 degrees of
freedom
Multiple R-squared:  0.7189,    Adjusted R-squared:
0.7159
F-statistic: 235.3 on 4 and 368 DF,  p-value: < 2.2e-16
```

According to the result, all the variables in this model are statistically significant. Given that this is a multiple regression, we examine the F-Statistic and associated p-value. The null hypothesis that all estimated betas are equal to zero can be rejected if the p-value corresponding to the F-statistics is less than 0.05. All predictors have some connection to the response variable. The increased F-statistics also suggest that the model was improved by deleting the irrelevant variables and including the Visit variable. With an adjusted R2 value of roughly 0.72, this model's error variance is roughly 72% lower than the response variable. Comparing the model with other models (1,2,3,4,5), model 6 is best based on the R2 being higher than other models.



~~~
```{r}
confint(lm(CDR ~ Group + Visit + nWBV , alzheimer))
```
~~~

```
                        2.5 %      97.5 %
(Intercept)         0.3070792  1.18601689
GroupDemented       0.3470799  0.49260725
GroupNondemented -0.3070594 -0.16457266
Visit               0.0154430  0.06061219
nWBV               -1.3698087 -0.19672707
```

Other Multiple Regression.

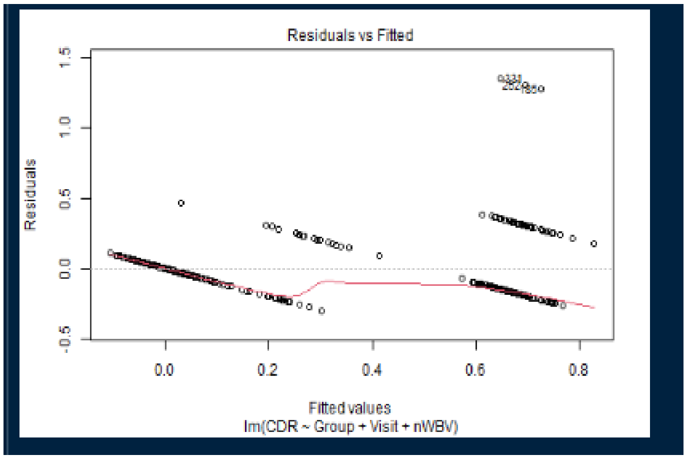
