Understanding the Correlation of Parkinson's dataset with Multivariate Analysis Packages

Rajha Priya. A[1,2*] and Preenon Bagchi[2,3]

[1]Padmashree Institute of Management and Sciences, Bengaluru, India.

[2]Vasishth Academy of Advanced Studies and Research (Sarvasumana Association), Bengaluru, India.

[3]Institute of Biosciences and Technology, MGM University, Aurangabad, India

[*]Corresponding author: Email: rajhapriya24@gmail.com

ABSTRACT: Parkinson's dataset available in open source was used in this work.Inferential statistical analysis is used here to inspect each unit from this dataset and to test a hypothesis depending on the sample data. From the analysis inferences could be extracted by applying probability and make generalizations about the whole data. This method can also be used in experimental and quasi-experimental research design or in program outcome evaluation. During the analysis, the whole population of the sample data was considered while making deductions. Specifically, multivariate analysis was used which involves more than two dependent variables, this method helps in reduction and simplification of data without losing its details. Packages like ggplot2, psych, corrgram, performance analysis, ggcorrplot, ggpubr and RColorBrewer Packages were used for interpreting the data.

Keywords: Parkinson's disease, multivariate analysis, interpretation, graph, inferential statistics, comparative data analysis

R. Somashekhar et al. (eds.), *Proceedings of the International Conference on Advances in Nano-Neuro-Bio-Quantum (ICAN 2023)*,
Advances in Intelligent Systems Research 69,
https://doi.org/10.2991/978-94-6463-294-1_2

1. INTRODUCTION

Parkinson's dataset which is available as open access at UCI Machine Learning Repository by the work "Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection" was utilized for this study. The dataset used for this study has a range of biomedical voice measurements of 31 people done by Little *et.al.* 2007 [1, 2]. With inferential statistics the dataset was interpreted to study each unit and further method involves the sampling theory, various tests of significance and statistical control. The goal of this inferential statistics to provide measurements that can describe the overall population of a project by studying a smaller sample of it. This is a statistical study of data where multiple measurements are made on each experimental units and where the relationship among multivariate measurements and their structures are important.

1.1: Multivariate data analysis

Data analysis is one of the most useful methods to understand the vast amount of data presented to them and synthesize evidence from it [3]. Multivariate analysis is used to interpret the dataset in order to avoid common pitfalls. The greatest virtue of such a model is that it considers as many factors into consideration as possible and results in reduction of bias and gives a result closer to reality [4].

R-Statistical packagesare used here for data analytics and visualization[5].R- Packages used in this work are ggplot2, psych, corrgram, performance analysis, RColorBrewer, ggcorrplot, ggpubr.

Ggplot2 package [6] (stands for The Grammar of Graphics) used here greatly improves the quality and aesthetics of graphs and gives more focuses to almost every section of the data and it also gives efficient commands to create complex plots from data.

Pearson correlation "psych package" [7] is used for scale construction using factor analysis, principal component analysis, cluster analysis and reliability analysis.By using psych correlation which is an extension of the pairs function that allows to easily add regression lines, confidence intervals and several additional arguments and it creates a graph of a correlation matrix colouring the regions by the level of correlation.

Performance Analytics package [8] is an econometric tool for performance and risk analysis of particular data and it works along with correlation factor using chart. Correlation as a function and the graph shows the diagonal bivariate scatter plots with a fitted line. The significant level is also associated to a symbol which represents the p-values in the graph. Chart correlation function of performance analytics package is one of the shortcuts to create a correlation plot in R with histograms, density functions, smoothed regression lines and correlation coefficients with the corresponding significance levels.In Performance analysis package, the significant level is also associated to a symbol which represents the p-values in the graph.

RColorBrewer package [9] helps to choose any suitable colour with the three group of data knows as sequential, diverging and qualitative.

Package ggcorrplot [10] helps in visualizing a correlation matrix with significance level on the correlogram. It also holds the correlating p-values in the correlation matrix

Preleminary test usingggpubr package [11] is carried over to check the test assumptions whether it is linear or non-linear. Along with this, Shapiro wilk test can be performed to state the

Null or Alternative hypothesis. The results could be predicted with the comparison of two p-values significant level which implies that the distribution of data.

2. MATERIALS AND METHODS

Parkinson biomedical voice measurement open-source dataset - https://archive.ics.uci.edu/ml/datasets/parkinsonsfrom Max Little University of Oxford in collaboration with the National Centre for Voice and Speech, Denver, Colorado who recorded the speech signals was taken for this work.

