# Table of Contents

[Table of Contents 2](#_Toc84200116)

[Descriptive Analysis – Parkinson 3](#_Toc84200117)

[Descriptive Analysis 3](#_Toc84200118)

[Identify data quality and apply Data cleansing 4](#_Toc84200119)

[Correlation of the Dataset 6](#_Toc84200120)

[Standardization – Z Score, Robust Scalar and Logarithmic Data Transformation 9](#_Toc84200121)

[Transformation on Jitter.PPQ5(Histogram) 11](#_Toc84200122)

[Apply Transformation on Shimmer.APQ3 12](#_Toc84200123)

[Apply Transformation on RPDE 13](#_Toc84200124)

[Apply Transformation on Total\_UPDRS 14](#_Toc84200125)

[Transformation on Jitter.PPQ5 vs Total\_UPDRS (Linear Regression) 15](#_Toc84200126)

[Transformation on Shimmer.APQ3 vs Total\_UPDRS (Linear Regression) 16](#_Toc84200127)

[Transformation on RPDE vs Total\_UPDRS (Linear Regression) 17](#_Toc84200128)

[Conclusion 18](#_Toc84200129)

[Reference 18](#_Toc84200130)

[Appendix – R Code 19](#_Toc84200131)

# Descriptive Analysis – Parkinson

Parkinson's disease is a kind of nervous system disorder. It starts from a barely noticeable tremor but gradually becomes stiffness or slowing of movement. Tremors are common, but the disorder also commonly causes stiffness or slowing of movement.

This report analyses the given data set from Parkinson's Disease and provides different views after applying standardization.

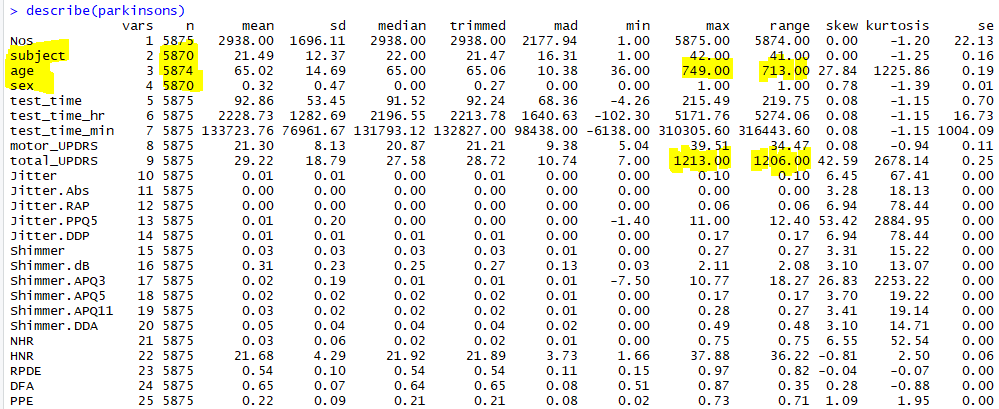
## Descriptive Analysis

The given dataset has 5,875 observations, a collection of biomedical voice data taken over a period from 42 individuals with Parkinson's disease. There are 27 columns in the data frame as below:

* NOS – Numeric variable. the row number of an observation
* Subject ID – Categorical variable. ID is given on test subject
* Age – Categorical variable. Age of the test subject
* Sex – Categorical variable. Gender of the subject, zero is male, and one is female
* Test time, Test time hour and Test time min – Numeric variable. Test time measurement
* Motor\_UPDRS – Numeric variable. Motor UPDRS score of the subject
* Total\_UPDRS – Numeric variable. Total UPDRS score of the subject
* Jitter (%, Abs, RAP, PPQ5, DDP) – Numeric variable. a measure of frequency in a variety of test
* Shimmer (dB, APQ3, APQ5, APQ11, DDA) – Numeric variable. a measure of amplitude in a variety of test
* NHR, HNR – Numeric variable. Measurement of noise to tonal components invoice
* RPDE – Numeric variable. Dynamical complexity measurement
* DFA – Numeric variable. the signal fractal scaling exponent
* PPE – Numeric variable. Fundamental frequency variation measurement.

## Identify data quality and apply Data cleansing

Let's look at the data quality; the overall quality is good, only 13 observations have missing values. In other words, 99.98% observations with value in all reading.

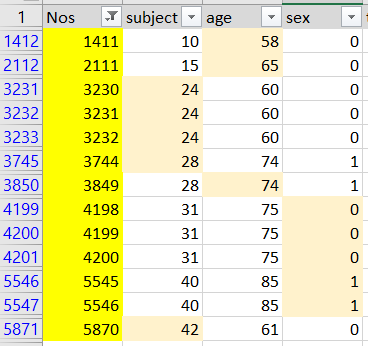
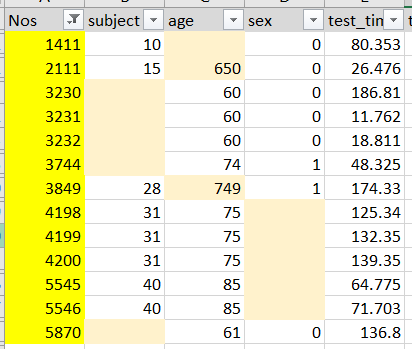


