

DataScanR

Quick Guide

Ilona Szczot¹, Jyotirmoy Das²

*¹Centrum för social och affektiv neurovetenskap, Linköping University,
Linköping, Sweden,*

*²Core Facility, Faculty of Medicine and Health Sciences, Linköping
University, Linköping, Sweden and Clinical Genomics Linköping, Science for
Life Laboratory, Sweden*

Table of Contents

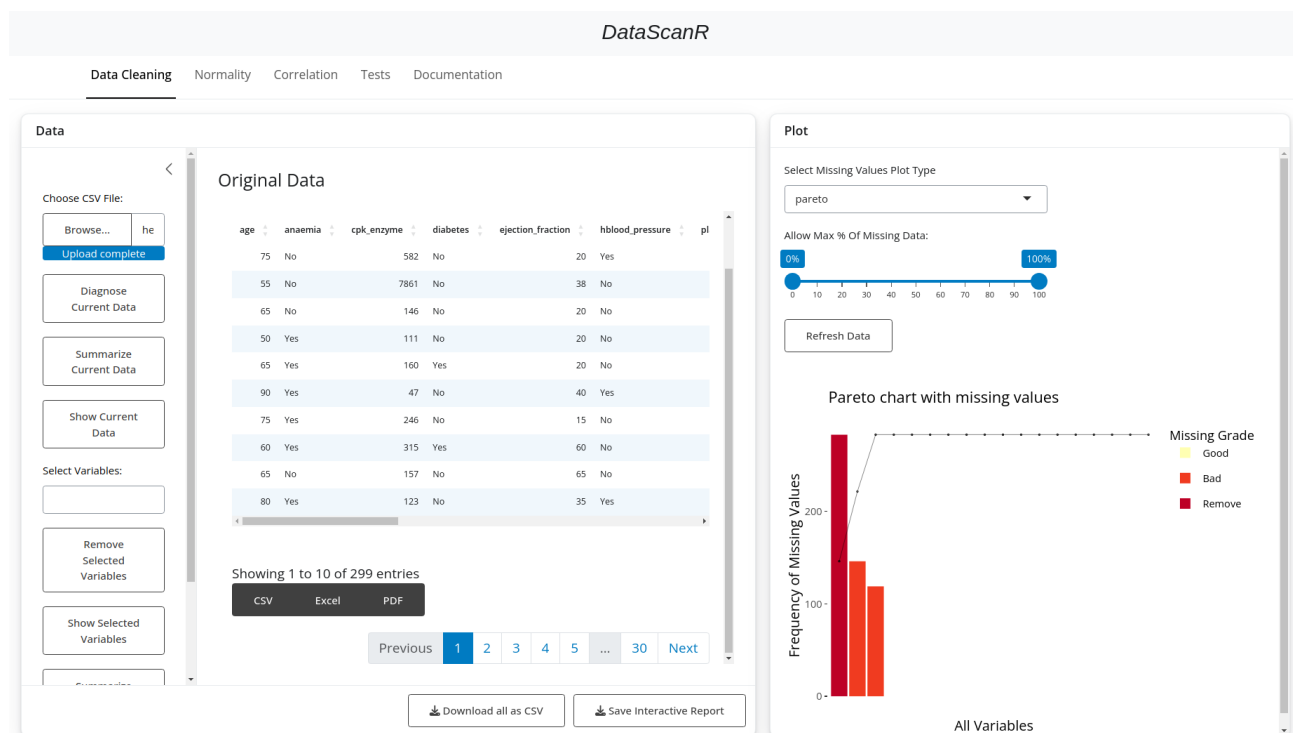
Description.....	3
1. Data Cleaning.....	3
1.1. Upload Data	3
1.2. Diagnose Current Data.....	4
1.3. Summarize Current Data	5
1.4. Show Current Data.....	6
1.5. Select Variables	6
1.6. Remove Selected Variables.....	6
1.7. Show Selected Variables	6
1.8. Summarize Selected Data	6
1.9. Restore Original Data	7
1.10. Download all as csv	7
1.11. Save interactive Report	7
1.12. Plot.....	8
2.Normality	9
2.1. Data	9
2.2. Plot.....	11
3. Correlation.....	12
3.1. Correlation Results.....	12
3.2. Plot Settings	13
3.2.1. Advanced options	14
4. Tests.....	16
4.1. Parametric (Normal-distribution tests).....	16
4.1.1. One-sample t-test.....	16
4.1.2. Independent two-sample t-test.....	17
4.1.3. Paired t-test.....	18
4.2. Non-parametric	20
4.2.1. Wilcoxon rank-sum test.....	20
4.2.2. Wilcoxon signed-rank test.....	21
4.2.3. Kruskal-Wallis test.....	23
6. Example Data	24
References	25

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.)



Copyright © 2024-2025. Ilona Szczot and Jyotirmoy Das. Created with R, Shiny. v1.0.0

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).

Choose CSV File:

Browse...

he

Upload complete

Diagnose Current Data

Summarize Current Data

Show Current Data

Select Variables:

Remove Selected Variables

Show Selected Variables

Original Data

age	anaemia	cpk_enzyme	diabetes	ejection_fraction	hblood_pressure	pl
75	No	582	No	20	Yes	
55	No	7861	No	38	No	
65	No	146	No	20	No	
50	Yes	111	No	20	No	
65	Yes	160	Yes	20	No	
90	Yes	47	No	40	Yes	
75	Yes	246	No	15	No	
60	Yes	315	Yes	60	No	
65	No	157	No	65	No	
80	Yes	123	No	35	Yes	

Showing 1 to 10 of 299 entries

a) CSV Excel PDF

b) Previous 1 2 3 4 5 ... 30 Next

c) Download all as CSV Save Interactive Report

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. [1] (Ryu, u.d.)

Choose CSV File:

Browse... he

Upload complete

Diagnose Current Data

Summarize Current Data

Show Current Data

Select Variables:

Remove Selected Variables

Show Selected Variables

Diagnostics

variables	types	missing_count	missing_percent	unique_count	unique_rate
age	integer	0	0	46	0.15
anaemia	character	0	0	2	0.00
cpk_enzyme	integer	0	0	208	0.69
diabetes	character	0	0	2	0.00
ejection_fraction	integer	0	0	17	0.05
hblood_pressure	character	0	0	2	0.00
platelets	numeric	0	0	176	0.58
creatinine	numeric	0	0	40	0.13
sodium	integer	0	0	27	0.0
sex	character	0	0	2	0.00

Showing 1 to 10 of 18 entries

CSV Excel PDF

Previous 1 2 Next

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. [1] (Ryu, u.d.)

