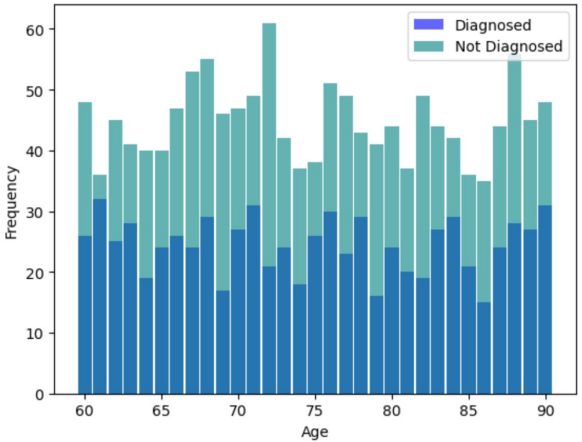
Final Project: Investigation of Alzheimer's Disease

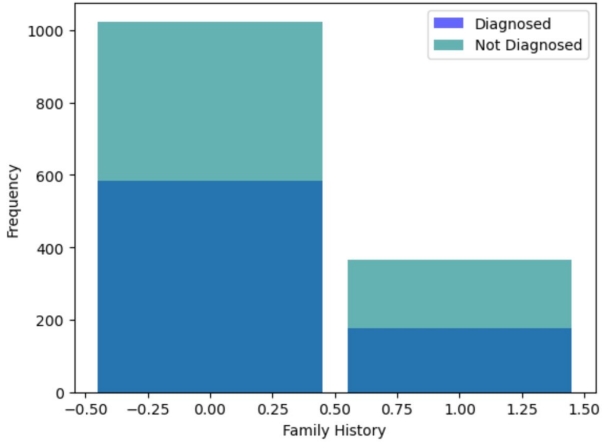
Kathryn Howe
DSC530-T301
Prof. Cary Jim
10 August 2024

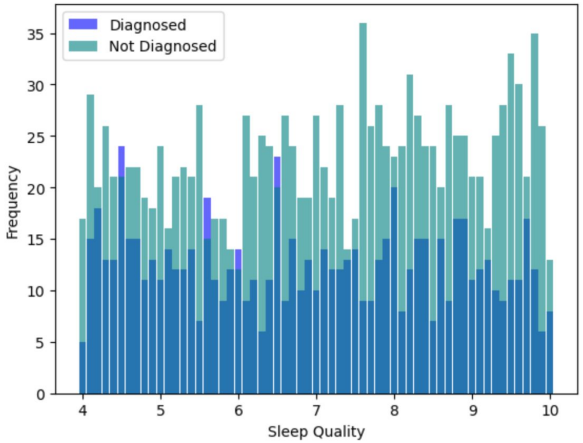
Variables

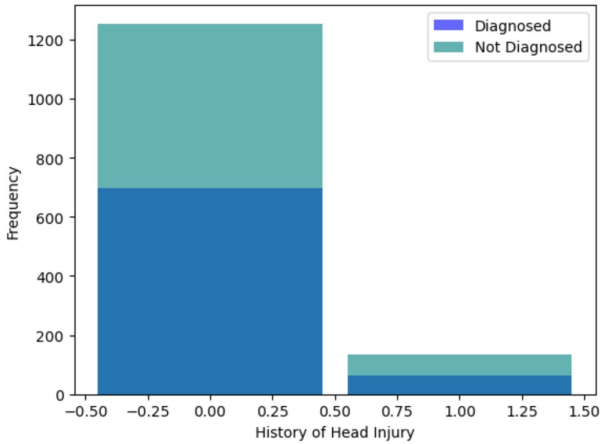
- PatientID: this is the variable assigned to each patient
- Diagnosis: this is the diagnosis that the patient recieved, 0 indicates no, 1 indicates yes
- FamilyHistoryAlzheimers: family history, 0 indicates no, 1 indicates yes
- Age: the age of the patient, ranging from 60-90 years old
- SleepQuality: sleep quality scores, ranging from 4-10
- HeadInjury: history of head injury, 0 indicates no, 1 indicates yes
- DietQuality: Diet quality score, ranging from 0 to 10
- MMSE: Mini-Mental State Examination score, ranging from 0 to 30. Lower scores indicate cognitive impairment
    - This variable was brought in later on in the investigation as a way to better quantify the diagnosis