For those missing information, we realize there are:

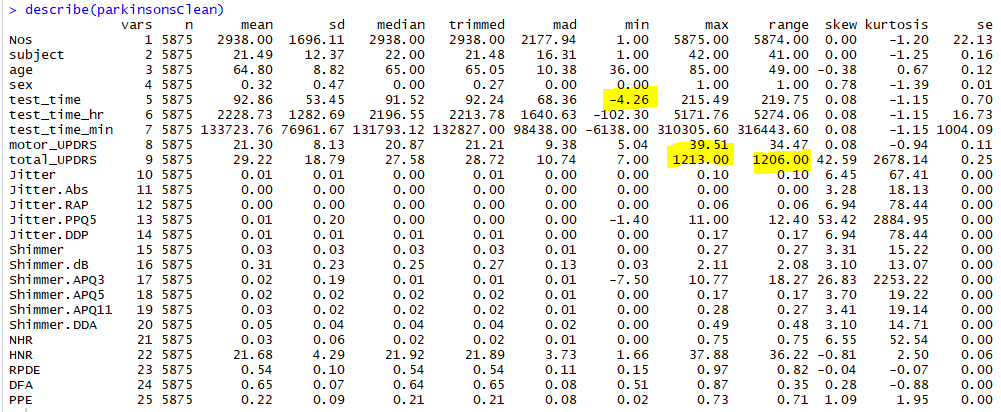
* Five observations have no value in Subject
* One observation has no value in Age, and two invalid values (650, 749) are found and believe it belongs to a typo
* Five observations have no value in Sex

**Resolution** Since this data was collected from the subject over many tests. We can clean up the data with the pattern; for example, in observation #1411, we can look at the range from #1400 to #1420 and identify whether those observations belong to the same subject. If true, we can repair the missing value. By applying this method to all fields, we've got:

Original Data Clean Data

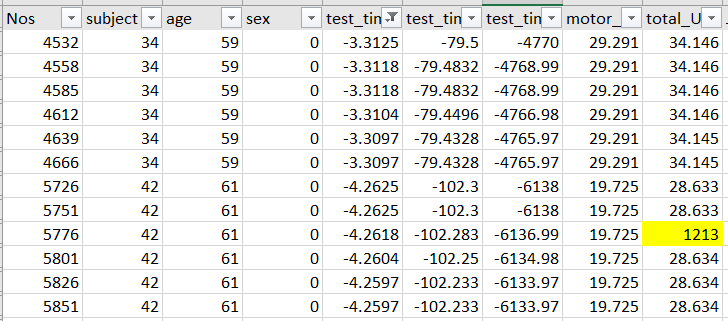


After data cleansing and looking into the dataset's summary again - the total number of variables (Subject, Age and Sex) were aligned to 5,875. However, taking another look at the test\_time does not make any sense with a negative timestamp. Also, there is an outliner (1213.00) on total\_UPDRS, with the mean value at 29.22 and median at 28.72.

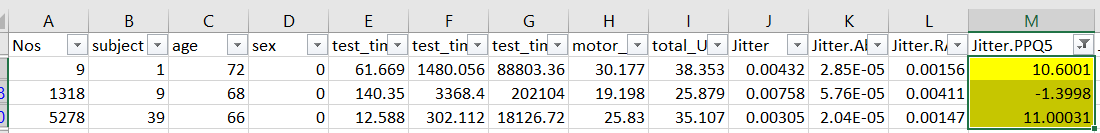
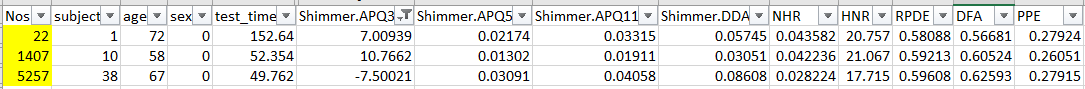


There are 12 observations found with negative test\_time, which includes the outliner of total\_UPDRS. Since 12 observations are only a tiny percent of the population, removing these records will not significantly impact the analysis.

