VK-DWMR Final

2024-04-28

Dementia Prediction

Introduction

Dementia refers to a range of cognitive declines that hinder daily functioning, with memory loss being a common example. Rather than a single disease, it encompasses various symptoms that affect memory and thinking skills, ultimately impacting a person's ability to carry out everyday activities. As a Cognitive Science major specializing in Neuroscience, I find it fitting to do an analysis on Dementia, since it has been fascinating the world of neuroscience to this day. In this project, I aim to identify key features associated with dementia onset and progression, ultimately constructing a robust predictive model capable of accurately classifying individuals at risk of developing dementia.

Understanding the Data

For the data, I used two publicly available datasets on OASIS (https://sites.wustl.edu/oasisbrains/ (https://sites.wustl.edu/oasisbrains/)): - oasis_cross-sectional.csv: Cross-sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults - oasis_longitudinal.csv: Longitudinal MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults

What do variables stand for

- **Subject.ID**
- **MRI.ID**
- **Group** *(Converted / Demented / Nondemented)*
- **Visit** - Number of visits
- **MR.Delay** ???

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)

Interpretations

Scores of 24 or higher out of 30 indicate normal cognition, while lower scores can suggest varying degrees of cognitive impairment: severe (≤9 points), moderate (10–18 points), or mild (19–23 points). Adjustment for education and age may be needed. Even a perfect score doesn't exclude dementia. Low scores often indicate dementia, but other mental disorders can also affect results. Physical issues like hearing or vision problems, or motor deficits, can interfere with interpretation if not properly noted.

Clinical Dementia Rating (CDR)

The CDR™ in one aspect is a 5-point scale used to characterize six domains of cognitive and functional performance applicable to Alzheimer disease and related dementias: Memory, Orientation, Judgment & Problem Solving, Community Affairs, Home & Hobbies, and Personal Care. This score is useful for characterizing and tracking a patient's level of impairment/dementia: * 0 = Normal * 0.5 = Very Mild Dementia * 1 = Mild Dementia * 2 = Moderate Dementia * 3 = Severe Dementia

Estimated total intracranial volume (eTIV)

The ICV measure, sometimes referred to as total intracranial volume (TIV), refers to the estimated volume of the cranial cavity as outlined by the supratentorial dura matter or cerebral contour when dura is not clearly detectable. ICV, along with age and gender are reported as covariates to adjust for regression analyses in investigating progressive neurodegenerative brain disorders, such as Alzheimer's disease, aging and cognitive impairment.

I uploaded these files in the file pane on the right and read them as follows.

```
data1 <- read.csv("oasis_longitudinal.csv")
data2 <- read.csv("oasis_cross-sectional.csv")
print(sample_n(data1, 5))
```

```
##   Subject.ID        MRI.ID       Group Visit MR.Delay M.F Hand Age EDUC SES
## 1  OAS2_0101 OAS2_0101_MR3 Nondemented     3     1631   F    R  76   18   2
## 2  OAS2_0066 OAS2_0066_MR1    Demented     1        0   M    R  61   18   1
## 3  OAS2_0161 OAS2_0161_MR1 Nondemented     1        0   M    R  77   16   1
## 4  OAS2_0073 OAS2_0073_MR2 Nondemented     2      580   F    R  72   14   3
## 5  OAS2_0028 OAS2_0028_MR2    Demented     2      610   M    R  66   18   2
##   MMSE CDR eTIV  nWBV   ASF
## 1   30   0 1379 0.757 1.273
## 2   30   1 1957 0.734 0.897
## 3   29   0 1818 0.734 0.965
## 4   28   0 1512 0.777 1.161
## 5   21   1 1562 0.717 1.124
```

```
print(sample_n(data2, 5))
```

