Homework 5

Motivation

Dementia refers to a decline in mental ability that is significant enough to disrupt daily life, with memory loss being just one example. Rather than being a specific illness, it is a general term used to describe a set of symptoms that arise due to a decline in memory or other cognitive skills, which can hamper a person's ability to carry out daily tasks.

Doctors cannot diagnose dementia through a single test. Instead, they rely on a combination of factors such as a thorough medical history, physical examination, laboratory tests, and observation of changes in thinking, behavior, and day-to-day function. While doctors can usually determine that a person has dementia, identifying the specific type can be difficult since different forms of dementia can share symptoms and brain changes. Sometimes, a doctor may diagnose dementia without specifying the exact type, and in such cases, a specialist such as a neurologist or gero-psychologist may need to be consulted.

This dataset contains longitudinal data from 150 individuals aged between 60 and 96, with each subject having been scanned at least twice, with a gap of at least one year between scans. The total number of imaging sessions included in the collection is 373, and each subject has 3 or 4 individual T1-weighted MRI scans obtained in a single scan session. Both men and women are included, and all subjects are right-handed. Of the 150 subjects, 72 were considered to be free of dementia throughout the study, while 64 were initially diagnosed with dementia and remained so during subsequent scans, with 51 of these individuals having mild to moderate Alzheimer's disease. Finally, 14 individuals were initially considered to be free of dementia, but were later diagnosed with dementia during subsequent visits.

Explanation of Variables:

Subject.ID MRI.ID Group (Converted / Demented / Nondemented) Visit - Number of visit

Demographics Info M.F - Gender Hand - Handedness (actually all subjects were right-handed so I will drop this column) Age EDUC - Years of education SES - Socioeconomic status as assessed by the Hollingshead Index of Social Position and classified into categories from 1

(highest status) to 5 (lowest status) Clinical Info MMSE - Mini-Mental State Examination score (range is from 0 = worst to 30 = best) CDR - Clinical Dementia Rating (0 = no dementia, 0.5 = very mild AD, 1 = mild AD, 2 = moderate AD) Derived anatomic volumes eTIV - Estimated total intracranial volume, mm3 nWBV - Normalized whole-brain volume, expressed as a percent of all voxels in the atlas-masked image that are labeled as gray or white matter by the automated tissue segmentation process ASF - Atlas scaling factor (unitless). Computed scaling factor that transforms native-space brain and skull to the atlas target (i.e., the determinant of the transform matrix)

Mini–Mental State Examination MMSE : The Mini-Mental State Examination (MMSE), also known as the Folstein test, is a widely-used questionnaire with 30 points that helps to assess cognitive impairment in clinical and research settings. It is commonly used in the medical and allied health fields as a screening tool for dementia and to measure the severity and progression of cognitive impairment in an individual over time. This makes it a useful way to track an individual's response to treatment. However, it is important to note that the MMSE is not intended to provide a diagnosis for any specific medical condition on its own. A score of 24 points or higher on the MMSE indicates normal cognitive function. Scores below this range can suggest mild (19-23 points), moderate (10-18 points), or severe (9 points or less) cognitive impairment. Educational attainment and age may need to be taken into account when interpreting the raw score. It should be noted that a score of 30 points on the MMSE does not necessarily rule out the presence of dementia. Low scores on the MMSE are strongly associated with dementia, but abnormal findings on the test can also indicate the presence of other mental disorders. Additionally, physical problems such as hearing or vision impairment or motor deficits can interfere with test interpretation if not properly accounted for.

Clinical Dementia Rating (CDR) :The CDR™ is a 5-point scale that assesses cognitive and functional performance in six areas related to Alzheimer's disease and similar dementias. These areas are Memory, Orientation, Judgment & Problem Solving, Community Affairs, Home & Hobbies, and Personal Care. To determine a rating for each area, a semi-structured interview is conducted with the patient and a reliable informant, typically a family member. This process is known as the CDR™ Assessment Protocol.

To guide clinicians in making appropriate ratings based on interview data and clinical judgment, the CDR™ Scoring Table provides descriptive anchors. In addition to ratings for each domain, an overall CDR™ score can be calculated using the CDR™ Scoring Algorithm. This score is useful for characterizing and monitoring the patient's level of impairment or dementia. 0 = Normal 0.5 = Very Mild Dementia 1 = Mild Dementia 2 = Moderate Dementia 3 = Severe Dementia

Loading Libraries

```
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(Hmisc)
```

```
Loading required package: lattice

Loading required package: survival

Loading required package: Formula

Attaching package: 'Hmisc'

The following objects are masked from 'package:dplyr':

    src, summarize

The following objects are masked from 'package:base':

    format.pval, units
```

```
library(PerformanceAnalytics)
```

```
Loading required package: xts

Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric


Attaching package: 'xts'

The following objects are masked from 'package:dplyr':

    first, last


Attaching package: 'PerformanceAnalytics'

The following object is masked from 'package:graphics':

    legend
```

```
library(cowplot)
library(caret)
```

```
Attaching package: 'caret'

The following object is masked from 'package:survival':

    cluster
```

```
library(rpart)
library(rpart.plot)
```

```
library(e1071)
```

```
Attaching package: 'e1071'

The following objects are masked from 'package:PerformanceAnalytics':

    kurtosis, skewness

The following object is masked from 'package:Hmisc':

    impute
```

```
library(randomForest)
```

```
randomForest 4.7-1.1

Type rfNews() to see new features/changes/bug fixes.


Attaching package: 'randomForest'

The following object is masked from 'package:dplyr':

    combine

The following object is masked from 'package:ggplot2':

    margin
```

```
library(gbm)
```

```
Loaded gbm 2.1.8.1
```

```
library(Metrics)
```

```
Attaching package: 'Metrics'

The following objects are masked from 'package:caret':

    precision, recall
```

```
library(vtreat)
```

```
Loading required package: wrapr


Attaching package: 'wrapr'

The following object is masked from 'package:dplyr':

    coalesce
```

```
library(AUC)
```

```
AUC 0.3.2

Type AUCNews() to see the change log and ?AUC to get an overview.


Attaching package: 'AUC'

The following objects are masked from 'package:Metrics':

    accuracy, auc

The following objects are masked from 'package:caret':

    sensitivity, specificity
```

```r
library(DataExplorer)
set.seed(123)
data <- read.csv("Dementia.csv")
print(sample_n(data, 10))
```

```
   Subject.ID        MRI.ID       Group Visit MR.Delay M.F Hand Age EDUC SES
1   OAS2_0081 OAS2_0081_MR2    Demented     2      659   F    R  84   12   4
2   OAS2_0008 OAS2_0008_MR1 Nondemented     1        0   F    R  93   14   2
3   OAS2_0092 OAS2_0092_MR1   Converted     1        0   F    R  83   12   2
4   OAS2_0149 OAS2_0149_MR1 Nondemented     1        0   F    R  81   13   2
5   OAS2_0055 OAS2_0055_MR2 Nondemented     2      726   M    R  67   13   3
6   OAS2_0145 OAS2_0145_MR2   Converted     2     1707   F    R  73   16   3
7   OAS2_0108 OAS2_0108_MR2    Demented     2      883   M    R  79   18   1
8   OAS2_0117 OAS2_0117_MR3 Nondemented     3     1345   M    R  76   20   2
9   OAS2_0186 OAS2_0186_MR2 Nondemented     2      763   F    R  63   13   2
10  OAS2_0070 OAS2_0070_MR4 Nondemented     4     1870   M    R  85   17   1
   MMSE CDR eTIV  nWBV   ASF
1    26 0.5 1273 0.686 1.378
2    30 0.0 1272 0.698 1.380
3    28 0.0 1383 0.748 1.269
4    29 0.0 1345 0.737 1.305
5    27 0.0 1365 0.827 1.285
6    29 0.5 1287 0.771 1.364
7    27 0.5 1569 0.781 1.118
8    30 0.0 1823 0.739 0.963
9    30 0.0 1327 0.796 1.323
10   30 0.0 1724 0.704 1.018
```