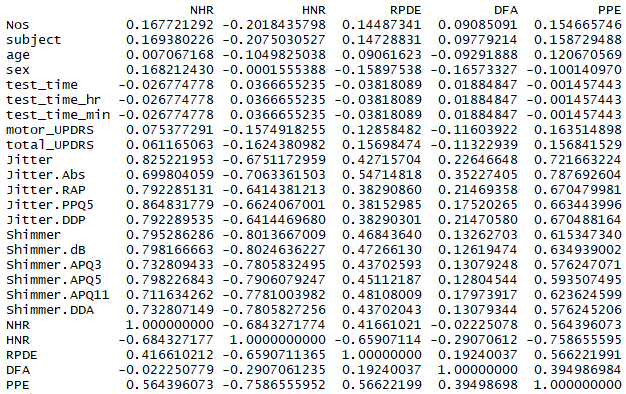
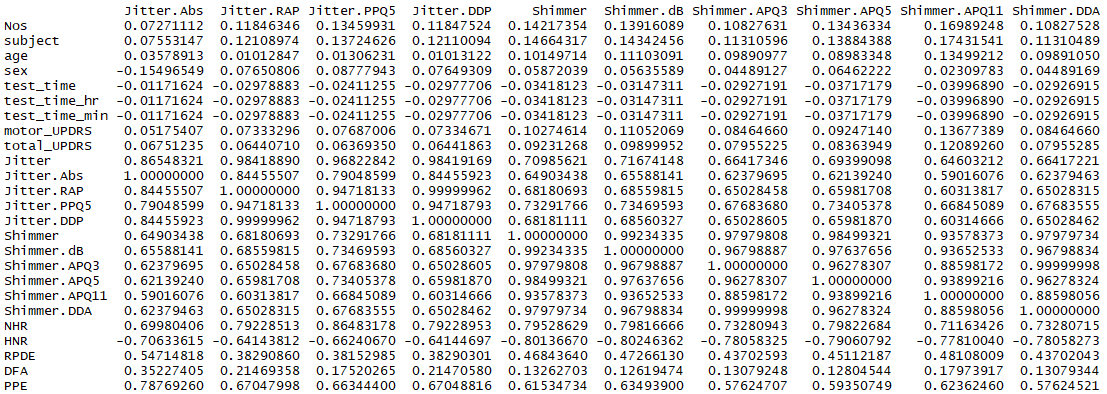
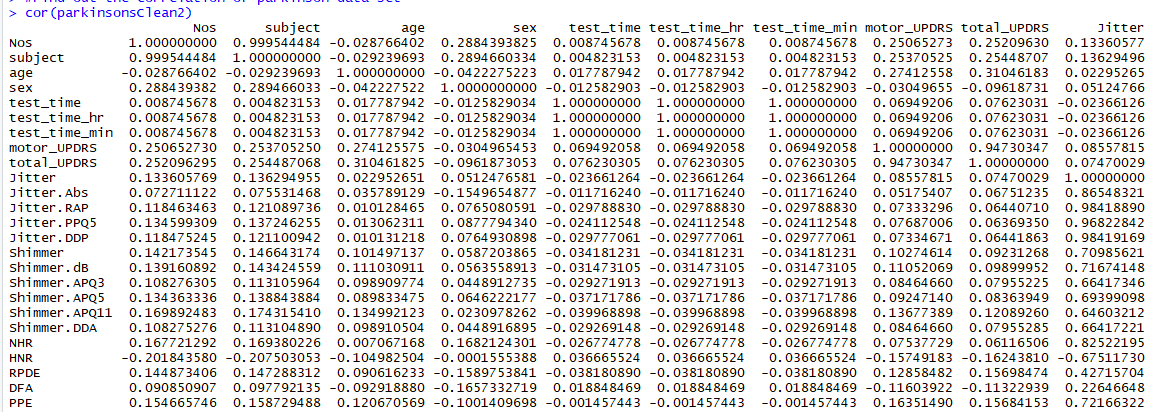
**Resolution**: remove these 12 records in the dataset.



There are three outliners in Shimmer APQ3 and three outliners in Jitter PPQ5, six records removed for the next section experiment. After data cleansing, the total observations are 5,857.



## Correlation of the Dataset



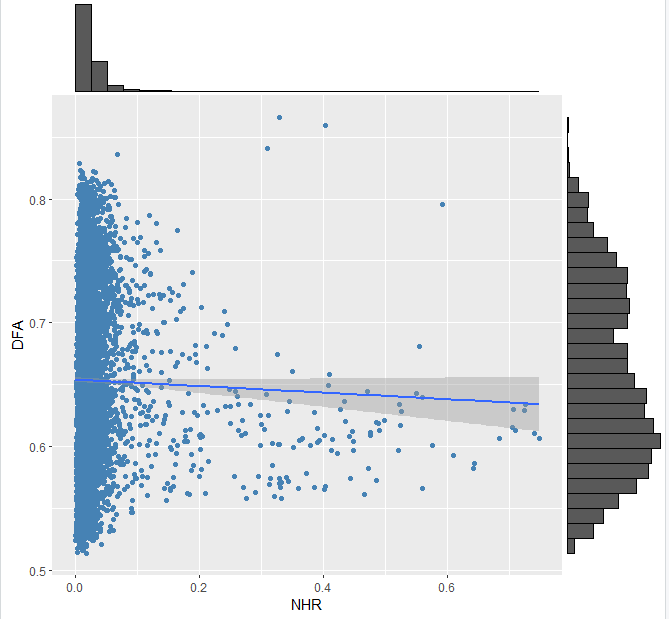
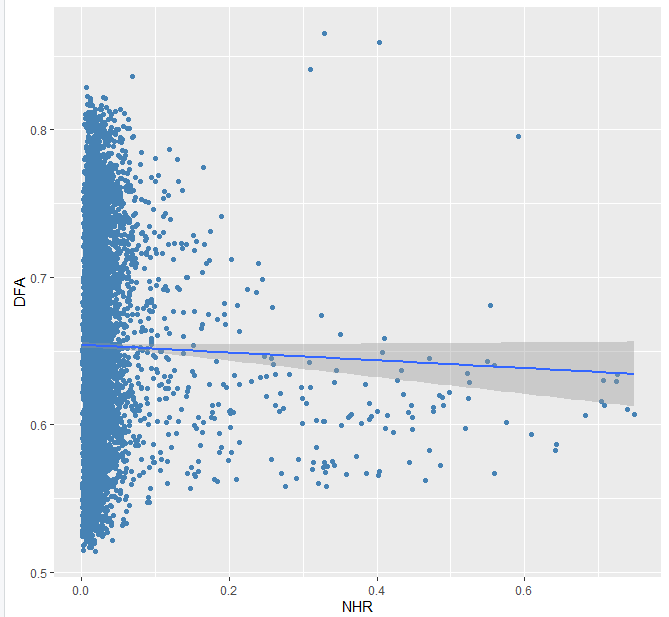
H0: There is no linear relationship between selected variables.