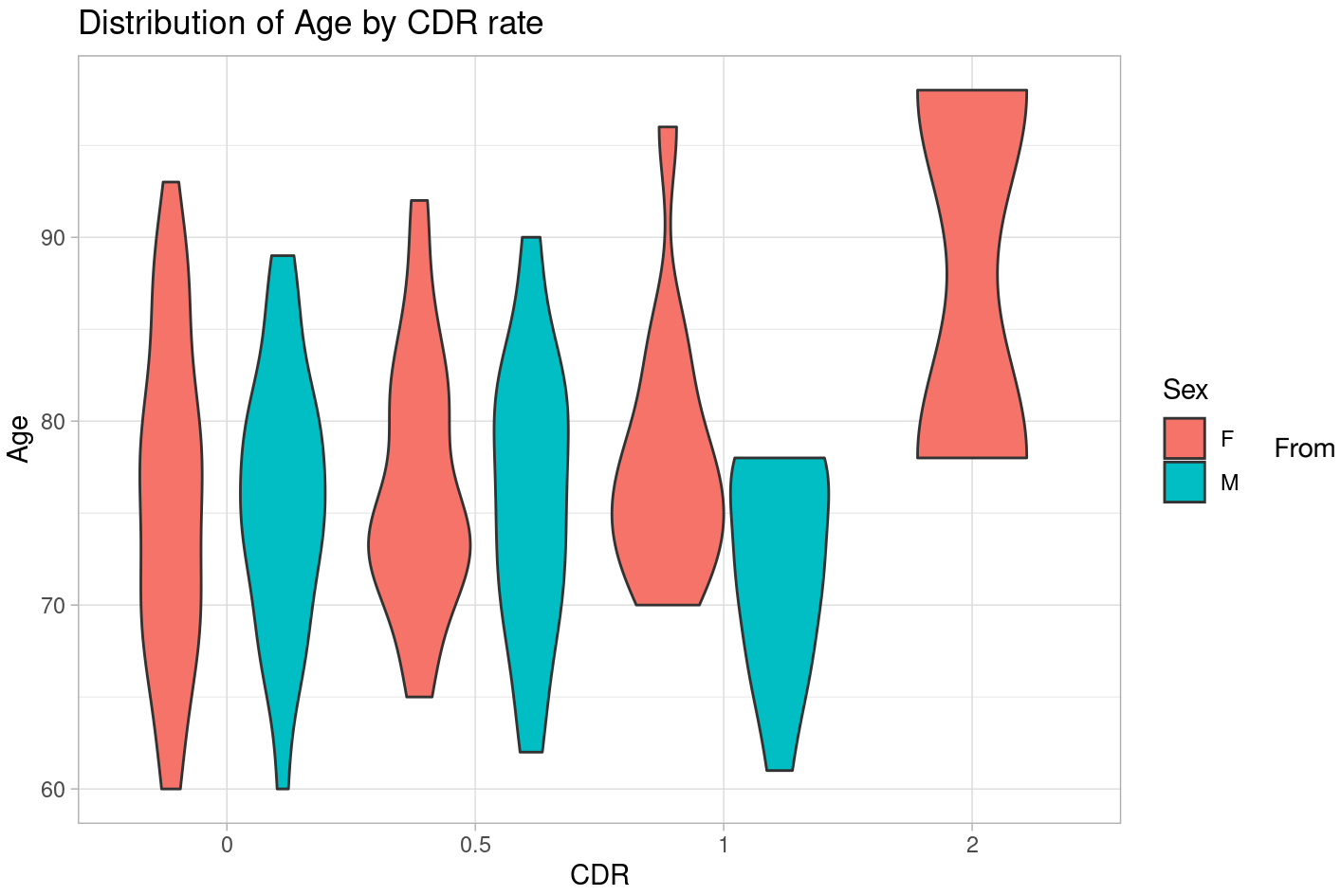
```
##             ID M.F Hand Age Educ SES MMSE CDR eTIV  nWBV   ASF Delay
## 1 OAS1_0425_MR1   F    R  78    1   4   23 1.0 1461 0.715 1.201   N/A
## 2 OAS1_0315_MR1   M    R  77    5   1   25 0.5 1604 0.773 1.094   N/A
## 3 OAS1_0314_MR1   M    R  27   NA  NA   NA  NA 1720 0.840 1.020   N/A
## 4 OAS1_0395_MR1   F    R  26   NA  NA   NA  NA 1295 0.834 1.356   N/A
## 5 OAS1_0332_MR1   M    R  72    1   3   29 0.0 1734 0.762 1.012   N/A
```

