Clemson University

Project Title: Detection of Alzheimer's disease using machine learning models.

CPSC 6300: Applied Data Science
Semester: Fall 2023
Check Point - 1

Course Instructor: **Dr. Nina Hubig**

Mentor : **Samaneh**

Akash Venugopal	C11430782
Bhavya Pulagam	C93153306
Michelle Sun	C88536352
Raghavendra Niteesh	C52595319
Sricharan Nibhanupudi	C17596944

Summary of the data set that, at a minimum, answers the following questions:

This dataset has 150 longitudinal participants, ranging in age from 60 to 96. For a total of 373 imaging sessions, each individual was scanned on two or more occasions, separated by at least a year. Three or four separate T1-weighted MRI scans performed during a single scan session are presented for each patient. There are men and women among the subjects, and they are all right-handed. Throughout the investigation, 72 of the individuals were classified as nondemented. Of the people scanned, 51 had mild to moderate Alzheimer's disease, and 64 were classified as demented at the time of their first visit and remained so for successive scans. After being classified as nondemented at their first visit, an additional 14 participants were later classified as demented at a later visit.

What is the unit of analysis?

The unit of analysis in this dataset is an individual subject. Each subject, characterized by age, gender, and dementia status, contributes multiple data points corresponding to different imaging sessions and the number of T1-weighted MRI scans obtained during those sessions. The dataset is longitudinal, capturing changes over time for each subject.

How many observations in total are in the data set?

Our experiment uses a CSV file from the Oasis Longitudinal Demographics, which has 374 rows and 15 columns of scan-related data on 150 participants that were recorded at least a year apart.

How many unique observations are in the data set?

There are no duplicate observations throughout the dataset, making all of the 374 rows unique. However, there are observations of the same participants, but over different time frames.

What time period is covered?

The time period differs here for every observation and each participant. While every participant has at least one observation present, not all the participants have data collected over many visits.

Column Descriptors

Column name	Full form
EDUC	Years Of Education
SES	Socio Economic Status
MMSE	Mini mental State Examination
CDR	Clinical Dementia Rating
eTIV	Estimated Total Intracranial Volume
nWBV	Normalize Whole Brain Volume
ASF	Atlas Scaling Factor

Brief summary of any data cleaning steps you have performed. For example, are there any particular observations / time periods / groups / etc. you have excluded?

Data Cleaning Steps:

Import csv file using pandas

```
df = pd.read_excel('oasis_longitudinal_demographics.xlsx')
df.head()
```

	Subject ID	MRI ID	Group	Visit	MR Delay	M/F	Hand	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
0	OAS2_0001	OAS2_0001_MR1	Nondemented	1	0	M	R	87	14	2.0	27.0	0.0	1986.550000	0.696106	0.883440
1	OAS2_0001	OAS2_0001_MR2	Nondemented	2	457	M	R	88	14	2.0	30.0	0.0	2004.479526	0.681062	0.875539
2	OAS2_0002	OAS2_0002_MR1	Demented	1	0	M	R	75	12	NaN	23.0	0.5	1678.290000	0.736336	1.045710
3	OAS2_0002	OAS2_0002_MR2	Demented	2	560	M	R	76	12	NaN	28.0	0.5	1737.620000	0.713402	1.010000
4	OAS2_0002	OAS2_0002_MR3	Demented	3	1895	M	R	80	12	NaN	22.0	0.5	1697.911134	0.701236	1.033623

In the following snippet, we tried to look for the number of null values in the dataset.

```
na_counts = df.isnull().sum()
print(na_counts)
```

```
Subject ID     0
MRI ID         0
Group          0
Visit          0
MR Delay       0
M/F            0
Hand           0
Age            0
EDUC           0
SES           19
MMSE           2
CDR            0
eTIV           0
nWBV           0
ASF            0
dtype: int64
```