I used multiple linear regression by switching the predictor variables. I compared the results of models 1 through 6. Based on the results of the numerous multiple regression analyses that were performed, the most effective model for this dataset is model 6. This is only the beginning of the model's analysis. As we continue with the data analysis, we examine the issues with model 6.

Potential Problems of the model

After deciding to use model 6, I further examine the model to determine whether it has no regression issues. The procedures used to determine whether the model is appropriate for use are as follows.

A. Non-linearity of the model

I created a residual graph to display the nonlinearity of the model. If there is a trend in the residuals of the plot against the predicted values, our linear model might be flawed in some way. My figure below has no discernible patterns when the residuals were plotted against the expected values. There appears to be no trend that would point to issues with the data's nonlinearity, as shown by the presence of both positive and negative residuals along the regression line. Therefore, the linear model that was used to evaluate the data is sufficient.



B. Correlation of Error Term

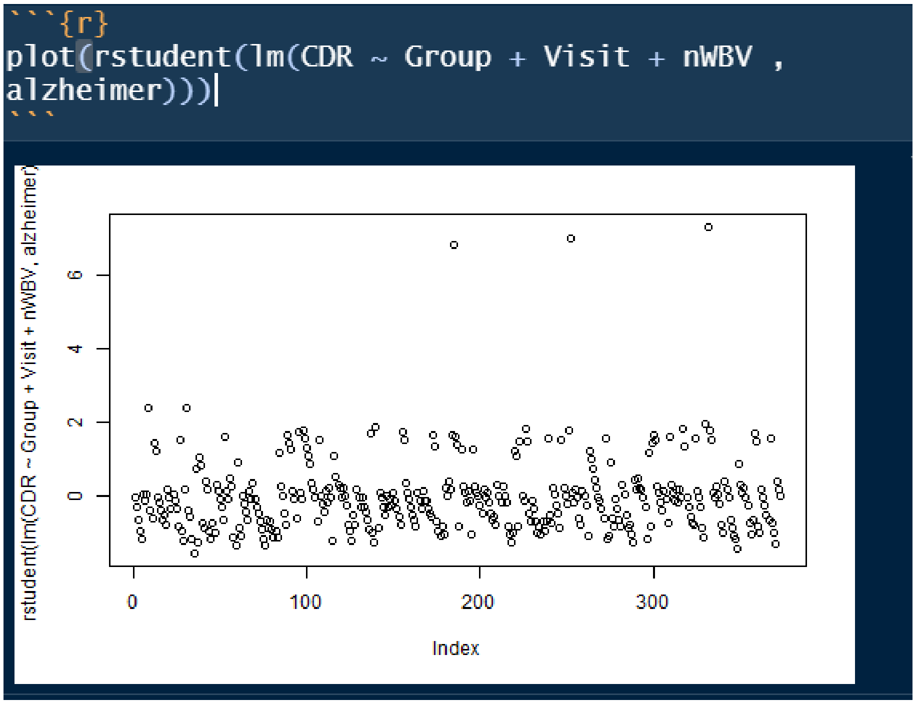
We can deduce that the dataset's error term is not correlated since the data were not collected at precise intervals. Therefore, the model's error terms do not appear to follow any pattern.

C. Non-constant Variance of Error Terms

