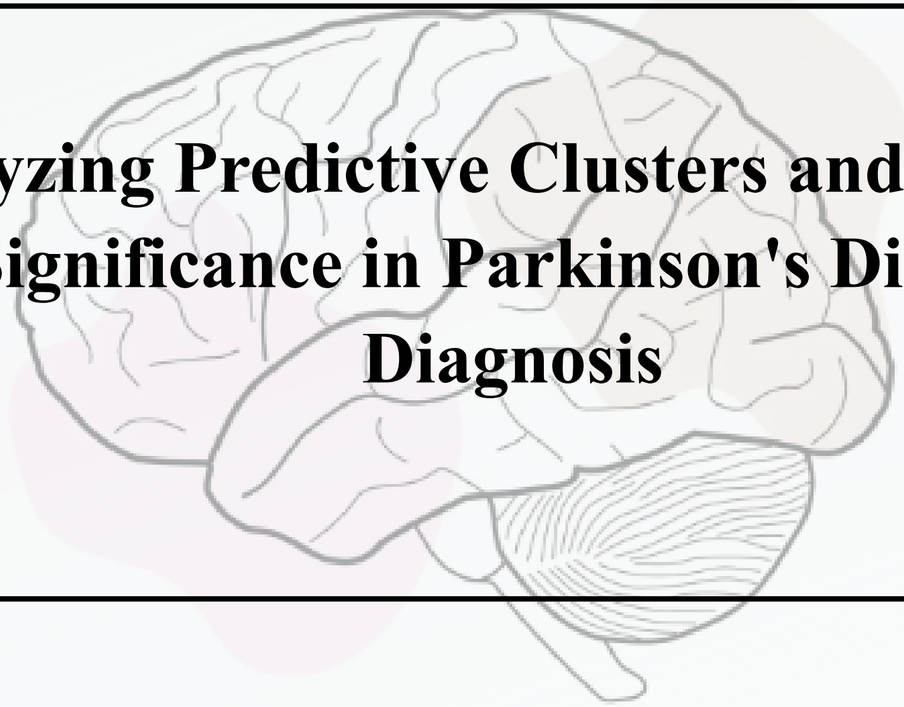
Analyzing Predictive Clusters and Variable Significance in Parkinson's Disease Diagnosis



Group 3 Team Members: Faizan Hussaini, Sri Ramya Panja, FNU Sahrash Fatima,Yesseswini Yenigandla

INTRODUCTION

HYPOTHESIS

• Null Hypothesis (H_0): None of the clusters (Demographic Details, Lifestyle Factors, Medical History, Clinical Measurements, Cognitive and Functional Assessments, Symptoms) are significantly associated with the diagnosis of Parkinson's Disease.

• Alternative Hypothesis (H_1): At least one of the clusters (Demographic Details, Lifestyle Factors, Medical History, Clinical Measurements, Cognitive and Functional Assessments, Symptoms) is significantly associated with the diagnosis of Parkinson's Disease.

OBJECTIVE: The objective of this project is to identify the cluster of variables most significantly associated with Parkinson's Disease diagnosis and determine the key variables within these clusters influencing the diagnosis.

SIGNIFICANCE: This project is important because Parkinson's Disease is usually diagnosed at a late stage, which makes the treatment less effective. Identifying key factors that predict the disease may help in early diagnosis, improvement of patient care, and personalized treatments. It will also support better healthcare planning and advance research to better understand and manage Parkinson's Disease(National Institute of Neurological Disorders & Stroke [US], 2004).

DATA AND VARIABLES

The dataset consists of 2,105 patients with six clusters of health information: Demographic Details (age, gender, ethnicity), Lifestyle Factors (BMI, smoking, physical activity), Medical History (hypertension, diabetes, family history), Clinical Measurements (blood pressure, cholesterol levels), Cognitive and Functional Assessments (UPDRS, MoCA scores), and Symptoms (tremor, rigidity, speech problems). It also contains a diagnosis indicator for Parkinson's versus non-Parkinson's. The dataset was derived from Kaggle(Parkinson's Disease Dataset Analysis, 2024).

Variable Name	Variable Type	Cluster
Age	Continuous	Demographic Details
Gender	Binary	Demographic Details
Ethnicity	Categorical	Demographic Details
BMI	Continuous	Lifestyle Factors
Smoking	Binary	Lifestyle Factors
Physical Activity	Continuous	Lifestyle Factors
Hypertension	Binary	Medical History
Diabetes	Binary	Medical History
Family History Parkinsons	Binary	Medical History
Systolic BP	Continuous	Clinical Measurements
Diastolic BP	Continuous	Clinical Measurements
Cholesterol Total	Continuous	Clinical Measurements
UPDRS (Unified Parkinson's Rating Scale)	Continuous	Cognitive & Functional Assessments
MoCA (Montreal Cognitive Assessment)	Continuous	Cognitive & Functional Assessments
Tremor	Binary	Symptoms
Rigidity	Binary	Symptoms
Speech Problems	Binary	Symptoms