Based on the above result, there are 21 null values in the dataset, 19 belong to SES and 2 belong to MMSE. We removed these values from the dataset in the following snippet and printed the null values to reassure the same.

```
df.dropna(subset=['SES','MMSE'], inplace=True)
na_counts = df.isnull().sum()
print(na_counts)
```

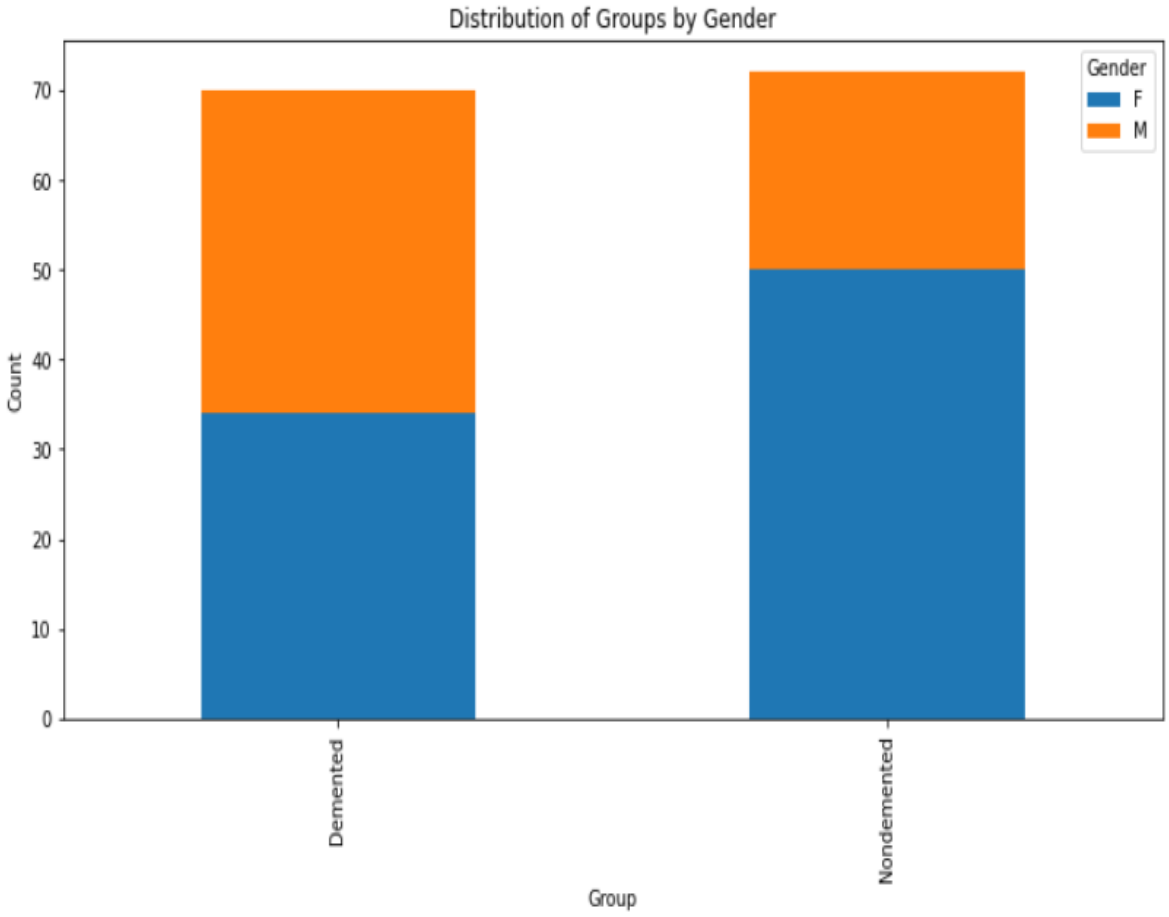
```
Subject ID    0
MRI ID        0
Group         0
Visit         0
MR Delay      0
M/F           0
Hand          0
Age           0
EDUC          0
SES           0
MMSE          0
CDR           0
eTIV          0
nWBV          0
ASF           0
dtype: int64
```

We considered only the visit 1 for each participant for the model, so we filtered that data and we have replaced the group name ‘Converted’ to ‘Demented’ as the person’s status has been converted from the state ‘Non-Demneted’.