I examine the residual plot for the presence of a funnel-shaped pattern or signs of heteroscedasticity to estimate the non-constant variance in mistakes of our model. However, I can infer from a residual plot of the model I have that no such pattern exists. There is no indication of heteroscedasticity, and the residuals seem to have a constant variance. The data wouldn't need to be transformed to create a new residual plot.

D. Outliers

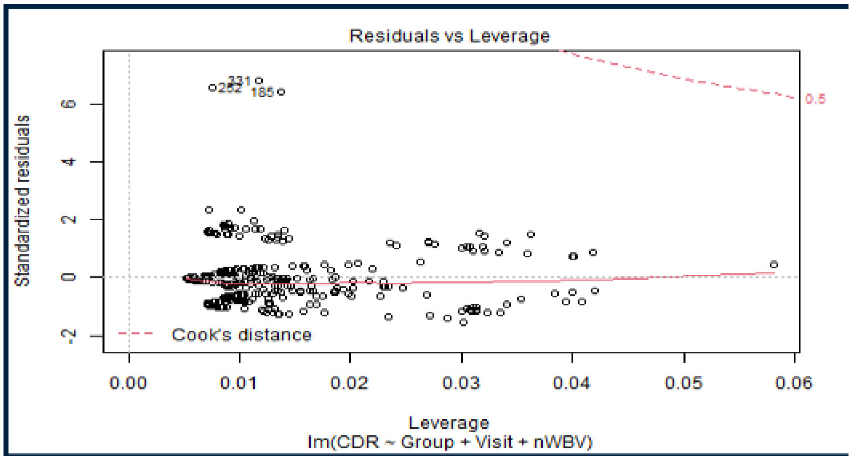
I continued my study by scanning the model for outliers. I plotted the Multiple Regression Model 6 studentized residuals to see if the model has any outliers. The plotted chart demonstrates that residuals are more than +2 standard deviation. This demonstrates that there are no outliers present.

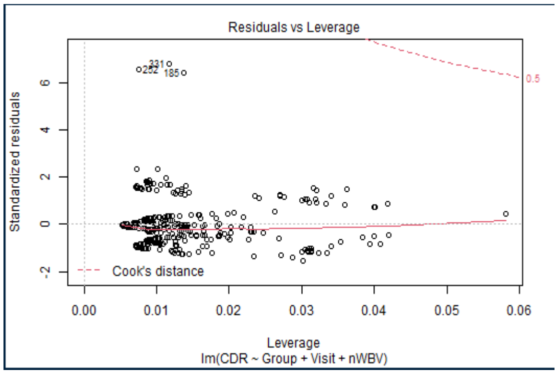


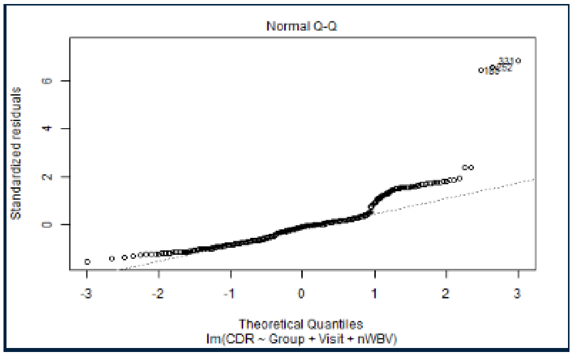
```{r}
plot(rstudent(lm(CDR ~ Group + Visit + nWBV ,
alzheimer)))
```

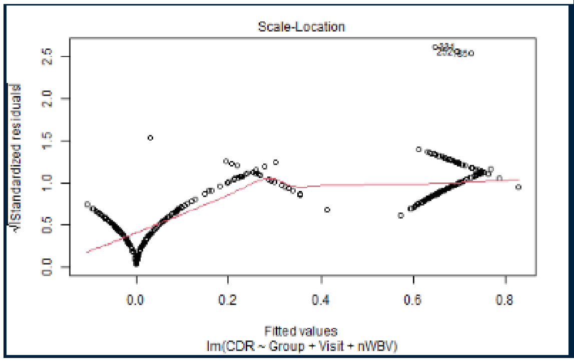
E. High Leverage Points

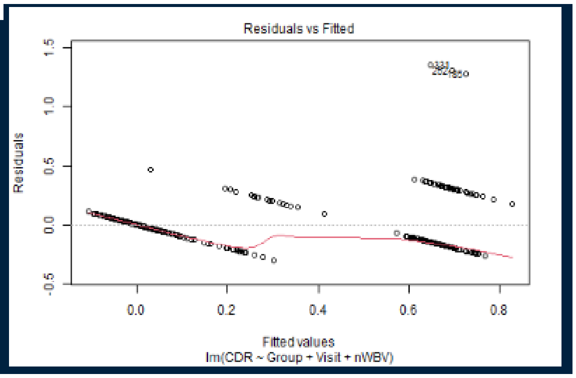
Residuals vs. Leverage plot shows no high leverage spots in my model. This means there is no need to change any variables to fit the model. Model 6 is excellent.





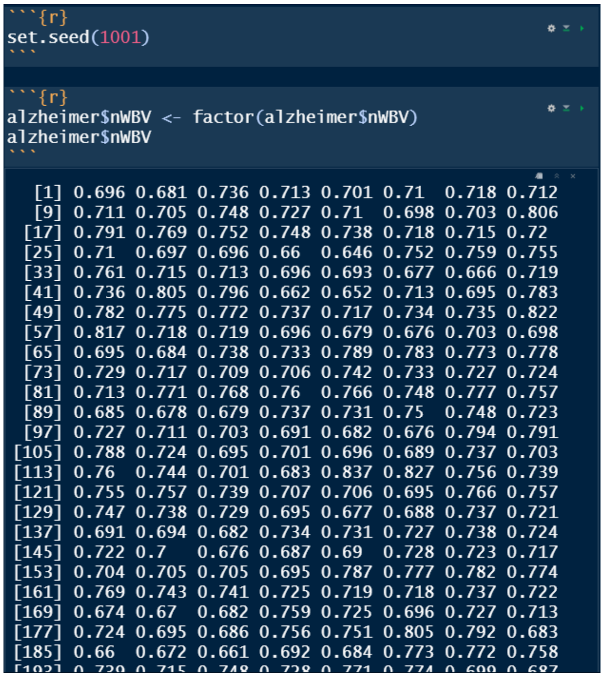






Clustering Techniques

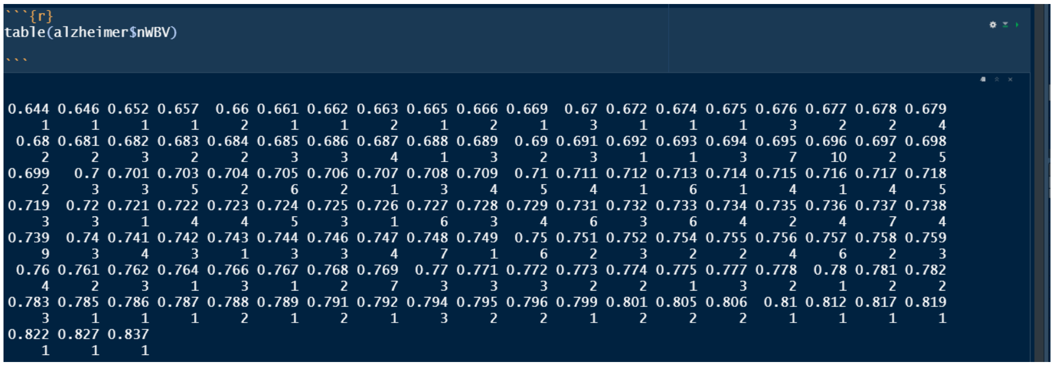
I chose Kmeans clustering to cluster instances of various dataset properties for my clustering approaches. The cluster codes and output that were used are listed below.