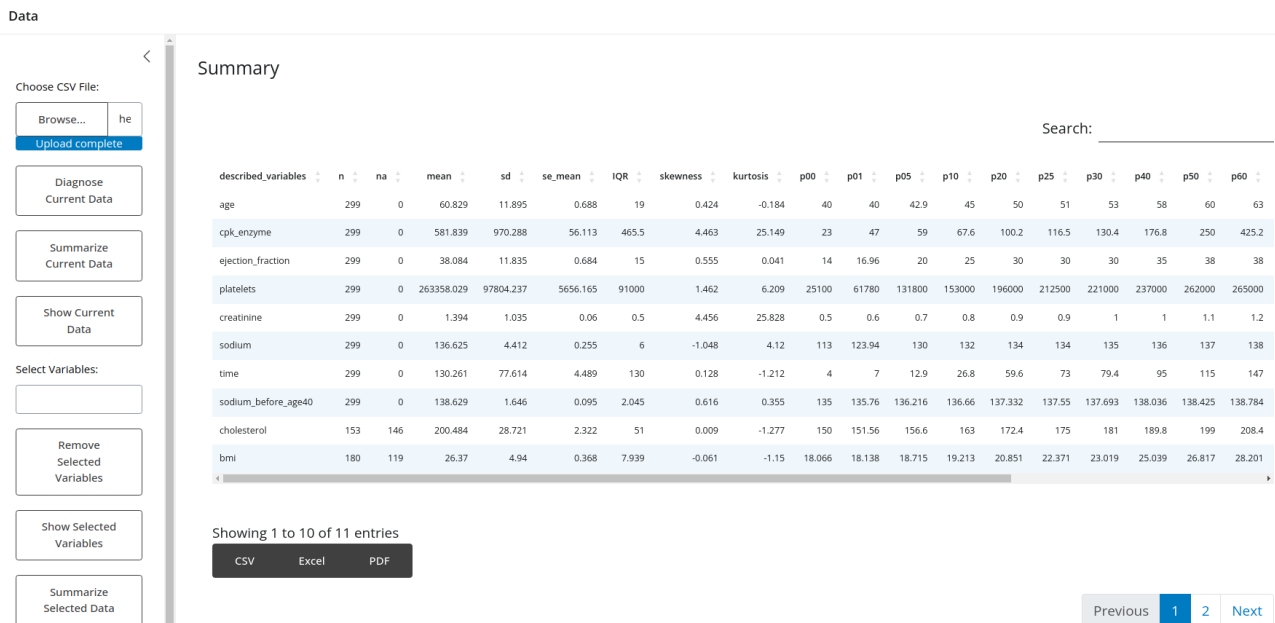


Fig4. Summary of the data

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 or pdf format, by clicking one of the buttons under the displayed table.

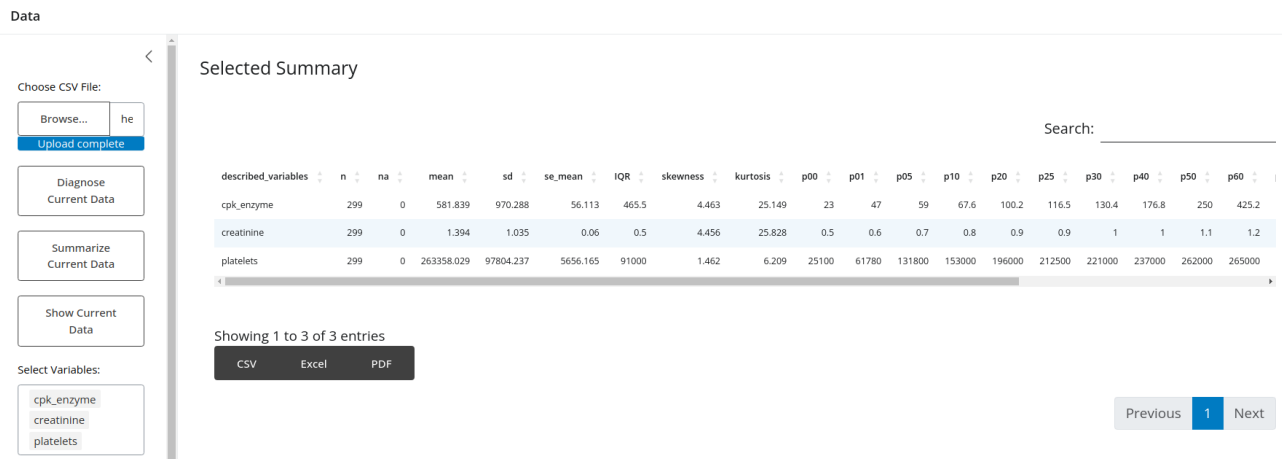


Fig5. Summary of selected data.

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.).