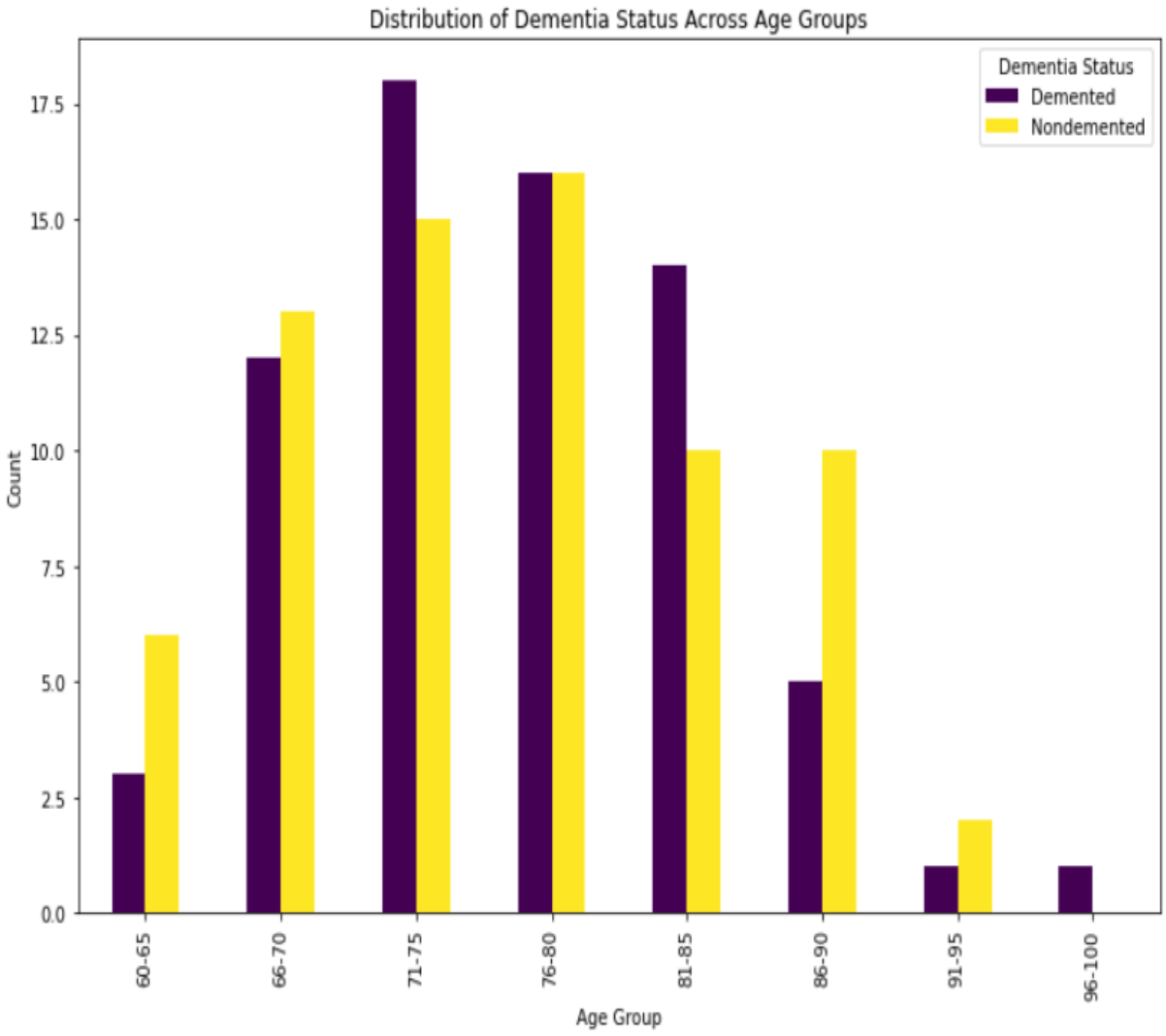
```
df = df.loc[df['Visit']==1] # use first visit data only because of the analysis we're doing
df = df.reset_index(drop=True) # reset index after filtering first visit data
df['Group'] = df['Group'].replace(['Converted'], ['Demented'])

group_counts = df['Group'].value_counts()
```