Data Manipulation

In this section of code I wanted to clean the data by removing unnecessary columns (Hand and Delay), handle missing values, and create a new column Dementia based on the CDR variable.

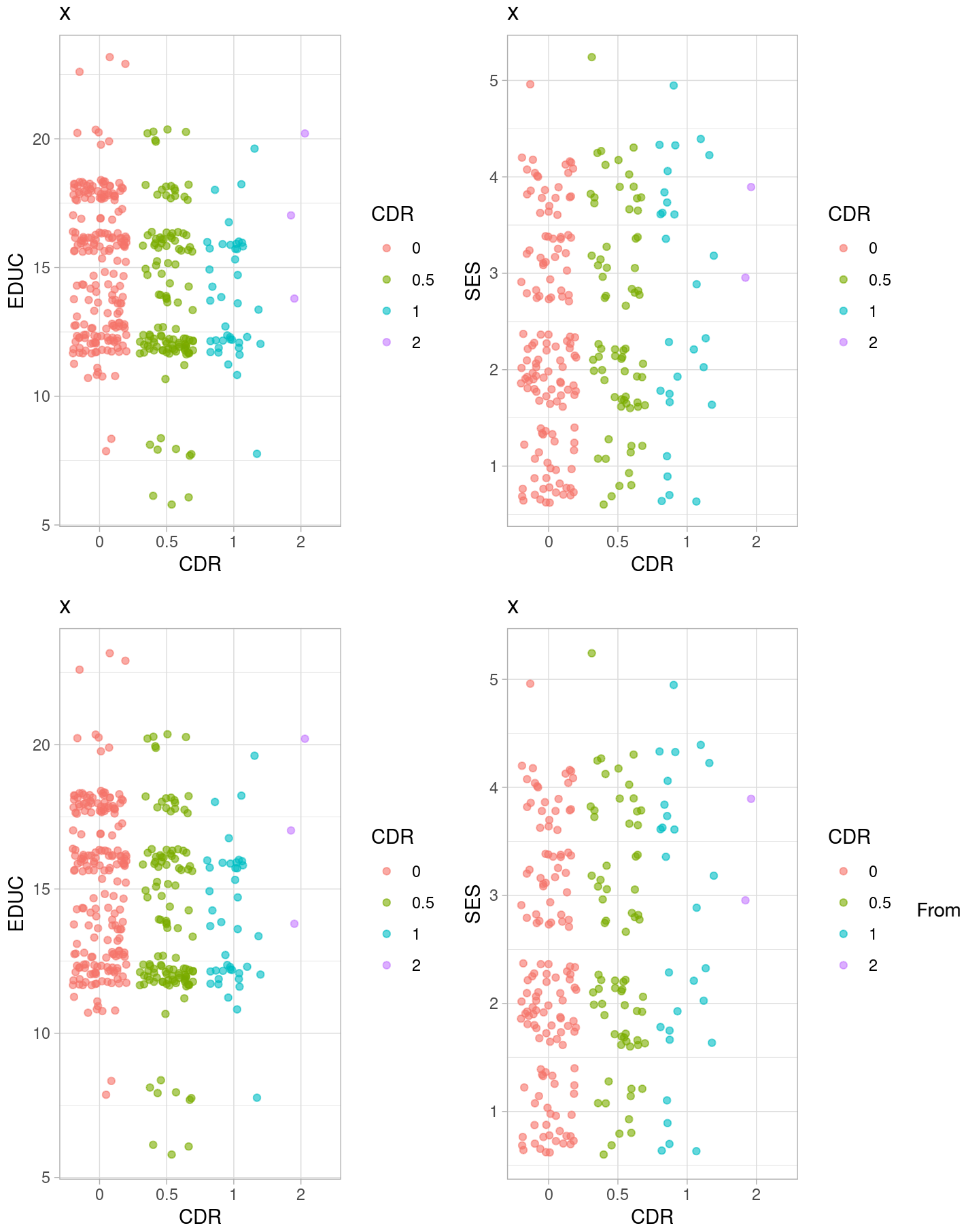
Exploratory Data Analysis

In this section I wanted to include violin plots and jitter plots to visualize the distribution and relationships between variables such as age, gender, education