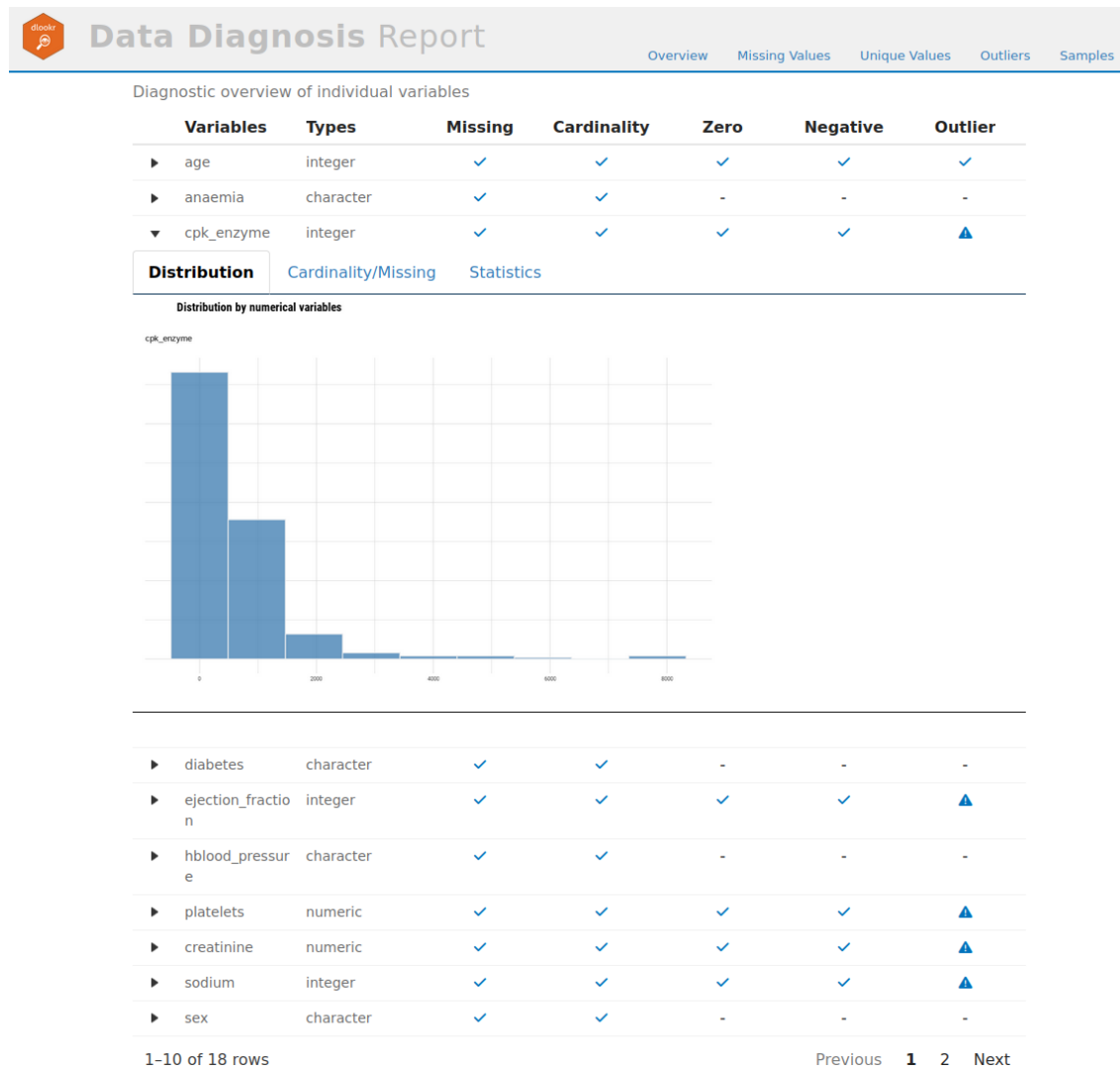


Fig6. Example of interactive report.

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.

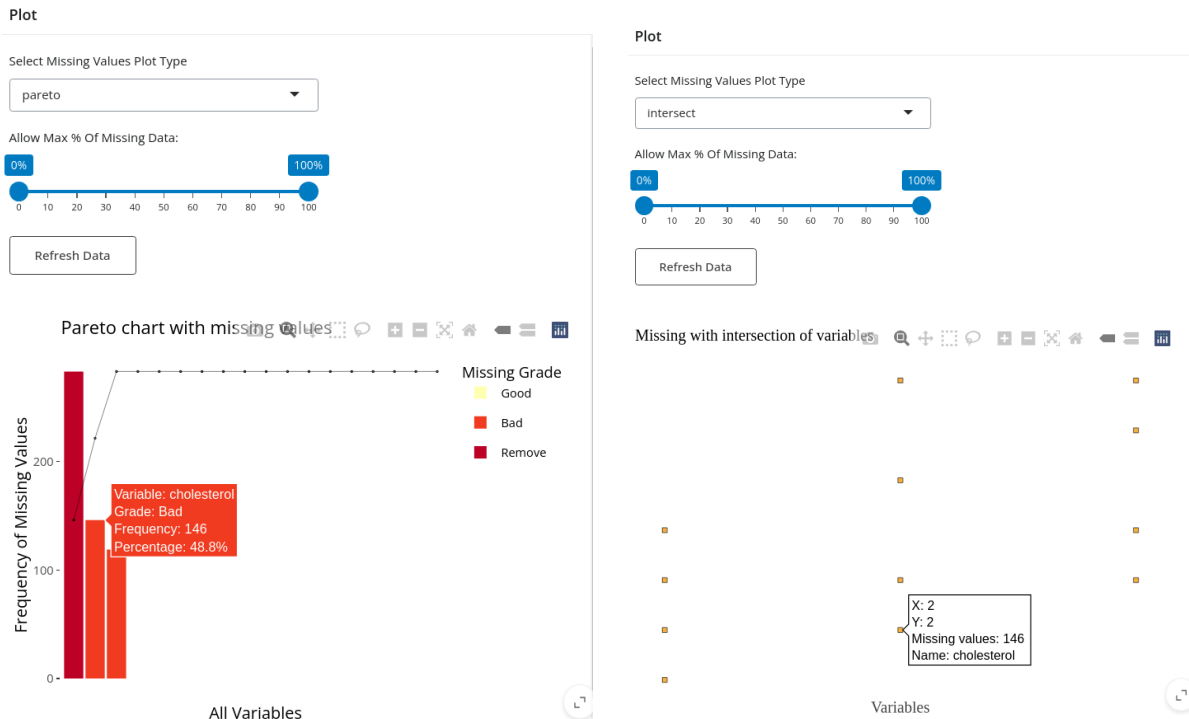


Fig7. The example of visualizing missing data in the dataset.

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".

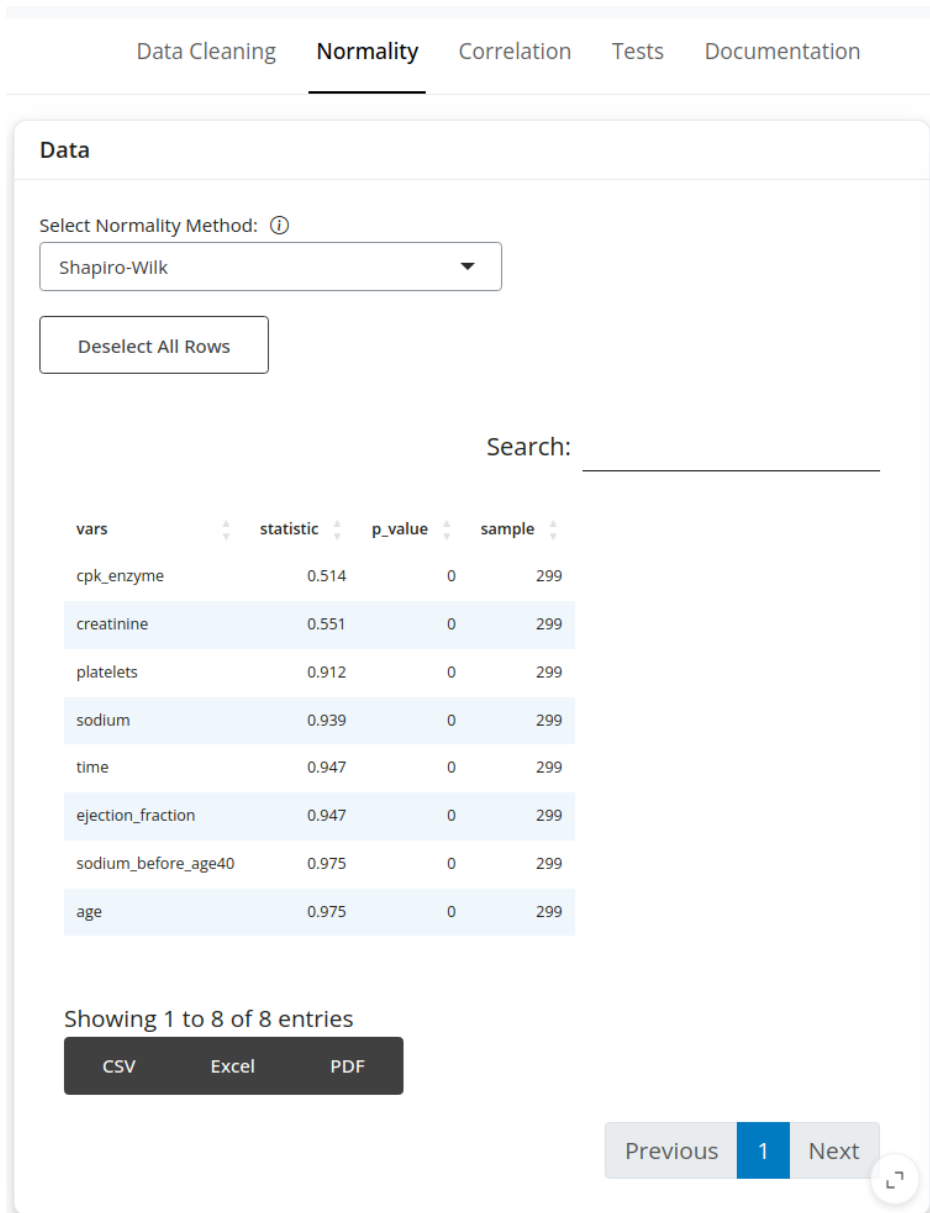


Fig8. Normality results table.

