Parkinson's Disease Data Analysis

Le Quynh Nhu Doan

*SP Jain School of Global Management, BDS22

Abstract- After Alzheimer's disease, Parkinson's disease (PD) is the second most common neurological health condition among the elderly population. It is estimated that approximately 10 million people worldwide, including approximately 100,000 people in Turkey, are affected by PD [1]. PD is typically observed in about 1% of individuals over the age of 65 [2]. Unfortunately, there is currently no known cure for PD [3]. Although there are drug therapies available to alleviate some of the symptoms caused by the disease, the diagnosis and treatment of PD often involve invasive methods [4]. This can make the process of identifying and treating patients with PD more complicated. Therefore, my main motivation for working with this dataset is to utilize multivariate analysis techniques to identify individuals who may be suffering from Parkinson's disease.

I. Introduction

In this analysis, I will analyze the Parkinson dataset from the UCI Machine Learning Repository. There are 2 sub-datasets in this dataset. The first dataset contains several measures related to the voice and motor symptoms of Parkinson's disease, including the clinician's total UPDRS score, various measures of vocal tremors, and several measures of variation in amplitude and frequency. The second dataset is about voice measurements of people who have Parkinson and who do not. My objective is to perform a thorough statistical analysis of the first dataset to gain insights into the relationships between the variables and identify important predictors of the clinician's total UPDRS score.

II. Methodology

This data analysis paper consists of five major sections. The number of pages may vary depending upon the topic of research work but generally comprises up to 5 to 7 pages. These are:

1) Abstract – Introduction
2) Data Collection (Source: https://archive.ics.uci.edu/ml/datasets/parkinsons)
3) Detailed EDA (Exploratory Data Analysis)
4) Modelling: Multiple Regression Analysis
5) Conclusions
6) References

III. Data Collection and Data Preprocessing

1. *Data Collection:*

The dataset was created by Athanasios Tsanas and Max Little of the University of Oxford, in collaboration with 10 medical centers in the US and Intel Corporation who developed the telemonitoring device to record the speech signals. The original study used a range of linear and nonlinear regression methods to predict the clinician's Parkinson's disease symptom score on the UPDRS scale.
Source: https://archive.ics.uci.edu/ml/datasets/parkinsons

2. *Variable and measures used in data analysis:*

a. *First dataset:*

Columns in the dataset contain subject number, subject age, subject gender, time interval from baseline recruitment date, motor UPDRS, total UPDRS, and 16 biomedical voice measures. Each row corresponds to one of 5,875 voice recording from these individuals.

Subject number: Integer that uniquely identifies each subject
Subject Age
Subject gender '0' - male, '1' - female
Test time - Time since recruitment into the trial. The integer part is the number of days since recruitment.

UPDRS: This is a clinician's scale for recording symptoms related to Parkinson's disease. The UPDRS metric consists of 44 sections, where each section addresses different symptoms in different parts of the body. Summing up these 44 sections gives

rise to the total-UPDRS score, which spans the range 0-176, with 0 representing perfectly healthy individual and 176 total disability.

Motor UPDRS - Clinician's motor UPDRS score, linearly interpolated - this forms sections 18-44 from the UPDRS sections
Total_UPDRS - Clinician's total UPDRS score, linearly interpolated - this includes all 44 sections
Jitter Percentage - measure of variation in fundamental frequency
Jitter (Absolute) - measure of variation in fundamental frequency
Jitter (RAP) - measure of variation in fundamental frequency
Jitter (PPQ5) - measure of variation in fundamental frequency
Jitter (DDP) - measure of variation in fundamental frequency

Shimmer - measures of variation in amplitude
Shimmer(dB)- measure of variation in amplitude
Shimmer:APQ3- measure of variation in amplitude
Shimmer:APQ5- measure of variation in amplitude
Shimmer:APQ11- measure of variation in amplitude
Shimmer:DDA- measure of variation in amplitude

NHR: measures of ratio of noise to tonal components in the voice
HNR: measures of ratio of noise to tonal components in the voice

RPDE - A nonlinear dynamical complexity measure
DFA - Signal fractal scaling exponent
PPE - A nonlinear measure of fundamental frequency variation

b. Second dataset:

There are a total of 195 rows of data or patients with 23 feature columns. The voice measurements are from 31 people, 23 with Parkinson's disease and 8 without. In this data set, 147 of the rows have Parkinson's whereas 48 do not. The features themselves are related to the frequency if a patient's vocalizations. The attributes are listed below:

Name - ASCII subject name and recording number
MDVP:Fo(Hz) - Average vocal fundamental frequency
MDVP:Fhi(Hz) - Maximum vocal fundamental frequency
MDVP:Flo(Hz) - Minimum vocal fundamental frequency
MDVP:Jitter(%) - Several measures of variation in fundamental frequency
MDVP:Jitter(Abs) - Several measures of variation in fundamental frequency
MDVP:RAP - Several measures of variation in fundamental frequency
MDVP:PPQ - Several measures of variation in fundamental frequency
Jitter:DDP - Several measures of variation in fundamental frequency
MDVP:Shimmer - Several measures of variation in amplitude
MDVP:Shimmer(dB) - Several measures of variation in amplitude
Shimmer:APQ3 - Several measures of variation in amplitude
Shimmer:APQ5 - Several measures of variation in amplitude
MDVP:APQ - Several measures of variation in amplitude
Shimmer:DDA - Several measures of variation in amplitude
NHR - Two measures of ratio of noise to tonal components in the voice
HNR - Two measures of ratio of noise to tonal components in the voice
status - Health status of the subject (one) - Parkinson's, (zero) - healthy
RPDE - Two nonlinear dynamical complexity measures
D2 - Two nonlinear dynamical complexity measures
DFA - Signal fractal scaling exponent
spread1 - Three nonlinear measures of fundamental frequency variation
spread2 - Three nonlinear measures of fundamental frequency variation
PPE - Three nonlinear measures of fundamental frequency variation

3. *Data Preprocessing*: Use Python to check for missing values.