DESCRIPTIVE STATISTICS AND VISUALIZATION

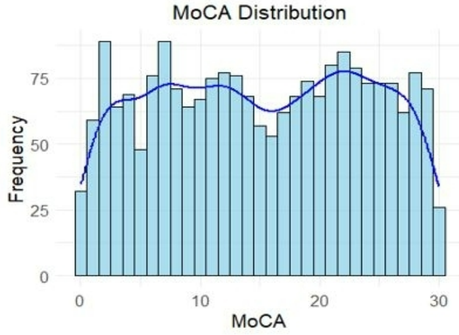
Categorical Variables

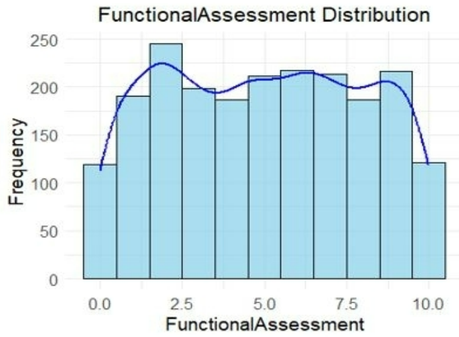
```
$Gender
  Var1 Freq Proportion
1    0 1068  0.5073634
2    1 1037  0.4926366

$Smoking
  Var1 Freq Proportion
1    0 1481  0.7035629
2    1  624  0.2964371

$FamilyHistoryParkinsons
  Var1 Freq Proportion
1    0 1798  0.8541568
2    1  307  0.1458432

$TraumaticBrainInjury
  Var1 Freq Proportion
1    0 1881  0.8935867
2    1  224  0.1064133

$Hypertension
  Var1 Freq Proportion
1    0 1798  0.8541568
2    1  307  0.1458432

$Diagnosis
  Var1 Freq Proportion
1    0  801  0.3805226
2    1 1304  0.6194774
```

Continuous Variables

```
print(summary_stats)

##   Age_Mean Age_Median Age_StdDev BMI_Mean BMI_Median BMI_StdDev
## 1  69.6019         70   11.59451 27.20949   27.18457   7.208099
##   AlcoholConsumption_Mean AlcoholConsumption_Median
AlcoholConsumption_StdDev
## 1                10.04041                  10.07034
5.687014
##   PhysicalActivity_Mean PhysicalActivity_Median PhysicalActivity_StdDev
## 1              5.016674                 5.03155                2.890919
##   DietQuality_Mean DietQuality_Median DietQuality_StdDev SleepQuality_Mean
## 1         4.912901           4.825187           2.872115          6.996639
##   SleepQuality_Median SleepQuality_StdDev SystolicBP_Mean
SystolicBP_Median
## 1            6.929819           1.753065        133.7197
133
##   SystolicBP_StdDev DiastolicBP_Mean DiastolicBP_Median DiastolicBP_StdDev
## 1          26.50236         90.24988                91           17.06149
##   CholesterolTotal_Mean CholesterolTotal_Median CholesterolTotal_StdDev
## 1              226.8608                228.5283               43.58941
##   CholesterolLDL_Mean CholesterolLDL_Median CholesterolLDL_StdDev
## 1            126.1479              126.8846              43.40704
##   CholesterolHDL_Mean CholesterolHDL_Median CholesterolHDL_StdDev
## 1            59.67035              59.34336              23.37092
##   CholesterolTriglycerides_Mean CholesterolTriglycerides_Median
## 1                      222.9405                         222.8025
##   CholesterolTriglycerides_StdDev UPDRS_Mean UPDRS_Median UPDRS_StdDev
## 1                        101.8958   101.4153      102.561     56.59145
##   MoCA_Mean MoCA_Median MoCA_StdDev FunctionalAssessment_Mean
## 1  15.09431    14.96357    8.643014                  4.989694
##   FunctionalAssessment_Median FunctionalAssessment_StdDev
## 1                    4.983227                    2.933877
```

Data distribution





EXPLORATORY DATA ANALYSIS:

Data upload-

Overall, we have 2,105 Rows,

and 35 columns.

```
# dataset
data <- read.csv("D:/iupui/2rd sem/R- stats/parkinsons_disease_data.csv")
# View the first few rows to confirm the data is loaded
head(data)
```

```
##    PatientID Age Gender Ethnicity EducationLevel      BMI Smoking
## 1       3058  85      0         3              1 19.61988       0
## 2       3059  75      0         0              2 16.24734       1
## 3       3060  70      1         0              0 15.36824       0
## 4       3061  52      0         0              0 15.45456       0
## 5       3062  87      0         0              1 18.61604       0
## 6       3063  68      1         2              1 39.42331       1
##    AlcoholConsumption PhysicalActivity DietQuality SleepQuality
## 1            5.108241         1.3806599    3.893969     9.283194
## 2            6.027648         8.4098041    8.513428     5.602470
```

Outliers :

We have no outliers for categorical variables.

```
##                                      Variable Outliers
## PatientID                           PatientID        0
## Age                                       Age        0
## Gender                                 Gender        0
## EducationLevel                  EducationLevel        0
## BMI                                       BMI        0
## Smoking                                 Smoking        0
## AlcoholConsumption           AlcoholConsumption        0
## PhysicalActivity               PhysicalActivity        0
## DietQuality                         DietQuality        0
## SleepQuality                       SleepQuality        0
## SystolicBP                           SystolicBP        0
## DiastolicBP                         DiastolicBP        0
## CholesterolTotal               CholesterolTotal        0
## CholesterolLDL                   CholesterolLDL        0
## CholesterolHDL                   CholesterolHDL        0
## CholesterolTriglycerides CholesterolTriglycerides        0
```

The dataset does not have a null value or duplicate.