With inferential statistics the dataset was interpreted to study each unit and further method involves the sampling theory, various tests of significance and statistical control.

Multivariate dataset analysis was carried over the dataset using various packages like ggplot2, psych, RColorBrewer, performance analysis, ggcorrplot, ggpubr.

Interpretation of dataset and study of relationship between variables without losing its details was done using inferential statistics techniques.

Ggplot2 package was used here toproduce quality and aesthetic graphs. This gave more focuses every section of the dataset. Further it created complex plots from data.

Psych packageas used toascertain the correlation and confidence intervals between the variables. Further, scatterplot was generated to understand the correlation.

Performance analysis used here to show the relationship between the datasets with the corresponding significance levels. This was ascertained using histograms plots, density functions, smoothed regression lines and correlation coefficients with significance levels.

Ggcorrplot packagecreates easy ready with significance level correlogram plots.

RColorBrewer packagechoose various colour schemes for mean and standard deviation value representation.

Ggpubr package used here to visualize the normal distribution pattern of the dataset.

The packages were installed and further graphical multivariate analysis using RStudio was performed. Scatterplots are used here since there are more than two variables and correlation between one variable versus the remaining ones is needed

3**. RESULTS AND DISCUSSION**

Parkinson's dataset available at https://archive.ics.uci.edu/ml/datasets/parkinsons was downloaded for this work.

As per the inferential multivariate statistics the data was analysed using various packages

Scatterplot was performed for the several measures of variation in amplitude MDVP.Shimmer, minimum vocal fundamental frequency MDVP.Flo.Hz., average vocal fundamental frequency MDVP.FO.Hz., and between the ratio of noise to tonal components in the voice HNR as seen in Fig. 1. Further in fig. 1 it is understood that there is increase in correlation except in the several measures of variation in amplitude MDVP.Shimmer.

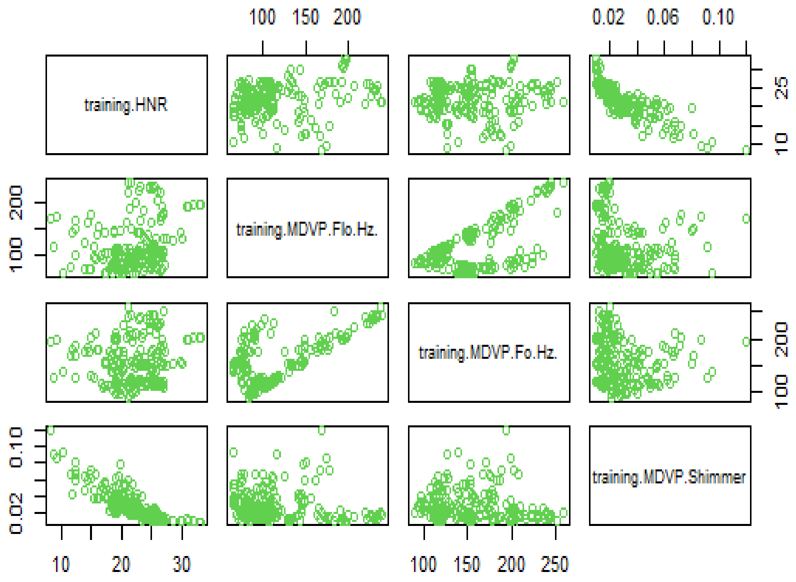

Fig. 1: Plot generated using “ggplot2 package”

Also, as per fig. 2 positive correlation is seen in all the distribution like maximum, minimum and average vocal fundamental frequency using ggplot2 package (MDVP.Fhi.Hz., MDVP.Flo.Hz., MDVP.Fo.Hz.