Ha: There is a linear relationship between selected variables.

There is a p-value of the correlation test, which is less than 0.05. Parkinson clean data is highlighted with a blue box. With the 5% significance level, we do not reject the null hypothesis that there is no linear relationship between those variables.

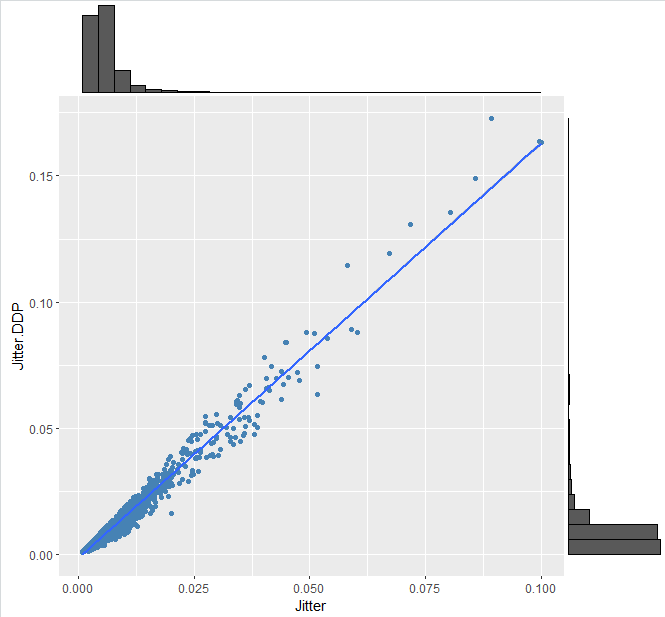
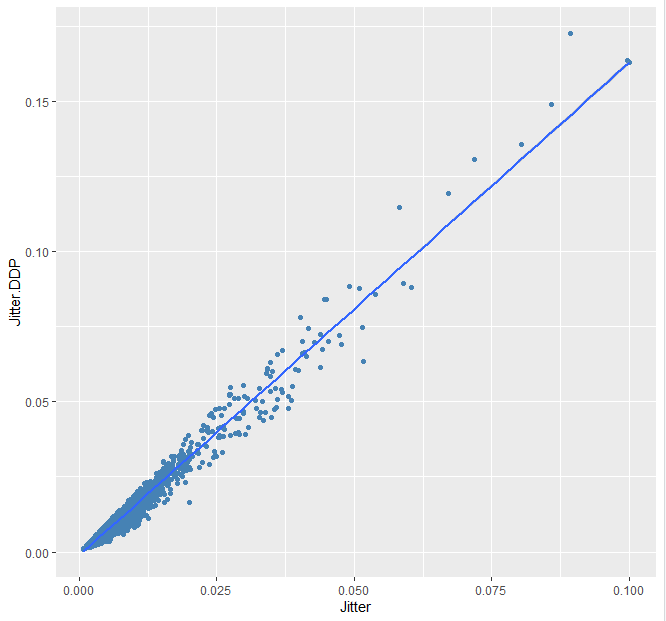
***Example 1 – NHR and DFA***

The correlation reading on variable NHR and DFA has a p-value at -0.0222, which is less than 0.05 significant level, and we reject the null hypothesis. In statistical speaking, there is no linear relationship between NHR and DFA variables.



