DataScanR

Quick Guide

Ilona Szczot[1], Jyotirmoy Das[2]
[1]Centrum för social och affektiv neurovetenskap, Linköping University, Linköping, Sweden,
[2]Core Facility, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden and Clinical Genomics Linköping, Science for Life Laboratory, Sweden

Table of Contents

Description

This application allows the user to preview large data files and perform basic data exploration. Supported file format is a standard, comma separated file format (.csv). The original file is never modified by the application. The data is not saved on the server.

1. Data Cleaning

Data Cleaning section of this application allows the user to upload data file and review the data quality as well as look into the basic statistical summary of the data. (Fig1.)

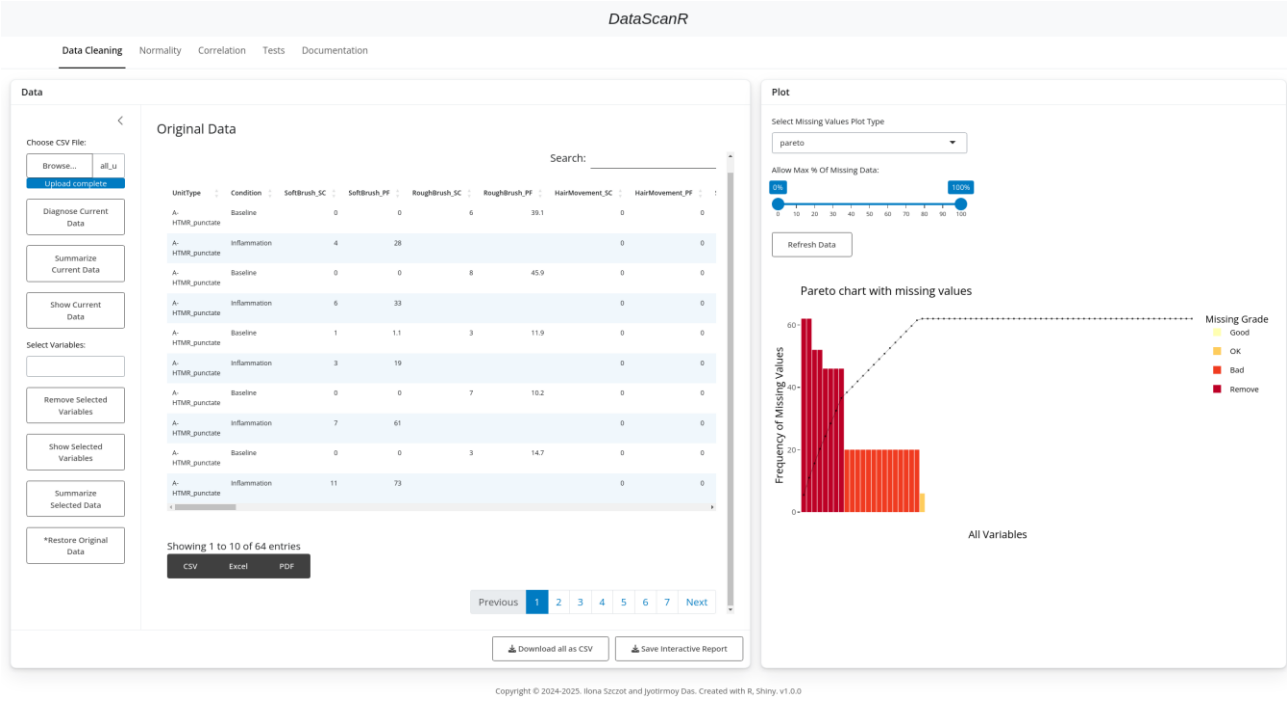


Fig1. Data Cleaning Section

1.1. Upload Data

Click on the “Browse...” button to select your data.csv file. This will allow you to see the data in the form of a table. Use “Previous” and “Next” buttons (Fig2. b), to see all pages of your data. You can easily export to csv, excel or pdf each displayed table (Fig2. a) or export all of the pre-processed data to csv file (Fig2. c).

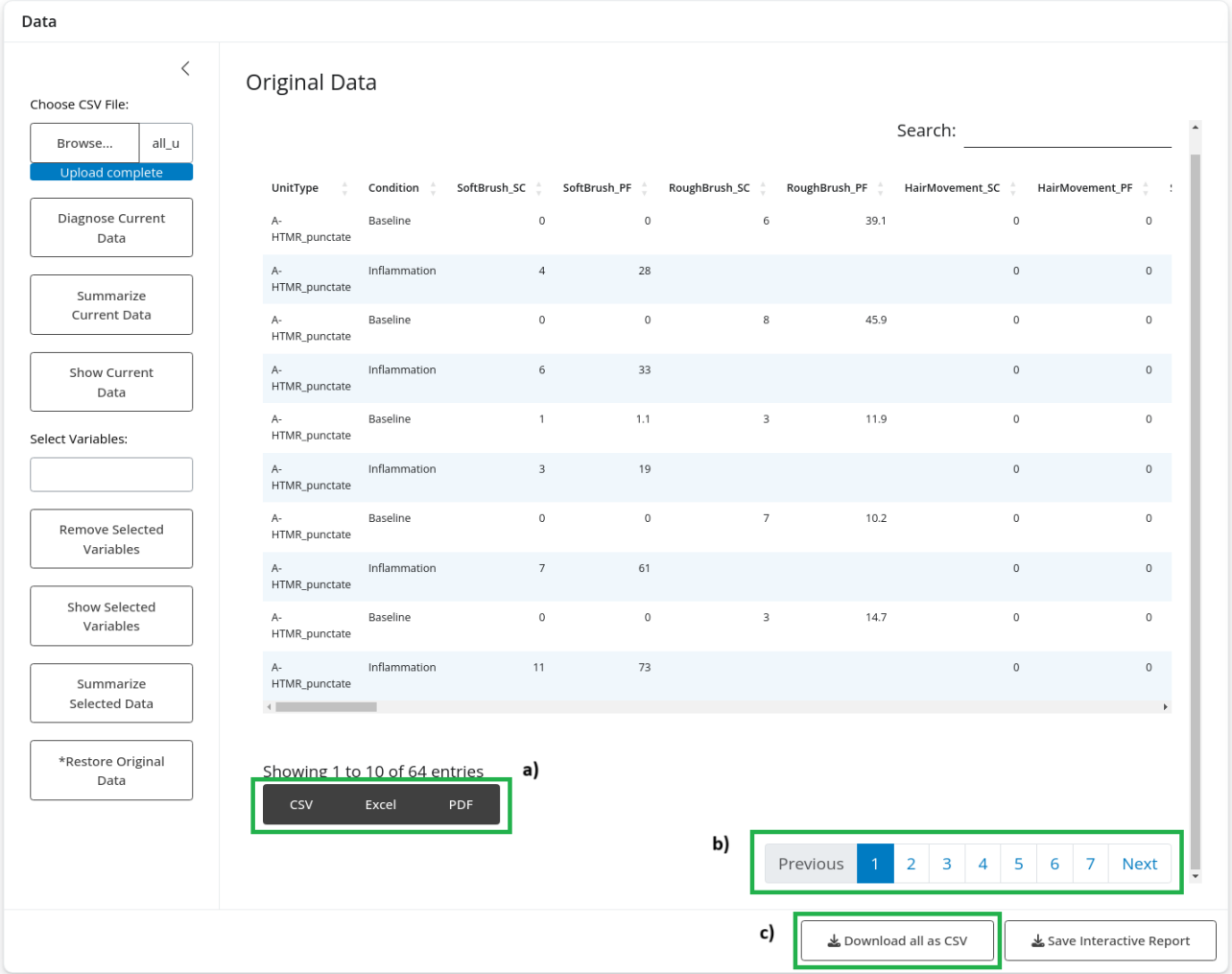


Fig2. Uploaded data file in the “Data Cleaning” section. a) Buttons to export currently displayed table to csv, excel or PDF. b) Buttons to change between the pages of the data. c) Button to export all data to csv.

1.2. Diagnose Current Data

This option allows to see what kind of variable types are there, how much data is missing as well as how many unique values are there for each variable (Fig3.).
This option uses diagnose function from dlookr package.