*Checking for Null or Missing values:

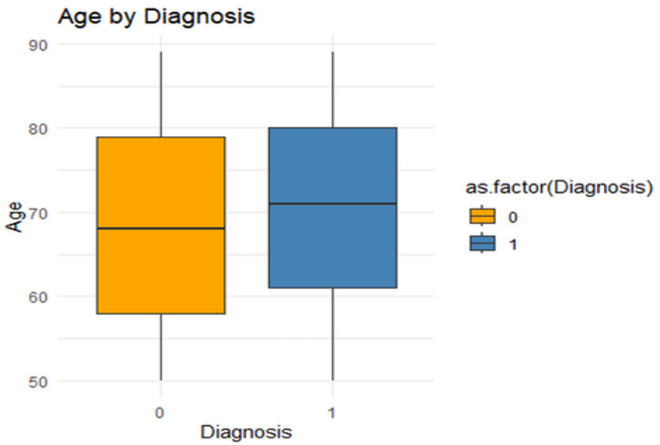
```
total_null_values <- sum(is.na(data))
print(total_null_values)
```

```
## [1] 0
```

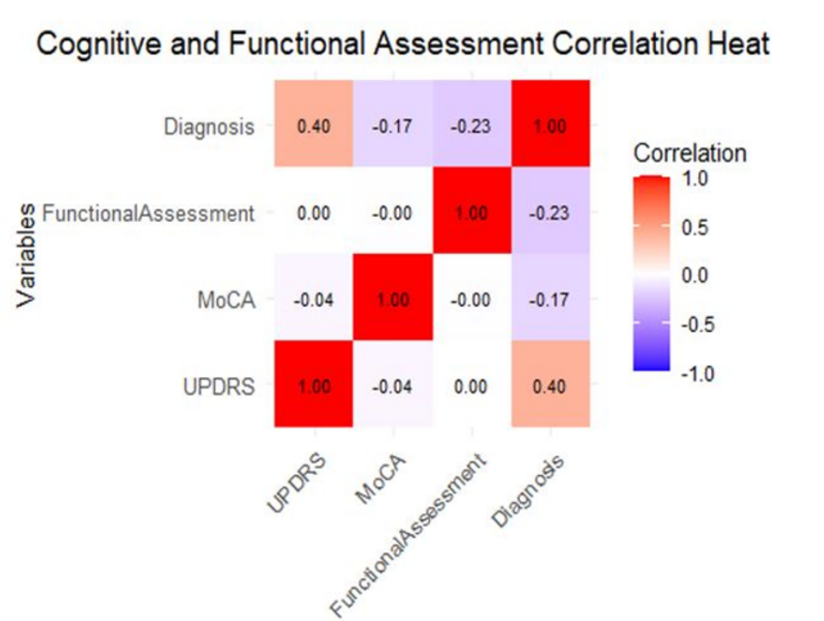
No Null values found.

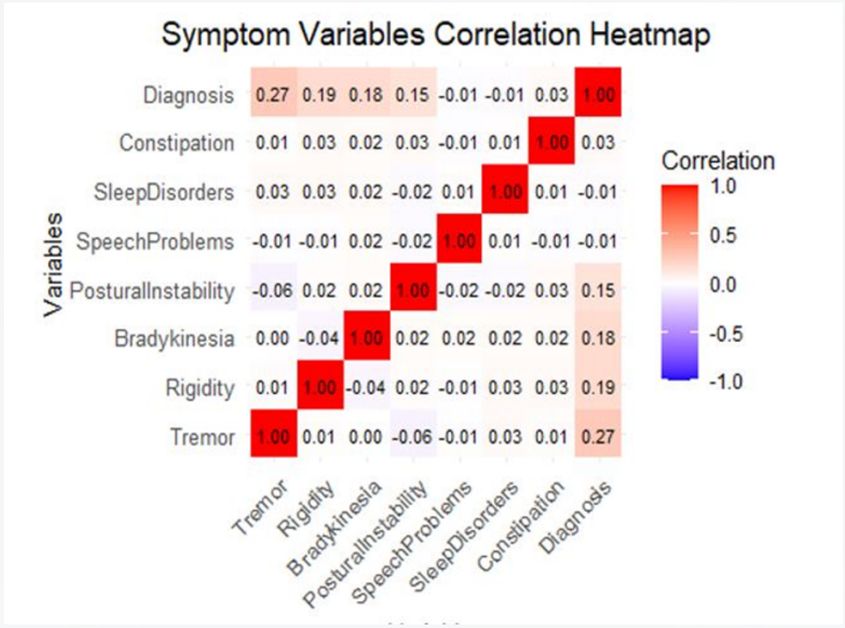
```
## [1] "Updated shape: 2105 rows and 35 columns"
```

The datasets doesnot have any duplicate values. As the shape is same before and after update.