Obtain details on every variable in the dataset.

```r
describe(data)
```

```
data

 15  Variables      373  Observations
--------------------------------------------------------------------------------
Subject.ID
       n  missing distinct
     373        0      150

lowest : OAS2_0001 OAS2_0002 OAS2_0004 OAS2_0005 OAS2_0007
```

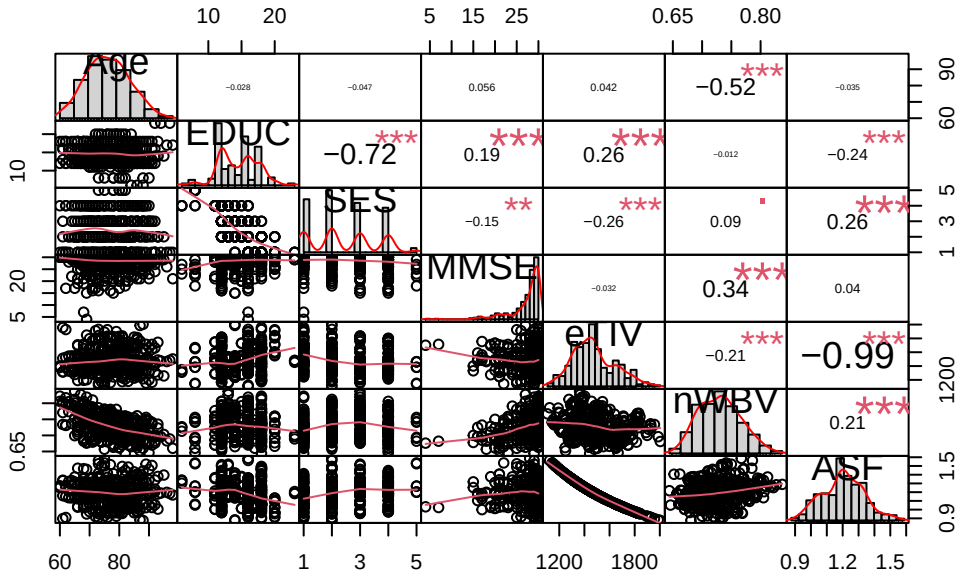
```
highest: OAS2_0182 OAS2_0183 OAS2_0184 OAS2_0185 OAS2_0186
---------------------------------------------------------------------------
MRI.ID
       n  missing distinct
     373        0      373

lowest : OAS2_0001_MR1 OAS2_0001_MR2 OAS2_0002_MR1 OAS2_0002_MR2 OAS2_0002_MR3
highest: OAS2_0185_MR2 OAS2_0185_MR3 OAS2_0186_MR1 OAS2_0186_MR2 OAS2_0186_MR3
---------------------------------------------------------------------------
Group
       n  missing distinct
     373        0        3

Value        Converted    Demented Nondemented
Frequency           37         146         190
Proportion       0.099       0.391       0.509
---------------------------------------------------------------------------
Visit
       n  missing distinct     Info     Mean      Gmd
     373        0        5    0.874    1.882   0.9552

lowest : 1 2 3 4 5, highest: 1 2 3 4 5

Value          1     2     3     4     5
Frequency    150   144    58    15     6
Proportion 0.402 0.386 0.155 0.040 0.016
---------------------------------------------------------------------------
MR.Delay
       n  missing distinct     Info     Mean      Gmd      .05      .10
     373        0      201    0.935    595.1    682.6        0        0
     .25      .50      .75      .90      .95
       0      552      873     1561     1828

lowest :    0  182  212  248  352, highest: 2386 2400 2508 2517 2639
---------------------------------------------------------------------------
M.F
       n  missing distinct
     373        0        2

Value          F     M
Frequency    213   160
Proportion 0.571 0.429
---------------------------------------------------------------------------
```

```
Hand
       n  missing distinct    value
     373        0        1        R

Value        R
Frequency  373
Proportion   1
---------------------------------------------------------------------------
Age
       n  missing distinct     Info     Mean      Gmd      .05      .10
     373        0       39    0.998    77.01    8.703     65.0     67.2
     .25      .50      .75      .90      .95
    71.0     77.0     82.0     87.8     90.0

lowest : 60 61 62 63 64, highest: 94 95 96 97 98
---------------------------------------------------------------------------
EDUC
       n  missing distinct     Info     Mean      Gmd      .05      .10
     373        0       12    0.962     14.6    3.183       11       12
     .25      .50      .75      .90      .95
      12       15       16       18       18

lowest :  6  8 11 12 13, highest: 16 17 18 20 23

Value          6     8    11    12    13    14    15    16    17    18    20
Frequency      3     9    11   103    27    33    17    81     9    64    13
Proportion 0.008 0.024 0.029 0.276 0.072 0.088 0.046 0.217 0.024 0.172 0.035

Value         23
Frequency      3
Proportion 0.008
---------------------------------------------------------------------------
SES
       n  missing distinct     Info     Mean      Gmd
     354       19        5    0.938     2.46    1.266

lowest : 1 2 3 4 5, highest: 1 2 3 4 5

Value          1     2     3     4     5
Frequency     88   103    82    74     7
Proportion 0.249 0.291 0.232 0.209 0.020
---------------------------------------------------------------------------
MMSE
```

```
       n  missing distinct     Info     Mean      Gmd      .05      .10
     371        2       18    0.954    27.34    3.417       20       22
     .25      .50      .75      .90      .95
      27       29       30       30       30

lowest :  4  7 15 16 17, highest: 26 27 28 29 30

Value          4     7    15    16    17    18    19    20    21    22    23
Frequency      1     1     2     3     5     2     3     7    11     7    11
Proportion 0.003 0.003 0.005 0.008 0.013 0.005 0.008 0.019 0.030 0.019 0.030

Value         24    25    26    27    28    29    30
Frequency      4    12    20    32    45    91   114
Proportion 0.011 0.032 0.054 0.086 0.121 0.245 0.307
---------------------------------------------------------------------------
CDR
       n  missing distinct     Info     Mean      Gmd
     373        0        4    0.794   0.2909   0.3683

Value        0.0   0.5   1.0   2.0
Frequency    206   123    41     3
Proportion 0.552 0.330 0.110 0.008
---------------------------------------------------------------------------
eTIV
       n  missing distinct     Info     Mean      Gmd      .05      .10
     373        0      286        1     1488    197.7     1234     1289
     .25      .50      .75      .90      .95
    1357     1470     1597     1731     1817

lowest : 1106 1123 1143 1151 1154, highest: 1928 1931 1957 1987 2004
---------------------------------------------------------------------------
nWBV
       n  missing distinct     Info     Mean      Gmd      .05      .10
     373        0      136        1   0.7296  0.04232   0.6746   0.6822
     .25      .50      .75      .90      .95
  0.7000   0.7290   0.7560   0.7796   0.7940

lowest : 0.644 0.646 0.652 0.657 0.660, highest: 0.817 0.819 0.822 0.827 0.837
---------------------------------------------------------------------------
ASF
       n  missing distinct     Info     Mean      Gmd      .05      .10
     373        0      265        1    1.195   0.1563   0.9656   1.0134
     .25      .50      .75      .90      .95
```

```
  1.0990    1.1940    1.2930    1.3618    1.4222

lowest : 0.876 0.883 0.897 0.909 0.910, highest: 1.521 1.525 1.535 1.563 1.587
------------------------------------------------------------------------------
```

```r
chart.Correlation(select(data, Age, EDUC, SES, MMSE, eTIV, nWBV, ASF), histogram = TRUE, m
```