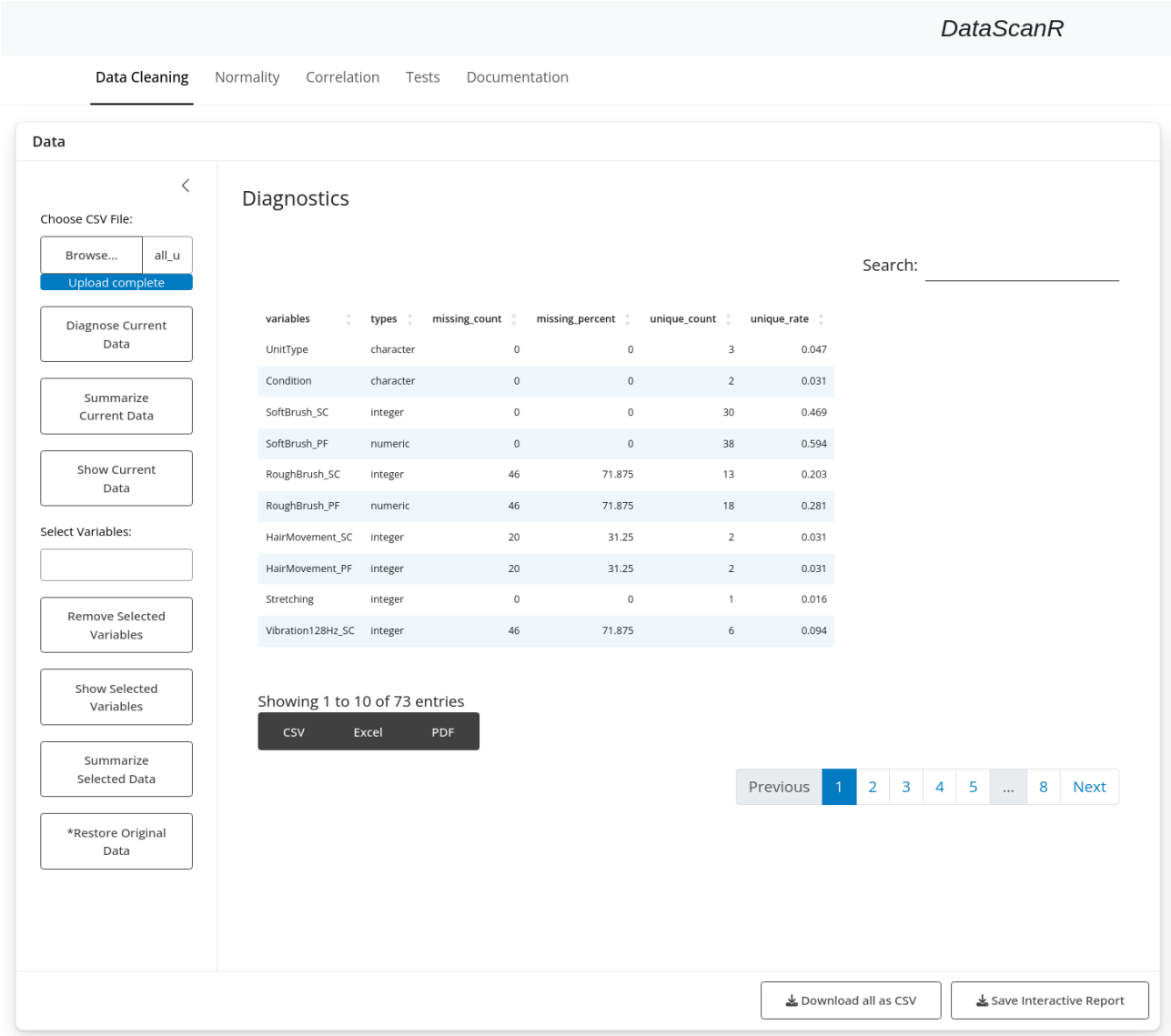


Fig3. Data diagnostics.

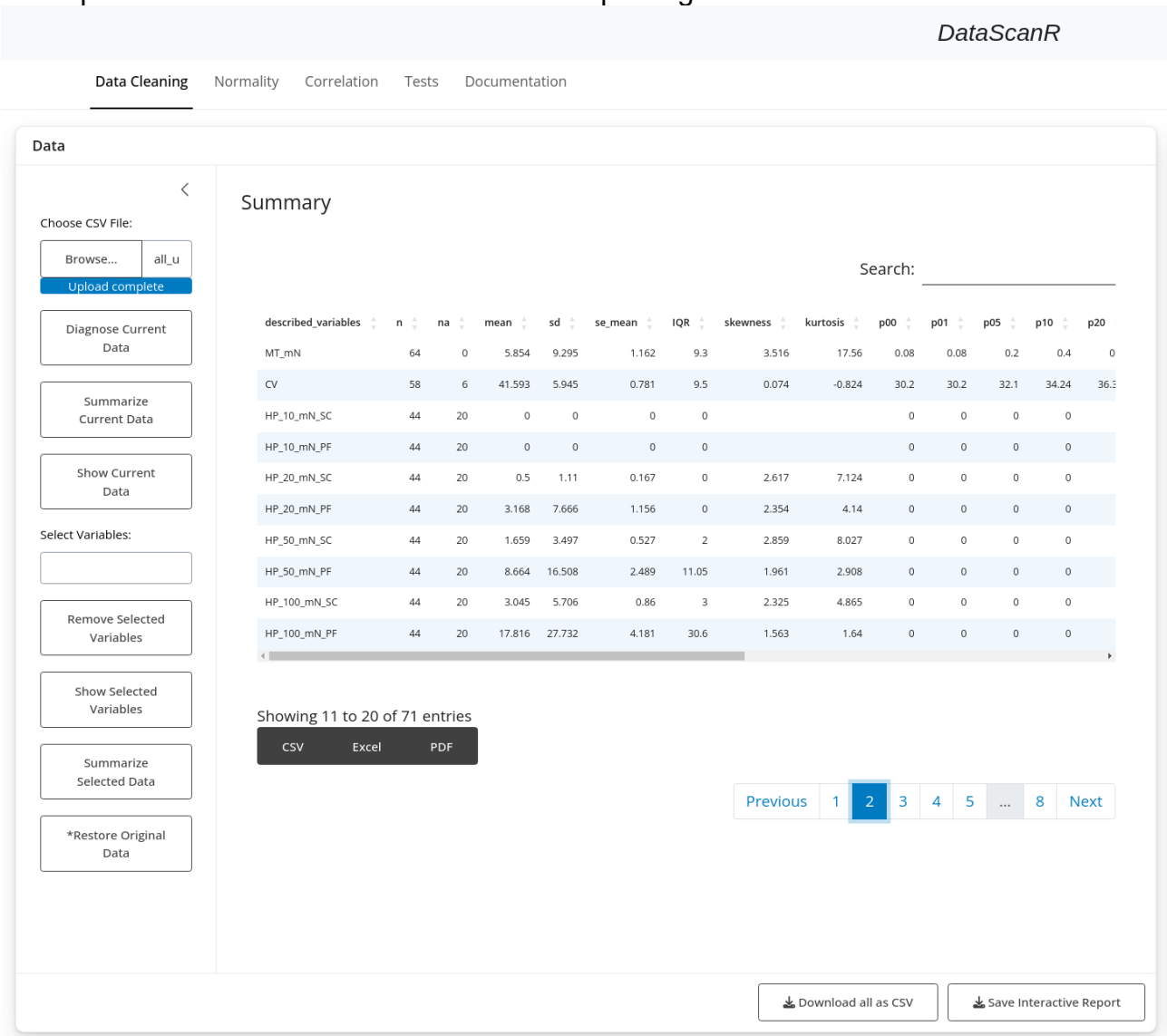
1.3. Summarize Current Data

This option calculates descriptive statistics for each variable (Fig4.):

- n : number of observations excluding missing values
- na : number of missing values
- mean : arithmetic average
- sd : standard deviation
- se_mean : standard error mean. sd/sqrt(n)
- IQR : interquartile range (Q3-Q1)
- skewness : skewness
- kurtosis : kurtosis
- p25 : Q1. 25% percentile
- p50 : median. 50% percentile
- p75 : Q3. 75% percentile
- p01, p05, p10, p20, p30 : 1%, 5%, 20%, 30% percentiles
- p40, p60, p70, p80 : 40%, 60%, 70%, 80% percentiles

- p90, p95, p99, p100 : 90%, 95%, 99%, 100% percentiles

This option uses describe function from dlookr package.



described_variables	n	na	mean	sd	se_mean	IQR	skewness	kurtosis	p00	p01	p05	p10	p20
MT_mN	64	0	5.854	9.295	1.162	9.3	3.516	17.56	0.08	0.08	0.2	0.4	0
CV	58	6	41.593	5.945	0.781	9.5	0.074	-0.824	30.2	30.2	32.1	34.24	36.3
HP_10_mN_SC	44	20	0	0	0	0			0	0	0	0	
HP_10_mN_PF	44	20	0	0	0	0			0	0	0	0	
HP_20_mN_SC	44	20	0.5	1.11	0.167	0	2.617	7.124	0	0	0	0	
HP_20_mN_PF	44	20	3.168	7.666	1.156	0	2.354	4.14	0	0	0	0	
HP_50_mN_SC	44	20	1.659	3.497	0.527	2	2.859	8.027	0	0	0	0	
HP_50_mN_PF	44	20	8.664	16.508	2.489	11.05	1.961	2.908	0	0	0	0	
HP_100_mN_SC	44	20	3.045	5.706	0.86	3	2.325	4.865	0	0	0	0	
HP_100_mN_PF	44	20	17.816	27.732	4.181	30.6	1.563	1.64	0	0	0	0	

Fig4. Summary of the data

1.4. Show Current Data

This option displays the current data as a data table.

1.5. Select Variables

Here, the user can find and select any of the variables from the data set. Based on previous data diagnostics and summary, one might want to exclude single variables from further analysis, preview only a subset of the variables, or summarize and export only a subset of the variables.