Dataset 1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5875 entries, 0 to 5874
Data columns (total 22 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   subject#        5875 non-null   int64
 1   age             5875 non-null   int64
 2   sex             5875 non-null   int64
 3   test_time       5875 non-null   float64
 4   motor_UPDRS     5875 non-null   float64
 5   total_UPDRS     5875 non-null   float64
 6   Jitter(%)       5875 non-null   float64
 7   Jitter(Abs)     5875 non-null   float64
 8   Jitter:RAP      5875 non-null   float64
 9   Jitter:PPQ5     5875 non-null   float64
 10  Jitter:DDP      5875 non-null   float64
 11  Shimmer         5875 non-null   float64
 12  Shimmer(dB)     5875 non-null   float64
 13  Shimmer:APQ3    5875 non-null   float64
 14  Shimmer:APQ5    5875 non-null   float64
 15  Shimmer:APQ11   5875 non-null   float64
 16  Shimmer:DDA     5875 non-null   float64
 17  NHR             5875 non-null   float64
 18  HNR             5875 non-null   float64
 19  RPDE            5875 non-null   float64
 20  DFA             5875 non-null   float64
 21  PPE             5875 non-null   float64
dtypes: float64(19), int64(3)
memory usage: 1009.9 KB
```

Dataset 2

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 195 entries, 0 to 194
Data columns (total 24 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   name              195 non-null    object
 1   MDVP:Fo(Hz)       195 non-null    float64
 2   MDVP:Fhi(Hz)      195 non-null    float64
 3   MDVP:Flo(Hz)      195 non-null    float64
 4   MDVP:Jitter(%)    195 non-null    float64
 5   MDVP:Jitter(Abs)  195 non-null    float64
 6   MDVP:RAP          195 non-null    float64
 7   MDVP:PPQ          195 non-null    float64
 8   Jitter:DDP        195 non-null    float64
 9   MDVP:Shimmer      195 non-null    float64
 10  MDVP:Shimmer(dB)  195 non-null    float64
 11  Shimmer:APQ3      195 non-null    float64
 12  Shimmer:APQ5      195 non-null    float64
 13  MDVP:APQ          195 non-null    float64
 14  Shimmer:DDA       195 non-null    float64
 15  NHR               195 non-null    float64
 16  HNR               195 non-null    float64
 17  status            195 non-null    int64
 18  RPDE              195 non-null    float64
 19  DFA               195 non-null    float64
 20  spread1           195 non-null    float64
 21  spread2           195 non-null    float64
 22  D2                195 non-null    float64
 23  PPE               195 non-null    float64
dtypes: float64(22), int64(1), object(1)
memory usage: 36.7+ KB
```

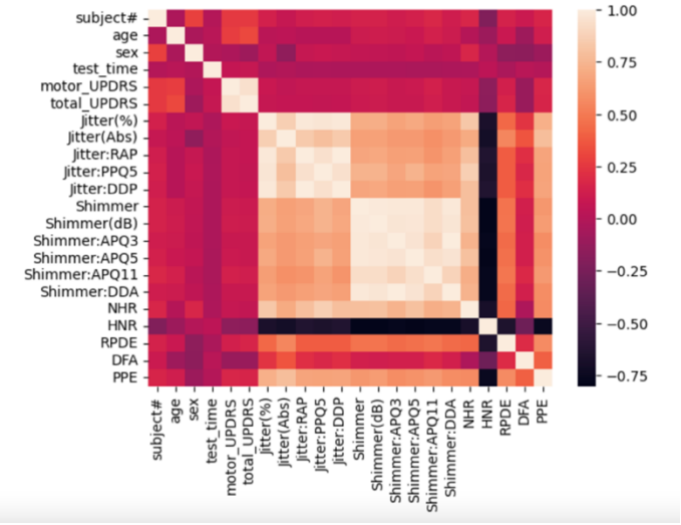
IV. Exploratory Data Analysis (EDA)

a. *Tools*: Using graphs and Tables conducted with Excel (Pivot table), JMP and Python.

b. *Notes*: There are some overlapping attributes between two sub-datasets. As a result, the data analysis will be performed on one sub-dataset at a time to ensure accurate and reliable results.

c. *Results*:

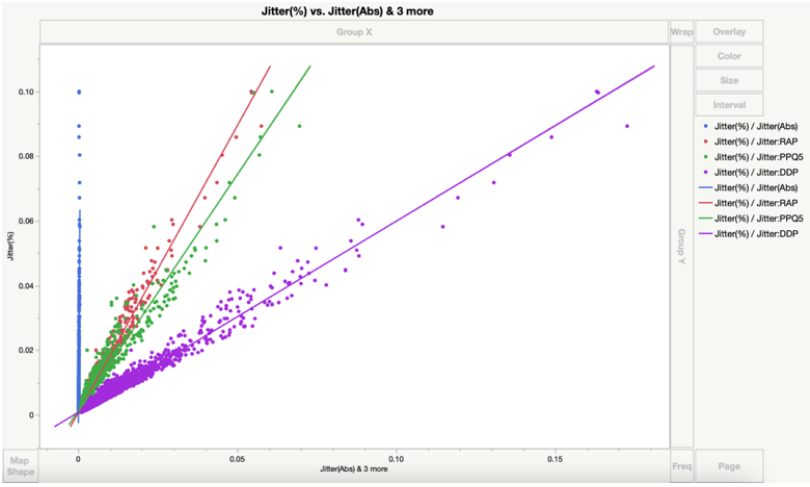
Insight 1 (Correlation) – Dataset 1)