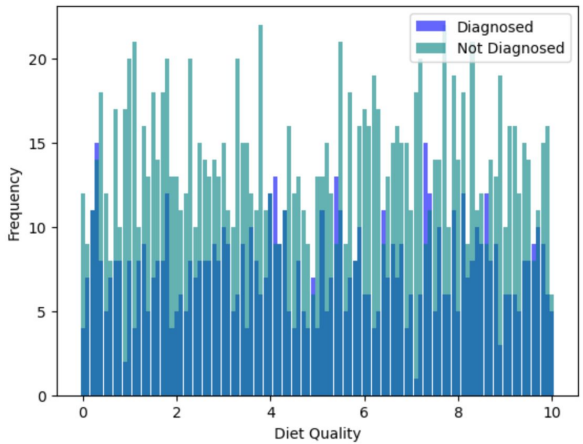
Histograms for Each Variable











Histogram Analysis

- The graphs for each variable seem to make the same shape, however the undiagnosed patients occur at a higher frequency.
- This is likely just due to more undiagnosed patients being sampled in this.
- This can be corroborated by utilizing the count function which returned

```
alzheimers.Diagnosis.value_counts()
```

```
Diagnosis
0    1389
1     760
Name: count, dtype: int64
```

Summary Statistics

Diagnosed:

- Age
  - Mean: 74.84210526315789
  - Median: 75.0
  - Mode: 0 61
  - Name: Age, dtype: int64
  - Variance: 83.72207197836484
  - Standard Deviation: 9.149976610809714
- Family History
  - Mean: 0.23289473684210527
  - Median: 0.0
  - Mode: 0 0
  - Name: FamilyHistoryAlzheimers, dtype: int64
  - Variance: 0.1788901601830639
  - Standard Deviation: 0.42295408755923364
- Sleep Quality
  - Mean: 6.916292027001315
  - Median: 6.912163812499999
  - Mode: 0 4.006171
  - Name: SleepQuality, Length: 760, dtype: float64
  - Variance: 3.0821204810673564
  - Standard Deviation: 1.755596901645522

Undiagnosed:

- Age
  - Mean: 74.94528437724982
  - Median: 75.0
  - Mode: 0 72
  - Name: Age, dtype: int64
  - Variance: 79.29383401489265
  - Standard Deviation: 8.904708530597318
- Family History
  - Mean: 0.26277897768178543
  - Median: 0.0
  - Mode: 0 0
  - Name: FamilyHistoryAlzheimers, dtype: int64
  - Variance: 0.1938657587508271
  - Standard Deviation: 0.4403018950116239
- Sleep Quality
  - Mean: 7.124831737259899
  - Median: 7.238501023
  - Mode: 0 4.002629
  - Name: SleepQuality, Length: 1389, dtype: float64
  - Variance: 3.1123883004738118
  - Standard Deviation: 1.7641962193797525

Summary Statistics Continued…

Diagnosed:

- Diet Quality
  - Mean: 5.026581195803948
  - Median: 5.0825412745000005
  - Mode: 0 0.019940
  - Variance: 8.474704852229998
  - Standard Deviation: 2.9111346331336168
- Previously Having a Head Injury
  - Mean: 0.08421052631578947
  - Median: 0.0
  - Mode: 0 0
  - Name: HeadInjury, dtype: int64
  - Variance: 0.07722071978364957
  - Standard Deviation: 0.27788616335407845

Undiagnosed:

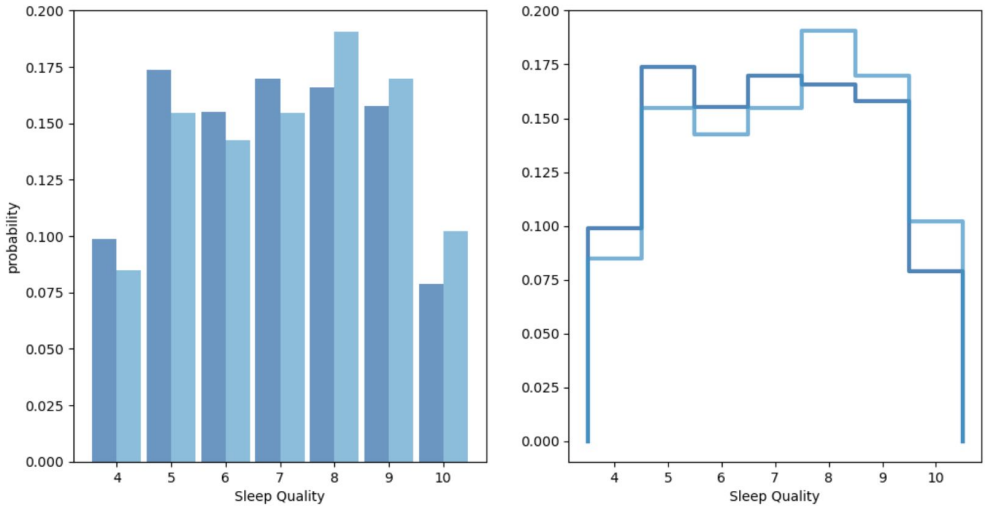
- Diet Quality
  - Mean: 4.9748392135658746
  - Median: 5.066292941
  - Mode: 0 0.009385
  - Variance: 8.46113172041315
  - Standard Deviation: 2.9088024546904436
- Previously Having a Head Injury
  - Mean: 0.09719222462203024
  - Median: 0.0
  - Mode: 0 0
  - Name: HeadInjury, dtype: int64
  - Variance: 0.08780911359944098
  - Standard Deviation: 0.2963260258557135