1.6. Remove Selected Variables

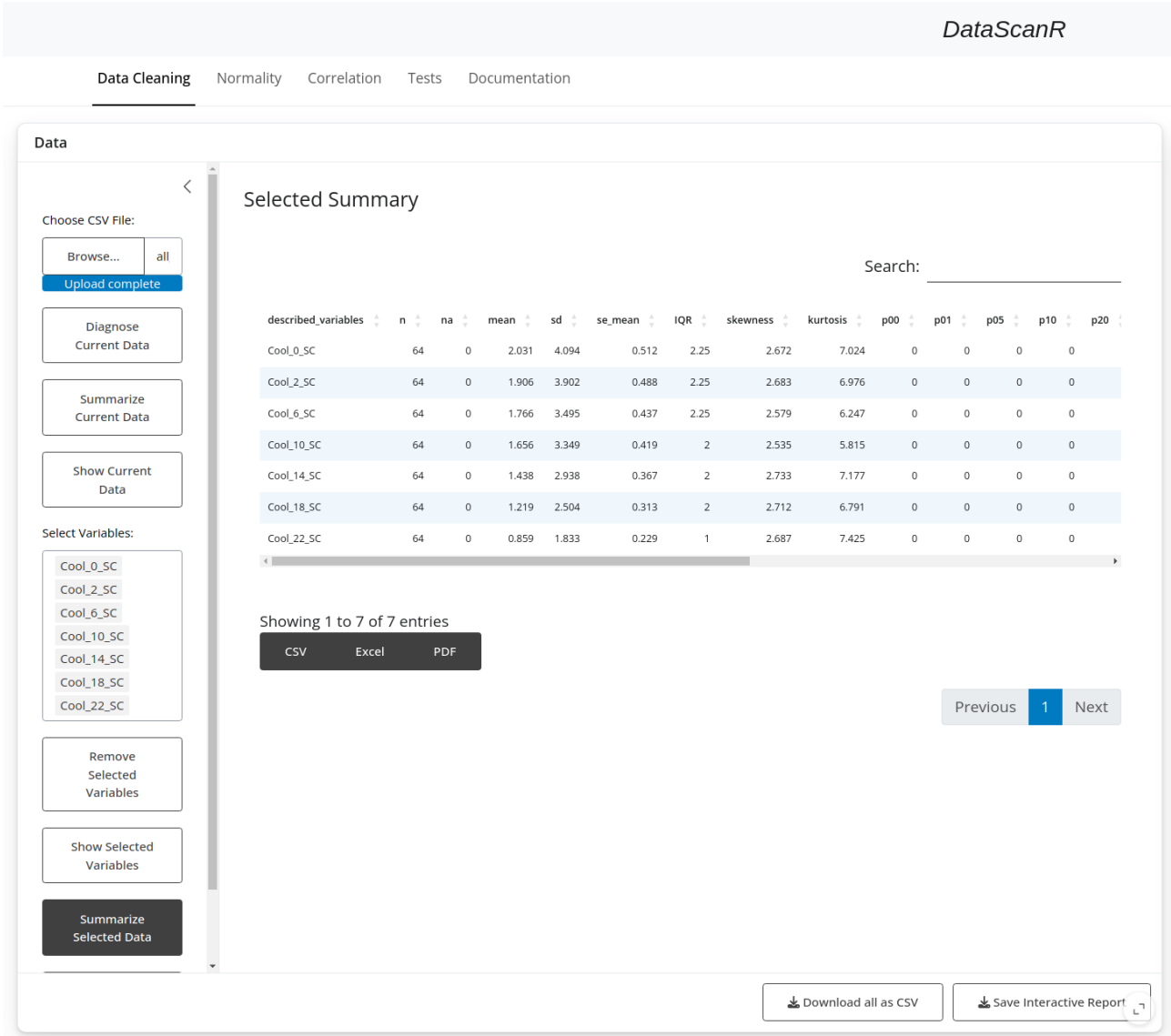
Removes from the dataset all variables that were selected under “Select Variables” section.

1.7. Show Selected Variables

Displays a data table with only the variables that were selected under “Select Variables” section.

1.8. Summarize Selected Data

Calculates descriptive statistics only for the variables that were selected under “Select Variables” section (Fig5.). This allows to focus on a subset of the data. One can export the summary statistics of such a subset into csv, excel of pdf format, by clicking one of the buttons under the displayed table.



described_variables	n	na	mean	sd	se_mean	IQR	skewness	kurtosis	p00	p01	p05	p10	p20
Cool_0_SC	64	0	2.031	4.094	0.512	2.25	2.672	7.024	0	0	0	0	
Cool_2_SC	64	0	1.906	3.902	0.488	2.25	2.683	6.976	0	0	0	0	
Cool_6_SC	64	0	1.766	3.495	0.437	2.25	2.579	6.247	0	0	0	0	
Cool_10_SC	64	0	1.656	3.349	0.419	2	2.535	5.815	0	0	0	0	
Cool_14_SC	64	0	1.438	2.938	0.367	2	2.733	7.177	0	0	0	0	
Cool_18_SC	64	0	1.219	2.504	0.313	2	2.712	6.791	0	0	0	0	
Cool_22_SC	64	0	0.859	1.833	0.229	1	2.687	7.425	0	0	0	0	

Fig5. Summary of selected data.

1.9. Restore Original Data

At any time, after removing some variables from the working dataset, the user can “go back” to working on all original data. Note that the original data file is never modified by the application!

1.10. Download all as csv

This button allows to save all pages with all current variables to a new csv file, while buttons visible directly under the table export only the currently displayed data table.

1.11. Save interactive Report

When the data table is displayed, it will generate html report with data diagnostics. That allows to store and share the information about the data quality without the need of using DataSanR. The report shows the general summary of data types, missing, data, outliers and possible duplicates (Fig6.).

Data Diagnosis Report

Overview | Missing Values | Unique Values | Outliers | Samples

Warnings

Checks	Judgements	Removes
56	74	14

Warnings	Types	Recommends
HP_800_mN_SC has 62 (96.9%) missing values	missing	judgement
HP_800_mN_PF has 62 (96.9%) missing values	missing	judgement
HP_600_mN_SC has 52 (81.2%) missing values	missing	judgement
HP_600_mN_PF has 52 (81.2%) missing values	missing	judgement
RoughBrush_SC has 46 (71.9%) missing values	missing	judgement
RoughBrush_PF has 46 (71.9%) missing values	missing	judgement
Vibration128Hz_SC has 46 (71.9%) missing values	missing	judgement
Vibration128Hz_PF has 46 (71.9%) missing values	missing	judgement
HairMovement_SC has 20 (31.2%) missing values	missing	judgement
HairMovement_PF has 20 (31.2%) missing values	missing	judgement

1–10 of 144 rows Previous 1 2 3 4 5 ... 15 Next

Variables

Diagnostic overview of individual variables

Variables	Types	Missing	Cardinality	Zero	Negative	Outlier
UnitType	character	✓	✓	-	-	-
Condition	character	✓	✓	-	-	-
SoftBrush_SC	integer	✓	✓	⚠	✓	⚠
SoftBrush_PF	numeric	✓	✓	⚠	✓	⚠

Fig6. Example of interactive report.

1.12. Plot

In this section, one can look at a graphical view of missing data for different variables (Fig7.). It allows to zoom in on the groups of variables with missing data and make more informed decisions about excluding the data or deciding on the acceptable threshold of missing data.
Under “Allow Max % Of Missing Data”, use the slider to select the acceptable threshold. Then, click on “Refresh Data” button to remove all variables that exceed the selected threshold. Note that, if you change your mind about the threshold, you can move the slider again and click “Refresh Data” button to apply new settings or click on “Restore Original Data” to restore all variables.

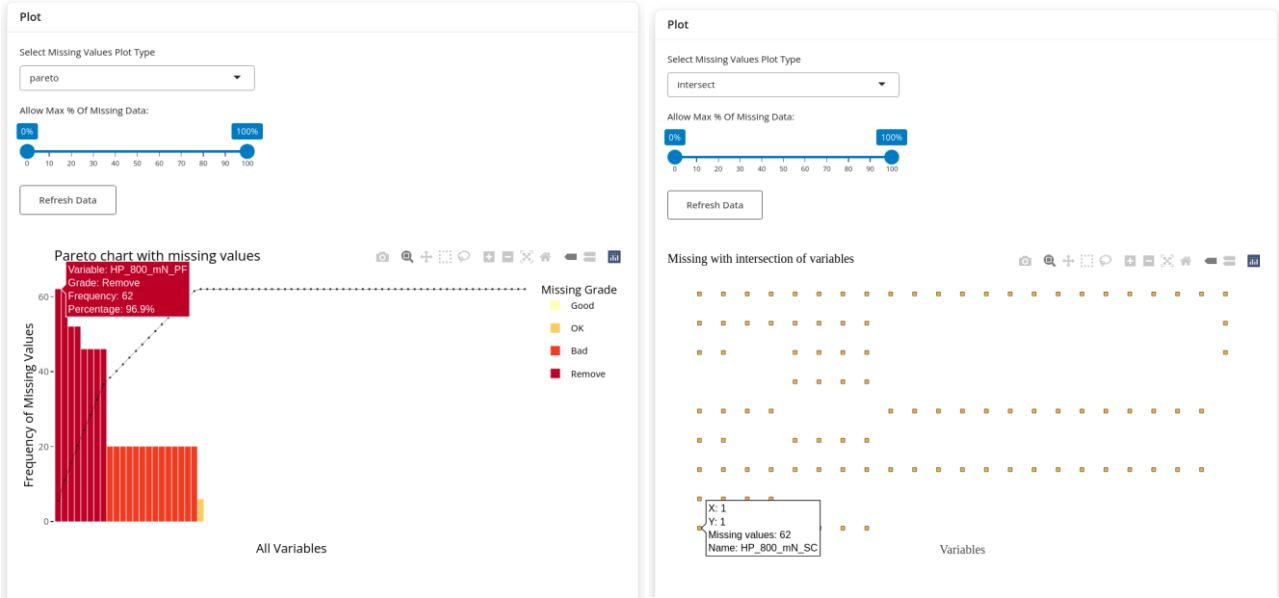


Fig7. The example of visualizing missing data in the dataset.