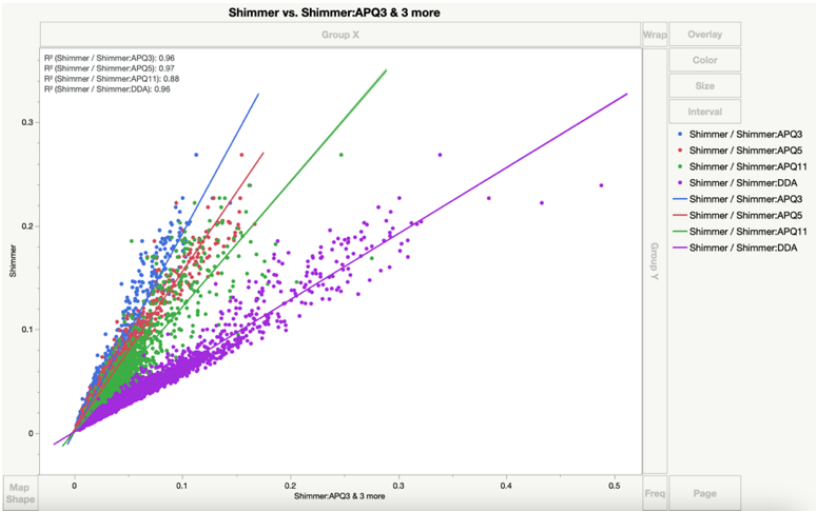
Correlation using heatmap.

OVERALL

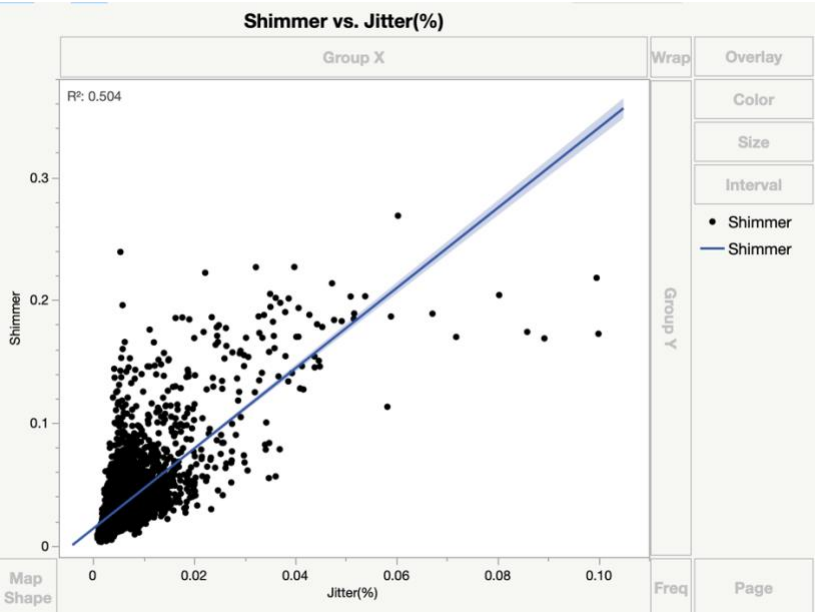
From this heatmap, Jitter and Shimmer variables are highly correlated with each other.
HNR are strongly uncorrelated with Jitter and Shimmer variables.
NHR, PPE are correlated with Jitter and Shimmer.
Motor_UPDRS and total_UPDRS are highly correlated.



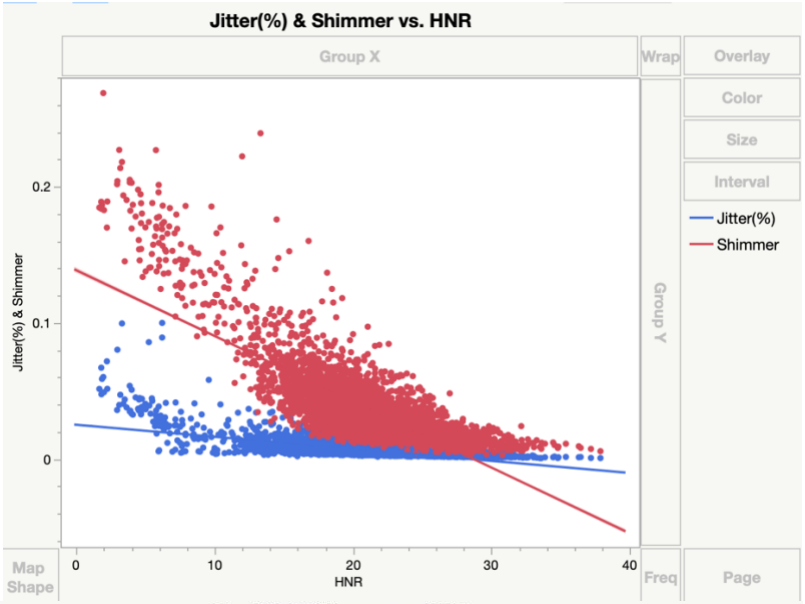
Jitter variables are highly correlated with each other