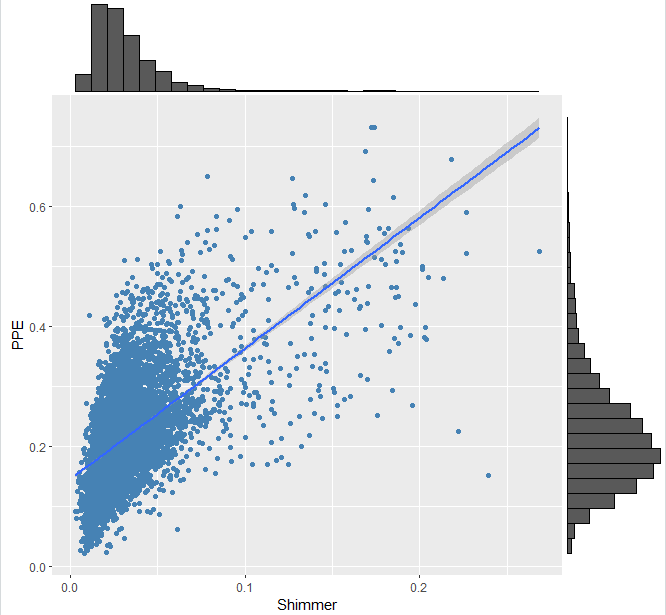
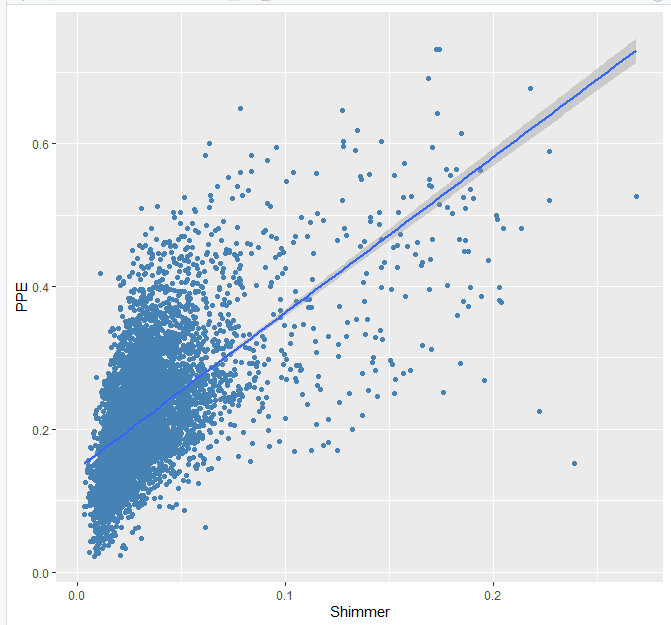
***Example 2 – Jitter and Jitter.DDP***

The correlation reading on variables 'Jitter' and 'Jitter.DDP' has a p-value of 0.984, greater than 0.05 significant level, and we accept the null hypothesis. In statistical speaking, there is a linear relationship between Jitter and Jitter.DDP variables.



***Example 3 – Shimmer and PPE***

The correlation reading on variables 'Jitter' and 'Jitter.DDP' has a p-value at 0.6153, greater than 0.05 significant level, and we accept the null hypothesis. In statistical speaking, there is a linear relationship between Jitter and Jitter.DDP variables.

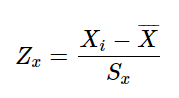


## Standardization – Z Score, Robust Scalar and Logarithmic Data Transformation

**Z-Score standardization**

Z-scores also refers to standardized scores as a common standard gives the scores: a mean of zero and a standard deviation of 1. In general, standardizing the score may not normally be distributed[[1]](#footnote-1). These are the steps to calculate the z-score

1. Subtract the mean from each measurement
2. Divide by standard deviation.

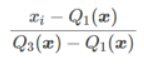


**Robust Scalar Standardization**

Robust Scaling can be calculated with the lower quartile (25th percentile), the median (50th percentile) and the upper quartile (75th percentile)[[2]](#footnote-2). General use to **ignore outliers**

The calculation is:

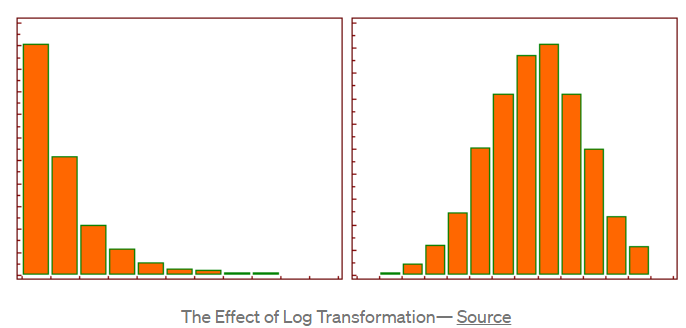
1. Subtract the median from each measurement
2. Divide by the Inter Quartile Range (IQR)



**Logarithmic Data Transformation**

Log transformation transforms each variable with log(variable); it aims to reduce or remove skewness after the change. Logarithmic Data transformation is generally used on **skewed** data and the sample transformation as below.

**Comparison on difference standardization**



|  |  |  |
| --- | --- | --- |
|  | Pro | Cons |
| Z-Score Transformation | This transformation considers the variability of a group or raw scores and the value of the mean.[[3]](#footnote-3)  This method allows raw data comparison from different experiments.[[4]](#footnote-4) | Assuming Z Score is normally distributed and if data is skewed, the distribution of the left and right of the original line is not equal. |
| Robust Scalar Standardization | More focus on the outliner by using the calculation via the interquartile range. | Focus on loose data rather than the median |
| Logarithmic Data Transformation | Can convert skewed data to distributed normally. | Only works in log-normal distribution. |

## Transformation on Jitter.PPQ5(Histogram)

|  |  |
| --- | --- |
| Original – Do not standardized | Z-Score Transformation |
|  |  |
| Robust Scalar Tranformation | Logarithmic Data Transformation |
|  |  |

Jitter.PPQ5 is left skewness distributed, Z-Score and Robust Scalar transformation can not convert it into a more statistical model. Logarithmic Data Transformation is the best fit for this variable.