~~~
```{r}
set.seed(1001)
~~~

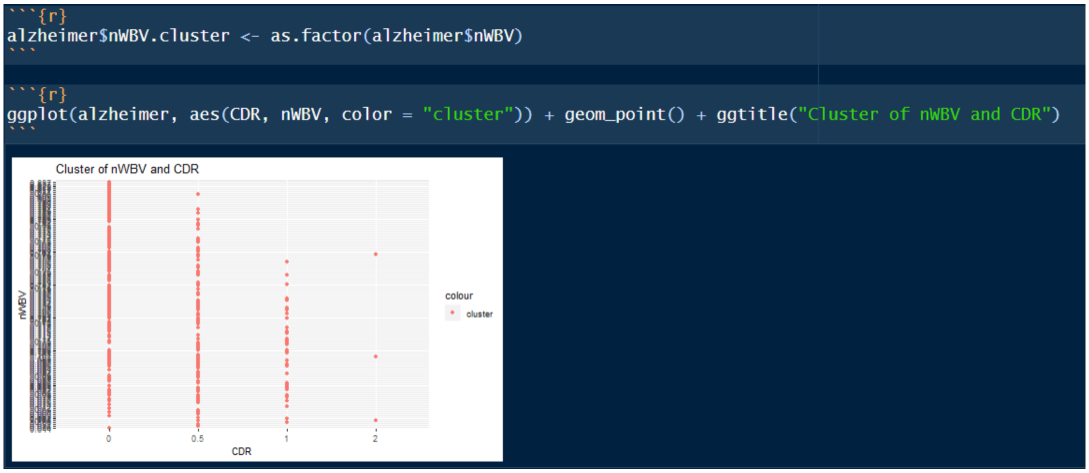
```{r}
alzheimer$nWBV <- factor(alzheimer$nWBV)
alzheimer$nWBV
```
```