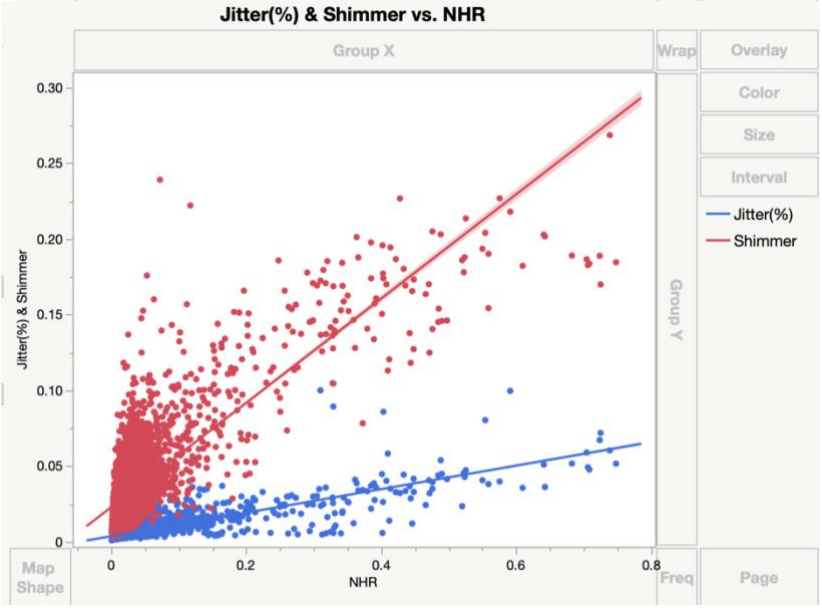
Shimmer variables are highly correlated with each other.



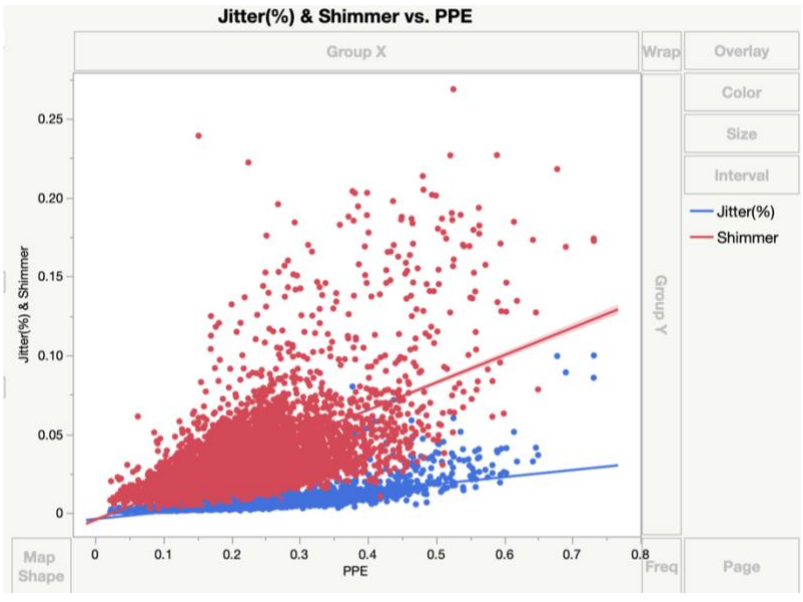
Jitter and Shimmer are moderately correlated with each other.



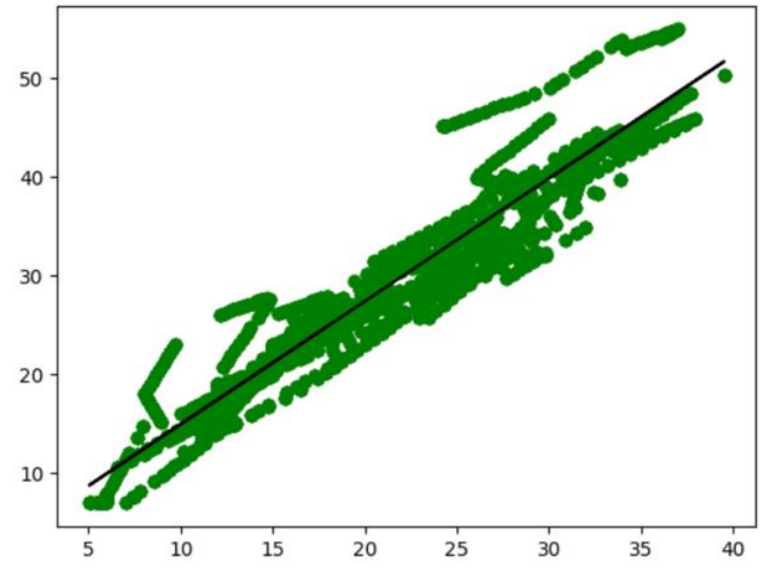
(Jitter, Shimmer) and HNR are uncorrelated with each other.