## Apply Transformation on Shimmer.APQ3

|  |  |
| --- | --- |
| Original – Do not standardized | Z-Score Transformation |
|  |  |
| Robust Scalar Tranformation | Logarithmic Data Transformation |
|  |  |

Similar to the previous one, the 'Shimmer. APQ3' variable is also a left skewness distributed. Z-Score and Robust Scalar transformation can not convert it into a more statistical model. Logarithmic Data Transformation is the best fit for this variable.

## Apply Transformation on RPDE

|  |  |
| --- | --- |
| Original – Do not standardized | Z-Score Transformation |
|  |  |
| Robust Scalar Tranformation | Logarithmic Data Transformation |
|  |  |

Variable RPDE has normally distributed data in the original dataset. Z-score and robust scalar transformation make the distribution more to the centre, more normally distributed. Logarithmic Data Transformation turns the distribution to the right skewness, which is not fit this variable.

## Apply Transformation on Total\_UPDRS

|  |  |
| --- | --- |
| Original – Do not standardized | Z-Score Transformation |
|  |  |
| Robust Scalar Tranformation | Logarithmic Data Transformation |
|  |  |

The distribution of Total\_UPDRS observations is likely normal distributed. Z- score can transform the model into normally distributed by the mean value. When focusing on the median and IQR, Robust Scalar transformation spreads the model along, and Logarithmic Data Transformation turns the model to right skewness. Thus Z-score is the best fit for this.

## Transformation on Jitter.PPQ5 vs Total\_UPDRS (Linear Regression)

|  |  |
| --- | --- |
| Original – Do not standardized | Z-Score Transformation |
|  |  |
| Robust Scalar Tranformation | Logarithmic Data Transformation (Best) |
|  |  |

## Transformation on Shimmer.APQ3 vs Total\_UPDRS (Linear Regression)

|  |  |
| --- | --- |
| Original – Do not standardized | Z-Score Transformation |
|  |  |
| Robust Scalar Tranformation | Logarithmic Data Transformation (Best) |
|  |  |

## Transformation on RPDE vs Total\_UPDRS (Linear Regression)

|  |  |
| --- | --- |
| Original – Do not standardized | Z-Score Transformation |
|  |  |
| Robust Scalar Tranformation | Logarithmic Data Transformation |
|  |  |

## Conclusion

In general, we investigate the dataset and perform cleansing based on experience and knowledge, then apply central theorem and graphical presentation to understand the data model.

There may not always have a normally distributed data model. There are three transformation methods introduced and experimented with within this report. For the correlation test on Jitter.PPQ5 and Total\_UPDRS (Page.15) and Shimmer.APQ3 and Total\_UPDRS (P.16) both data was left skewness in the original dataset. Logarithmic transformation indicates the distribution are statistically normal distributed; Residual standard error on Jitter.PPQ5 and Total\_UPDRS are 0.5727 (was 0.003729 in original) and Shimmer.APQ3 is 0.6131 (was 0.0132 in original) after logarithmic, which means the average amount of actual values differ from the predictions (the line). By looking in the Adjusted R square value, Jitter.PPQ5 is 0.0245, which was 0.003887 and Shimmer.APQ3 is 0.01546, which was 0.006159. These show the percentage of variation within the dependent variables that all predictors explained. In other words, the value closer to one is better. Thus, Logarithmic Data Transformation is the best fit for Total\_UPDRS.

## Reference

"Data Preprocessing Using Python Sklearn | by Kesari Mohan Reddy | Medium." Accessed October 2, 2021. https://medium.com/@kesarimohan87/data-preprocessing-6c87d27156.

"Machine Learning - Pros and Cons of Using the Zscore of a Dataset before Normalizing It during Feature Engineering? - Data Science Stack Exchange." Accessed October 2, 2021. https://datascience.stackexchange.com/questions/48562/pros-and-cons-of-using-the-zscore-of-a-dataset-before-normalizing-it-during-feat.

"What Are Z-Scores? Quick Tutorial with Examples." Accessed October 2, 2021. https://www.spss-tutorials.com/z-scores-what-and-why/.

"Why Do Researchers Use Z-Scores to Determine Probabilities? Are There Advantages to Using This Tool? Provide Examples of the z-Score in use. | Study.Com." Accessed October 2, 2021. https://study.com/academy/answer/why-do-researchers-use-z-scores-to-determine-probabilities-are-there-advantages-to-using-this-tool-provide-examples-of-the-z-score-in-use.html.

# Appendix – R Code

#Descriptive statistic in R

# The easiest way to get ggplot2 is to install the whole tidyverse:

# install.packages("tidyverse")

#set

setwd("C:/Users/user/Documents/2. Lanagara College/DANA\_4810/Assignment1\_Pakinson")

#read in the data

library("readr")

parkinsons <- read\_csv("Parkinsons.csv")

#install.packages("psych")

library("psych")

describe(parkinsons)

#Read the Clean Data set

parkinsonsClean <- read\_csv("ParkinsonsClean.csv")

describe(parkinsonsClean)

#Read the Clean Data set

parkinsonsClean2 <- read\_csv("ParkinsonsClean2.csv")

describe(parkinsonsClean2)

#Find out the correlation of Parkinson data set

cor(parkinsonsClean2)