Description of outcome with an appropriate visualization technique

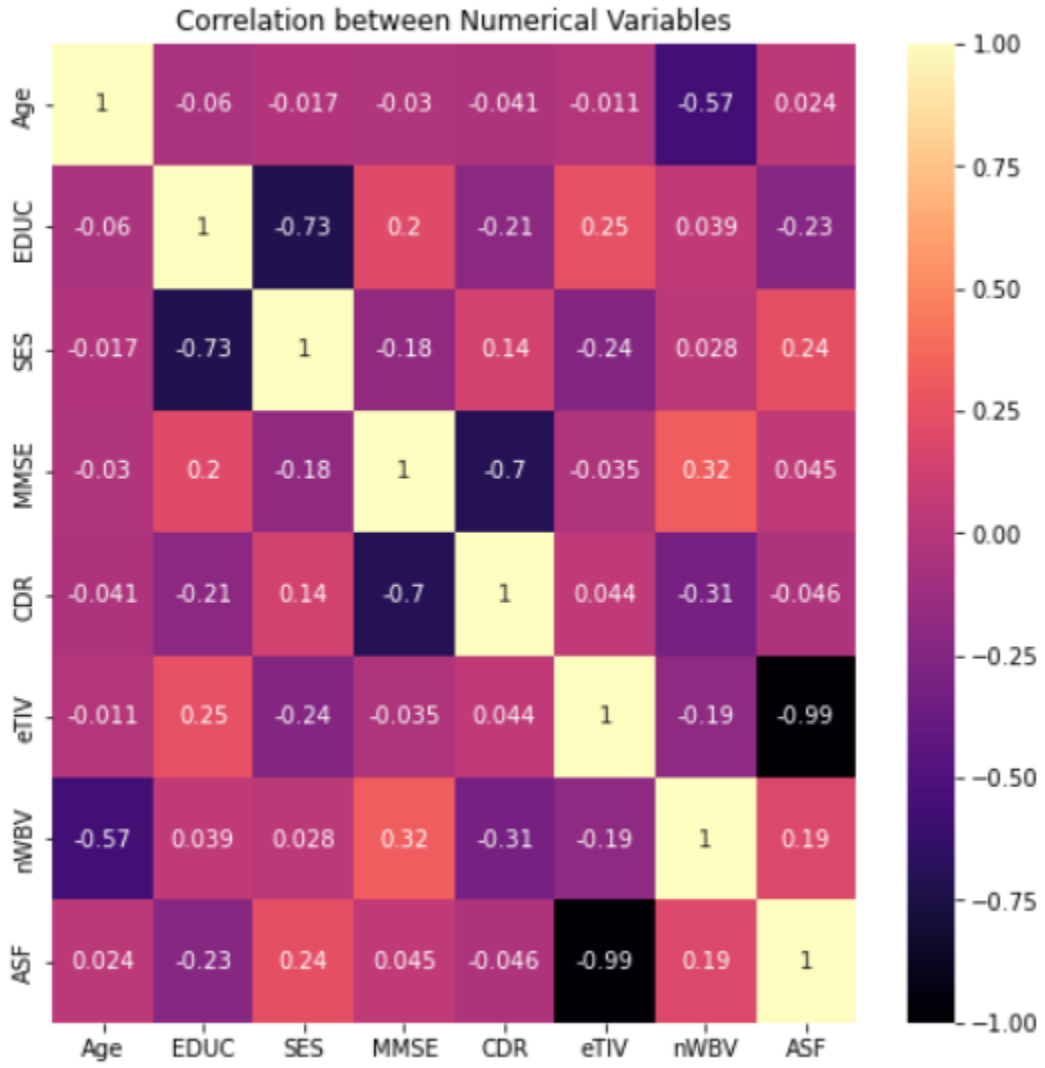
- Men are more likely to have dementia, an Alzheimer's Disease, than women. In the following stacked bar graph, we can observe that the distribution of men is more in the ‘Demented’ group and vice-versa for the ‘Nondemented’ group.