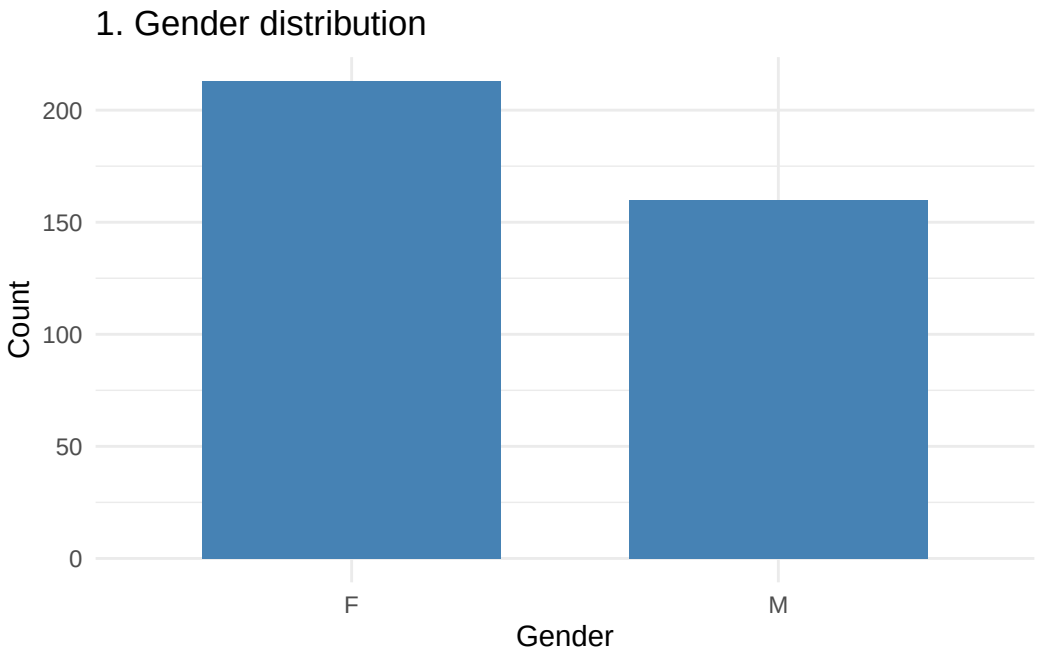
Earlier, it was observed that certain columns in the dataset contain null values. Therefore, the next step would be to substitute those missing values with the median value for the respective column.

```r
data <- select(data, -Hand) #drop Hand column since all objects were right-handed
data$SES[is.na(data$SES)] <- median(data$SES, na.rm = TRUE)
data$MMSE[is.na(data$MMSE)] <- median(data$MMSE, na.rm = TRUE)

#creating new column with Dementia diagnosis
#data$Dementia <- 0
#data$Dementia[data$CDR == 0] <- 0
#data$Dementia[data$CDR > 0] <- 1
#data$Dementia <- as.factor(data$Dementia)
```

Exploratory Data Analysis

Class of CDR will be our predicted value. Let's see how it depends on other variables.

```
ggplot(data, aes(x=factor(M.F)))+
  geom_bar(width=0.7, fill="steelblue")+
  theme_minimal() + labs(title = "1. Gender distribution",
       x = "Gender",
       y = "Count")
```



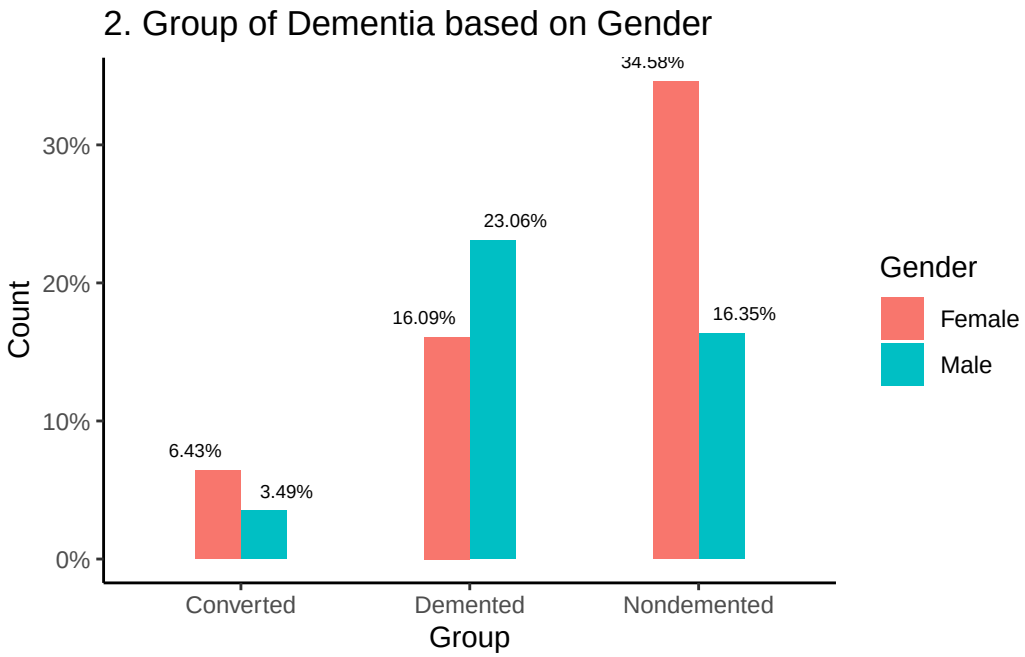
More females than males in this scenario.

```
ggplot(data = data,
       aes(
         x = Group,
         y = prop.table(stat(count)),
         fill = factor(data$M.F), width = -6,
         label = scales::percent(prop.table(stat(count)))
       )) +
  geom_bar(position = position_dodge(), width = 0.4) + theme(axis.text = element_text(size
  geom_text(
```

```
    stat = "count",
    position = position_dodge(.8),
    vjust = -1,
    size = 2.5
  ) + scale_y_continuous(labels = scales::percent) +
  labs(title = "2. Group of Dementia based on Gender",
       x = "Group",
       y = "Count") +
  theme_classic() +
  scale_fill_discrete(
    name = "Gender",
    labels = c("Female", "Male")
  )
```

```
Warning: `stat(count)` was deprecated in ggplot2 3.4.0.
i Please use `after_stat(count)` instead.
```

```
Warning: Use of `data$M.F` is discouraged.
i Use `M.F` instead.
Use of `data$M.F` is discouraged.
i Use `M.F` instead.
```