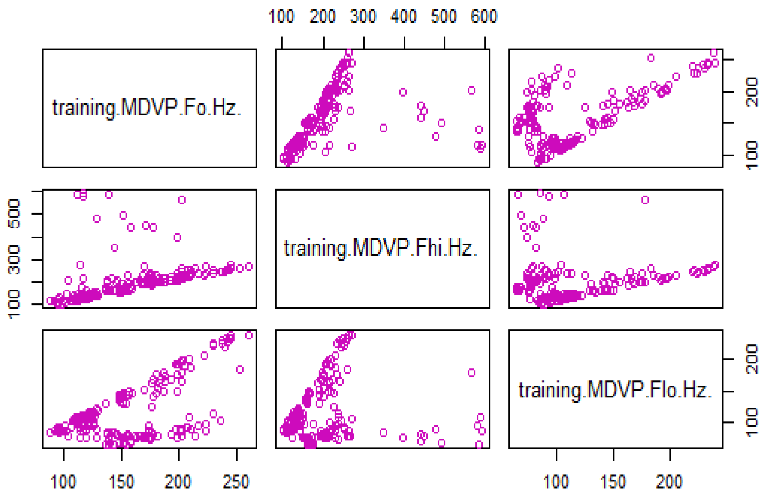

Fig. 2: Correlation representation in the datasetusing ggplot2 package.

As per fig. 3 it is positive correlation is seen. The 3d scatterplot shows the increase in overall range of the data.

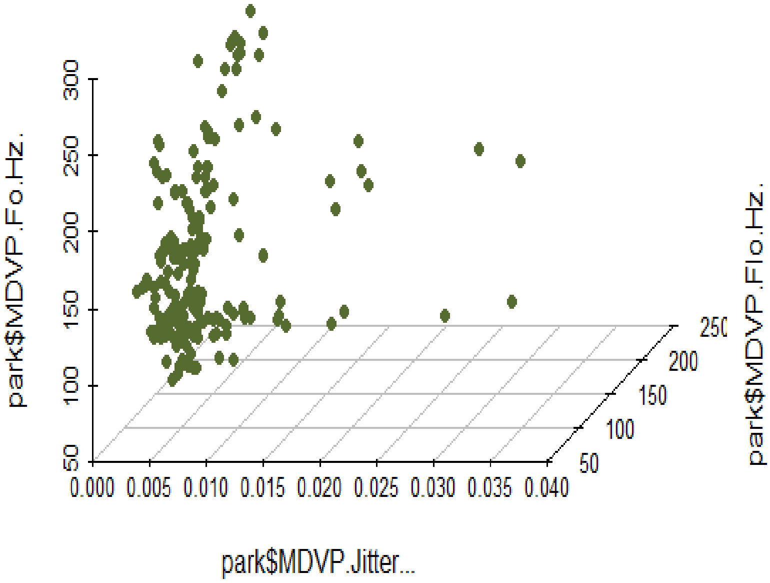

Fig. 3: Three-dimensional scatterplot.

Fig. 4 depictsmoderate positive (Pearson) correlation amongst the dataset. The graphs depict positive correlation for principal component analysis, cluster analysis and reliability analysis. The correlation was performed for maximum, minimum and average voice frequency and the graph shows the strength of linear relationship between the MDVP.Fhi.Hz., MDVP.Fo.Hz., MDVP.Flo.Hz., Hz. MDVP.Jitter all shows moderate correlation since the value lies between -1 to +1.

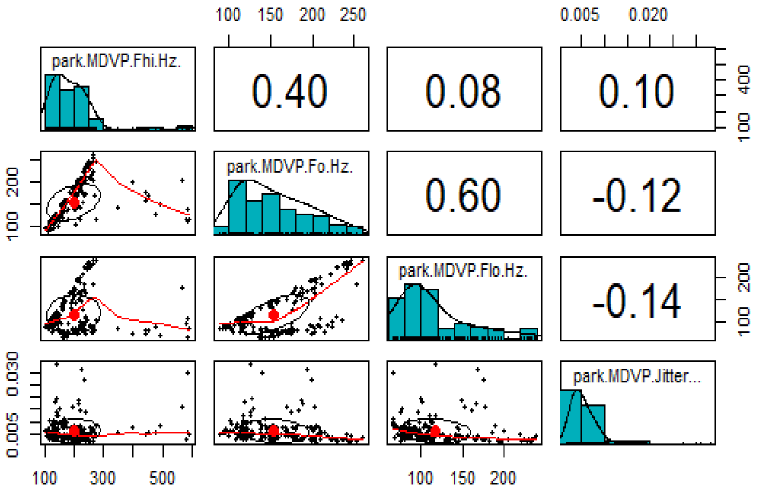

Fig. 4: psych plot.

Fig. 5 shows bivariate scatter plots with a fitted lines displayed at the bottom of the diagonal and at the top of diagonal, the value of the correlation plus the significance level as stars. Each significance level is associated to a symbol of p-values. And this p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true. Here, the values are all statistically significant.