The above interactive, plotly visualizations are based on missing functions from dlookr package

2.Normality

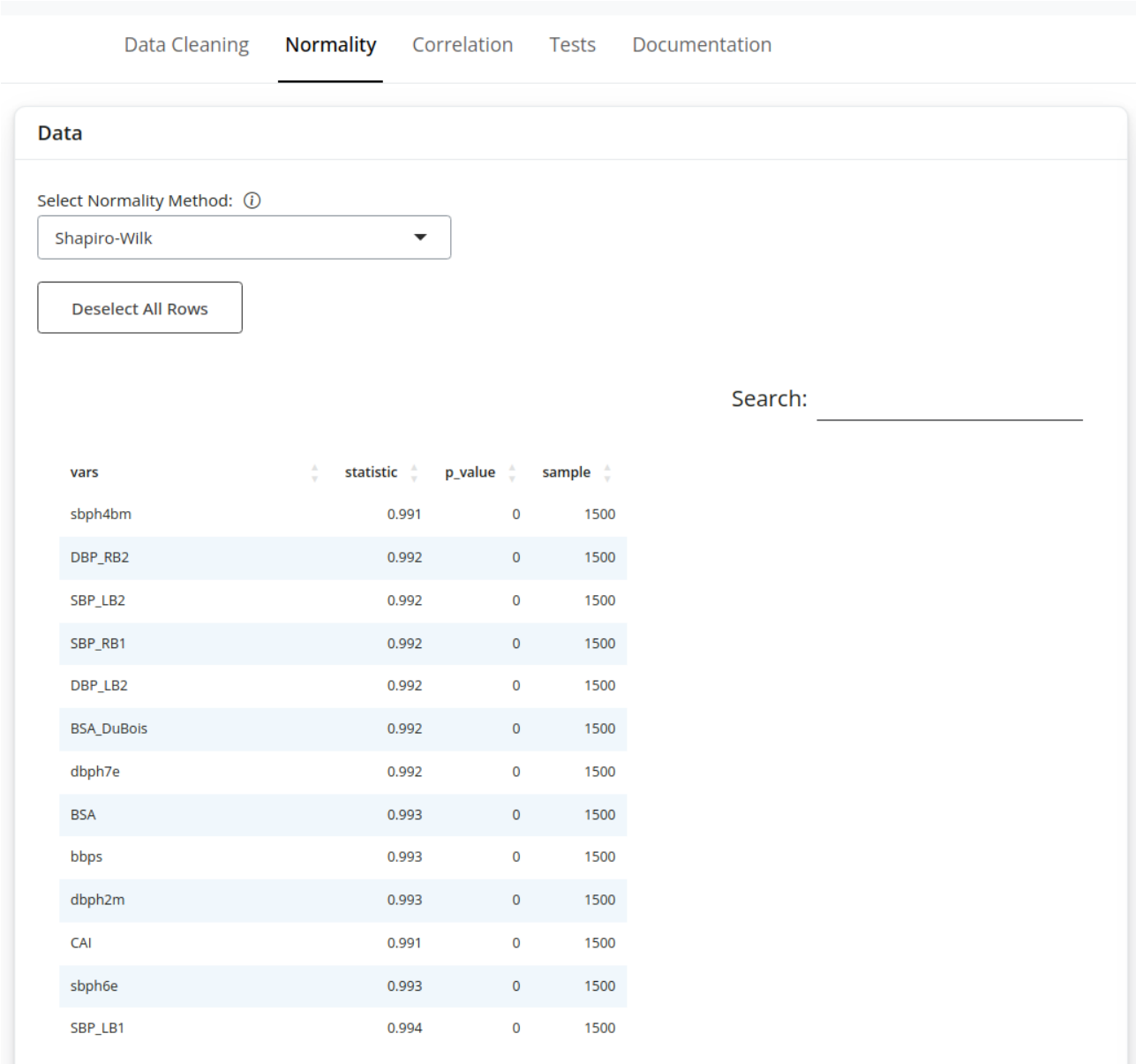
Normality section of DataScanR application allows to check whether the data is normally distributed.

2.1. Data

The application will automatically apply one of the two methods based on the following criteria:

- Shapiro-Wilk: for dataset < 2000 observations.
- Kolmogorov-Smirnov: for dataset > 2000 observations.

In both cases we can interpret the result in the following manner:
p-value < 0.05 and statistic close to 1 tell us that we can
reject the null hypothesis of the normally distributed data.
The users can choose the normality method themselves, by changing the method in the drop-down menu for “Select Normality Method”.



vars	statistic	p_value	sample
sbph4bm	0.991	0	1500
DBP_RB2	0.992	0	1500
SBP_LB2	0.992	0	1500
SBP_RB1	0.992	0	1500
DBP_LB2	0.992	0	1500
BSA_DuBois	0.992	0	1500
dbph7e	0.992	0	1500
BSA	0.993	0	1500
bbps	0.993	0	1500
dbph2m	0.993	0	1500
CAI	0.991	0	1500
sbph6e	0.993	0	1500
SBP_LB1	0.994	0	1500

Fig8. Normality results table.

2.2. Plot

In this part of “Normality” section, one can visualize the distributions of up to 6 variables at a time.
Find and select the variables in the table on the left, and the plots will be automatically updated on the right, in the “Plot” section. The user can choose between box, violin, histogram, box_distribution and violin_box. All displayed plots are available to download as the most popular file formats. Here are some of the examples of distribution visualizations (Fig9.):

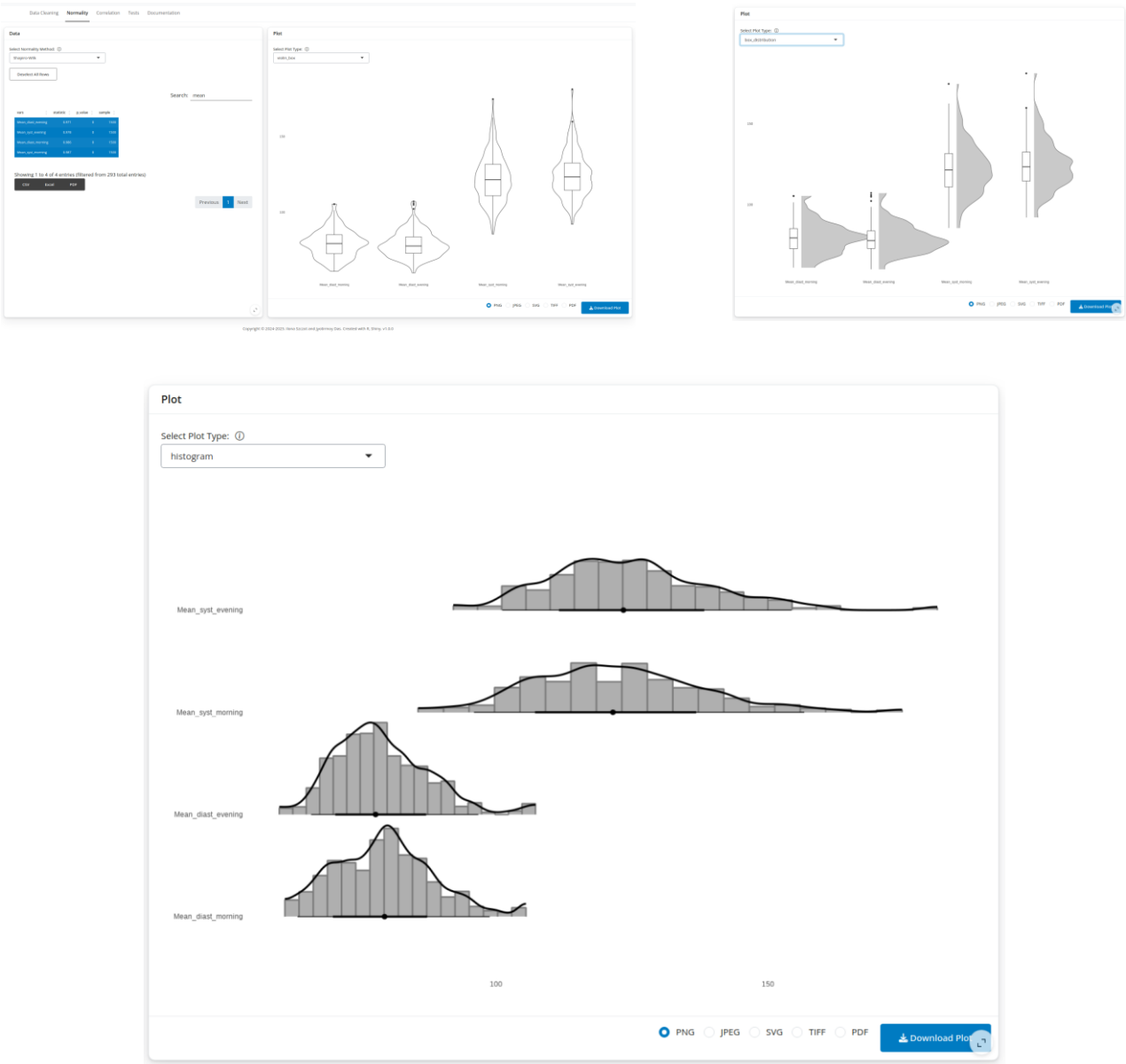


Fig9. The examples of distribution visualizations.

“normality_diagnosis” (Fig10.) is a special plot type that uses plot_normality function from dlookr package to visualize:

- Histogram of original data
- Q-Q plot of original data
- histogram of log transformed data
- Histogram of square root transformed data

This plot type allows to only visualize one variable at a time.

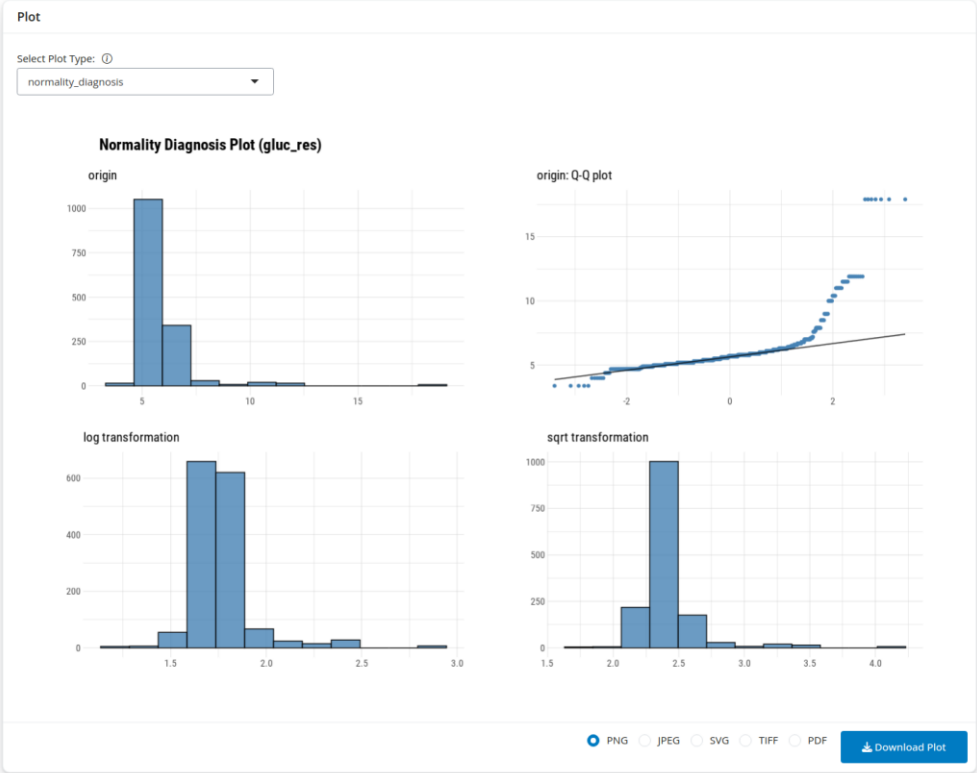


Fig10. Normality diagnosis plot.