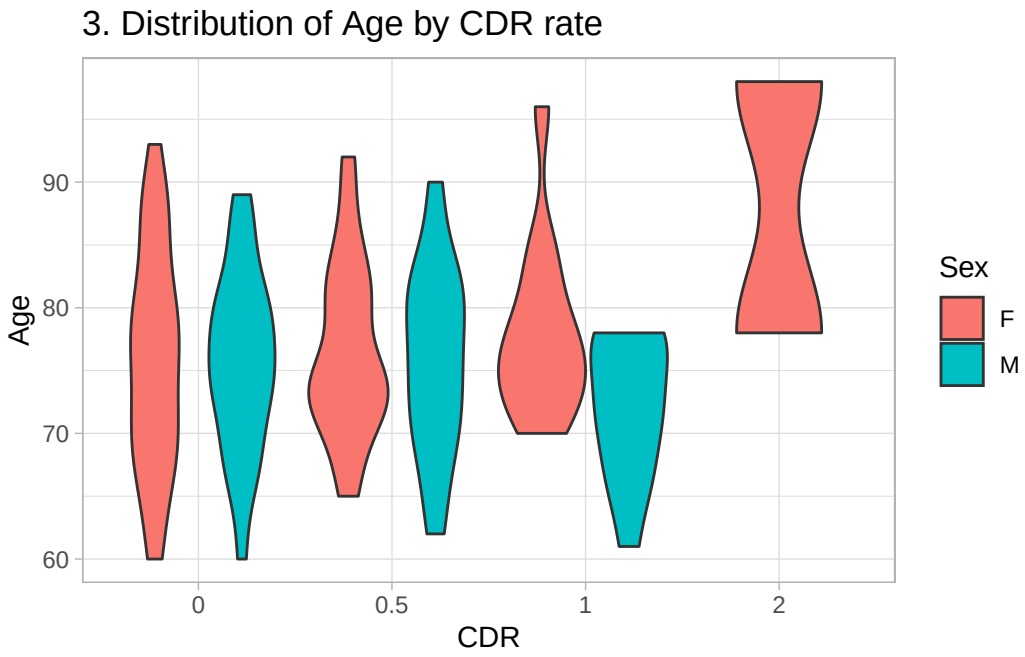


We can see that the highest is Non-demented people and the highest is Females.

A violin plot is a graphical representation that shows the distribution of numerical data for one or multiple groups using density curves. The thickness of each curve represents the approximate frequency of data points in that region. It is a type of data visualization that is effective for comparing the distribution of numeric data across one or more groups. It is a useful tool for identifying differences and similarities between groups and for observing the shape and density of each distribution.

```
data %>%
    select(Subject.ID, Age, CDR, M.F) %>%
    group_by(Subject.ID, CDR, M.F) %>%
    summarise_all(funs(min)) %>%
    as.data.frame() %>%
    mutate(CDR = as.factor(CDR)) %>%
ggplot(aes(x = CDR, y = Age, fill = M.F)) +
    geom_violin() +
    labs(title = "3. Distribution of Age by CDR rate",
         fill = "Sex") +
    theme_light()
```

```
Warning: `funs()` was deprecated in dplyr 0.8.0.
i Please use a list of either functions or lambdas:

# Simple named list: list(mean = mean, median = median)

# Auto named with `tibble::lst()`: tibble::lst(mean, median)

# Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))

Warning: Groups with fewer than two data points have been dropped.
```

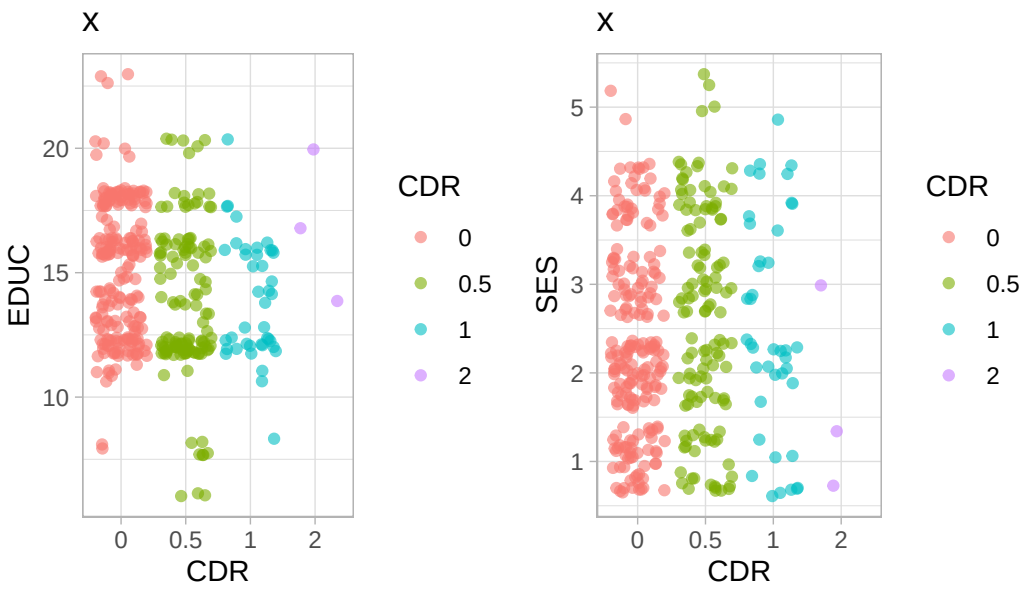


There does not appear to be a clear correlation between age, sex, and the diagnosis of dementia.

The jitter geom is a convenient shortcut for geom_point(position = "jitter"). It introduces a slight amount of random variation to the placement of each point, which can be beneficial in addressing the issue of overplotting that arises from the limited amount of data points in smaller datasets.

```
a <- data %>%
    select(EDUC, CDR, M.F) %>%
    mutate(CDR = as.factor(CDR)) %>%
ggplot(aes(x = CDR, y = EDUC)) +
    geom_jitter(aes(col = CDR), alpha = 0.6) +
    labs(title = "x") +
    theme_light()

b <- data %>%
    select(SES, CDR, M.F) %>%
    mutate(CDR = as.factor(CDR)) %>%
ggplot(aes(x = CDR, y = SES)) +
    geom_jitter(aes(col = CDR), alpha = 0.6) +
    labs(title = "x") +
```