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.):

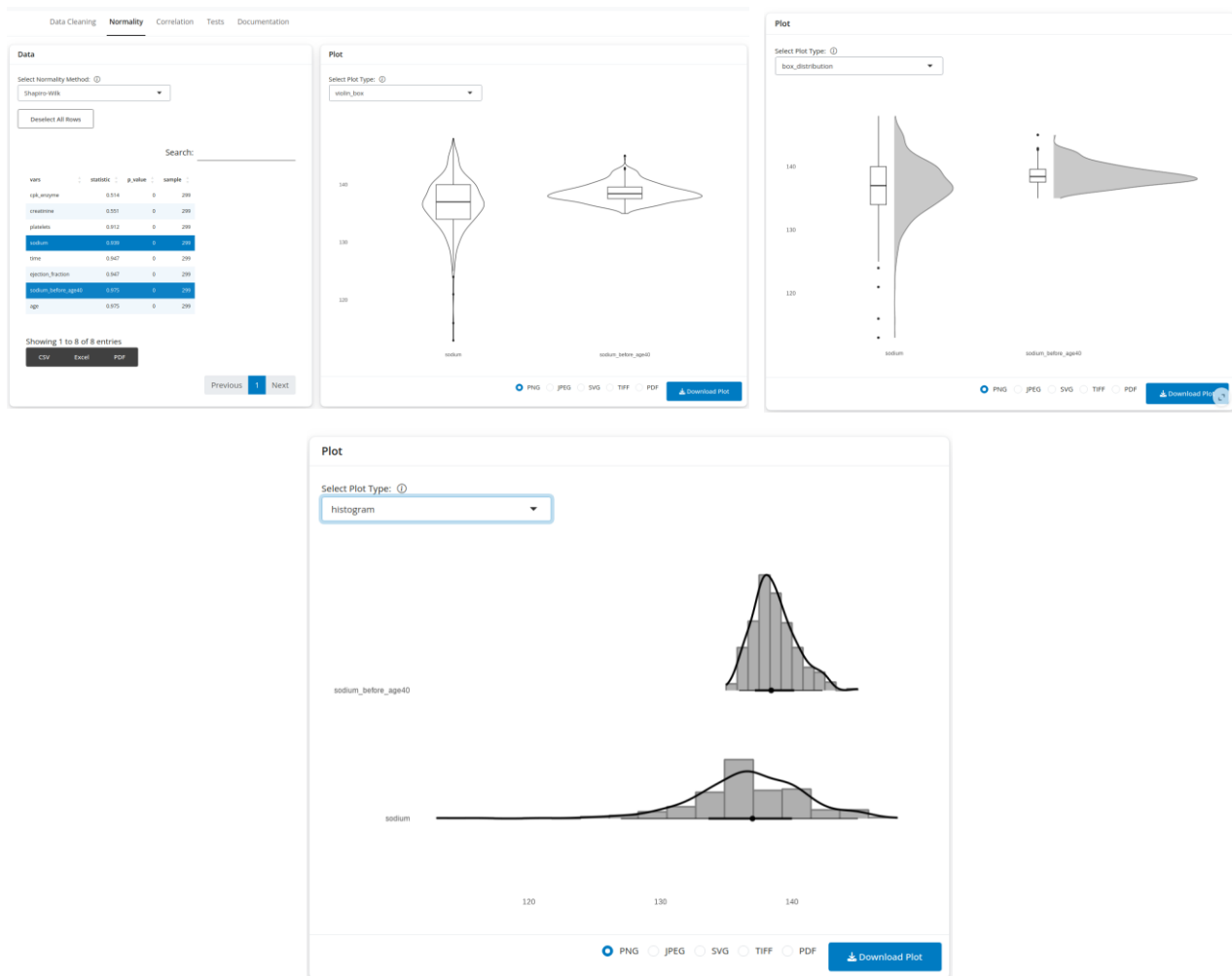


Fig9. The examples of distribution visualizations.

“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.