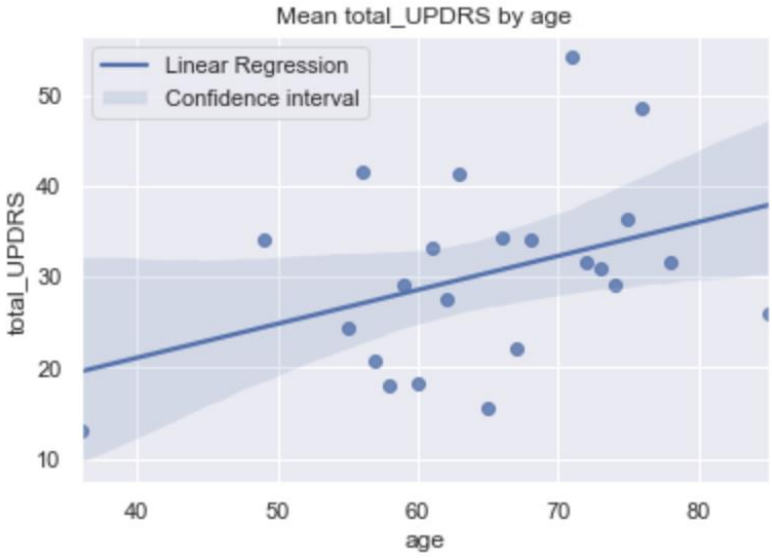
(Jitter, Shimmer) are correlated with NHR



(Jitter, Shimmer) are correlated with PPE.



Total_UPDRS and motor_UPDRS are highly correlated with each other. Therefore, motor_UPDRS can be removed to be avoid redundancy.



There is a slightly correlation between age and total_UPDRS.

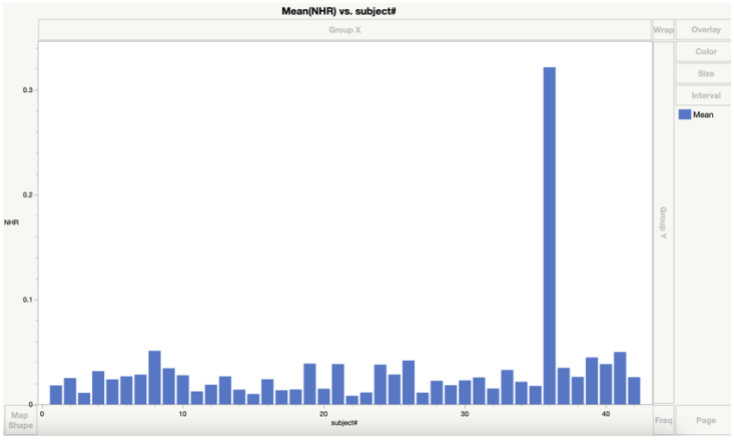
Insight 1:

i. There is a strong correlation between amplitude values (Jitter) and frequency values (Shimmer). This suggests that changes in amplitude values are related to changes in frequency values in the voice data.
ii. (Amplitude and frequency values) are correlated with the ratio of noise to tonal components in the voice (NHR) and measure of fundamental frequency variation (PPE), which means changes in amplitude and frequency values are associated with changes in the ratio of noise to tonal components and the measure of fundamental frequency variation in the voice data as well.
iii. HNR (the ratio of noise to tonal components in the voice) are negatively correlated with other features.
iv. Total_UPDRS might be affected with age.

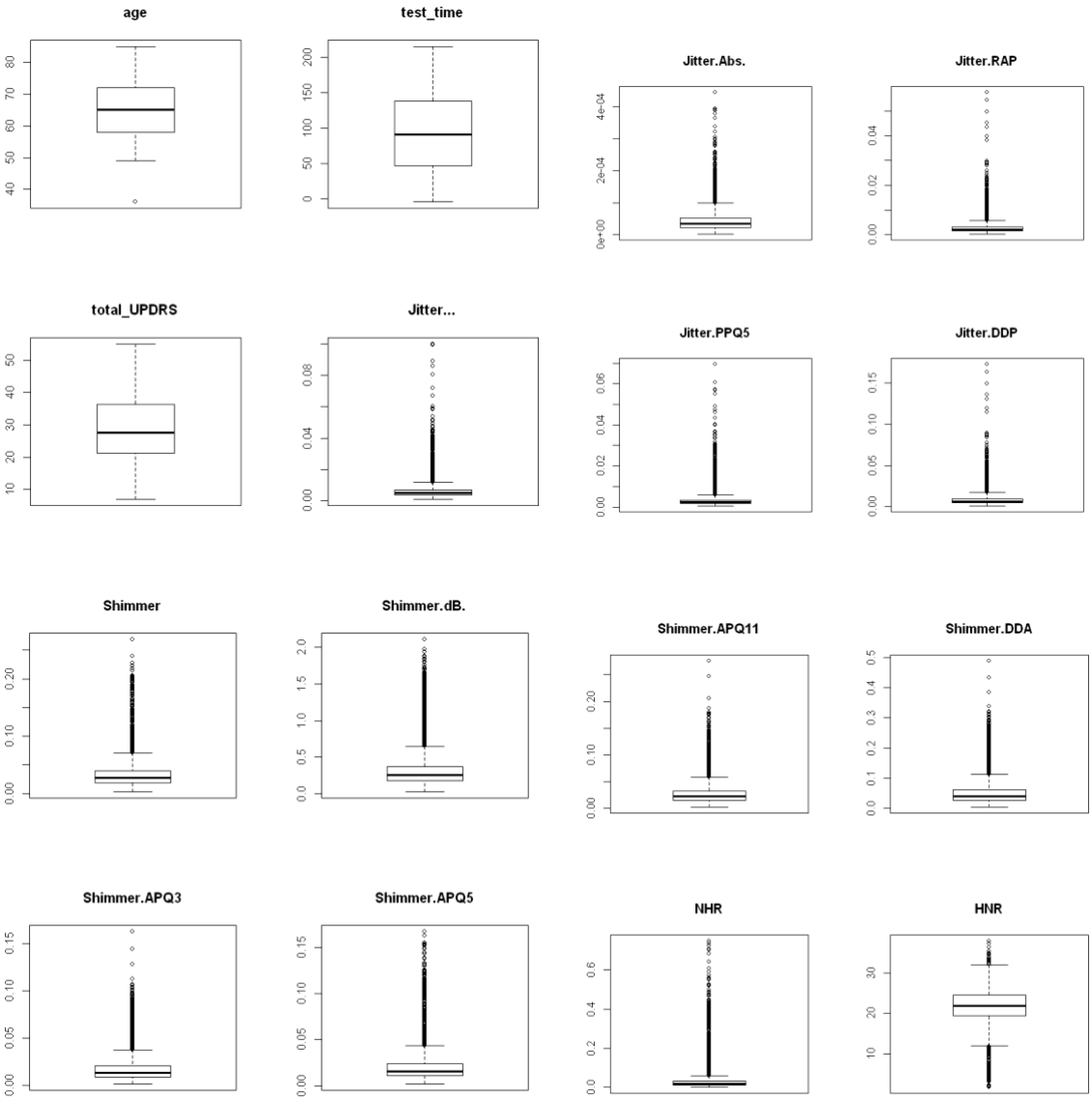
➔ The ratio of noise to tonal components **is** a significant factor **in** identifying the presence of Parkinson's disease.