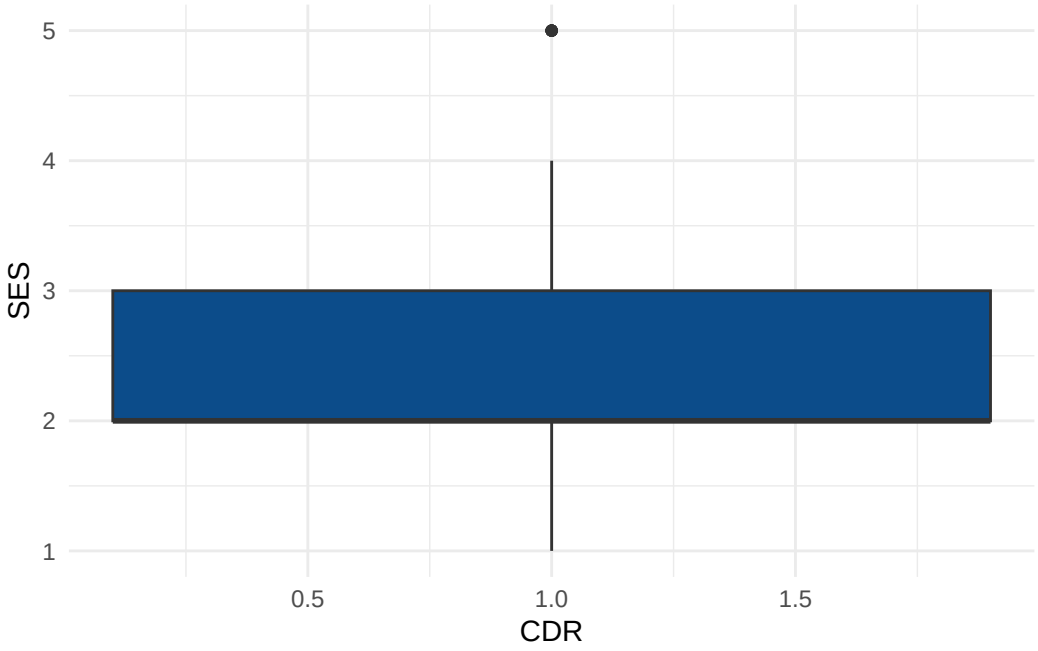
```
    theme_light()

p <- plot_grid(a, b)
title <- ggdraw() + draw_label("4. Distribution of Education and Social Economic Status",
plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))
```



There is still no clear association observed between the level of education and socioeconomic status, and the diagnosis of dementia.

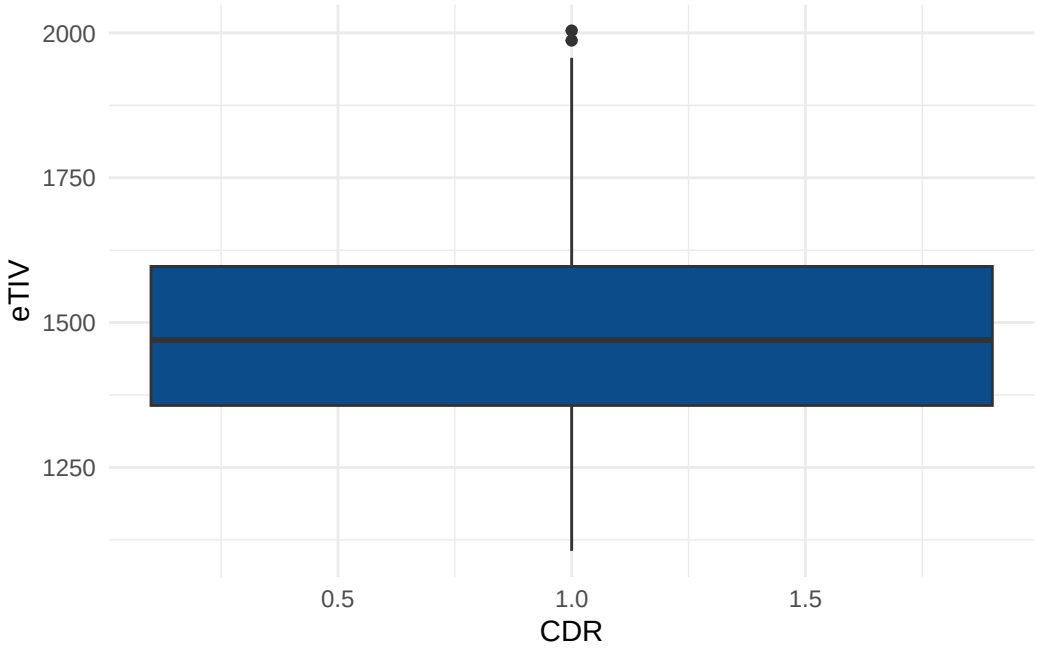
```
ggplot(data) + aes(x = CDR, y = SES) + geom_boxplot(fill = "#0c4c8a") + theme_minimal()
```

```
Warning: Continuous x aesthetic
i did you forget `aes(group = ...)`?
```



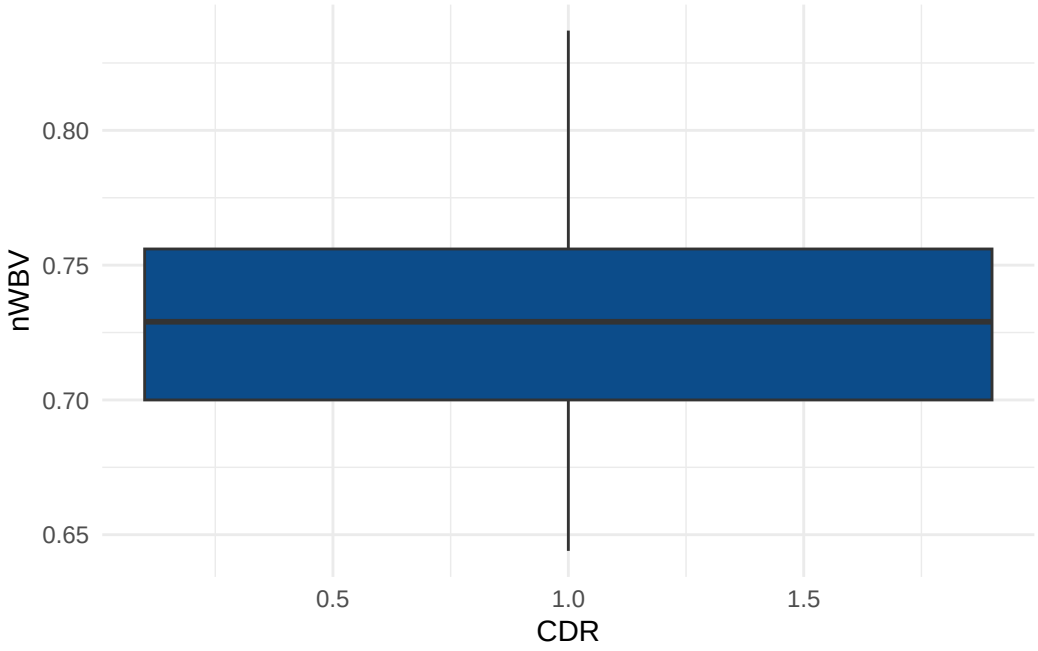

```
ggplot(data) + aes(x = CDR, y = eTIV) + geom_boxplot(fill = "#0c4c8a") + theme_minimal()
```

```
Warning: Continuous x aesthetic
i did you forget `aes(group = ...)`?
```