level, socioeconomic status, MMSE score, whole-brain volume, and dementia diagnosis (CDR). The analysis aims to identify any potential associations or patterns that may help in understanding dementia indicators



this violin plot, there seems to be no obvious connection between Age/Sex and Dementia Diagnosis. Moving on…

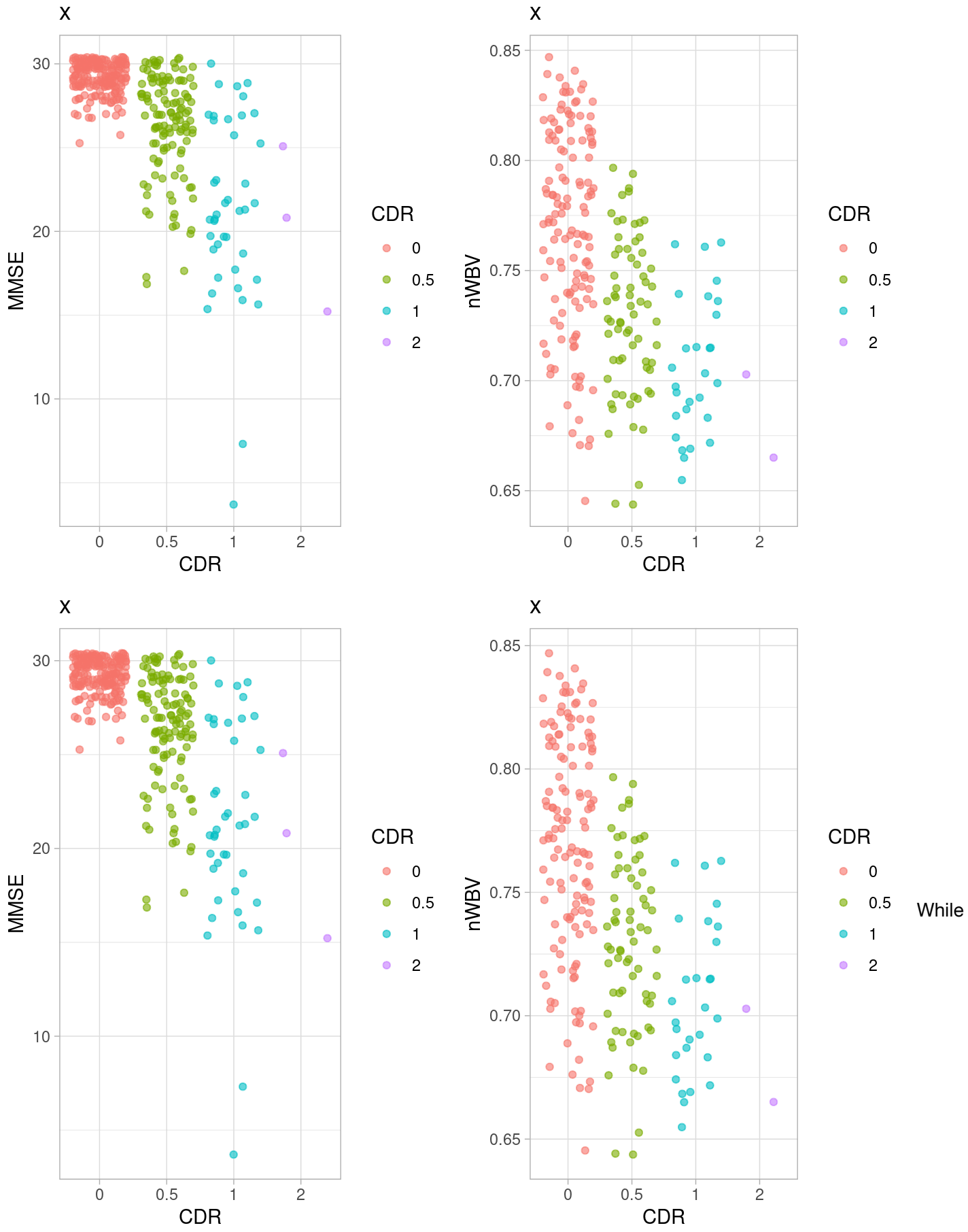
Distribution of Education and Social Economic Status