3. Correlation

Correlation section of this application allows to check the correlations between different numerical variables in the dataset.

3.1. Correlation Results

In this part, one can select the numerical variables of interest, the method to use for calculating the correlations (pearson, kendall, spearman), the hypothesis to test (less, greater, two-sided), and the level of confidence. After clicking on “Calculate Correlations” button, the results are displayed in a form of the table, which can be saved to a file.

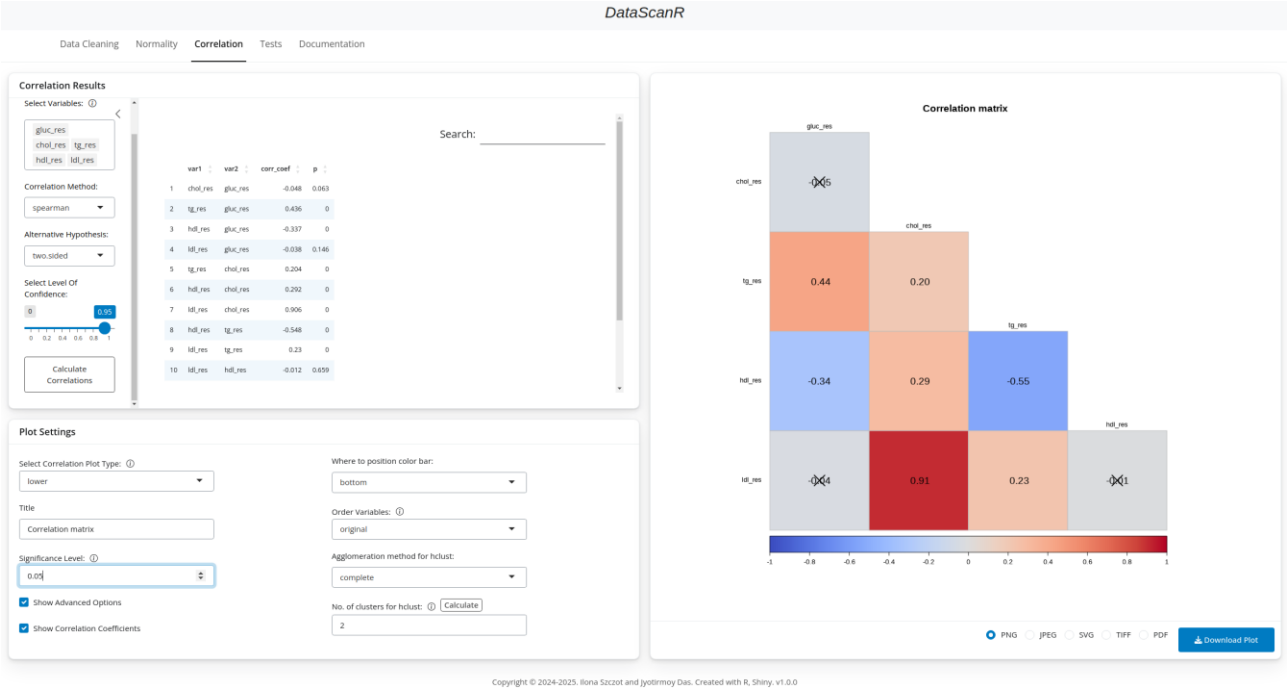


Fig11. The example of Correlation section.

3.2. Plot Settings

The test results are visualized as a correlogram. The user can choose one of the following layouts:

- “full”: display full correlation matrix. This plot type will show clustering squares only if hclust is selected in Order Variables drop down menu (Advanced Options).
- “upper”: display upper triangular of the correlation matrix
- “lower”: display lower triangular of the correlation matrix
- “confidence_interval”: display confidence intervals. The plot will show confidence intervals only if they were calculated in the result table.

3.2.1. Advanced options

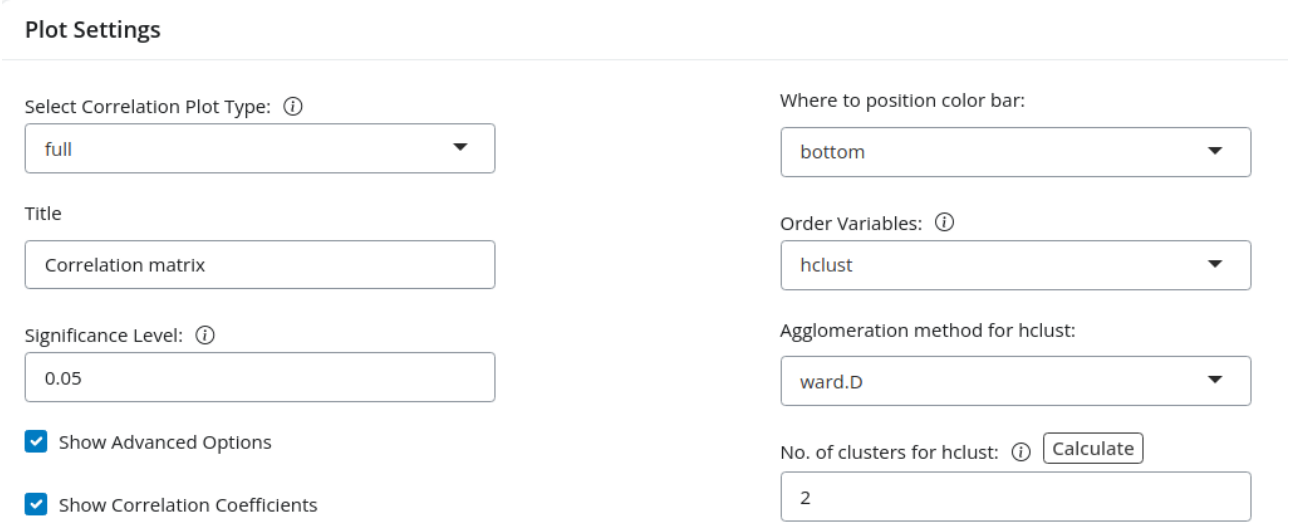


Fig12. Plot settings available for customization.

3.2.1.1. Show Correlation Coefficient

Correlation coefficients can be either shown on the plot or hidden. Use “Show Correlation Coefficient” check box to decide. Under “Significance Level” option selecting value 1, will show all correlation coefficients values on the plot. Regardless of whether they are significant or not. Set this value to your own significance level to see on the plot which correlation coefficients are not significant. Those will appear as crossed out.

3.2.1.2. Order Variables

Determines how the variables should be grouped on the plot:

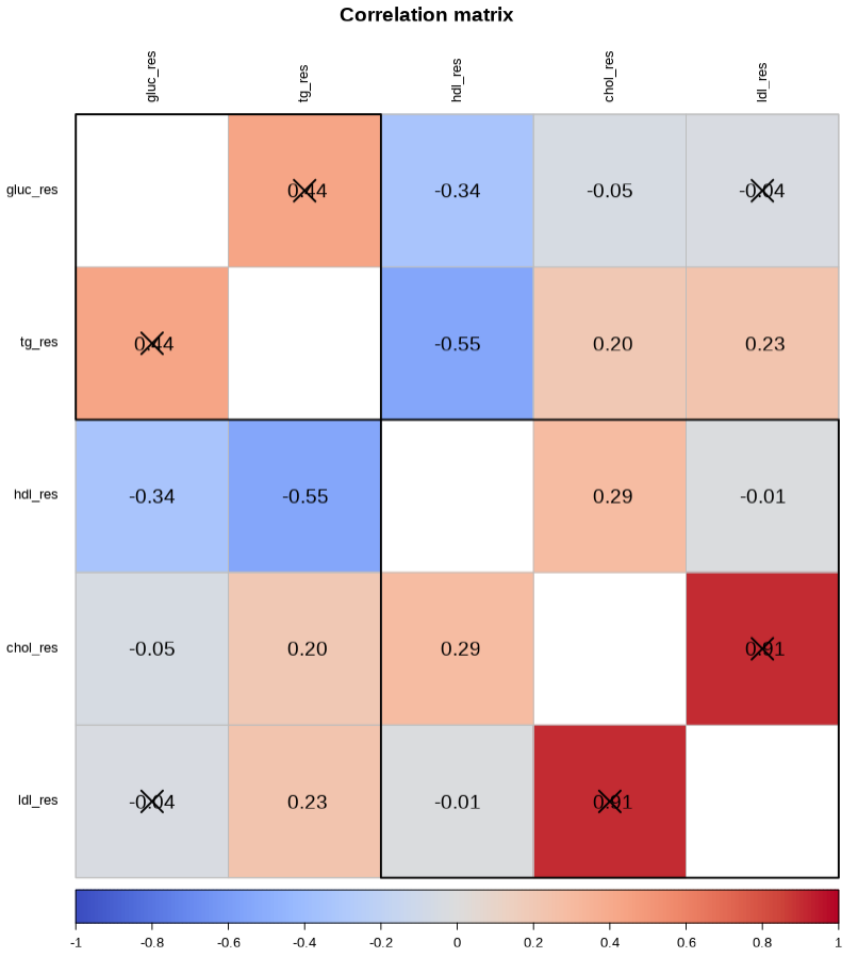
- original: the original order
- AOE: the angular order of the eigenvectors
- FPC: the first principal component order
- hclust: the hierarchical clustering order. This will show clustering squares on a “full” plot type
- alphabet: alphabetical order

3.2.1.3. Agglomeration method for hclust

The agglomeration method to be used when “Order” is hclust (hierarchical clustering order). This should be one of: 'ward.D', 'ward.D2', 'single', 'complete', 'average', 'mcquitty', 'median' or 'centroid'.