#---------------Scattorplot and best line------------------

library("tidyverse")

library("ggplot2")

#install.packages("ggExtra")

library("ggExtra")

#Graphical presentation NHR and DFA

p1 <- ggplot(parkinsonsClean2, aes(x = NHR, y = DFA)) +

geom\_point(color= "steelblue") +

geom\_smooth(method = "lm")

p1

#second Graph

ggMarginal(p1, parkinsonsClean2, NHR, DFA, type = c("histogram"))

#Graphical presentation Jitter and Jitter.DDP

p2 <- ggplot(parkinsonsClean2,aes(x = Jitter, y = Jitter.DDP)) +

geom\_point(color= "steelblue") +

geom\_smooth(method = "lm")

p2

#second Graph

ggMarginal(p2, parkinsonsClean2, Jitter, Jitter.DDP, type = c("histogram"))

#Graphical presentation Shimmer and PPE

p3 <- ggplot(parkinsonsClean2, aes(x = Shimmer,y = PPE)) +

geom\_point(color= "steelblue") +

geom\_smooth(method = "lm")

p3

#second Graph

ggMarginal(p3, parkinsonsClean2, Shimmer, PPE, type = c("histogram"))

#------------------- Z Score --------------------------------------------------------------

library("readr")

parkinsons\_ZScore <- read\_csv("ParkinsonsClean2.csv")

#Z transform

zScore\_Parkinsons <- as.data.frame(scale(parkinsons\_ZScore, center = TRUE, scale = TRUE))

#histogram of variables - Original Dataset

hist(parkinsonsClean2$Jitter.PPQ5)

hist(parkinsonsClean2$Shimmer.APQ3)

hist(parkinsonsClean2$RPDE)

hist(parkinsonsClean2$total\_UPDRS)

#histogram of variables - After Z-score transformed Dataset

hist(zScore\_Parkinsons$Jitter.PPQ5)

hist(zScore\_Parkinsons$Shimmer.APQ3)

hist(zScore\_Parkinsons$RPDE)

hist(zScore\_Parkinsons$total\_UPDRS)

#------------------- Robust Scalar --------------------------------------------------------------

library("readr")

parkinsons\_Robust <- read\_csv("ParkinsonsClean2.csv")

print(parkinsons\_Robust)

#ROBUST TRANSFORM

# Store Robust calculation formula into robust\_transformation variable

robust\_transformation<- function(x){(x- median(x)) /(quantile(x,probs = .75)-quantile(x,probs = .25))}

# Apply robust transformation logic onto Partkins\_Robust

robust\_Parkinsons <- as.data.frame(lapply(parkinsons\_Robust, robust\_transformation))

#histogram of variables - Original Dataset

hist(parkinsonsClean2$Jitter.PPQ5)

hist(parkinsonsClean2$Shimmer.APQ3)

hist(parkinsonsClean2$RPDE)

hist(parkinsonsClean2$total\_UPDRS)

#histogram of variables - After Z-score transformed Dataset

hist(robust\_Parkinsons$Jitter.PPQ5)

hist(robust\_Parkinsons$Shimmer.APQ3)

hist(robust\_Parkinsons$RPDE)

hist(robust\_Parkinsons$total\_UPDRS)

#------------------- Log --------------------------------------------------------------

library("readr")

parkinsons\_log <- read\_csv("ParkinsonsClean2.csv")

#Checking for Skewness

#install.packages("e1071")

library("e1071")

library("dplyr")

skewness(parkinsons\_log$Jitter.PPQ5)

skewness(parkinsons\_log$Shimmer.APQ3)

skewness(parkinsons\_log$RPDE)

skewness(parkinsons\_log$total\_UPDRS)

# log in R example - Perform log calculation on whole Dataset

logTransform\_Parkinsons = log(parkinsons\_log)

#histogram of variables - After log transformed Dataset

hist(logTransform\_Parkinsons$Jitter.PPQ5)

hist(logTransform\_Parkinsons$Shimmer.APQ3)

hist(logTransform\_Parkinsons$RPDE)

hist(logTransform\_Parkinsons$total\_UPDRS)

#------------------- Linear Regression - Original Dataset --------------------------------------------------------------

lmHeight = lm(Jitter.PPQ5~total\_UPDRS, data = parkinsonsClean2) #Create the linear regression

summary(lmHeight) #Review the results

library('tidyverse')

library('ggplot2')

ggplot(data= parkinsonsClean2,aes(x = Jitter.PPQ5, y = total\_UPDRS)) +

geom\_point(color='blue') +

geom\_smooth(method="lm")

lmHeight = lm(Shimmer.APQ3~total\_UPDRS, data = parkinsonsClean2) #Create the linear regression

summary(lmHeight) #Review the results

library('tidyverse')

library('ggplot2')

ggplot(data= parkinsonsClean2,aes(x = Shimmer.APQ3, y = total\_UPDRS)) +

geom\_point(color='blue') +

geom\_smooth(method="lm")

lmHeight = lm(RPDE~total\_UPDRS, data = parkinsonsClean2) #Create the linear regression

summary(lmHeight) #Review the results

library('tidyverse')

library('ggplot2')

ggplot(data= parkinsonsClean2,aes(x = RPDE, y = total\_UPDRS)) +

geom\_point(color='blue') +

geom\_smooth(method="lm")