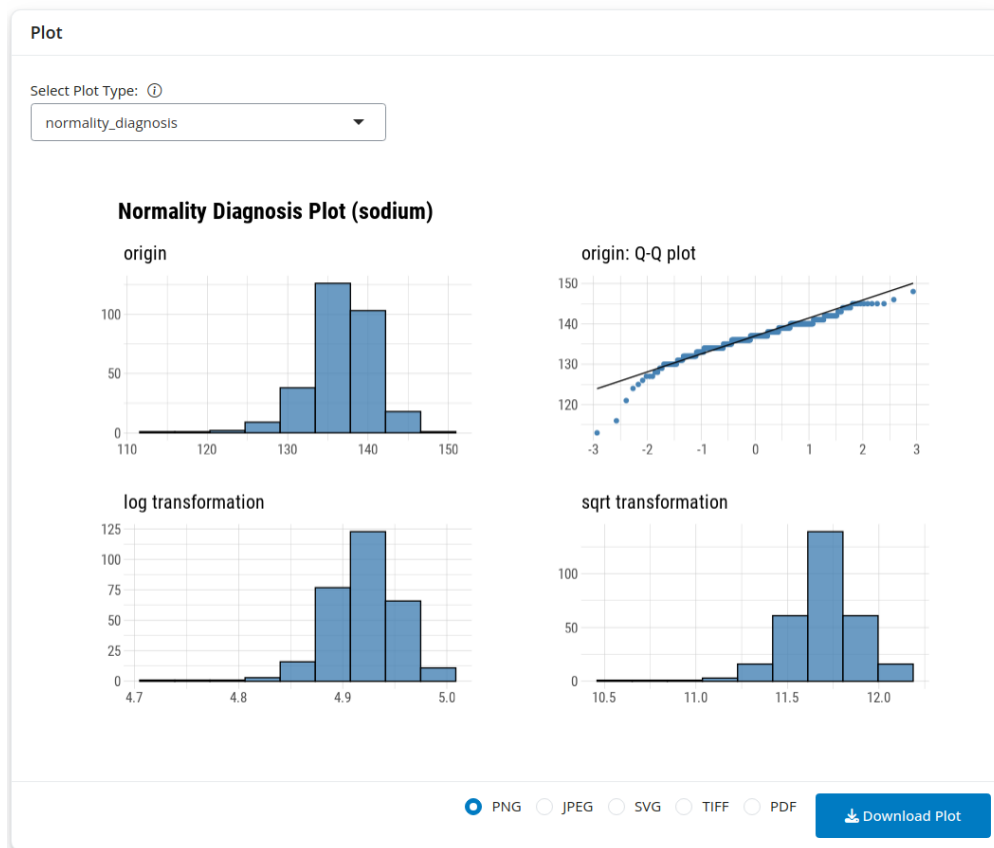


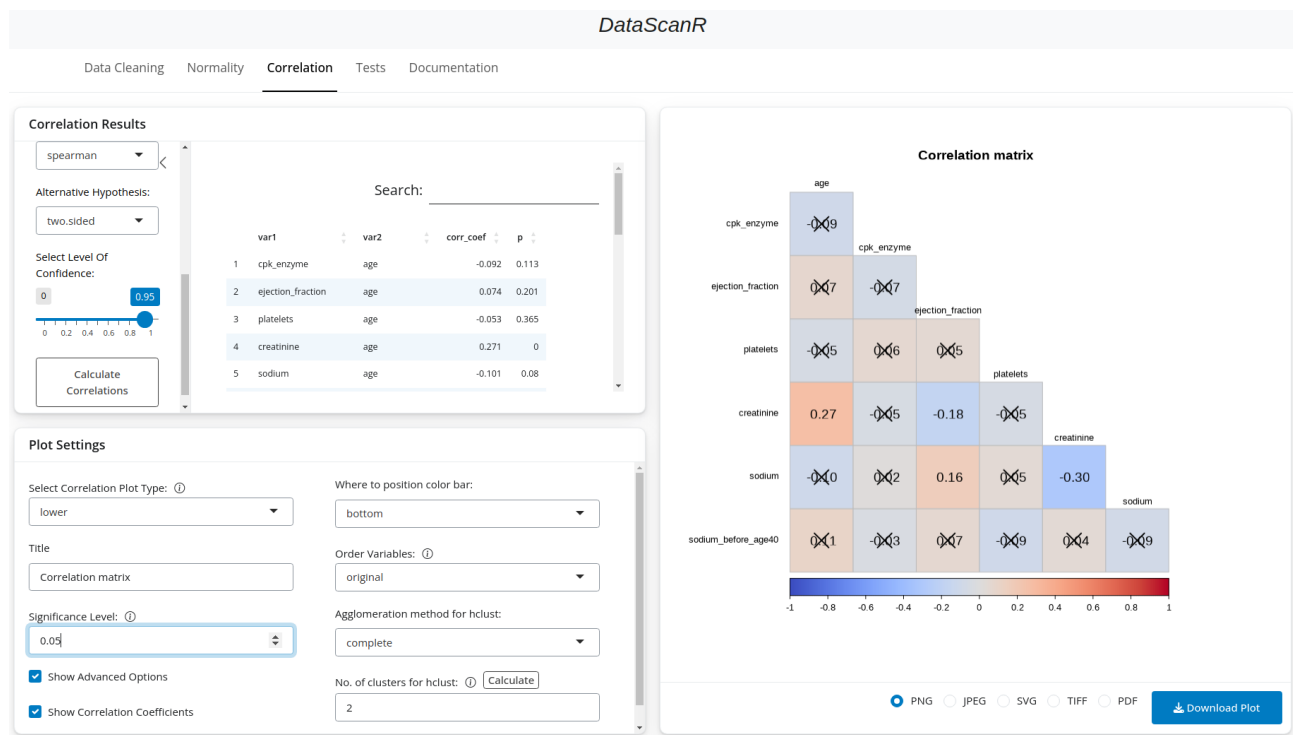
Fig10. Normality diagnosis plot.

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.



Copyright © 2024-2025. Ilona Szczot and Jyotirmoy Das. Created with R, Shiny. v1.0.0

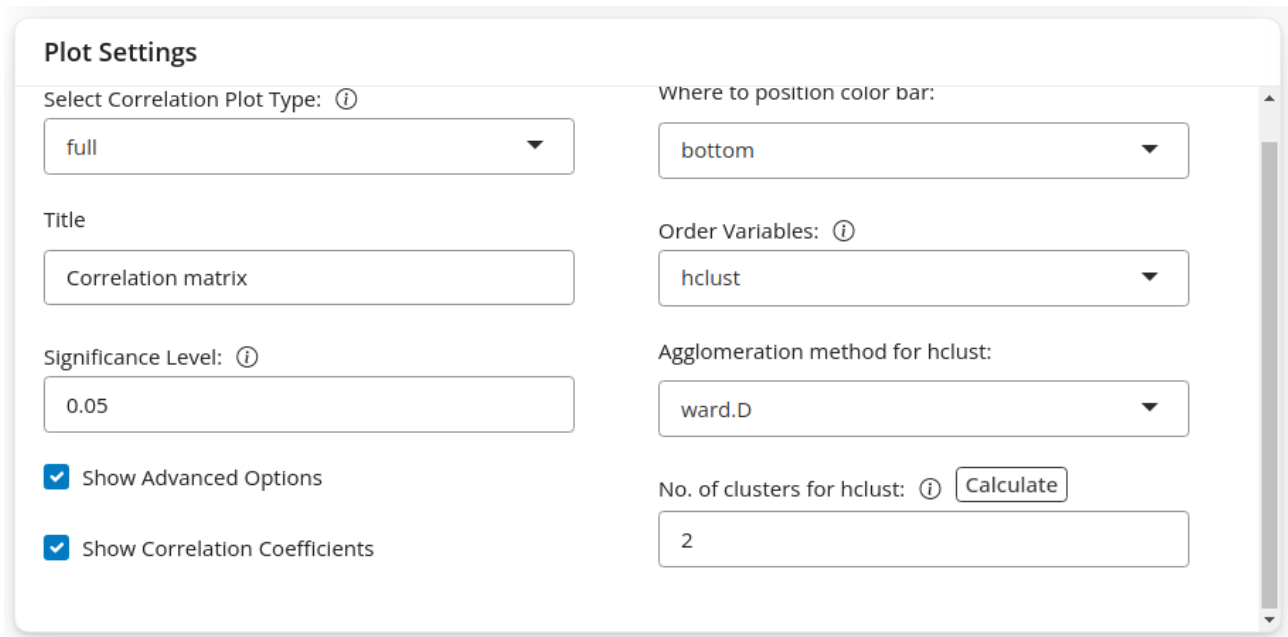
Fig11. The example of Correlation section.

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



The screenshot shows a 'Plot Settings' dialog box with the following options:

- Select Correlation Plot Type:** A dropdown menu with 'full' selected.
- Where to position color bar:** A dropdown menu with 'bottom' selected.
- Title:** A text input field containing 'Correlation matrix'.
- Order Variables:** A dropdown menu with 'hclust' selected.
- Significance Level:** A text input field containing '0.05'.
- Agglomeration method for hclust:** A dropdown menu with 'ward.D' selected.
- No. of clusters for hclust:** A text input field containing '2', with a 'Calculate' button next to it.
- Show Advanced Options:** A checked checkbox.
- Show Correlation Coefficients:** A checked checkbox.