HEATMAP FOR EACH CLUSTER





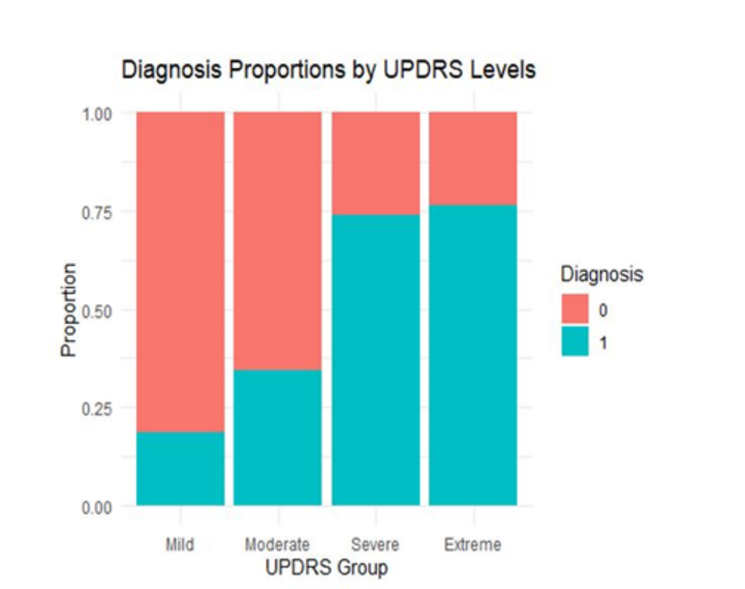
Highly Significant Variables: Age, Tremor, Bradykinesia, Postural Instability, UPDRS, MoCA. Moderately Significant Variables: Physical Activity, Sleep Quality, Hypertension, Depression, LDL Cholesterol, Systolic BP, Functional Assessment, BMI. Insignificant Variables: Smoking, Ethnicity, Education Level, Diastolic BP, and others.

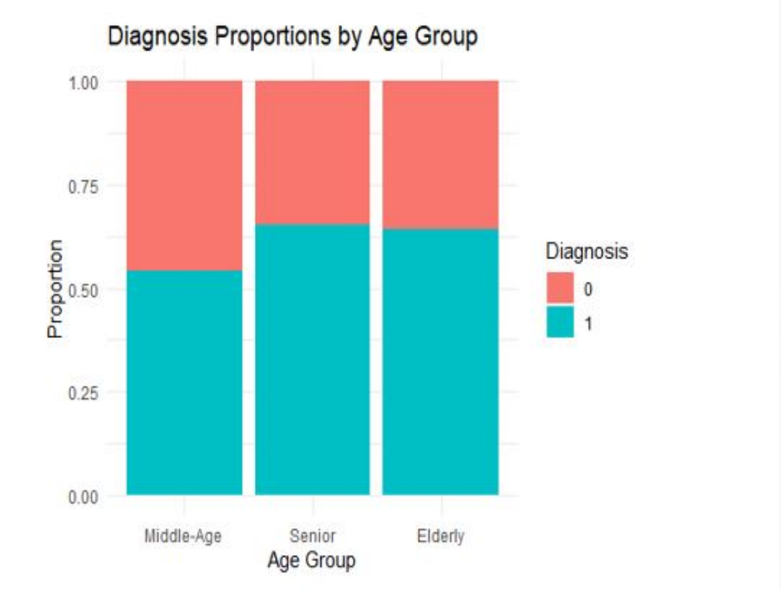
Interaction analysis:

Significant interactions were found in the Clinical Measurements, Cognitive and Functional Assessments, and Symptom clusters, revealing relationships between blood pressure, cholesterol, motor, cognitive, and functional evaluations, and various symptoms in Parkinson's Disease. However, no significant interactions were observed in the Demographic Details, Lifestyle Factors, or Medical History clusters.

```
##
## Demographic_Details: No significant interactions.
##
## Lifestyle_Factors: No significant interactions.
##
## Medical_History: No significant interactions.
##
## Clinical_Measurements:
##  - SystolicBP:DiastolicBP (p-value: 0.00634)
##  - SystolicBP:CholesterolTotal (p-value: 0.02850)
##
## Cognitive_and_Functional_Assessments:
##  - UPDRS:FunctionalAssessment (p-value: 0.03567)
##  - UPDRS:MoCA (p-value: 0.04754)
##
## Symptoms:
##  - Tremor:PosturalInstability (p-value: 0.00015)
##  - Tremor:Bradykinesia (p-value: 0.00985)
##  - PosturalInstability:SpeechProblems (p-value: 0.02696)
##  - Rigidity:Bradykinesia (p-value: 0.03888)
##  - Bradykinesia:SleepDisorders (p-value: 0.04146)
```

Subgrouping Faceted plots





Statistical Methods

Univariate Analysis

Mann-Whitney U Test and Chi-Square Test was performed

Rationale:

Continuous Variables (Mann-Whitney U Test):

Non-parametric test chosen because Shapiro-Wilk test indicated data was not normally distributed.

Categorical Variables (Chi-Square Test):

Assesses independence between variables; all expected frequencies met the criteria (≥ 5).