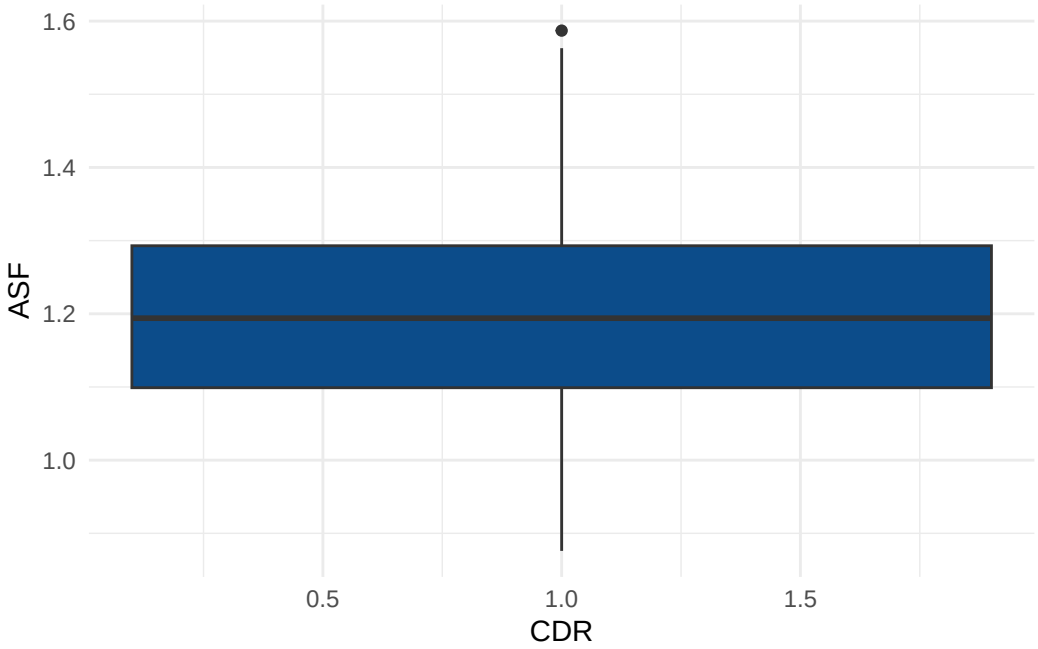

```
ggplot(data) + aes(x = CDR, y = nWBV) + geom_boxplot(fill = "#0c4c8a") + theme_minimal()
```

```
Warning: Continuous x aesthetic
i did you forget `aes(group = ...)`?
```



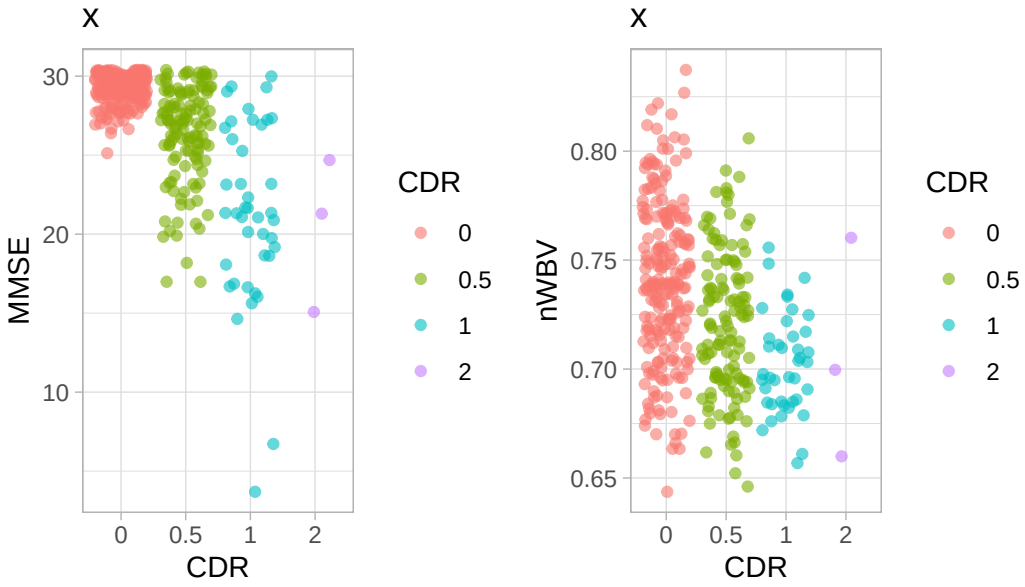

```
ggplot(data) + aes(x = CDR, y = ASF) + geom_boxplot(fill = "#0c4c8a") + theme_minimal()
```

```
Warning: Continuous x aesthetic
i did you forget `aes(group = ...)`?
```




```
x <- data %>%
    select(MMSE, CDR, M.F) %>%
    mutate(CDR = as.factor(CDR)) %>%
ggplot(aes(x = CDR, y = MMSE)) +
    geom_jitter(aes(col = CDR), alpha = 0.6) +
    labs(title = "x") +
    theme_light()

y <- data %>%
    select(nWBV, CDR, M.F) %>%
    mutate(CDR = as.factor(CDR)) %>%
ggplot(aes(x = CDR, y = nWBV)) +
    geom_jitter(aes(col = CDR), alpha = 0.6) +
    labs(title = "x") +
    theme_light()
p <- plot_grid(x, y)
title <- ggdraw() + draw_label("5. Distribution of MMSE Score and Wole-brain Volume", font
plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))
```

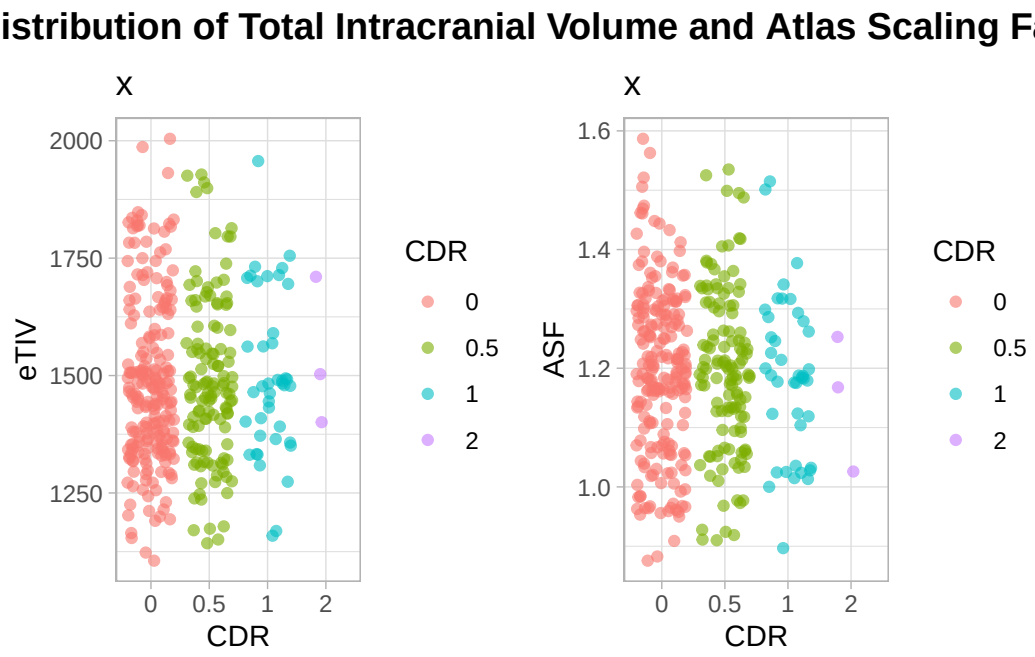
5. Distribution of MMSE Score and Wole−brain Volume