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

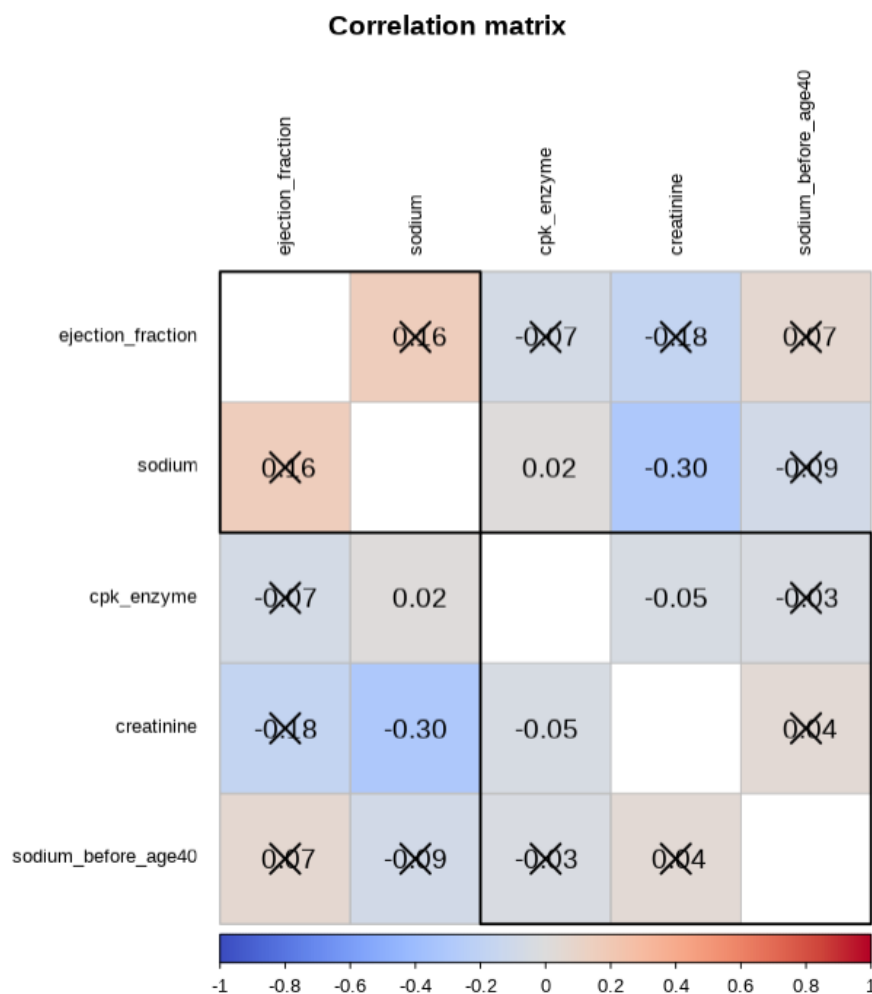
3.2.1.3. Agglomeration method for hclust

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 user 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.

This section uses corrplot package for visualizations. [2] (Wei, T, & Simko, V, 2017)

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 (μ), 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.).

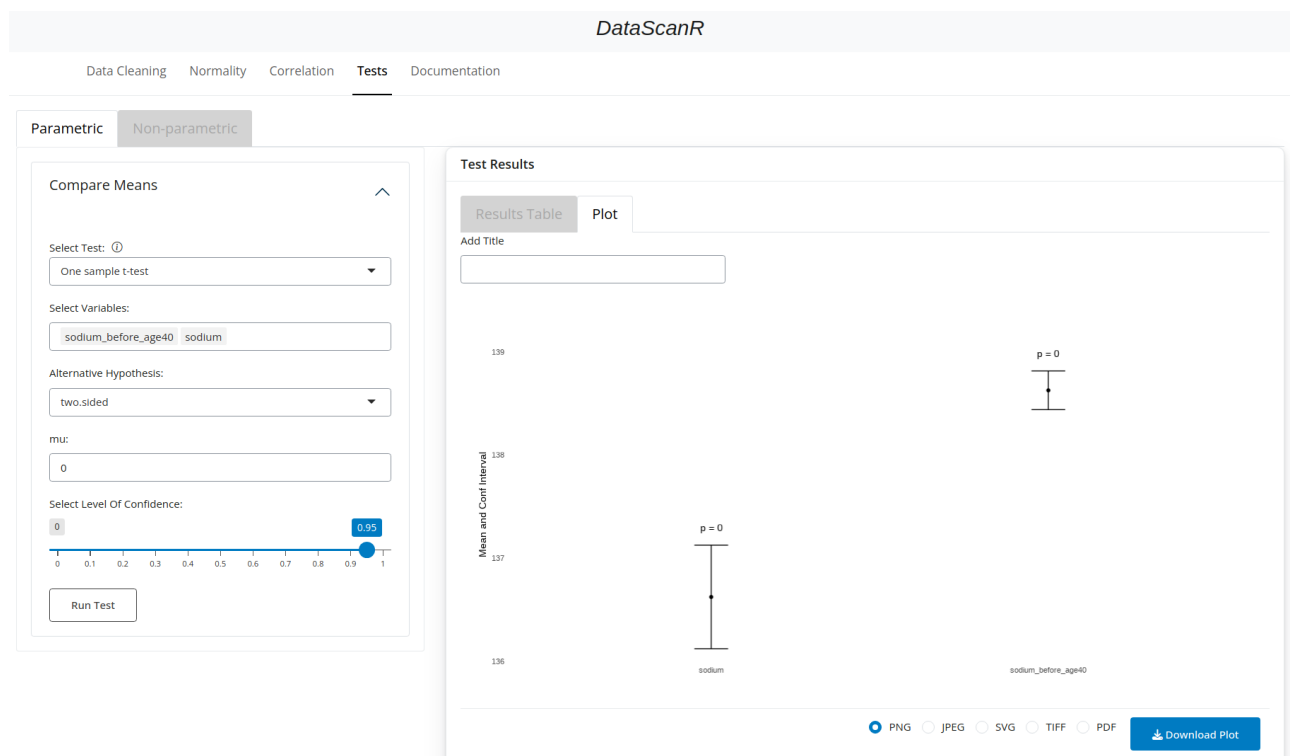


Fig14. One-sample t-test results.

Test Results

Results Table

Plot

One sample t-test

Search:

vars	null_val	estimate	alternative	p_value	lowCI	uppCI	conf_level	statistic	parameter	samples
sodium_before_age40	0	138.629	two.sided	0	138.442	138.817	0.95	1456.35	298	299
sodium	0	136.625	two.sided	0	136.123	137.128	0.95	535.408	298	299

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 (μ), 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.).