3.2.1.4. No. of clusters for hclust

By default, it will show 2 clustering squares on a “full” plot type if hclust order is selected (Fig13.). The used can decide themselves on the number of clusters or click on “Calculate” button to calculate the optimal number of clusters.
The calculation is using NbClust function and selected agglomeration method to calculate the optimal number of clusters. If there are any missing values, it will impute those missing values. Therefore, if there are many missing values, the number of clusters might vary between the calculation runs.



PNG JPEG SVG TIFF PDF Download Plot

Fig13. Correlogram with visible cluster squares.

4. Tests

Here, the user has a few basic statistical tests to choose from. Depending on the data distribution, the user can choose either “Parametric” tab with normal distribution tests or “Non-parametric” tab with other tests.

4.1. Parametric (Normal-distribution tests)

4.1.1. One-sample t-test

Compares the sample mean to a known or hypothesized mean.
The user can select one or more variables to test, from the “Select Variables” drop down menu.
One can choose what alternative hypothesis will be tested: two-sided, less or greater, what is the theoretical mean to be tested (mu), as well as the level of confidence.
The results will be visible in the form of the table under “Results Table” tab (Fig15.) or as a plot under “Plot” tab (Fig14.).

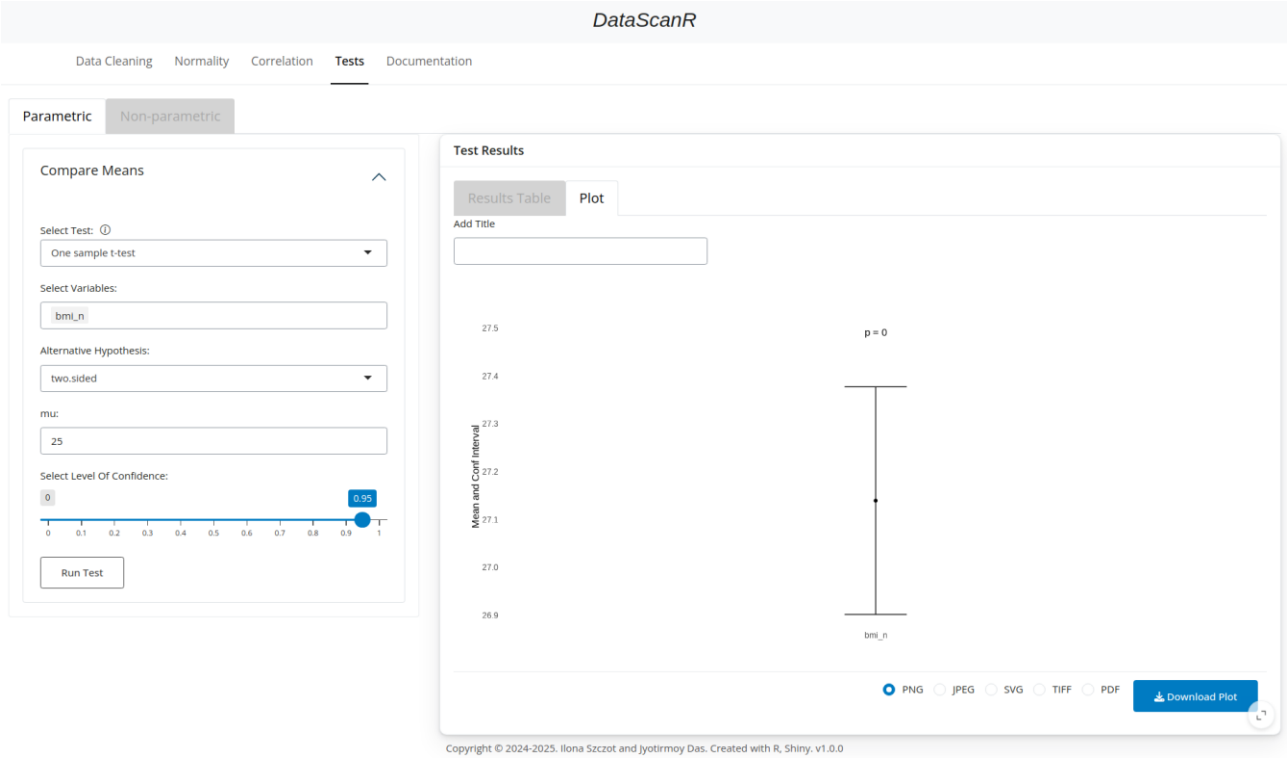


Fig14. One-sample t-test results.

vars	null_val	estimate	alternative	p_value	lowCI	uppCI	conf_level	statistic	parameter	samples
bmi_n	25	27.14	two.sided	0	26.902	27.378	0.95	17.646	1485	1500

Fig15. One-sample t-test results table.

4.1.2. Independent two-sample t-test

Compares the means of two independent groups.

- If the groups are defined in a variable:

Select one or more variables from the “Select Variables” drop down menu. Then check the check box “Run By Group”. When “Select Group Column” option appears, select one variable with the groups that you would like to compare.
One can choose what alternative hypothesis will be tested: two-sided, less or greater, what is hypothesized difference in means to be tested (mu), as well as the level of confidence.
The results will be visible in the form of the table under “Results Table” tab (Fig17.) or as a plot under “Plot” tab (Fig16.).

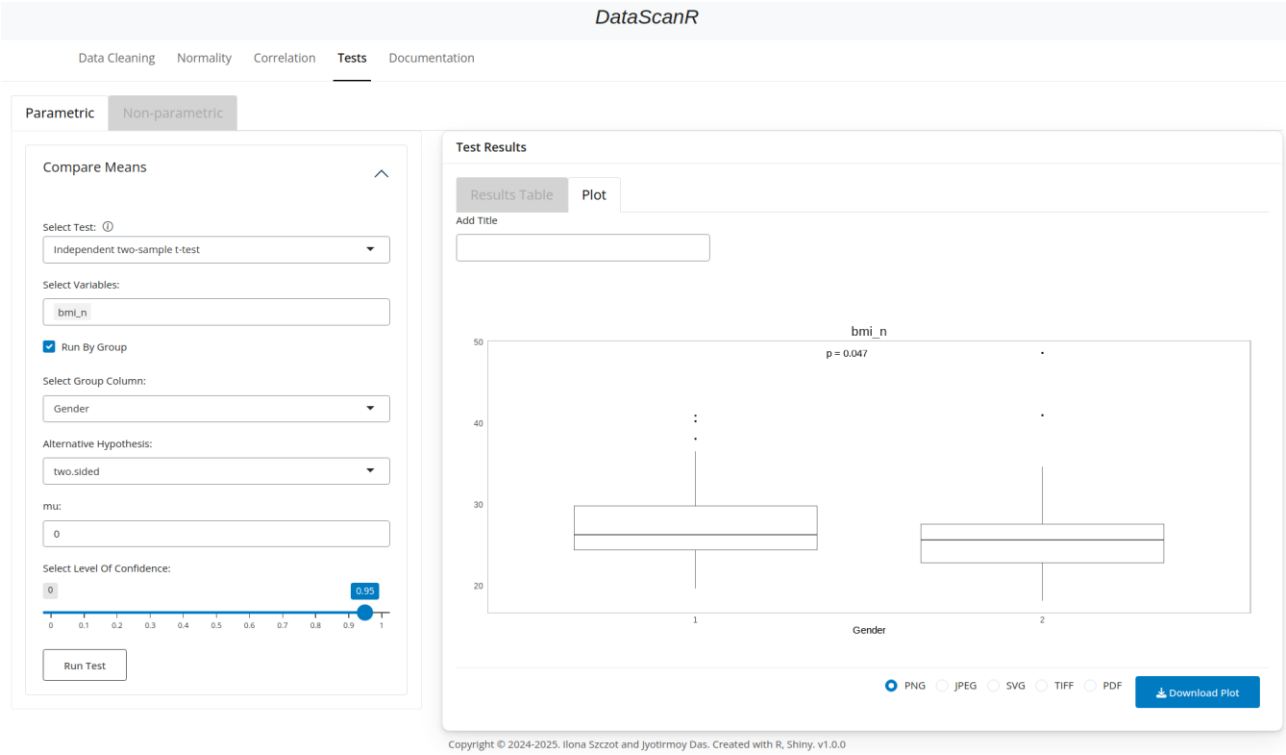


Fig16. Independent two-sample t-test results.

Test Results

Results Table | Plot

Independent two-sample t-test

Group: Gender

Search:

vars	null_val	estimate_1	estimate_2	alternative	p_value	lowCI	uppCI	conf_level	statistic	parameter	samples
bmi_n	0	27.299	26.66	two.sided	0.047	0.008	1.268	0.95	1.99	516.388	1131

Fig17. Independent two-sample t-test results table.

- If the groups are in two, separate variables:

Select two variables from the “Select Variables” drop down menu. Leave the check box: “Run By Group” empty. Then the test will be performed between the two selected variables.
One can choose what alternative hypothesis will be tested: two-sided, less or greater, what is hypothesized difference in means to be tested (mu), as well as the level of confidence.
The results will be visible in the form of the table under “Results Table” tab or as a plot under “Plot” tab.

4.1.3. Paired t-test

Compares means from the same group at two different times or under two different conditions.

- If the times/conditions are defined in a variable:

Select one or more variables from the “Select Variables” drop down menu. Then check the check box “Run By Group”. When “Select Group Column” option appears, select one variable with the times/conditions that you would like to compare.
One can choose what alternative hypothesis will be tested: two-sided, less or greater, as well as the level of confidence.
The results will be visible in the form of the table under “Results Table” tab (Fig19.) or as a plot under “Plot” tab (Fig18.).