Appropriateness:

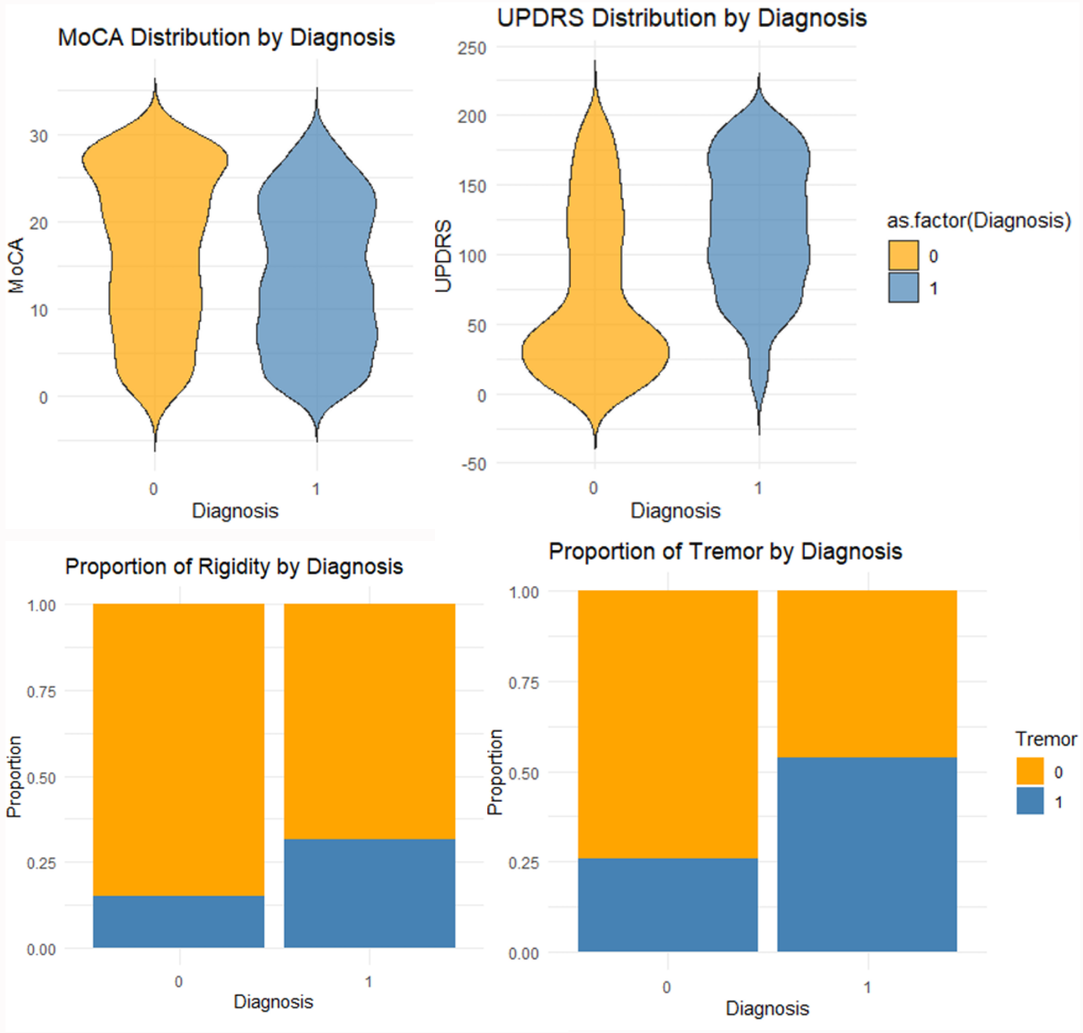
Mann-Whitney U Test is ideal for comparing medians of non-normally distributed data.

Chi-Square Test effectively identifies relationships in categorical data.

Limitations:

Mann-Whitney U Test: Does not provide information about effect size or direction beyond median differences.

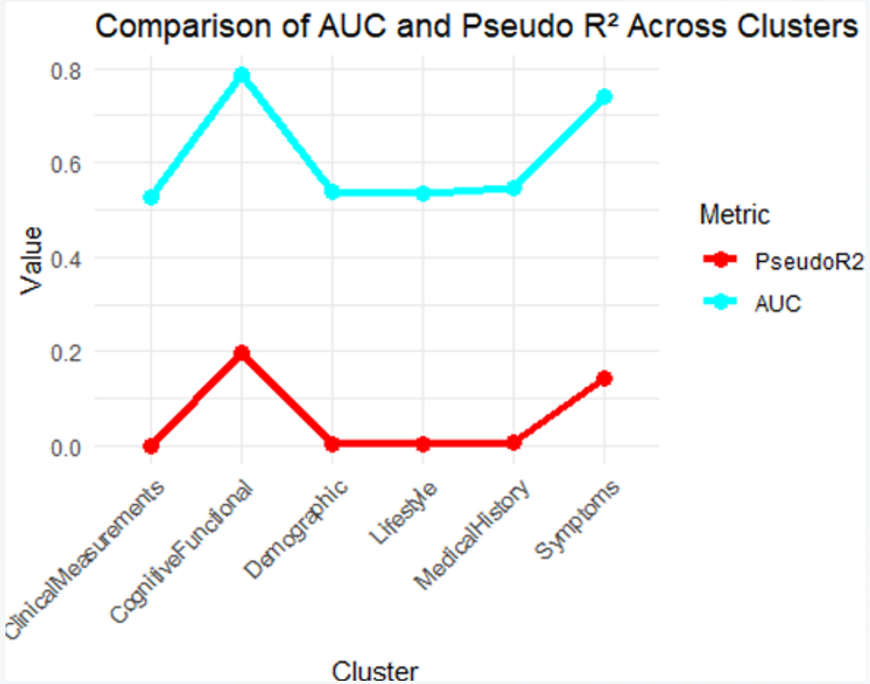
Chi-Square Test: Cannot measure the strength of association and assumes large sample sizes.