```
  [1] 0.696 0.681 0.736 0.713 0.701 0.71  0.718 0.712
  [9] 0.711 0.705 0.748 0.727 0.71  0.698 0.703 0.806
 [17] 0.791 0.769 0.752 0.748 0.738 0.718 0.715 0.72
 [25] 0.71  0.697 0.696 0.66  0.646 0.752 0.759 0.755
 [33] 0.761 0.715 0.713 0.696 0.693 0.677 0.666 0.719
 [41] 0.736 0.805 0.796 0.662 0.652 0.713 0.695 0.783
 [49] 0.782 0.775 0.772 0.737 0.717 0.734 0.735 0.822
 [57] 0.817 0.718 0.719 0.696 0.679 0.676 0.703 0.698
 [65] 0.695 0.684 0.738 0.733 0.789 0.783 0.773 0.778
 [73] 0.729 0.717 0.709 0.706 0.742 0.733 0.727 0.724
 [81] 0.713 0.771 0.768 0.76  0.766 0.748 0.777 0.757
 [89] 0.685 0.678 0.679 0.737 0.731 0.75  0.748 0.723
 [97] 0.727 0.711 0.703 0.691 0.682 0.676 0.794 0.791
[105] 0.788 0.724 0.695 0.701 0.696 0.689 0.737 0.703
[113] 0.76  0.744 0.701 0.683 0.837 0.827 0.756 0.739
[121] 0.755 0.757 0.739 0.707 0.706 0.695 0.766 0.757
[129] 0.747 0.738 0.729 0.695 0.677 0.688 0.737 0.721
[137] 0.691 0.694 0.682 0.734 0.731 0.727 0.738 0.724
[145] 0.722 0.7   0.676 0.687 0.69  0.728 0.723 0.717
[153] 0.704 0.705 0.705 0.695 0.787 0.777 0.782 0.774
[161] 0.769 0.743 0.741 0.725 0.719 0.718 0.737 0.722
[169] 0.674 0.67  0.682 0.759 0.725 0.696 0.727 0.713
[177] 0.724 0.695 0.686 0.756 0.751 0.805 0.792 0.683
[185] 0.66  0.672 0.661 0.692 0.684 0.773 0.772 0.758
[193] 0.729 0.715 0.748 0.728 0.771 0.774 0.699 0.687
```
```

+



~~~
```{r}
table(alzheimer$nWBV)

```
~~~

~~~
0.644 0.646 0.652 0.657  0.66 0.661 0.662 0.663 0.665 0.666 0.669  0.67 0.672 0.674 0.675 0.676 0.677 0.678 0.679 
    1     1     1     1     2     1     1     2     1     2     1     3     1     1     1     3     2     2     4 
 0.68 0.681 0.682 0.683 0.684 0.685 0.686 0.687 0.688 0.689  0.69 0.691 0.692 0.693 0.694 0.695 0.696 0.697 0.698 
    2     2     3     2     2     3     3     4     1     3     2     3     1     1     3     7    10     2     5 
0.699   0.7 0.701 0.703 0.704 0.705 0.706 0.707 0.708 0.709  0.71 0.711 0.712 0.713 0.714 0.715 0.716 0.717 0.718 
    2     3     3     5     2     6     2     1     3     4     5     4     1     6     1     4     1     4     5 
0.719  0.72 0.721 0.722 0.723 0.724 0.725 0.726 0.727 0.728 0.729 0.731 0.732 0.733 0.734 0.735 0.736 0.737 0.738 
    3     3     1     4     4     5     3     1     6     3     4     6     3     6     4     2     4     7     4 
0.739  0.74 0.741 0.742 0.743 0.744 0.746 0.747 0.748 0.749  0.75 0.751 0.752 0.754 0.755 0.756 0.757 0.758 0.759 
    9     3     4     3     1     3     3     4     7     1     6     2     3     2     2     4     6     2     3 
 0.76 0.761 0.762 0.764 0.766 0.767 0.768 0.769  0.77 0.771 0.772 0.773 0.774 0.775 0.777 0.778  0.78 0.781 0.782 
    4     2     3     1     3     1     2     7     3     3     3     2     2     1     3     2     1     2     2 
0.783 0.785 0.786 0.787 0.788 0.789 0.791 0.792 0.794 0.795 0.796 0.799 0.801 0.805 0.806  0.81 0.812 0.817 0.819 
    3     1     1     1     2     1     2     1     3     2     2     1     2     2     2     1     1     1     1 
0.822 0.827 0.837 
    1     1     1 
~~~



~~~
```{r}
alzheimer$nWBV.cluster <- as.factor(alzheimer$nWBV)
```

```{r}
ggplot(alzheimer, aes(CDR, nWBV, color = "cluster")) + geom_point() + ggtitle("Cluster of nWBV and CDR")
```
~~~

Future Works

It would be fascinating to do predictive modeling in the future to determine how CDR depends on other variables and whether a machine learning algorithm can be trusted to correctly diagnose Alzheimer's, a challenging mental condition.

SUMMARY

Using a box plot, scatterplot, heatmap, and correlation matrix, I produced various visual representations of my dataset. Additionally, I was able to identify the model for my dataset. Because the p-value was close to 0 and the R-square of the statistics was close to 0.72 (72% of the coefficient of variance), my hope that my research would reveal patterns and linkages between the variables presented in the dataset was realized.

According to my analysis, there was a relationship between the Group variable's predictor (Group) and the response variable (CDR) because the F-statistics' p-value was small and close to 0, allowing us to reject the null hypothesis.

I accept the null hypothesis because the M.F variable exhibited a low statistically significant connection with the response variable (CDR); the p-value was quite high but still less than 0.05.

We cannot rule out the null hypothesis because the visit variable (Predictor) showed a marginally weak relationship with the response variable (CDR) according to the somewhat high p-value of the F-statistics.

Due to the low p-value of the F-statistics, the nWBV variable (Predictor) exhibited a marginally statistically significant connection with the response variable (CDR). Therefore, the null hypothesis is rejected.

Model 6 had the greatest adjusted R squared value, which demonstrated that changes in the predictors were connected to changes in the response variable, making it the best model out of the six I tested. The model of choice, model 6, rather than model 5, was altered by the insertion of Visit back into the model with a marginally low significance and the removal of unrelated variables (EDUC, eTIV, ASF). Due to their lower adjusted R-squared values than model 6, the other models were disregarded.

Except for the three outliers in the residual plot, which were connected to the patients' nWBV results, model 6 did not appear to have any issues with it.