From this jitter plot, there still seems to be no obvious connection between Education Level/Social Economic Status

and Dementia Diagnosis. Sigh… we move on.

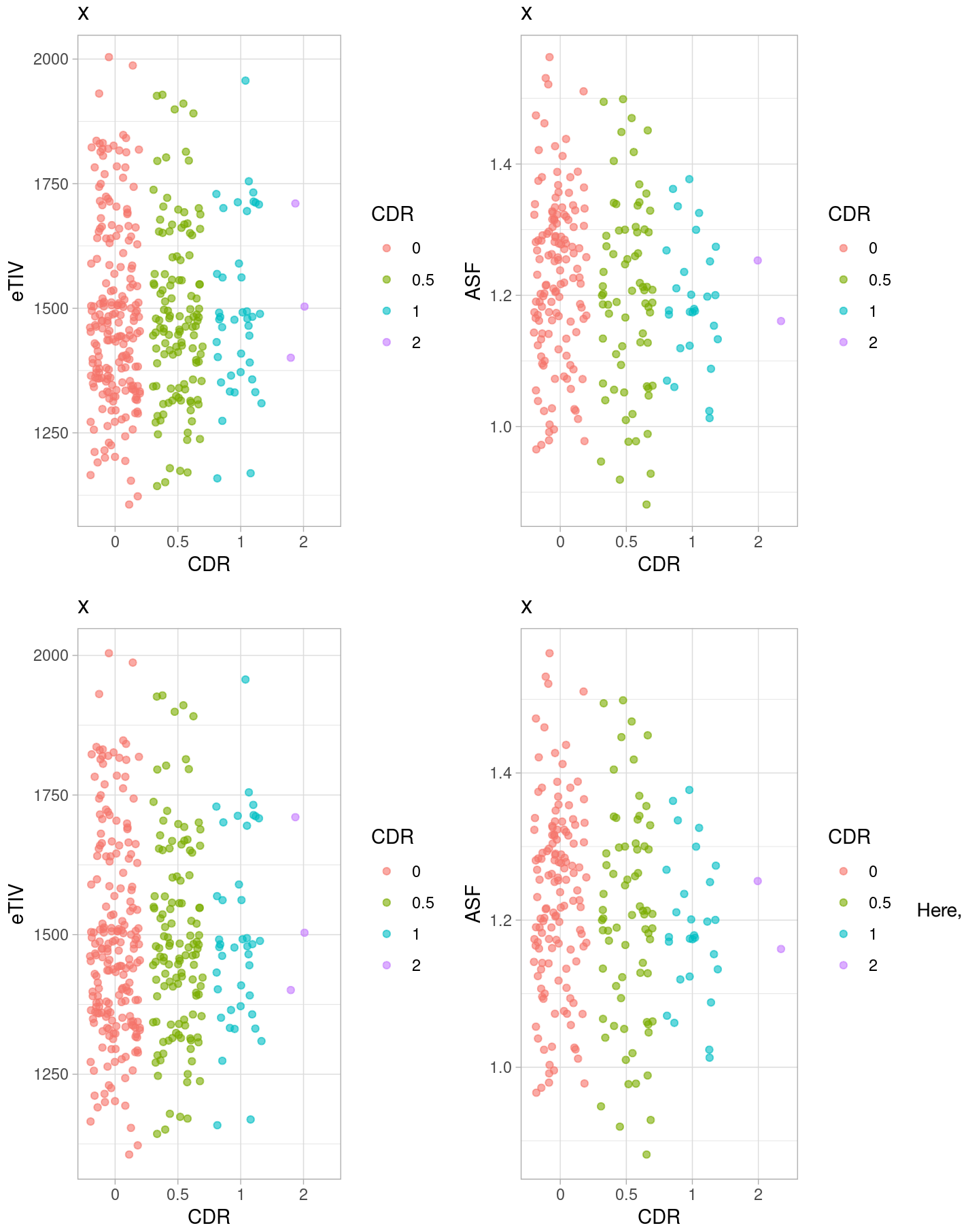
Distribution of MMSE Score and Whole-brain Volume