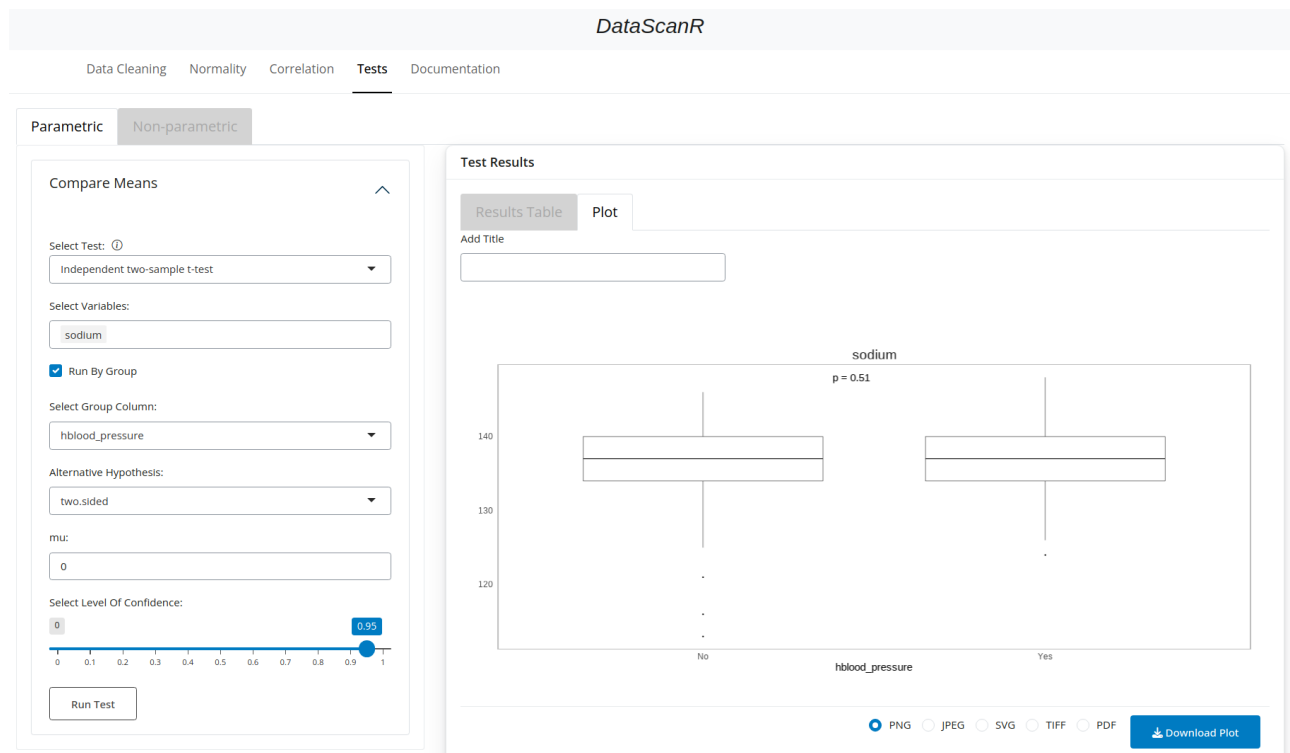


Fig16. Independent two-sample t-test results.

Results Table										
Plot										
Independent two-sample t-test										
Group: hblood_pressure										
Search: _____										
vars	null_val	estimate_No	estimate_Yes	alternative	p_value	lowCI	uppCI	conf_level	statistic	parameter
sodium	0	136.505	136.848	two.sided	0.511	-1.366	0.681	0.95	-0.659	231.956

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 (μ), 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.).

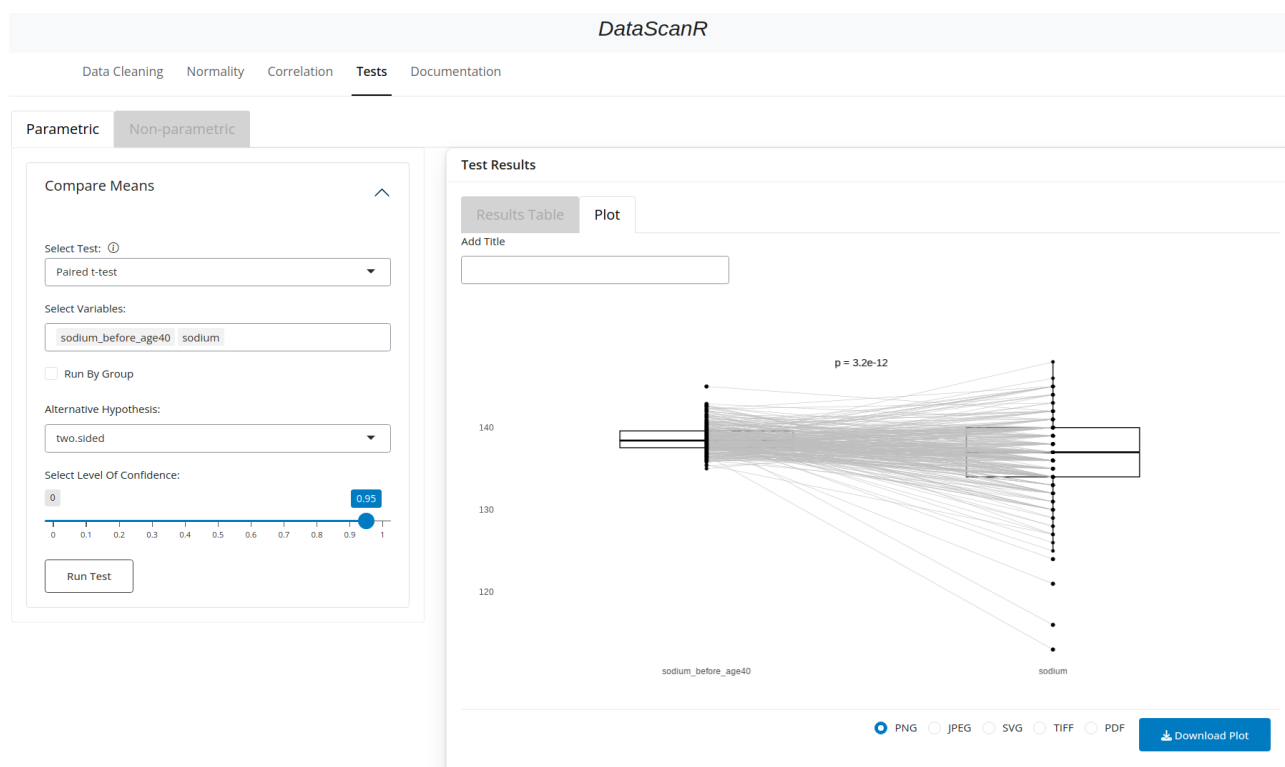


Fig18. Paired t-test results.

Test Results

Results Table Plot

Paired t-test

Search: _____

null_val	mean1	mean2	mean_difference	alternative	p_value	lowCI	uppCI	conf_level	statistic	parameter	sr
0	138.629	136.625	2.004	two.sided	0	1.461	2.546	0.95	7.27	298	

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 (μ), 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.).