Multivariate Analysis

Logistic Regression was performed to model predictors and Parkinson's diagnosis

Rationale:
Logistic regression models relationships between independent predictors (e.g., UPDRS, MoCA, Tremor, Rigidity) and a binary outcome (Parkinson's diagnosis).
Evaluates combined effects while controlling for confounding variables.
Appropriateness:
Well-suited for binary outcomes (e.g., diagnosis).
Pseudo R^2 and AUC metrics provide insight into model performance:
Combined model: Pseudo $R^2 = 0.261$, AUC = 0.826
Limitations:
Assumes linearity between predictors and log-odds.
Results depend on sample size; small sample sizes can reduce reliability.



CognitiveFunctional (AUC ~0.79, Pseudo R^2 ~0.20) and Symptoms (AUC ~0.74, Pseudo R^2 ~0.14) are the strongest predictors, while other clusters show weaker predictive power.

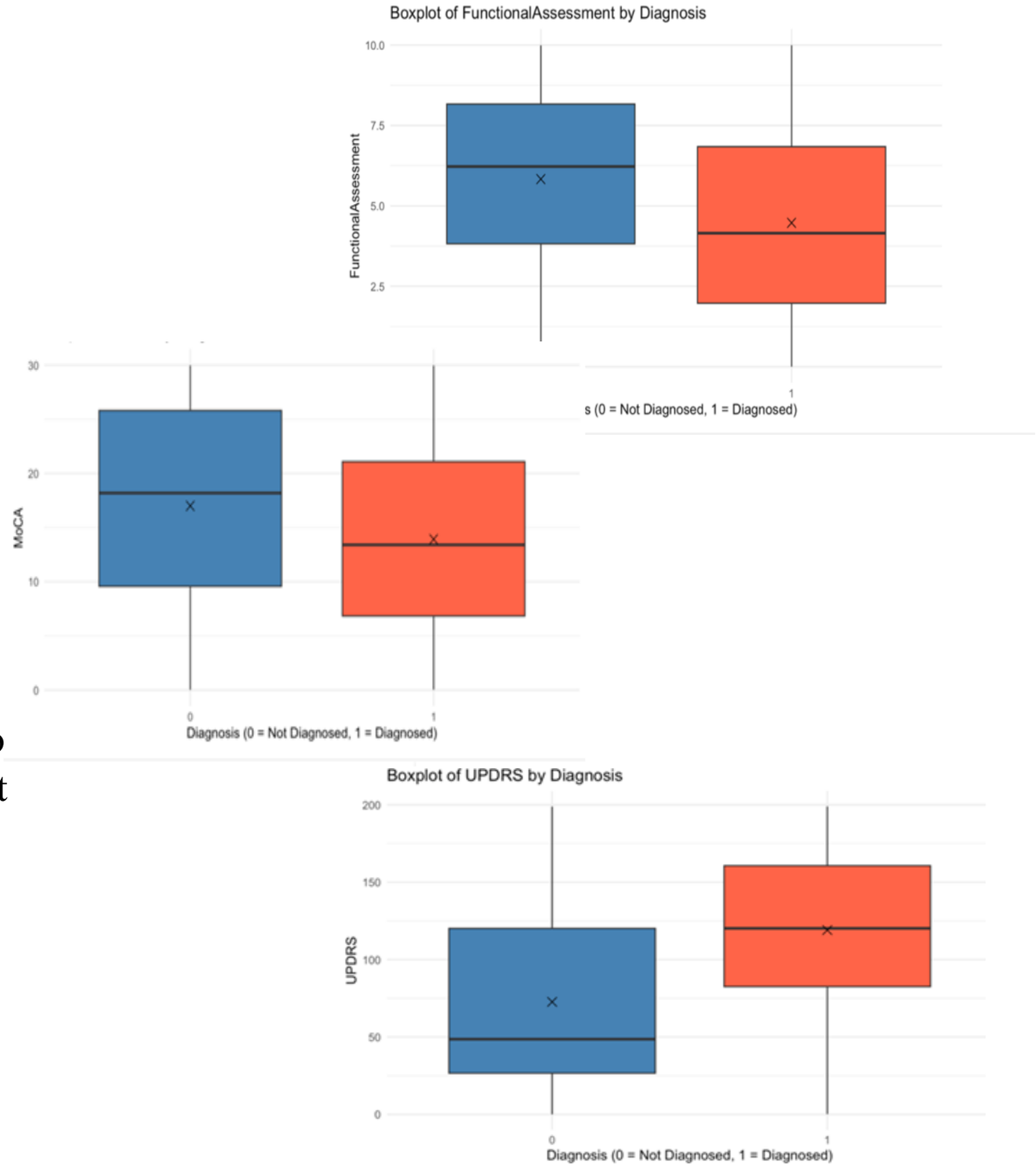
Welch Two-Sample T-Test Results

UPDRS:
Diagnosed group has significantly higher scores ($p<0.001$).
Indicates greater disease severity.