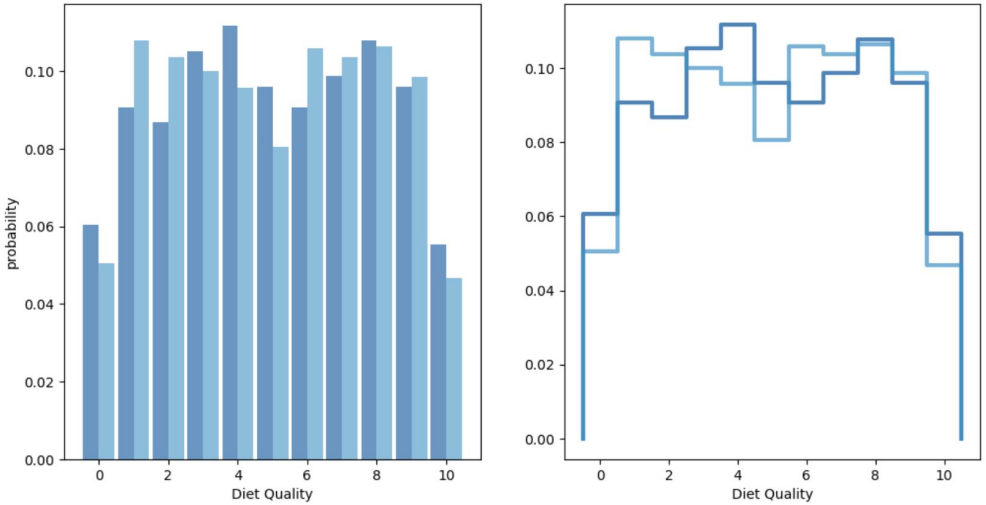
PMF

The results of the PMF analysis for Sleep Quality and Diet Quality for the diagnosed and undiagnosed patients is here.

Both graphs, as well as for both diagnoses, appear to make the same shape, with most of the distribution in the center, but a few values on the end.

This is what led to me doing normal distribution analysis later on, as it most closely resembles normal distribution.