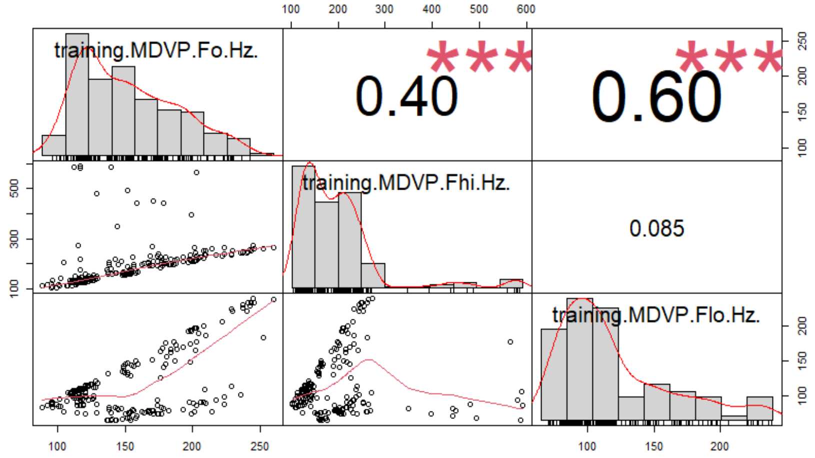

Fig. 5: The distribution of each variable is shown on the diagonal using performance analysis.

The fig. 6 graph displays the significance level on the correlogram. Also, it has the function computing a matrix of correlation p-values. Positive correlation is seen between 1 to 0.5.

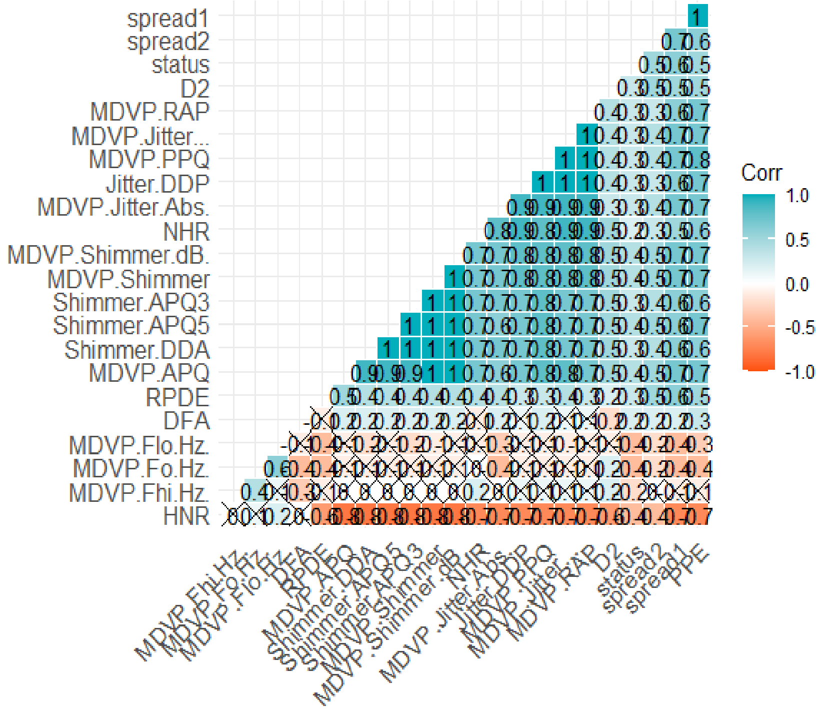

Fig. 6: ggcorrplot

The peak mean value in each variable in fig. 7 was noted. Blue represents the maximum frequency 600 whereas minimum vocal frequency has the peak value of 280 and the average has the value of 230 as shown in the graph in fig. 7.