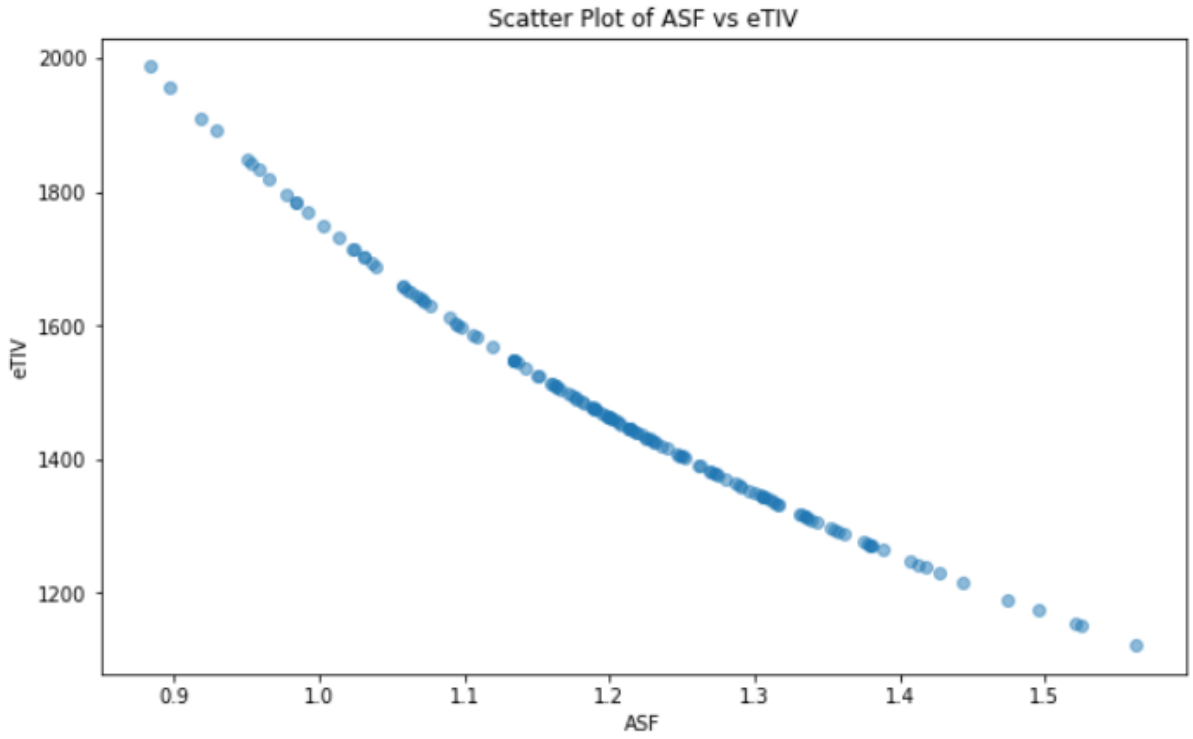
- Higher concentration of 71-85 years old in demented group than those in the nondemented patients. In the following grouped bar plot, we can observe that the count of the Demnted group in the groups of 71-75, 76-80 & 81-85 is significantly high when compared to that of the other age groups.