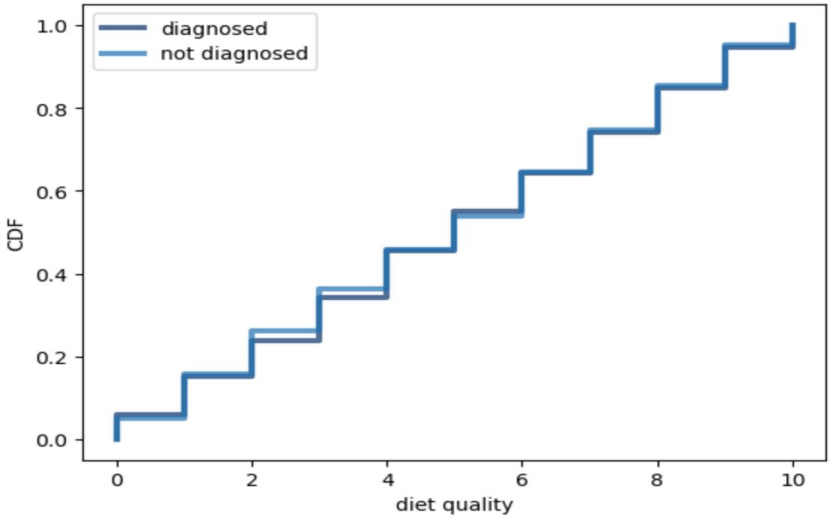


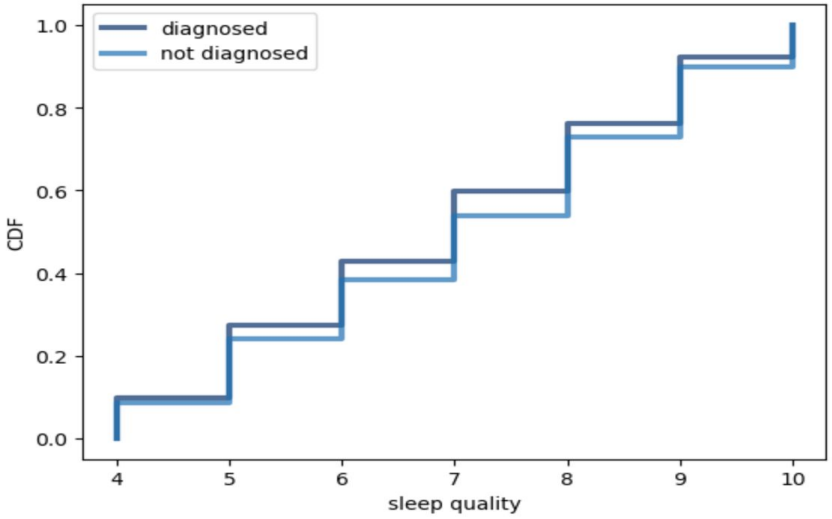
CDF

The same variables explored in our PMF were explored in the CDF.

In both of these CDFs, the steps are near identical in height, which indicates that these values are all occurring at relatively the same frequency for both those diagnosed with Alzheimer's as those without a diagnosis.

The question we are investigating is if there are any impacts or indicator of a future diagnosis, but thus far, the diagnosed patients and patients without a diagnosis are showing near identical results in most other categories.



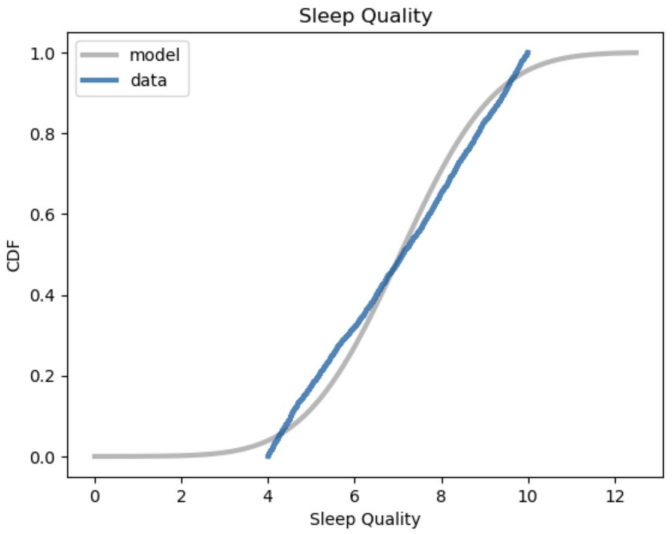


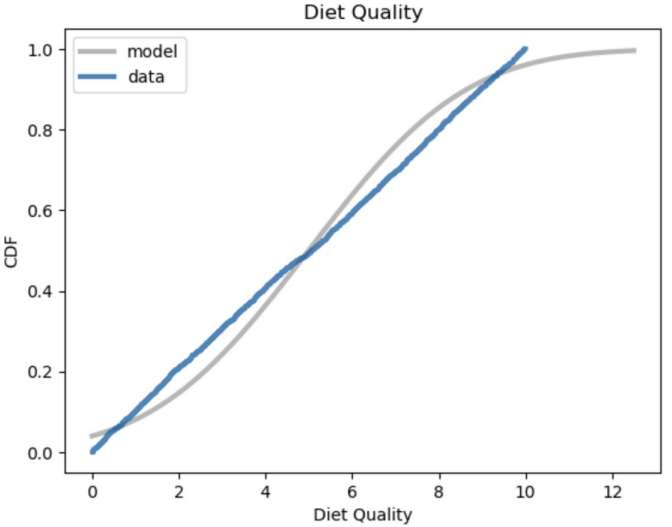
Analytical Distribution

As mentioned previously a normal distribution analysis was chosen for the variables based on the PMF.