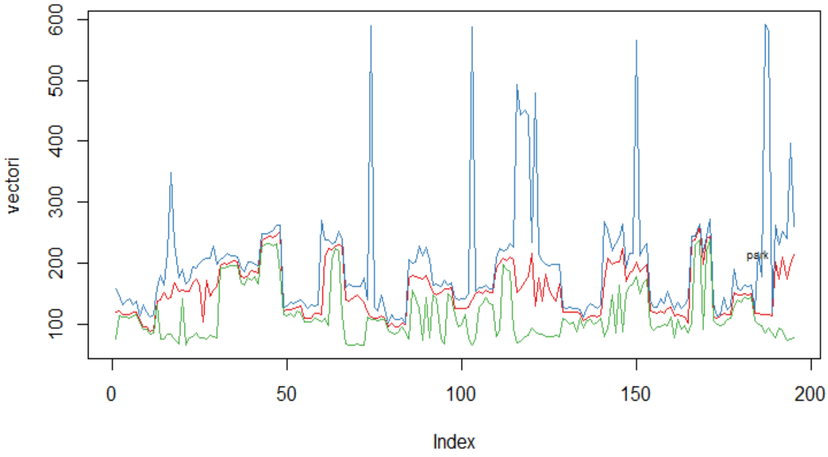

Fig. 7: RColorBrewer Plot

Preleminary test using ggpubr packageis carried to check the test assumptions. It is seen that the graph is almost linear. Fig. 8(a), (b) infers the Preleminary test outcome based on Shapiro-Wilk normality test performed using “ggpubr” package. Here both the maximum vocal fundamental frequency and minimum vocal fundamental frequency are analysed and both data of normality plots shows the normal distribution pattern.