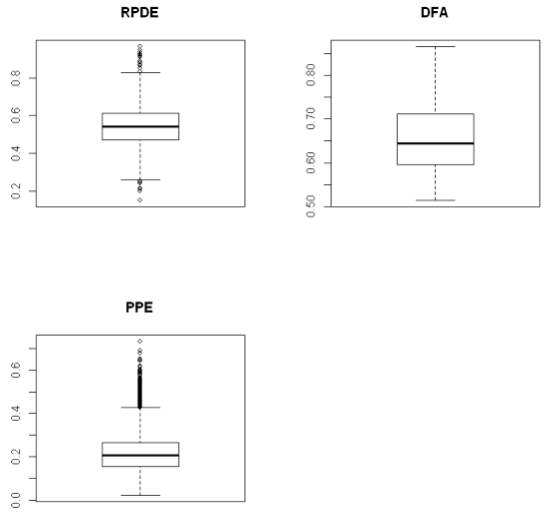
Insight 2 (Outliers) – Both datasets

First dataset:

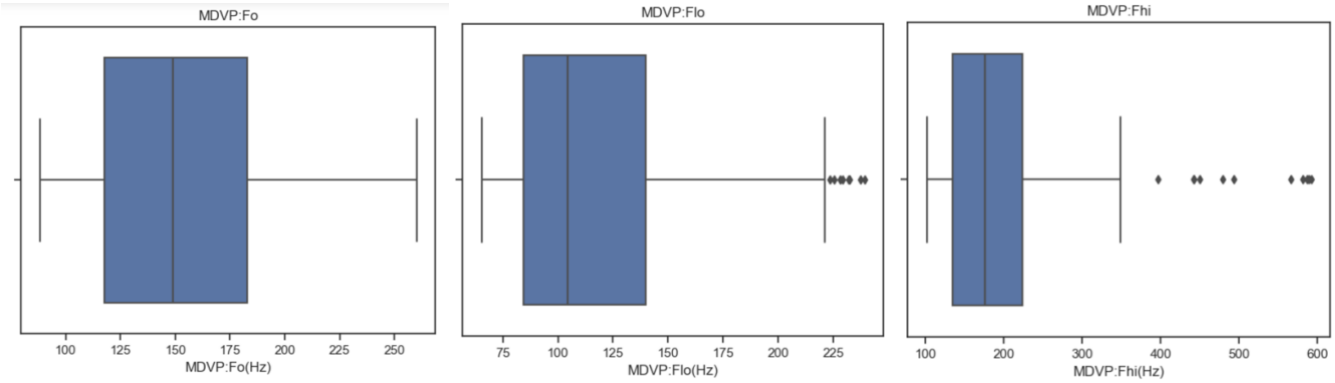


There is an abnormally high value of NHR at the person at age 62 while the average of test time of everyone is almost the same.





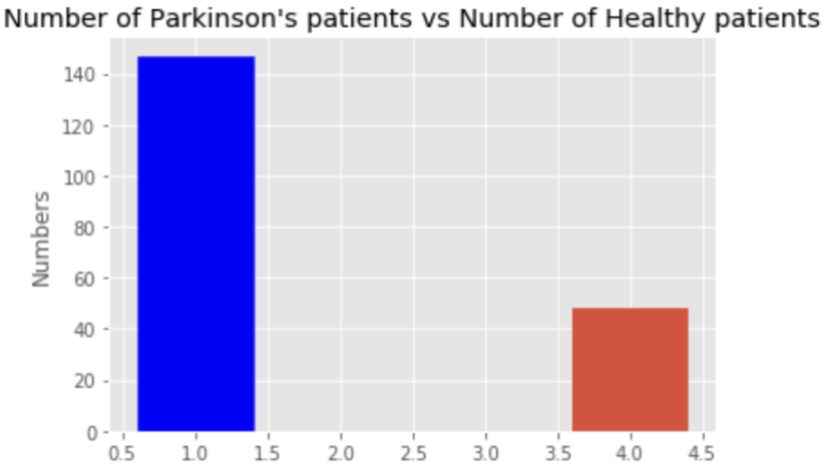
Second dataset:



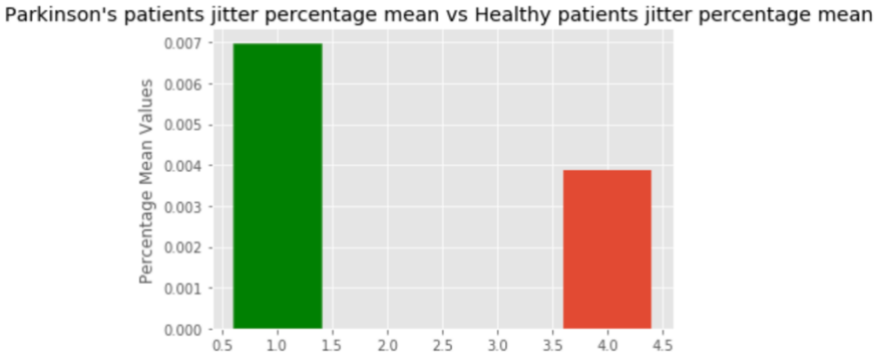
Insight 2:

i. From the visualization above, there are outliers in most of the attributes except for DFA, total_UPDRS, test_time and age.
ii. there are some outliners for the maximum and minimum vocal fundamental frequencies (i.e. MDVP:Fhi(Hz) and MDVP:Flo(Hz))
iii. As a result, it is necessary to perform feature rescaling as part of the pre-processing phase in the model building process.

Insight 3 (Univariate) – Dataset 2



The number of Parkinson's patients is approximately 2.7 times higher than healthy people.

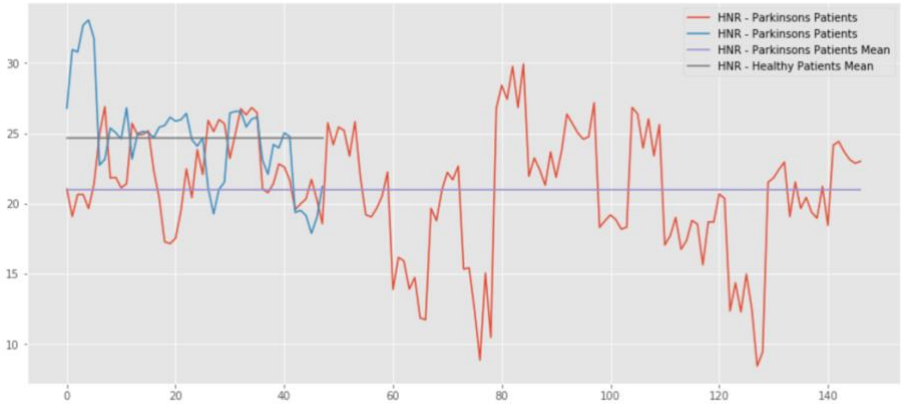


As can be seen, the mean values of jitter of Parkinson's patients are approximately 2 times higher than healthy people.

Insight 3:

Jitter (%) values can indicate people with Parkinson's disease or assisting with early stage diagnosis.

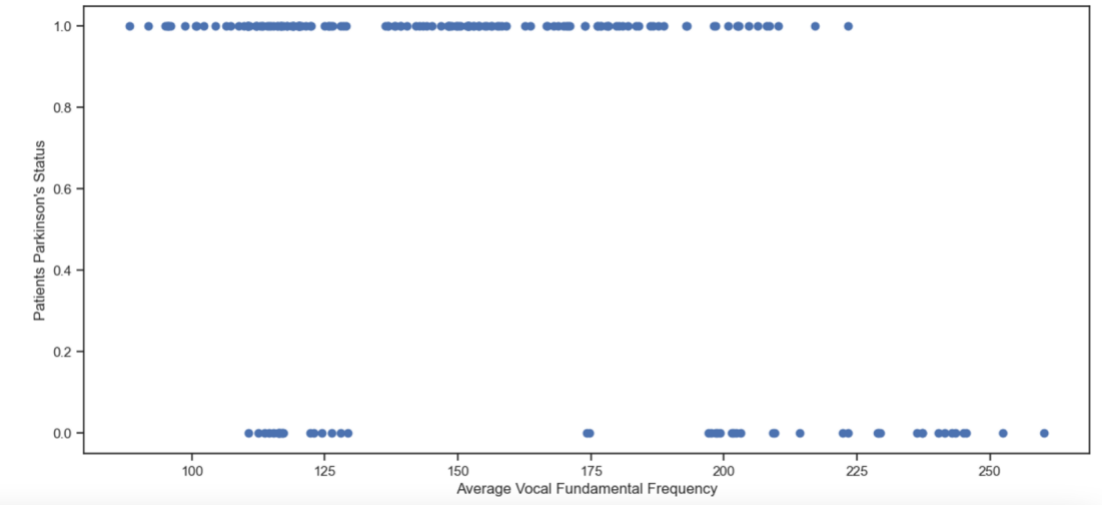
Insight 4 (Line graph) – Dataset 2



Insight 4: Based on the depicted plot, it is evident that the average mean Harmonic-to-Noise Ratio (HNR) value for healthy patients is higher compared to patients with Parkinson's disease.
Lower HNR values are indicative symptoms commonly associated with Parkinson's disease.

Insight 5 (Scatterplot) – Dataset2

MDVP: Fo(Hz) affects patient



Insight 5:

MDVP:Fo(Hz) (Average vocal fundamental frequency) and Status is negatively correlated, which means higher values of MDVP:Fo(Hz) indicate lower values of Status, which corresponds to healthier patients. It is noteworthy that patients with an Average Vocal Fundamental Frequency (Hz) higher than 190 tend to be healthy, while those below 190 are predominantly classified as patients with Parkinson's disease.

V. MULTIPLE REGRESSION ANALYSIS

Notes: I will perform and present a Multiple Regression Analysis only on the first dataset.