the MMS examination results of subjects not diagnosed with Dementia concentrate near 27-30 point rate, MMSE

results of subjects diagnosed with Dementia seems to be more spread out. We can see that subjects had the highest MMSE score but still have Clinical Dementia Rating of 0.5 or 1.

Distribution of Total Intracranial Volume and Atlas Scaling Factor



Here,

we cqn see that normalized whole-brain volume seems to have a bigger spread for subjects with CDR = 0 and

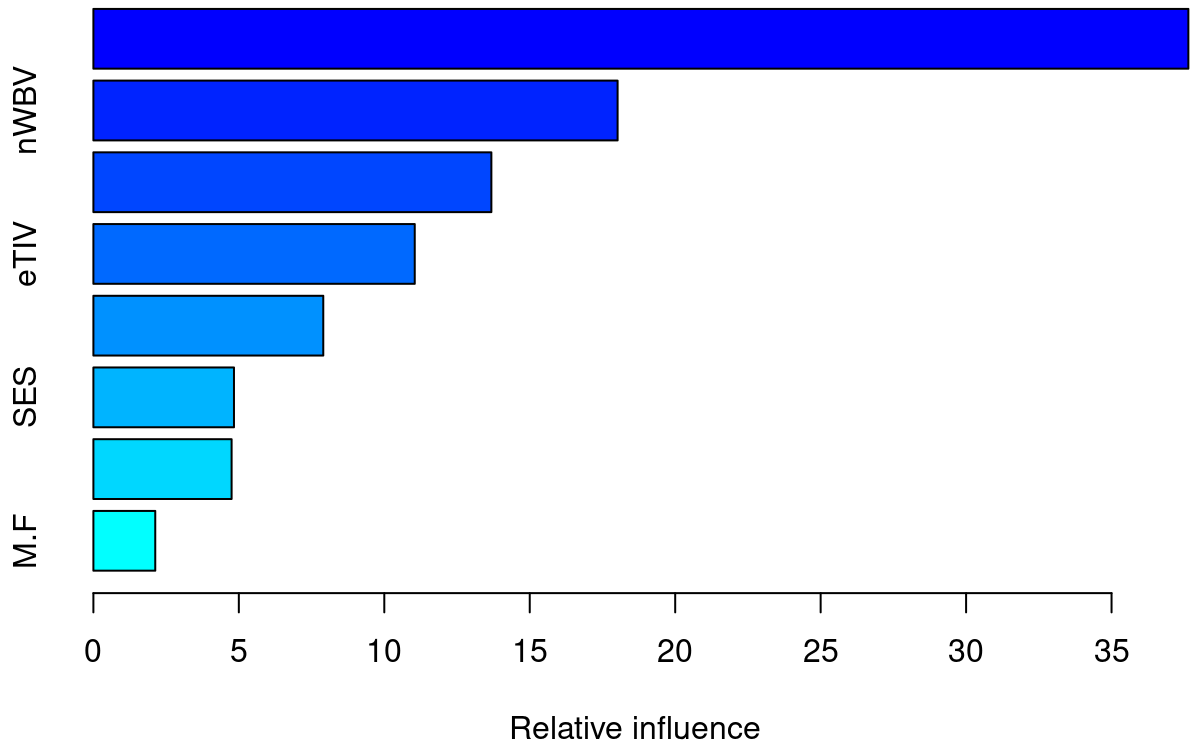
narrows as CDR increases.