Distribution of Dementia Status Across Age Groups
Dementia Status
Demented
Nondemented
Count
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
60-65
66-70
71-75
76-80
81-85
86-90
91-95
96-100
Age Group

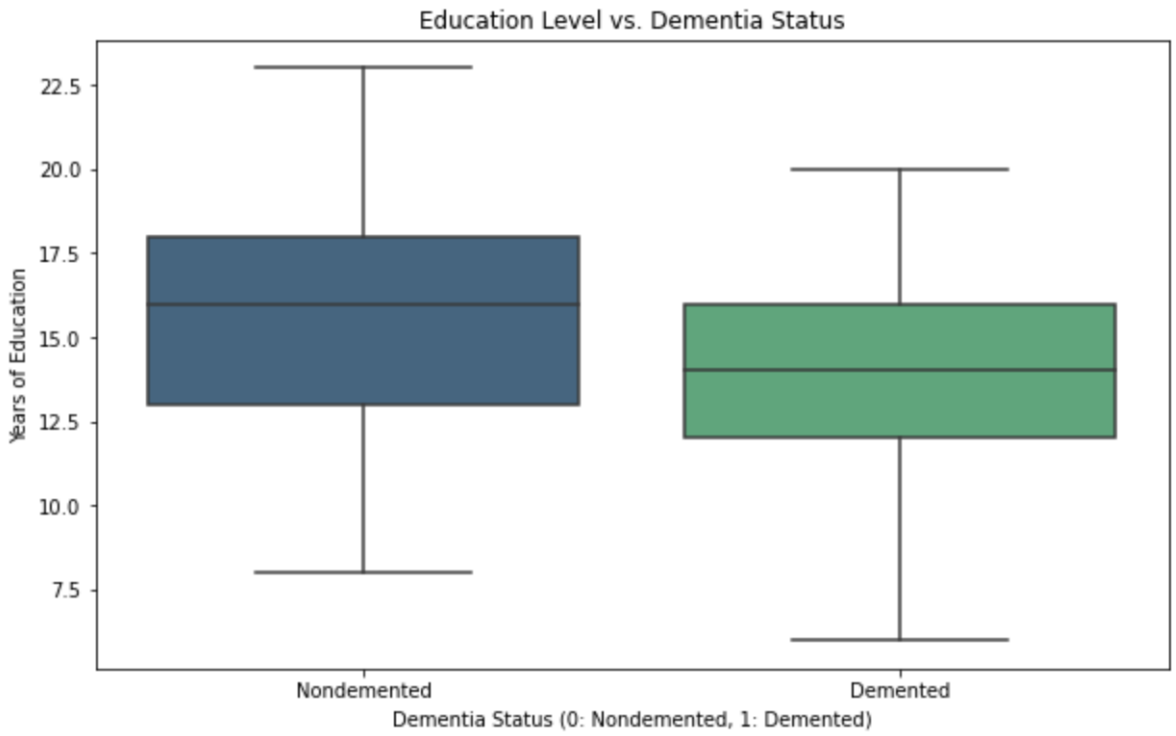
- Based on the following heat map that represents the correlation values between the different columns of the dataset , we can observe that ASF and eTIV have a negative correlation value of -0.99, suggesting that they are inversely related to one another implying that as the value of ASF goes up, the value of eTIV goes down and vice-versa.