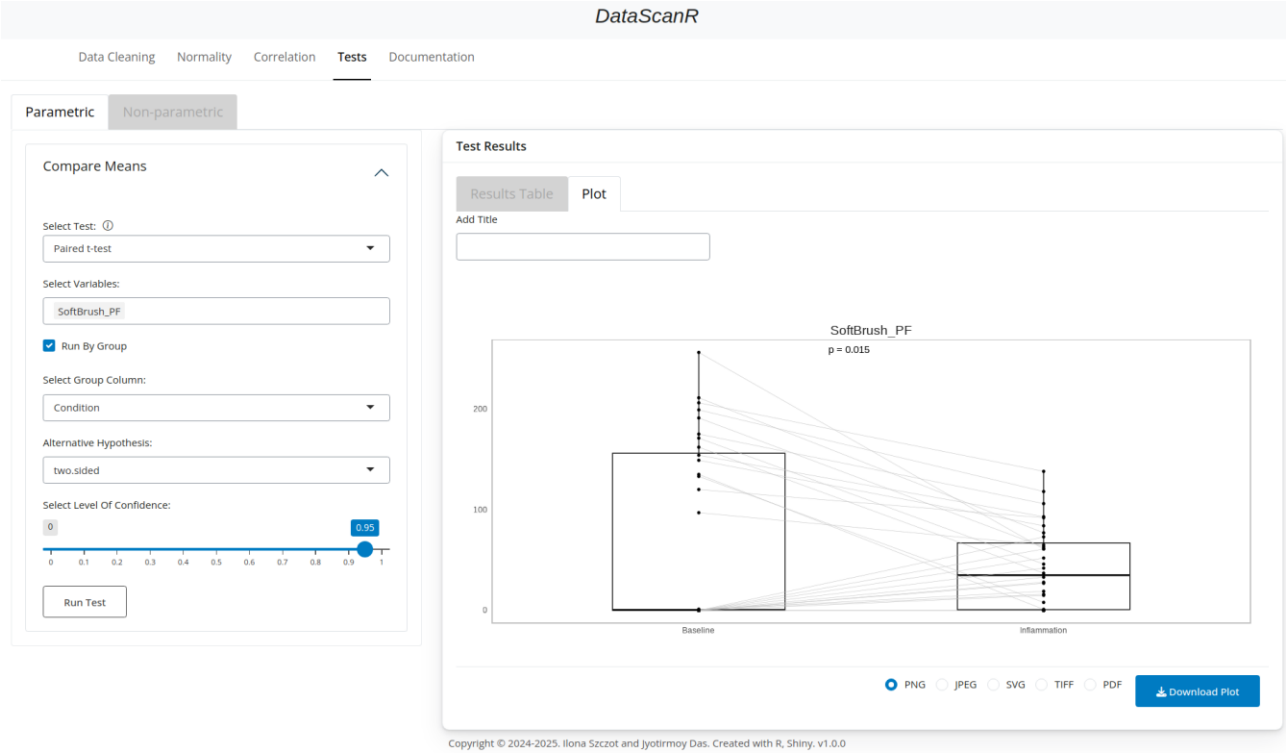


Fig18. Paired t-test results.

Results Table | Plot

Paired t-test

Group: Condition

vars	null_val	mean1	mean2	mean_difference	alternative	p_value	lowCI	uppCI	conf_level	statistic	parameter	samples
SoftBrush_PF_Baseline_vs_Inflammation	0	73.787	42.347	31.441	two.sided	0.015	6.519	56.363	0.95	2.573	31	32

Fig19. Paired t-test results table.

- If the times/conditions are in two, separate variables:

Select two variables from the “Select Variables” drop down menu. Leave the check box: “Run By Group” empty. Then the test will be performed between the two selected variables.
One can choose what alternative hypothesis will be tested: two-sided, less or greater, as well as the level of confidence.
The results will be visible in the form of the table under “Results Table” tab or as a plot under “Plot” tab.

4.2. Non-parametric

4.2.1. Wilcoxon rank-sum test

Compares the distributions of two independent groups.

- If the groups are defined in a variable:

Select one or more variables from the “Select Variables” drop down menu. Then check the check box “Run By Group”. When “Select Group Column” option appears, select one variable with the groups that you would like to compare.
One can choose what alternative hypothesis will be tested: two-sided, less or greater, what is hypothesized difference in medians to be tested (mu), as well as the level of confidence.
The results will be visible in the form of the table under “Results Table” tab (Fig21.) or as a plot under “Plot” tab (Fig20.).

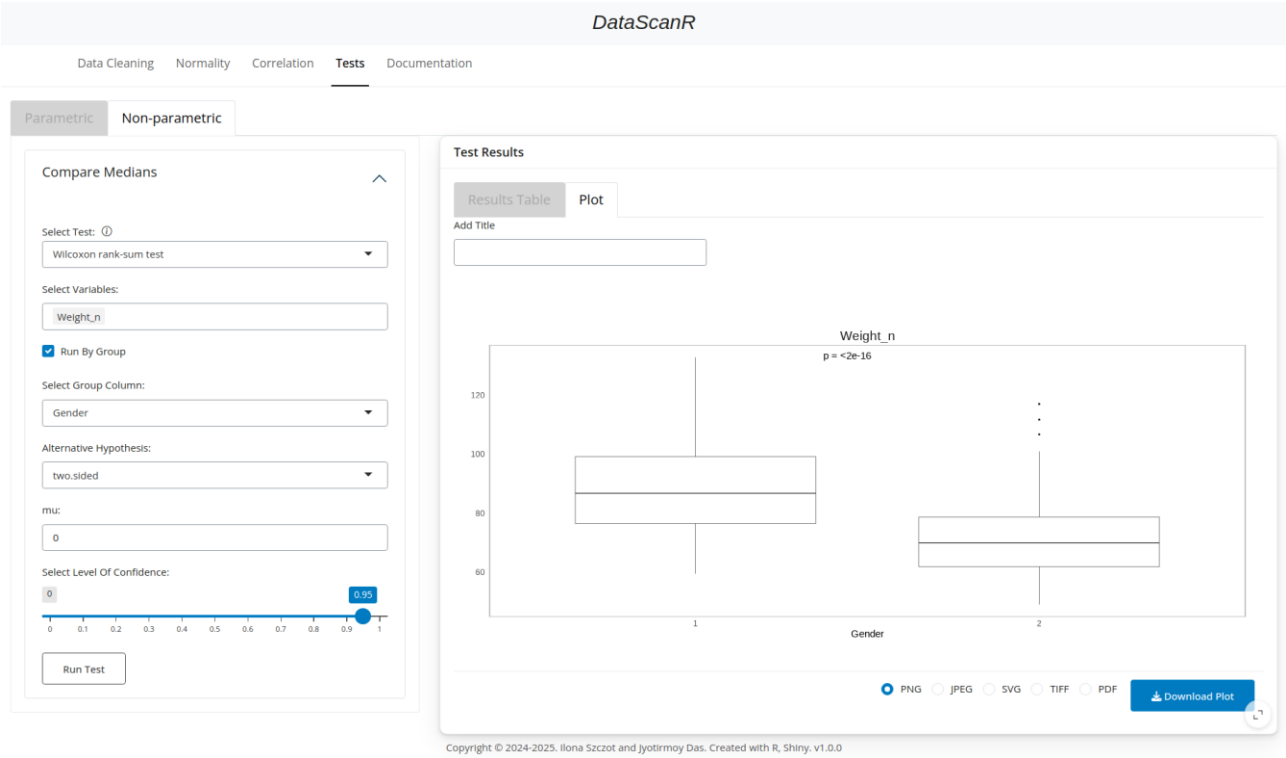
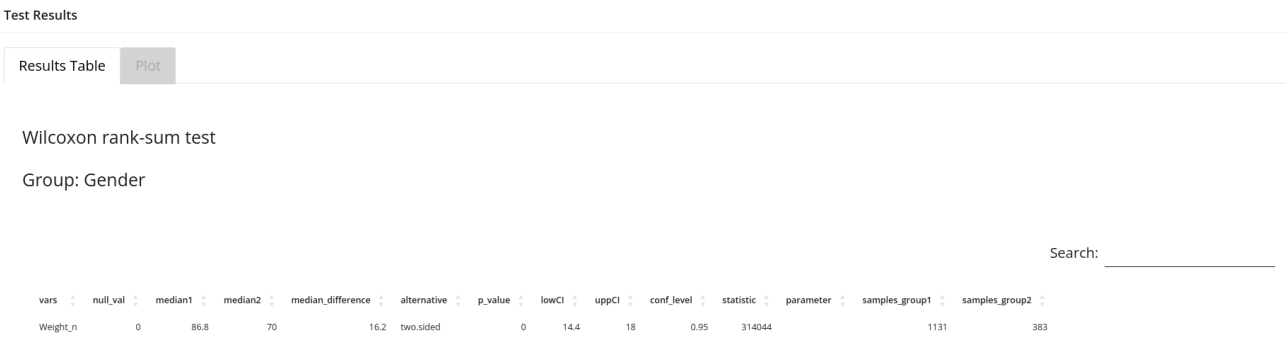


Fig20. Wilcoxon rank-sum test results.



vars	null_val	median1	median2	median_difference	alternative	p_value	lowCI	uppCI	conf_level	statistic	parameter	samples_group1	samples_group2
Weight_n	0	86.8	70	16.2	two.sided	0	14.4	18	0.95	314044		1131	383

Fig21. Wilcoxon rank-sum test results table.

- If the groups are in two, separate variables:

Select two variables from the “Select Variables” drop down menu. Leave the check box: “Run By Group” empty. Then the test will be performed between the two selected variables.
One can choose what alternative hypothesis will be tested: two-sided, less or greater, what is hypothesized difference in mediand to be tested (mu), as well as the level of confidence.
The results will be visible in the form of the table under “Results Table” tab or as a plot under “Plot” tab.

4.2.2. Wilcoxon signed-rank test

Compares paired data (two related samples or repeated measures on a single sample).