Preparing the data for GBM model

Gradient-Boosting Model

I wanted to use the Gradient-Boosting Model because it predicts dementia diagnosis using the trained model on the test dataset and evaluates the model's performance using confusion matrix analysis and area under the ROC curve (AUC) calculation. The AUC value is computed to assess the model's predictive performance, with higher values indicating better prediction accuracy.

```
## A gradient boosted model with multinomial loss function.
## 5000 iterations were performed.
## The best test-set iteration was 130.
## There were 8 predictors of which 8 had non-zero influence.
```



```
##       var   rel.inf
## MMSE MMSE 37.640364
## nWBV nWBV 18.020083
## Age   Age 13.680484
## eTIV eTIV 11.046407
## ASF   ASF  7.902432
## SES   SES  4.832698
## EDUC EDUC  4.750985
## M.F   M.F  2.126547
```

```
## Warning in predict.gbm(object = model_gbm, newdata = select(test, -CDR), : NAs
## introduced by coercion
```

```
## NULL
```

```
## [1] "0"   "0.5" "1"   "2"
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0 0.5  1  2
##        0    0   0  0  0
##        0.5  0   0  0  0
##        1   36  14  0  0
##        2    1  12  3  1
##
## Overall Statistics
##
##                Accuracy : 0.0149
##                  95% CI : (4e-04, 0.0804)
##     No Information Rate : 0.5522
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : -0.0231
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 0 Class: 0.5 Class: 1 Class: 2
## Sensitivity            0.0000     0.0000  0.00000  1.00000
## Specificity            1.0000     1.0000  0.21875  0.75758
## Pos Pred Value            NaN        NaN  0.00000  0.05882
## Neg Pred Value         0.4478     0.6119  0.82353  1.00000
## Prevalence             0.5522     0.3881  0.04478  0.01493
## Detection Rate         0.0000     0.0000  0.00000  0.01493
## Detection Prevalence   0.0000     0.0000  0.74627  0.25373
## Balanced Accuracy      0.5000     0.5000  0.10938  0.87879
```

What is AUC?

AUC is an abbrevation for *area under the curve*. It is used in classification analysis in order to determine which of the used models predicts the classes best. The closer AUC for a model comes to 1, the better it is. So models with higher AUCs are preferred over those with lower AUCs.

```
## [1] "AUC for GBM Model = 0.92"
```

I wanted to use the gradient boosting machine learning model because I believe it is useful for dementia prediction. Here, I wanted to combine the strengths of ensemble learning, stage-wise training, flexible loss function optimization, and regularization to create accurate and robust predictive models. Therefore, it is well-suited for handling the complexities and challenges of medical datasets.

Conclusion

We can see that the GBM model gives us an accuracy of prediction about ~70%. We could also see that Clinical Dementia Rating higly depends of result of Mini-Mental State Examination, while Age, Educational Level and Social-Economic Status have not great influence. Although, it is important to remember that Dementia and Alzheimer's desease is complex mental issue, so we can not fully rely on ML algorithms to make a diagnosis. However, it can help us interpret large amounts of medical data to find the overall bigger picture.