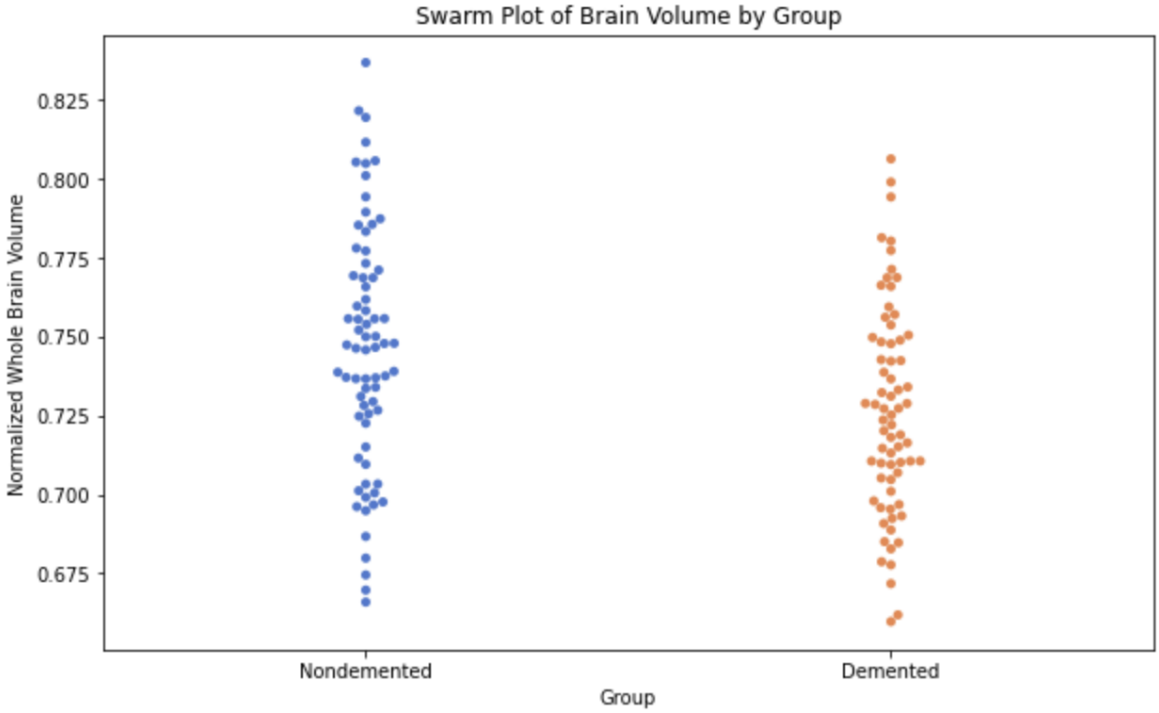
- The following scatterplot between ASF and eTIV also depicts the inverse relationship between these two variables.