For both of these, the data looks more like a fit line than data distribution. These appear to be almost uniformally distributed for the patients. The next part of this analysis will look into the differences between those diagnosed and those undiagnosed.

The analysis was reproduced and the distribution for those diagnosed and those undiagnosed are virtually identical. The means and variances for both are very similar and indicate there there is little difference in sleep quality between these groups of patients.

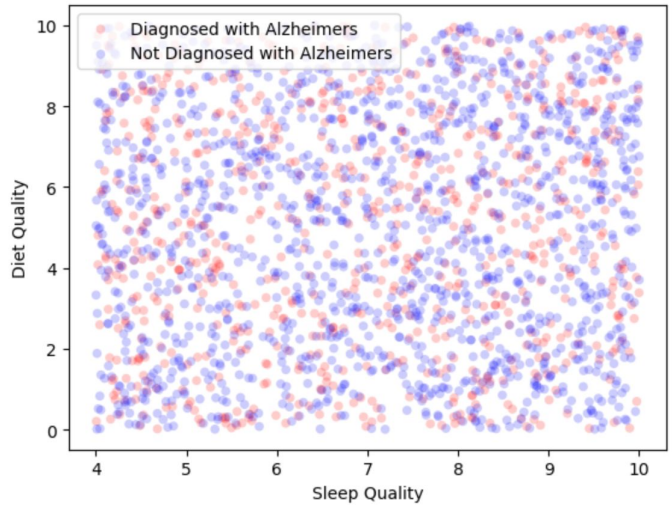


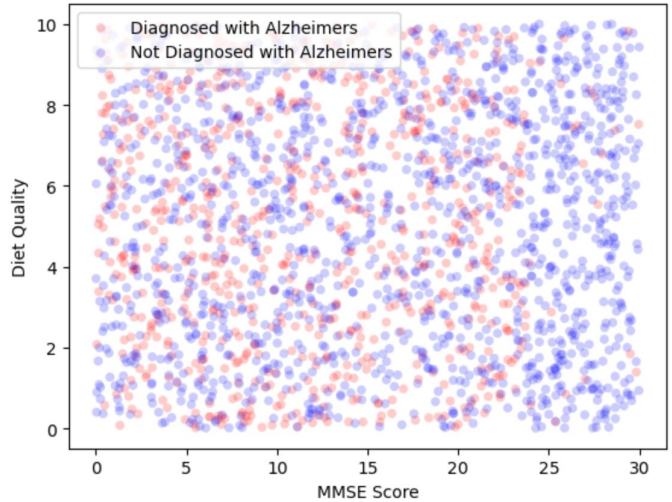


Scatter Plots

The ScatterPlot comparing the sleep and diet quality of those diagnosed with alzheimer's with those not diagnosed. This scatter plot shows no clear indication between these variables and the diagnosis.

Since diagnosis is a yes or no, I decided to introduce the MMSE as a method to compare the diagnosed and undiagnosed patients. In the scatter plot, there is a clear region where those undiagnosed have higher MMSE scores, however that seems uncorrelated to their quality of diet.





Hypothesis Testing

- For hypothesis testing, I utilized a modified CorrelationPermute function from the reading
- Hypothesis testing was performed comparing Age and the MMSE scores
- MMSE was utilized to give more variation to those experiencing cognitive impairment
- In the diagnosed dataset, the returned p value was 0.172
- In the undiagnosed patients, the p value returned was 0.277
- The values are higher which indicates that there is likely no correlation and any potential correlation is just due to chance

Regression Analysis

- I did OLS Regression Analysis comparing SleepQuality and MMSE with the diagnosis of the patient
- MMSE can give a more variable value associated with the diagnosis which is why it was used here
- The p-value of < 0.0001 indicates that the MMSE does likely have a correlation with diagnosis, which was expected.
- Interestingly, the p-value for SleepQuality was 0.009 which is slightly statistically significant, which could indicate a correlation with diagnosis.
    - However based on our other analysis, this correlation is likely due to chance

Overall

- None of the variable in our initial research seem to have a major correlation with diagnosis
- All of the distribution seem random and uniform
- The only variable that seemed to correlate with diagnosis was the MMSE, which was expected as it is part of the diagnosis process for patients
- Future research could entail performing these analyses on other variables present in this dataset such as:
  - Physical Activity
  - Alcohol Consumption
  - Etc
- It is important to find any correlation between these to help people who are predisposed to Alzheimer's an opportunity for early diagnosis and early treatment