- If the conditions are defined in a variable:

Select one or more variables from the “Select Variables” drop down menu. Then check the check box “Run By Group”. When “Select Group Column” option appears, select one variable with the conditions that you would like to compare.
One can choose what alternative hypothesis will be tested: two-sided, less or greater, as well as the level of confidence.
The results will be visible in the form of the table under “Results Table” tab (Fig23.) or as a plot under “Plot” tab (Fig22.).

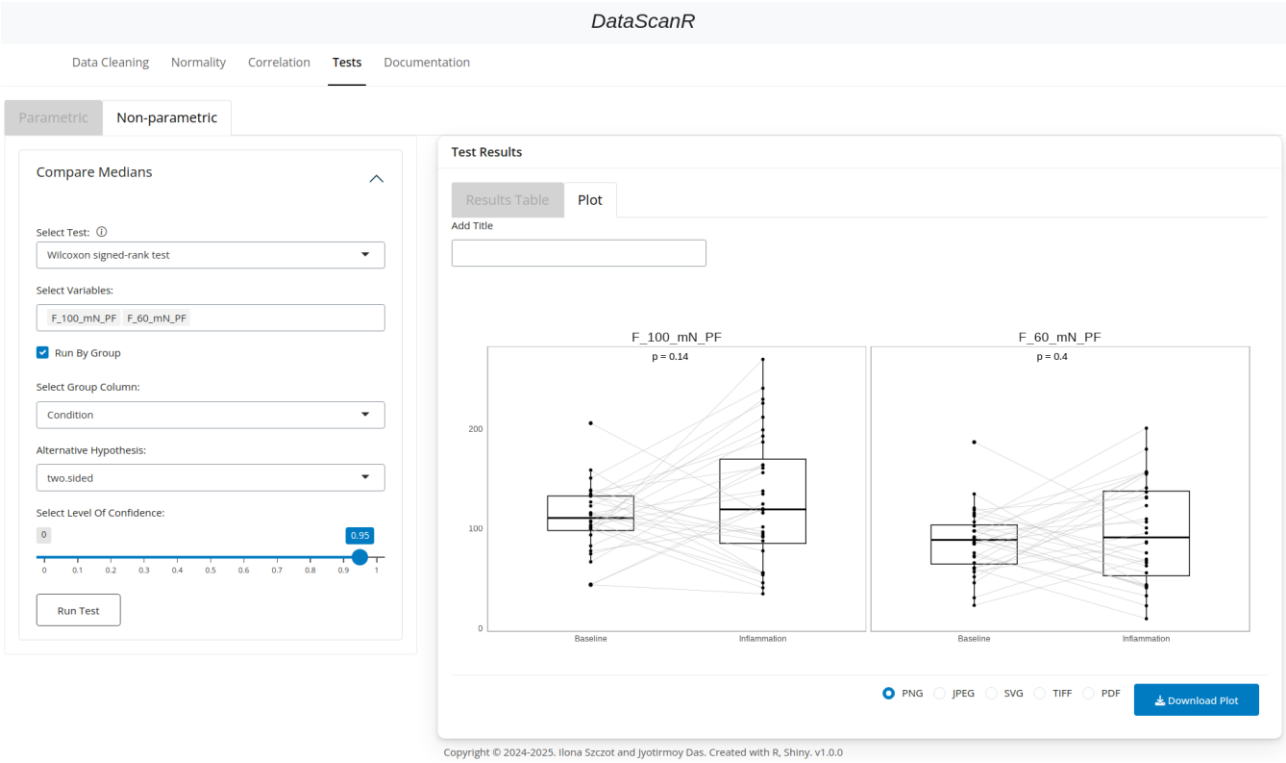


Fig22. Wilcoxon signed-rank test results.

Test Results

Results Table | Plot

Wilcoxon signed-rank test

Group: Condition

Search:

vars	null_val	median1	median2	median_difference	alternative	p_value	lowCI	uppCI	conf_level	statistic	parameter	samples_group1	samples_group2
F_100_mN_PF_Baseline_vs_Inflammation	0	111	119.6	-19.394	two.sided	0.14	-48.55	6.5	0.95	185		32	32
F_60_mN_PF_Baseline_vs_Inflammation	0	89	91.5	-6.895	two.sided	0.395	-29.65	12	0.95	218.5		32	32

Fig23. Wilcoxon signed-rank test results table.

- If the conditions are in two separate variables:

Select two variables from the “Select Variables” drop down menu. Leave the check box: “Run By Group” empty. Then the test will be performed between the two selected variables.
One can choose what alternative hypothesis will be tested: two-sided, less or greater, as well as the level of confidence.
The results will be visible in the form of the table under “Results Table” tab or as a plot under “Plot” tab.

4.2.3. Kruskal-Wallis test

Non-parametric alternative to one-way ANOVA, compares more than two independent groups.

- If the groups are defined in a variable:

Select one or more variables from the “Select Variables” drop down menu. Then check the check box “Run By Group”. When “Select Group Column” option appears, select one variable with the groups that you would like to compare.
The results will be visible in the form of the table under “Results Table” tab (Fig25.) or as a plot under “Plot” tab (Fig24.).

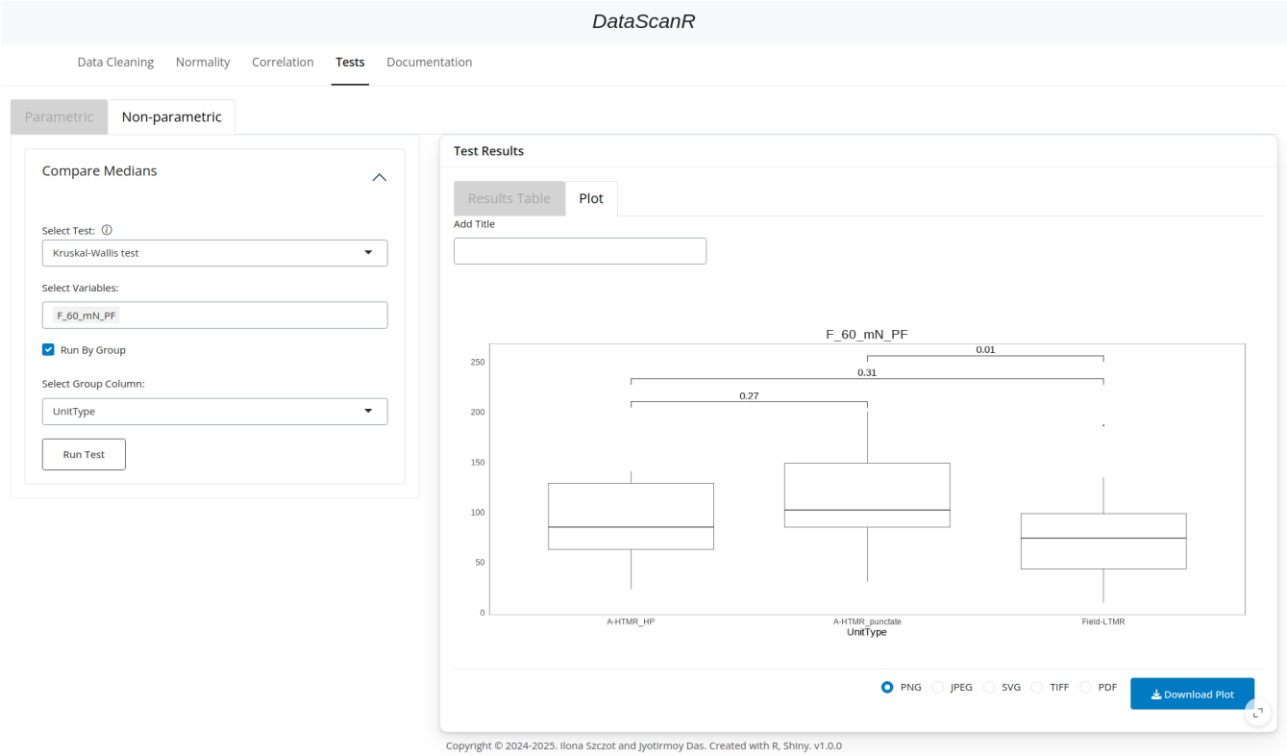
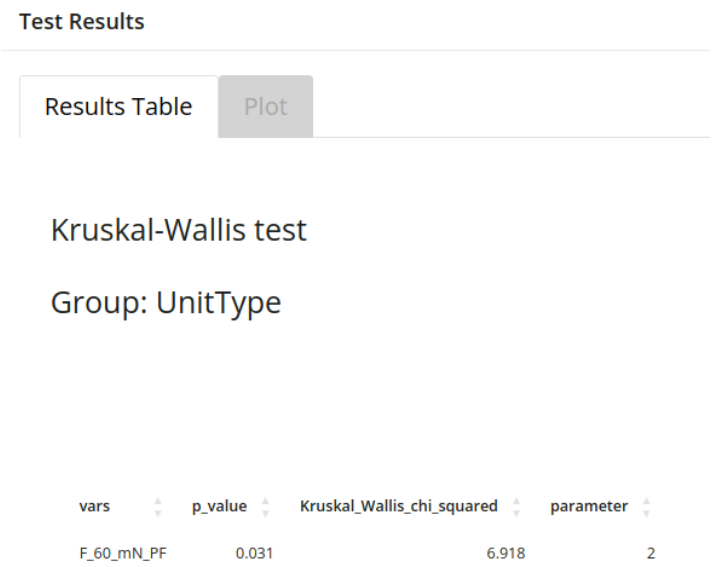


Fig24. Kruskal-Wallis test results.



vars	p_value	Kruskal_Wallis_chi_squared	parameter
F_60_mN_PF	0.031	6.918	2

Fig25. Kruskal-Wallis test results table.

- If the groups are in three or more separate variables:

Select three or more variables from the “Select Variables” drop down menu. Leave the check box: “Run By Group” empty. Then the test will be performed between the selected variables.
The results will be visible in the form of the table under “Results Table” tab or as a plot under “Plot” tab.