The MMS test scores of individuals without Dementia tend to cluster around 27-30 points, whereas the scores of those diagnosed with Dementia appear to be more widely distributed. We observe that some individuals have the highest MMSE scores, but still have a Clinical Dementia Rating of 0.5 or 1. There does not appear to be a clear relationship between Estimated total intracranial volume and Dementia Diagnosis.

```
a <- data %>%
    select(eTIV, CDR, M.F) %>%
    mutate(CDR = as.factor(CDR)) %>%
ggplot(aes(x = CDR, y = eTIV)) +
    geom_jitter(aes(col = CDR), alpha = 0.6) +
    labs(title = "x") +
    theme_light()

b <- data %>%
    select(ASF, CDR, M.F) %>%
    mutate(CDR = as.factor(CDR)) %>%
ggplot(aes(x = CDR, y = ASF)) +
    geom_jitter(aes(col = CDR), alpha = 0.6) +
    labs(title = "x") +
    theme_light()
p <- plot_grid(a, b)
```

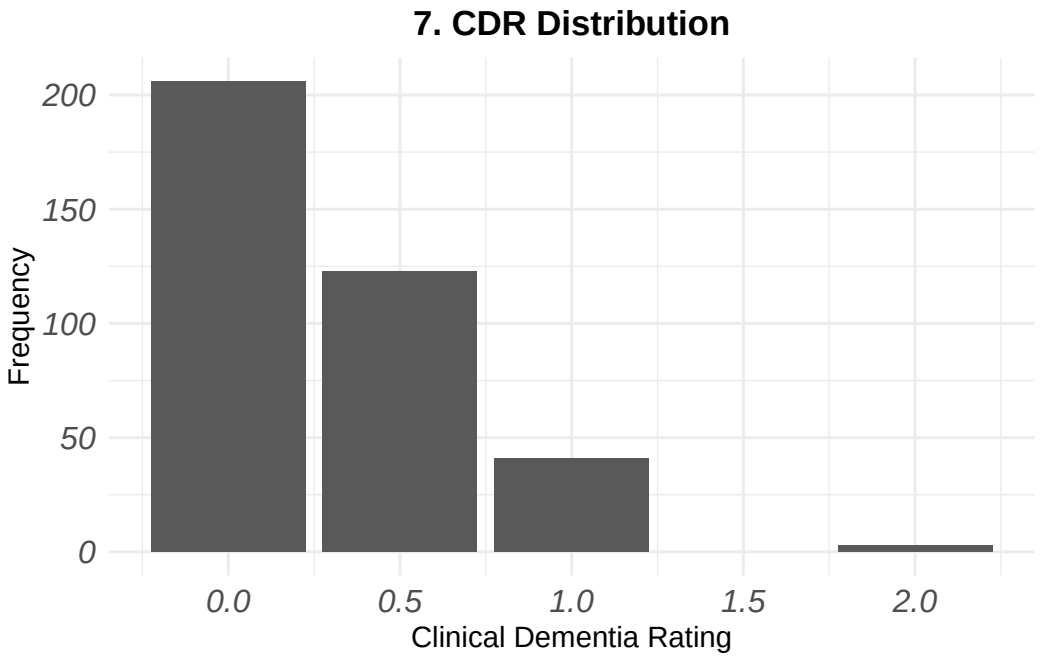
```
title <- ggdraw() + draw_label("6. Distribution of Total Intracranial Volume and Atlas Sca
plot_grid(title, p, ncol=1, rel_heights=c(0.1, 1))
```



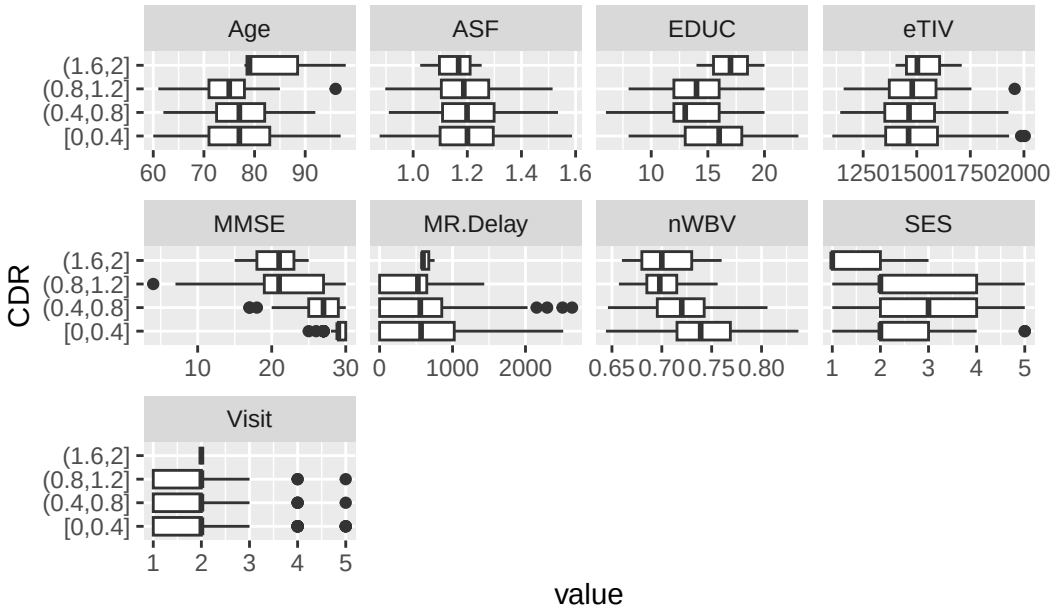
The whole-brain volume that has been normalized appears to have a wider range for subjects with a CDR score of 0, but narrows as the CDR score increases. However, there doesn't seem to be a clear relationship between the atlas scaling factor and dementia diagnosis.

```
data %>%
  ggplot(aes(CDR,fill=CDR))+geom_bar()+
  theme_minimal()+
  labs(x="Clinical Dementia Rating",y="Frequency",title="7. CDR Distribution")+
  theme(plot.title = element_text(hjust=0.5,color="black",face="bold"),
        axis.text = element_text(face="italic",size=12))+
  scale_fill_manual(values=c(RColorBrewer::brewer.pal(4,"PuRd")))
```

```
Warning: The following aesthetics were dropped during statistical transformation: fill
i This can happen when ggplot fails to infer the correct grouping structure in
  the data.
i Did you forget to specify a `group` aesthetic or to convert a numerical
  variable into a factor?
```




```
plot_boxplot(data, by = "CDR")
```



#Tree-based Models Tree-based learning algorithms are widely regarded as among the most effective and commonly employed techniques for supervised learning. These methods provide predictive models with impressive levels of accuracy, stability, and interpretability. Unlike linear models, they excel at capturing non-linear relationships. Furthermore, they can be applied to various problem types, such as classification and regression. Decision trees, random forests, and gradient boosting are highly favored and extensively utilized in a wide range of data science problems.