MoCA:
Non-diagnosed group has significantly higher scores ($p<0.001$).
Reflects better cognitive function in non-diagnosed individuals.

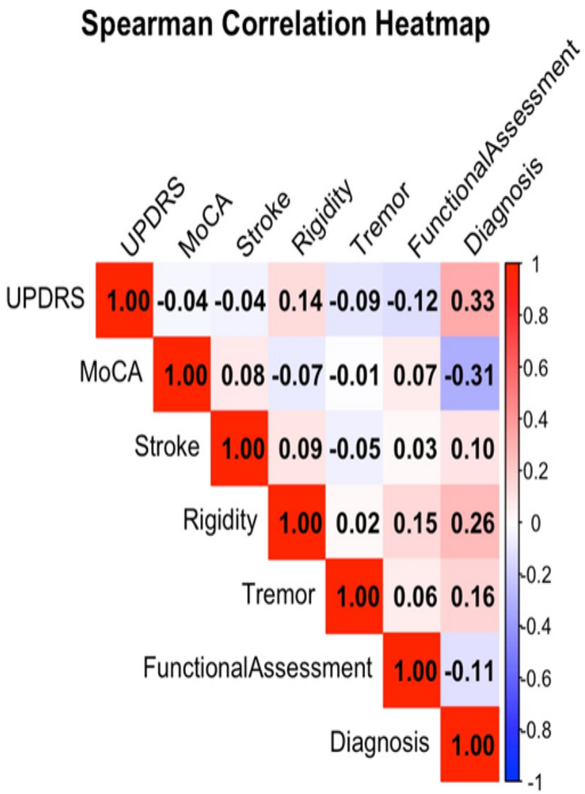
FunctionalAssessment:
Non-diagnosed group has significantly higher scores ($p<0.001$).
Lower functional capacity is associated with diagnosis

Rationale
We used Welch's t-test to compare the variances between our two independent groups (diagnosed vs. non-diagnosed), as it is robust to unequal variances and sample sizes. The results are visually represented using box plots.



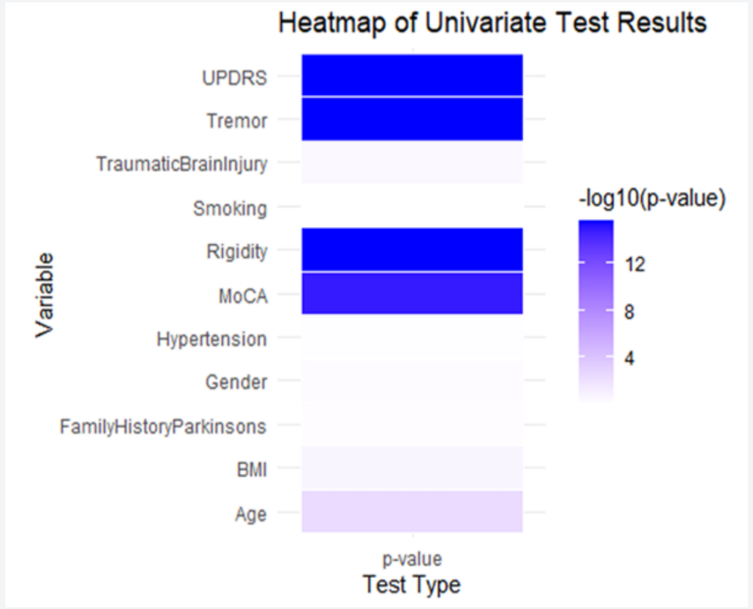
Spearman Correlation matrix Results

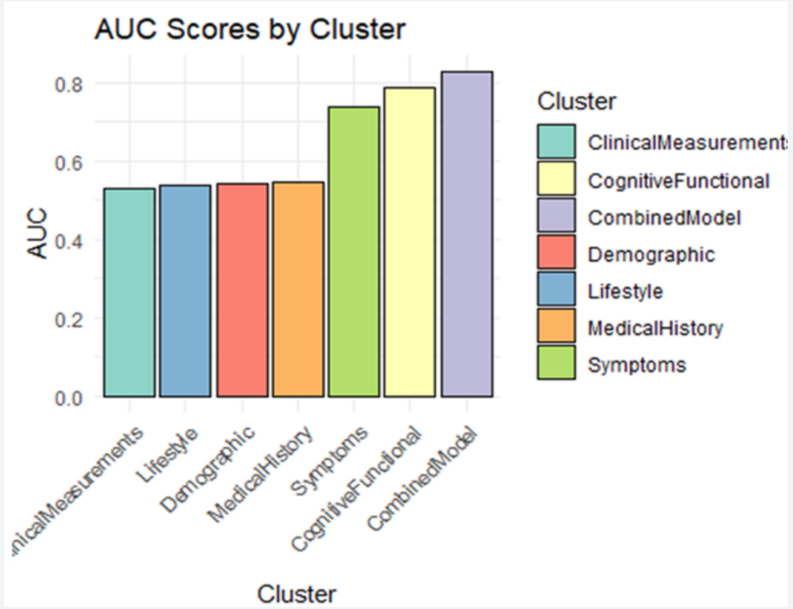
UPDRS and Diagnosis: Positive correlation (0.33) suggests that higher UPDRS scores are associated with being diagnosed.
MoCA and Diagnosis: Negative correlation (-0.31) indicates that lower MoCA scores are linked with being diagnosed.
Rigidity and Diagnosis: Positive correlation (0.26) shows that higher rigidity scores are more common in diagnosed individuals.
Tremor and Diagnosis: Positive correlation (0.16) reflects a weaker but present relationship between tremor severity and diagnosis.
Functional Assessment has a slight negative correlation (-0.11) with Diagnosis, suggesting lower functional scores are associated with diagnosis.