#------------------- Linear Regression - zScore Dataset --------------------------------------------------------------

lmHeight = lm(Jitter.PPQ5~total\_UPDRS, data = zScore\_Parkinsons) #Create the linear regression

summary(lmHeight) #Review the results

library('tidyverse')

library('ggplot2')

ggplot(data= zScore\_Parkinsons,aes(x = Jitter.PPQ5, y = total\_UPDRS)) +

geom\_point(color='blue') +

geom\_smooth(method="lm")

lmHeight = lm(Shimmer.APQ3~total\_UPDRS, data = zScore\_Parkinsons) #Create the linear regression

summary(lmHeight) #Review the results

library('tidyverse')

library('ggplot2')

ggplot(data= zScore\_Parkinsons,aes(x = Shimmer.APQ3, y = total\_UPDRS)) +

geom\_point(color='blue') +

geom\_smooth(method="lm")

lmHeight = lm(RPDE~total\_UPDRS, data = zScore\_Parkinsons) #Create the linear regression

summary(lmHeight) #Review the results

library('tidyverse')

library('ggplot2')

ggplot(data= zScore\_Parkinsons,aes(x = RPDE, y = total\_UPDRS)) +

geom\_point(color='blue') +

geom\_smooth(method="lm")

#------------------- Linear Regression - Robust Scalar Dataset --------------------------------------------------------------

lmHeight = lm(Jitter.PPQ5~total\_UPDRS, data = robust\_Parkinsons) #Create the linear regression

summary(lmHeight) #Review the results

library('tidyverse')

library('ggplot2')

ggplot(data= robust\_Parkinsons,aes(x = Jitter.PPQ5, y = total\_UPDRS)) +

geom\_point(color='blue') +

geom\_smooth(method="lm")

lmHeight = lm(Shimmer.APQ3~total\_UPDRS, data = robust\_Parkinsons) #Create the linear regression

summary(lmHeight) #Review the results

library('tidyverse')

library('ggplot2')

ggplot(data= robust\_Parkinsons,aes(x = Shimmer.APQ3, y = total\_UPDRS)) +

geom\_point(color='blue') +

geom\_smooth(method="lm")

lmHeight = lm(RPDE~total\_UPDRS, data = robust\_Parkinsons) #Create the linear regression

summary(lmHeight) #Review the results

library('tidyverse')

library('ggplot2')

ggplot(data= robust\_Parkinsons,aes(x = RPDE, y = total\_UPDRS)) +

geom\_point(color='blue') +

geom\_smooth(method="lm")

#------------------- Linear Regression - Log Dataset --------------------------------------------------------------

lmHeight = lm(Jitter.PPQ5~total\_UPDRS, data = logTransform\_Parkinsons) #Create the linear regression

summary(lmHeight) #Review the results

library('tidyverse')

library('ggplot2')

ggplot(data= logTransform\_Parkinsons,aes(x = Jitter.PPQ5, y = total\_UPDRS)) +

geom\_point(color='blue') +

geom\_smooth(method="lm")

lmHeight = lm(Shimmer.APQ3~total\_UPDRS, data = logTransform\_Parkinsons) #Create the linear regression

summary(lmHeight) #Review the results

library('tidyverse')

library('ggplot2')

ggplot(data= logTransform\_Parkinsons,aes(x = Shimmer.APQ3, y = total\_UPDRS)) +

geom\_point(color='blue') +

geom\_smooth(method="lm")

lmHeight = lm(RPDE~total\_UPDRS, data = logTransform\_Parkinsons) #Create the linear regression

summary(lmHeight) #Review the results

library('tidyverse')

library('ggplot2')

ggplot(data= logTransform\_Parkinsons,aes(x = RPDE, y = total\_UPDRS)) +

geom\_point(color='blue') +

geom\_smooth(method="lm")

1. “What Are Z-Scores? Quick Tutorial with Examples,” accessed October 2, 2021, https://www.spss-tutorials.com/z-scores-what-and-why/. [↑](#footnote-ref-1)
2. “Data Preprocessing Using Python Sklearn | by Kesari Mohan Reddy | Medium,” accessed October 2, 2021, https://medium.com/@kesarimohan87/data-preprocessing-6c87d27156. [↑](#footnote-ref-2)
3. “Why Do Researchers Use Z-Scores to Determine Probabilities? Are There Advantages to Using This Tool? Provide Examples of the z-Score in Use. | Study.Com,” accessed October 2, 2021, https://study.com/academy/answer/why-do-researchers-use-z-scores-to-determine-probabilities-are-there-advantages-to-using-this-tool-provide-examples-of-the-z-score-in-use.html. [↑](#footnote-ref-3)
4. “Machine Learning - Pros and Cons of Using the Zscore of a Dataset before Normalizing It during Feature Engineering? - Data Science Stack Exchange,” accessed October 2, 2021, https://datascience.stackexchange.com/questions/48562/pros-and-cons-of-using-the-zscore-of-a-dataset-before-normalizing-it-during-feat. [↑](#footnote-ref-4)