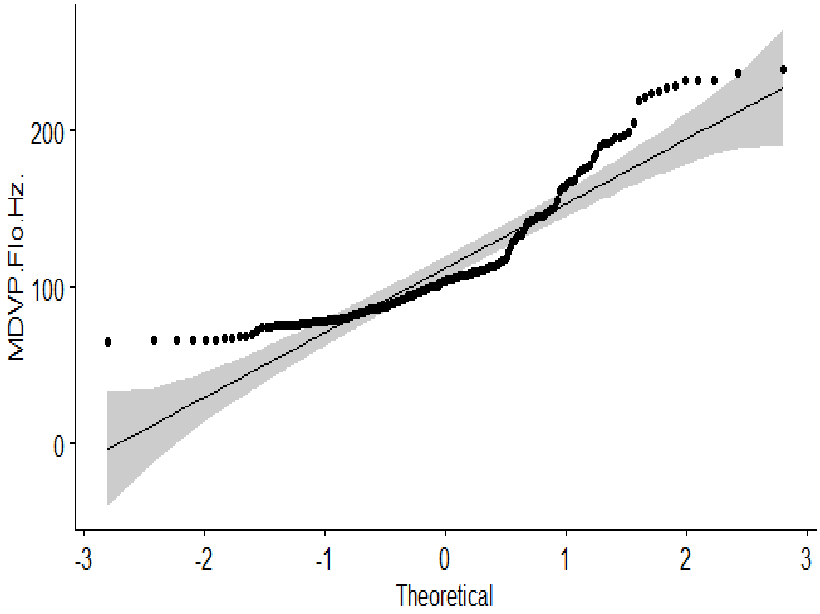

Fig. 8(a): ggpubr plot of MDVP.Flo.Hz

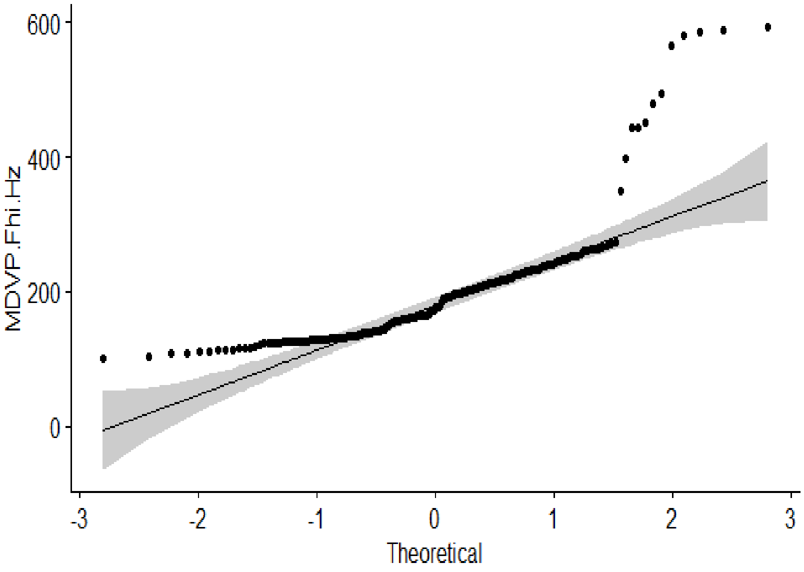

Fig. 8(b): ggpubr plot of MDVP.Fhi.Hz

CONCLUSION

Parkinson biomedical voice measurement open-source dataset was used in this study. Multivariate analysis of dataset was carried overusing various packages like ggplot2, psych, performance analysis, RColorBrewer, ggcorrplot, ggpubr. The inferential statistics was performed to interpret each unit of data and also the sampling theory, various tests of significance and overall positive correlation was obtained in the dataset. The linear relationship between the maximum, minimum and average voice frequency have moderate strength of correlation which is measured using the strength of straight line or linear relationship between two variables. And it clearly represents that MDVP.Flo.Hz. has a weak negative linear relationship via a shaky linear rule whereas MDVP.Fhi.Hz., MDVP.Fo.Hz., of maximum and average voice frequency has a moderate positive linear relationship via a fuzzy-firm linear rule. Further the scatter plots associated with p-values shows that null hypothesis is true because the values are not affecting each other and it shows significance. Followed by normal distribution pattern was observed using Shapiro-Wilk test with the curved patterns. By interpretating using correlation the data association between one another was studied in the manner of scatter plot to a linear regression line. And also, the scientific research data different units were considered to interpret the data to visual form here which permits the accuracy, summarize the large dataset and frequency distribution has clearly studied.

DECLARATION

1. Ethics approval and consent to participate: Not applicable
2. Consent for publication: Not applicable

Item. No.	Data	Resource
1.	Parkinson's dataset	https://archive.ics.uci.edu/ml/datasets/parkinsons
2.	ggplot2 package	Author of this package is Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, andmaintaineris Thomas Lin Pedersen.
3.	psych package	Author and maintainer areWilliam Revelle
4.	ggcorrplot package	Author and maintainer of this package are Alboukadel Kassambara.
5.	PerformanceAnalytics package	Authors of this package is Brian G. Peterson, Peter Carl, Kris Boudt, Ross Bennett, Joshua Ulrich, Eric Zivot, Dries Cornilly, Eric Hung, Matthieu Lestel, Kyle Balkissoon, Diethelm Wuertz, Anthony Alexander Christidis, R. Douglas Martin, Zeheng 'Zenith' Zhou, Justin M. Shea and this package and maintainer is Brian G. Peterson
6.	ggpubr package	Author and maintainer are Alboukadel Kassambara.
7.	RColorBrewer package	Author and maintainer of this package are Erich Neuwirth

3. Availability of data and materials:

4. Competing interests: None

5. Funding: None

6. Authors' contributions: All authors have equal contribution

7. Acknowledgement: We acknowledge the lab support provided by our institution.

REFERENCE

1. Little M, McSharry P, Roberts S, Costello D, Moroz I. 2007, Exploiting Nonlinear Recurrence and Fractal Scaling Properties of Voice Disorder Detection. 6:23
2. Little M, McSharry P, Hunter E, Spielman J, Ramig L. 2009,Suitability of Dysphonia Measurements for Telemonitoring of Parkinson’s Disease. Vol.56, No.4
3. Chan B, 2018, Data Analysis using R Programming. 1082:47-122, Doi: 10.1038/psp.2013.56
4. Ren P., Bayard J., Dong L., Chen J, Mao L, Ma D, Sanchez M, Morejon D, Bringas M, Yao D, Jahanshahi M, Sosa P2020, Multivariate Analysis of Joint Motion Data by Kinect: Application to Parkinson’s Disease. 28(1):181-190, DOI: 10.1109/TNSRE.2019.2953707
5. Giorgi F., Ceraolo C., Mercatelli D 2022, The R Language: An Engine for Bioinformatics and Data Science. 12(5):648, DOI: 10.3390/life12050648
6. Ito K, Murphy D 2013, Application of ggplot2 to Pharmacometric Graphics. 2(10): e79
7. Revelle W, 2016, psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, http://CRAN.R-project.org/package=psych Version = 1.6.

8. Peterson BG, Carl P, K Boudt, Bennett R, Ulrich J, Zivot E, Lestel M, Balkissoon K, Wuertz D, 2014, PerformanceAnalytics: Econometric tools for performance and risk analysis, R package version 1 (3)

9. Neuwirth E, 2022, RColorBrewer, https://cran.r-project.org/web/packages/RColorBrewer/index.html

10. Alboukadel A, 2019, ggcorrplot, https://cran.r-project.org/web/packages/ggcorrplot/index.html

11. Kassambara A 2020, ggpubr, https://cran.r-project.org/web/packages/ggpubr/index.html