Rationale
We used Spearman correlation to analyze the relationships between variables in our dataset because it is non-parametric measure that is ideal for non-normally distributed data.

FINDINGS





The results reveal that different clusters contribute variably to the prediction of Parkinson's Disease. Significant variables are identified in the heatmap. These findings validate the alternative hypothesis by showing that specific clusters and variables are indeed associated with Parkinson's Disease, providing evidence to reject the null hypothesis and support the alternative.

An unexpected finding is that Family History of Parkinson's does not appear to be significant in any of the models, despite existing literature suggesting it has an impact on the risk of developing Parkinson’s disease(Sellbach et al., 2006). This discrepancy could be due to limitations in the dataset, such as incomplete or inaccurate reporting of family history, or differences in the population studied compared to those in prior research

CONCLUSION

Our analysis therefore supports the rejection of H_0 in favor of H_1: There is at least one cluster significantly associated with the diagnosis of PD.
Of the clusters, the Cognitive and Functional Assessments cluster was the strongest in predicting the target variable. Among the top variables that contribute to the prediction are MoCA, UPDRS, Tremors, and Rigidity. While UPDRS, Tremors, and Rigidity positively correlate with the diagnosis, MoCA gives a negative correlation

The strengths of this study are in using a complete set of clusters and powerful statistical methods to identify key variables that affect the diagnosis of Parkinson's Disease. However, potential biases from missing or incomplete data include the underrepresentation of the Family History variable

Future research could investigate the temporal relationships between clusters and their variables, focusing on how they contribute to the progression and early detection of Parkinson’s Disease

REFERENCES

National Institute of Neurological Disorders, & Stroke (US). (2004). *Parkinson's disease: Challenges, progress, and promise*. National Institute of Neurological Disorders and Stroke, National Institutes of Health.

Parkinson's Disease Dataset Analysis. (2024, June 11). Kaggle. https://www.kaggle.com/datasets/rabieelkharoua/parkinsons-disease-dataset-analysis

Sellbach, A. N., Boyle, R. S., Silburn, P. A., & Mellick, G. D. (2006). Parkinson's disease and family history. *Parkinsonism & related disorders*, *12*(7), 399–409. https://doi.org/10.1016/j.parkreldis.2006.03.002

Appendix

We have overall 2,105 Rows and 35 Columns.

*Data Preprocessing:

Shape

```
dim(data)
```

```
## [1] 2105   35
```

*Checking for Null or Missing values:

```
total_null_values <- sum(is.na(data))
print(total_null_values)
```

```
## [1] 0
```

No Null values found.

```
# Calculate correlation matrix
correlation_matrix <- cor(cluster_data, use = "complete.obs")

# Melt the correlation matrix for plotting
melted_correlation <- melt(correlation_matrix)

# Create the heatmap
ggplot(melted_correlation, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  geom_text(aes(label = sprintf("%.2f", value)), size = 3, color = "black
+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation") +
  labs(
    title = title,
    x = "Variables",
    y = "Variables"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
    axis.text.y = element_text(size = 10),
    plot.title = element_text(hjust = 0.5, size = 14)
  )
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.4.2
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
# Select only numeric columns
numeric_columns <- names(data)[sapply(data, is.numeric)]

# Create histograms for all numeric columns
histograms <- lapply(numeric_columns, function(col) {
  ggplot(data, aes_string(x = col)) +
    geom_histogram(binwidth = 1, fill = "skyblue", color = "black", alpha =
0.7) +
    geom_density(aes(y = after_stat(count)), color = "blue", size = 1) + #
Overlay density plot
    labs(title = paste(col, "Distribution"), x = col, y = "Frequency") +
    theme_minimal() +
    theme(
      plot.title = element_text(hjust = 0.5, size = 16),
      axis.title = element_text(size = 14),
      axis.text = element_text(size = 12),
      plot.margin = margin(15, 15, 15, 15) # Add spacing around each plot
```

```
# Install and load the scatterplot3d package
if (!requireNamespace("scatterplot3d", quietly = TRUE)) {
  install.packages("scatterplot3d")
}
library(scatterplot3d)

# Create the 3D scatter plot
scatterplot3d(
  x = data$UPDRS,
  y = data$MoCA,
  z = data$Diagnosis,
  pch = 16,                # Solid points
  color = ifelse(data$Diagnosis == 1, "red", "blue"),  # Color by Dia
  xlab = "UPDRS",
  ylab = "MoCA",
  zlab = "Diagnosis",
  main = "3D Scatter Plot - UPDRS vs MoCA with Diagnosis"
)
```