Y variables: I chose total_UPDRS as Y variable because the purpose of this paper is focusing on the UPDRS score. Also, because total_UPDRS and motor_UPDRS are positively correlated, I decided to remove motor_UPDRS to avoid redundancy. (using drop() function)

Steps: Perform Multiple Regression Analysis

1. Set up Independent and dependent variables(total_UPDRS)
2. Split the data into train and test data sets

```
X Train shape  (4700, 20)
X Test shape   (1175, 20)
y Train shape  (4700, 1)
y Test shape   (1175, 1)
```

3. Shape of Train and Test Sets
4. Train a Multiple Linear Regression Model

	Attribute	Co-efficient
0	subject#	0.259850
1	age	0.308181
2	sex	-4.912115
3	test_time	0.014533
4	Jitter(%)	-385.496783
5	Jitter(Abs)	-43570.573402
6	Jitter:RAP	-66150.646550
7	Jitter:PPQ5	-123.786355
8	Jitter:DDP	22494.727442
9	Shimmer	37.837356
10	Shimmer(dB)	-0.795241
11	Shimmer:APQ3	7863.461688
12	Shimmer:APQ5	20.773232
13	Shimmer:APQ11	14.017016
14	Shimmer:DDA	-2675.285298
15	NHR	-22.712197
16	HNR	-0.488598
17	RPDE	1.116406
18	DFA	-35.735302
19	PPE	15.408529

5. Review the Regression coefficients

	total_UPDRS
2667	26.230
2560	26.230
2098	20.867
642	44.503
552	15.264
1777	31.936
1081	24.702
1501	21.323
5707	42.401
2353	36.022

	Predicted
0	26.238423
1	30.314382
2	27.167150
3	29.374134
4	30.590497
5	27.541697
6	24.147979
7	26.136897
8	32.442921
9	23.496818

6. Evaluate the model

```
Mean Absolute Error 7.759655268485181
Mean Squared Error 89.27805504319426
Root Mean Squared  Error 9.448706527519747
R Squared 0.2610396973037986
```

Findings as Results of Regression

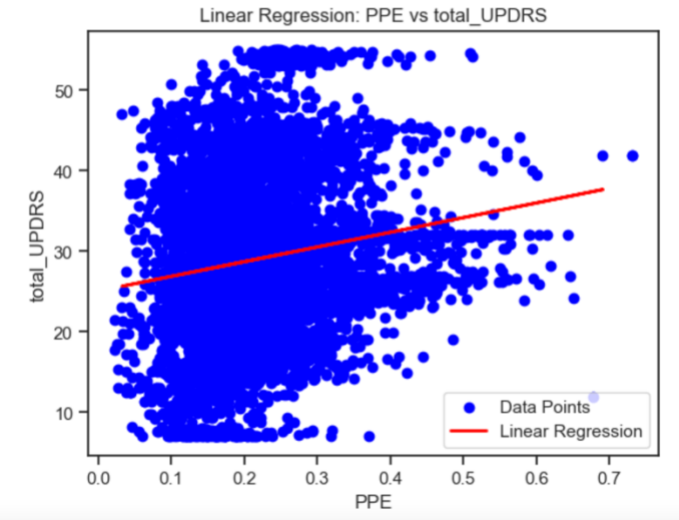
Typically, variables with higher absolute coefficient values are considered more important in predicting the Y variable, as they have a larger impact on the predicted outcome. A higher positive coefficient indicates a positive relationship, while a lower negative coefficient indicates a negative relationship with the dependent variable.

From the result, the attributes with higher absolute coefficients (indicating higher importance) in predicting the y variable are:

1. age (0.308181)
2. RPDE (1.116406)
3. PPE (15.408529)
4. Jitter(%) (-385.496783)
5. Jitter(Abs) (-43570.573402)
6. Jitter:RAP (-66150.646550)
7. Shimmer:APQ3 (7863.461688)
8. Shimmer:APQ5 (20.773232)
9. Shimmer:APQ11 (14.017016)
10. DFA (-35.735302)

As can be seen, age, RPDE, PPE, Jitter(%), Shimmer and DFA are some important variables that can help predicting the total_UPDRS variable.

```
Mean Squared Error: 108.15865925401134
Intercept: 25.059712773645497
Coefficient: 18.252010431845584
```



This is an example of linear regression of PPE vs total_UPDRS

VI. Conclusion

1. Higher values of HNR may be associated with a lower risk of Parkinson's disease, whereas lower values of NHR may also indicate a lower risk of Parkinson's disease.
2. Jitter(%) values may serve as an indicator for Parkinson's disease in certain cases.
3. Lower HNR values are commonly observed symptoms that are associated with Parkinson's disease.
4. Individuals with an Average Vocal Fundamental Frequency (Hz) above 190 are often considered healthy, while those below 190 are more likely to be classified as patients with Parkinson's disease.

VII. REFERENCES

[1]*Statistics*. (n.d.). Parkinson's Foundation. https://www.parkinson.org/understanding-parkinsons/statistics
[2]*Parkinson's Disease: Etiology, Neuropathology, and Pathogenesis - Parkinson's Disease - NCBI Bookshelf*. (n.d.). National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/books/NBK536722/
[3]*About Parkinson's | Garvan Institute of Medical Research*. (n.d.). Garvan Institute of Medical Research. https://www.garvan.org.au/research/diseases/parkinsons-disease/about
[4]*Parkinson's disease - Diagnosis and treatment - Mayo Clinic*. (2023, February 17). Mayo Clinic - Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/diagnosis-treatment/drc-20376062.