2. Demented patients were less educated in terms of years of education.

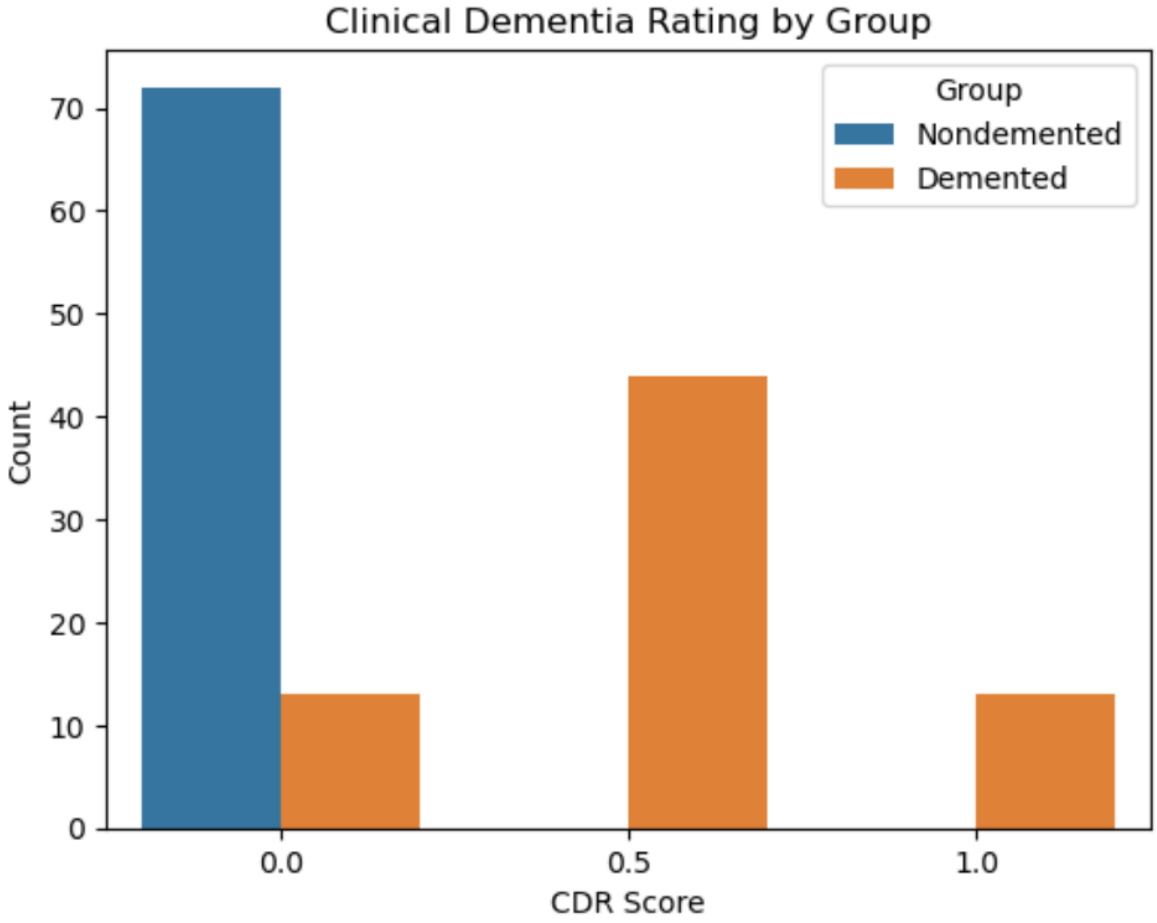


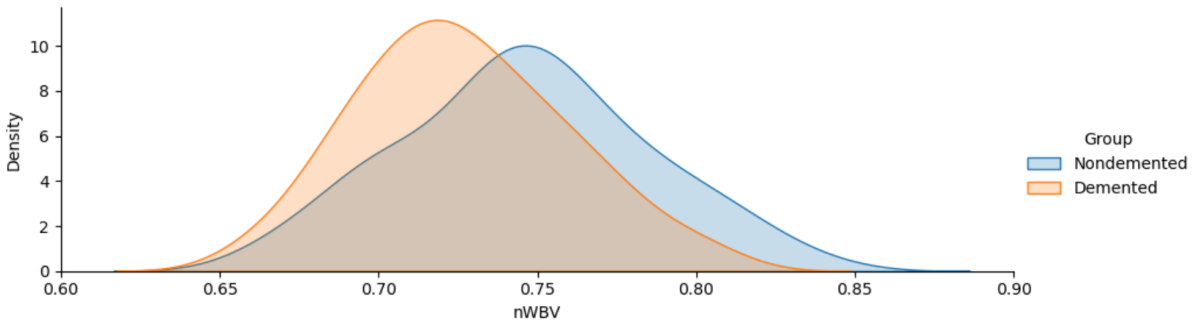
- The Nondemented group has higher brain volume than Demented group.



Description of key predictors with appropriate visualization techniques that compare predictors to the response. You should investigate all predictors in your data as part of your project. For the purpose of this assignment, pick the one or two predictors that you think are going to be most important in explaining the outcome. Your selection of predictors can either be guided by your domain knowledge or be the result of your EDA on all predictors.

nWBV, normalized whole brain volume, and CDR, clinical dementia rating, are the two predictors we expect to be most important in predicting whether a patient has dementia or not. nWBV scores on average are lower for individuals with dementia than those without, while CDR scores are higher for individuals with dementia. The CDR, or clinical dementia rating, is strongly correlated with whether a patient has dementia, with all non-demented patients having a CDR score of 0 and most demented patients having a CDR score of 0.5


Clinical Dementia Rating by Group
Group
Nondemented
Demented
Count
CDR Score
0
10
20
30
40
50
60
70
0.0
0.5
1.0


Density
nWBV
Group
Nondemented
Demented
0
2
4
6
8
10
0.60
0.65
0.70
0.75
0.80
0.85
0.90