##Preparation and splitting the data

```
#preparing data
df <- data %>%
  select(M.F, Age, EDUC, SES, MMSE, eTIV, nWBV, ASF, CDR) %>%
  mutate(CDR = as.factor(CDR))

n_train <- round(0.8 * nrow(df)) #80%  data set (length) as integer
train_index <- sample(1:nrow(df), n_train) #creating a vector with random indices
train <- df[train_index, ] #creating train data set
test <- df[-train_index, ] #creating test data set

formula <- CDR ~ M.F + Age + EDUC + SES + MMSE + eTIV + nWBV #CDR as response and all othe
k <- 5   #(k=5)cross validation
splitPlan <- kWayCrossValidation(nrow(df), k, NULL, NULL) #creating 5-folds cross validati
```

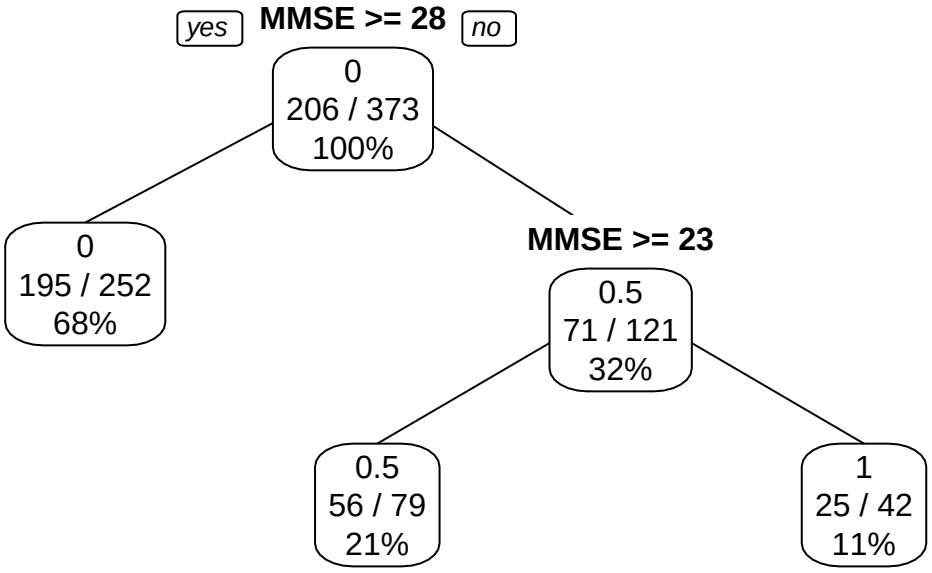
The training formula is: CDR is predicted by the variables M.F, Age, EDUC, SES, MMSE, eTIV, and nWBV. Atlas Scaling Factor (ASF) has been removed from the formula due to its linear dependancy leading to multicollinearity.

Decision Tree Model

A decision tree is a widely utilized supervised learning algorithm primarily employed in classification problems. It can handle both categorical and continuous input and output variables. This technique involves dividing the population or sample into two or more homogeneous subsets, known as sub-populations, based on the most significant differentiating factor among the input variables.

We will proceed to train a basic decision tree model and display the output to determine the optimal complexity parameter (CP) value using cross-validation. The complexity parameter is employed to regulate the size of the decision tree and select the most suitable tree size.

```r
opt_compar <- 0 #list with optimal parameters
for(i in 1:k) {
  split <- splitPlan[[i]]
  #training decision tree model
  model_crossv <- rpart(formula = formula,
               data = df[split$train,],
               method = "class")
  #get the best CP value
  opt_compar[i] <- model_crossv$cptable[which.min(model_crossv$cptable[,"xerror"]),"CP"]
}

#training the model with optimal CP parameter on whole data set
model_decisiontree <- rpart(formula = formula,
               data = df,
               method = "class",
               cp = mean(opt_compar))
#plotingt decision tree model
prp(x = model_decisiontree, type=1, extra = 102)
```



```r
#testing the model
prediction_dt <- predict(object = model_crossv,
                newdata = df,
                type = "class")

#print confusion matrix
confusionMatrix(data = prediction_dt,
                reference = df$CDR)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0 0.5   1   2
       0   190  40   4   0
       0.5  16  72  15   1
       1     0  11  22   2
       2     0   0   0   0

Overall Statistics

               Accuracy : 0.7614
                 95% CI : (0.7148, 0.8038)
    No Information Rate : 0.5523
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5672

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 0.5 Class: 1 Class: 2
Sensitivity            0.9223     0.5854  0.53659 0.000000
Specificity            0.7365     0.8720  0.96084 1.000000
Pos Pred Value         0.8120     0.6923  0.62857      NaN
Neg Pred Value         0.8849     0.8104  0.94379 0.991957
Prevalence             0.5523     0.3298  0.10992 0.008043
Detection Rate         0.5094     0.1930  0.05898 0.000000
Detection Prevalence   0.6273     0.2788  0.09383 0.000000
Balanced Accuracy      0.8294     0.7287  0.74871 0.500000
```

The confusion matrix presents the counts of predicted classes (rows) compared to the actual

reference classes (columns). For example, in the first row, the model predicted class 0 for 188 instances where the actual reference class was also 0. Similarly, the model predicted class 0.5 for 87 instances where the actual reference class was 0.5.

##Overall Statistics:

Accuracy: The overall accuracy of the model is 0.7828, indicating that approximately 78.28% of the predictions were correct. 95% CI: The 95% confidence interval for the accuracy ranges from 0.7375 to 0.8236. No Information Rate (NIR): The NIR represents the accuracy that could be achieved by always predicting the most frequent class in the data. In this case, the NIR is 0.5523, suggesting that the model performs significantly better than simply predicting the most common class. Kappa: The Kappa statistic measures the agreement between the predicted and actual classes, considering the possibility of agreement occurring by chance. A value of 0.6005 indicates moderate agreement.

Statistics by Class:

This section provides metrics such as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), prevalence, detection rate, detection prevalence, and balanced accuracy for each class. Sensitivity: Also known as the true positive rate or recall, it represents the proportion of correctly predicted instances of a particular class out of all instances of that class. For example, the sensitivity for class 0 is 0.9126, indicating that the model successfully predicted class 0 in 91.26% of instances where the actual class was 0. Specificity: This refers to the true negative rate, which is the proportion of correctly predicted instances that do not belong to a specific class out of all instances not belonging to that class. For instance, the specificity for class 1 is 0.99398, indicating that the model correctly identified 99.398% of instances not belonging to class 1. Pos Pred Value: The positive predictive value, also known as precision, represents the proportion of correctly predicted instances of a particular class out of all instances predicted as that class. For example, the PPV for class 0.5 is 0.7073, meaning that 70.73% of instances predicted as class 0.5 were actually class 0.5. Neg Pred Value: The negative predictive value represents the proportion of correctly predicted instances not belonging to a particular class out of all instances not predicted as that class. For instance, the NPV for class 2 is 0.991957, meaning that 99.1957% of instances predicted as not class 2 were indeed not class 2. Prevalence: This indicates the proportion of instances belonging to a particular class out of all instances. For example, the prevalence of class 0 is 0.5523, indicating that 55.23% of instances belong to class 0. Detection Rate: This represents the proportion of correctly predicted instances of a particular class out of all instances. For example, the detection rate for class 0.5 is 0.2332, meaning that the model successfully detected 23.32% of instances belonging to class 0.5. Detection Prevalence: This refers to the proportion of instances predicted as a particular class out of all instances. For example, the detection prevalence for class 1 is 0.05094, indicating that 5.094% of instances were predicted as class 1. Balanced Accuracy: This calculates the average of sensitivity.