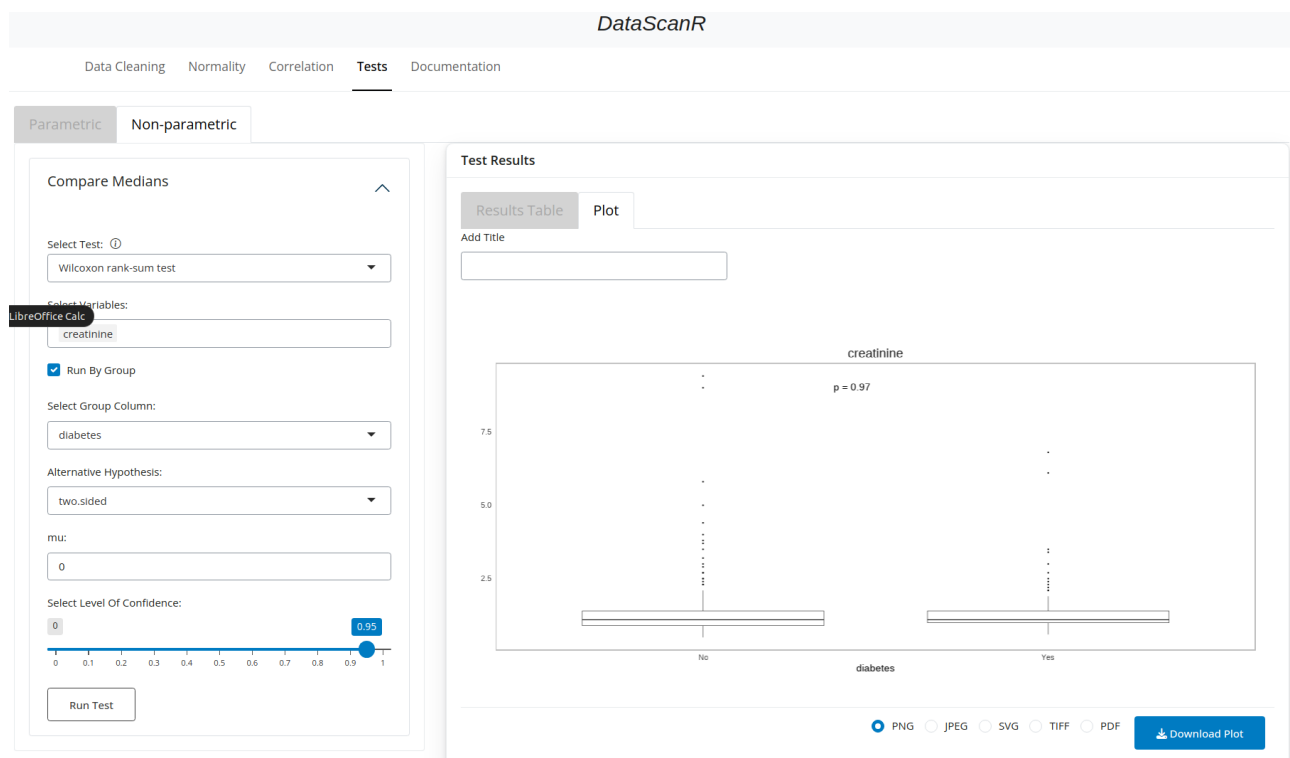


Fig20. Wilcoxon rank-sum test results.

Results Table										
Plot										
Wilcoxon rank-sum test										
Group: diabetes										
Search: _____										
vars	null_val	median1	median2	median_difference	alternative	p_value	lowCI	uppCI	conf_level	statistic
creatinine	0	1.1	1.1	0	two.sided	0.973	-0.1	0.1	0.95	10850.5

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 median to be tested (μ), 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.).

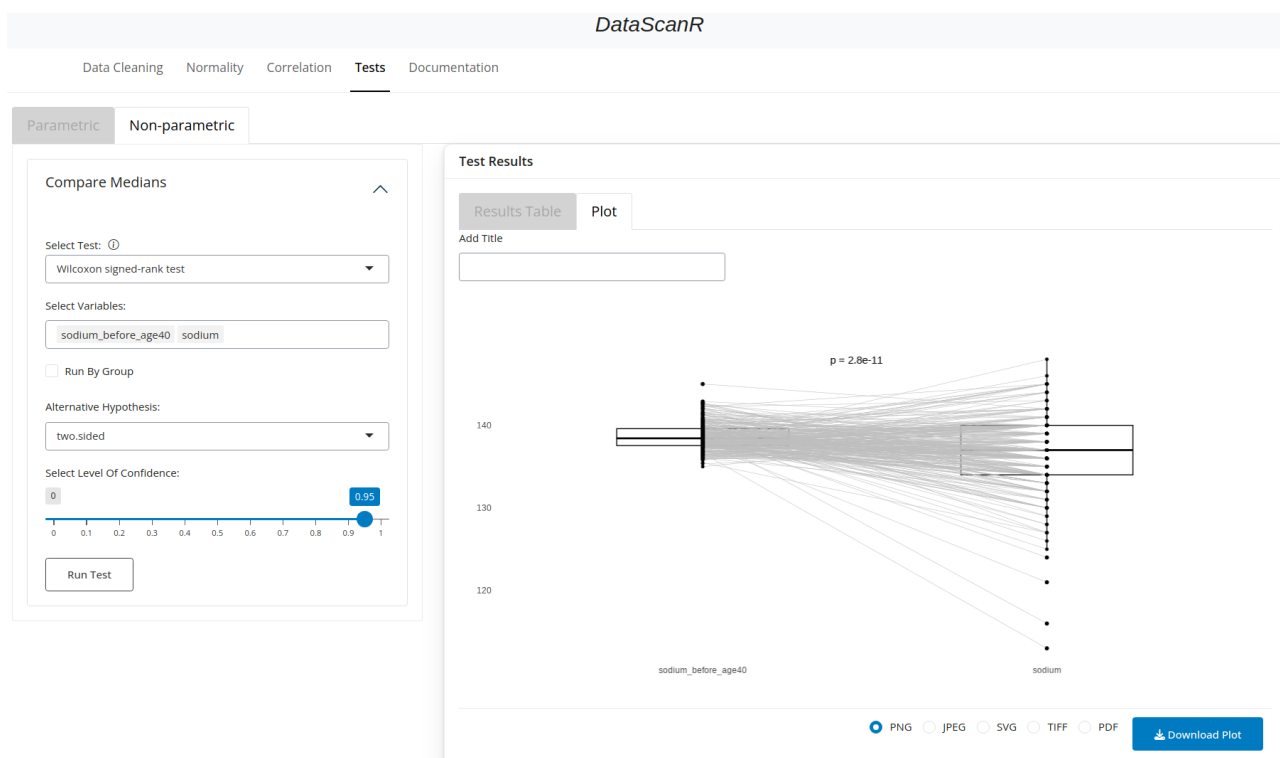


Fig22. Wilcoxon signed-rank test results.

Test Results

Results Table **Plot**

Wilcoxon signed-rank test

Search: _____

vars	null_val	median1	median2	median_difference	alternative	p_value	lowCI	uppCI	conf_level
sodium_before_age40_vs_sodium	0	138.425	137	1.778	two.sided	0	1.265	2.295	(

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.).

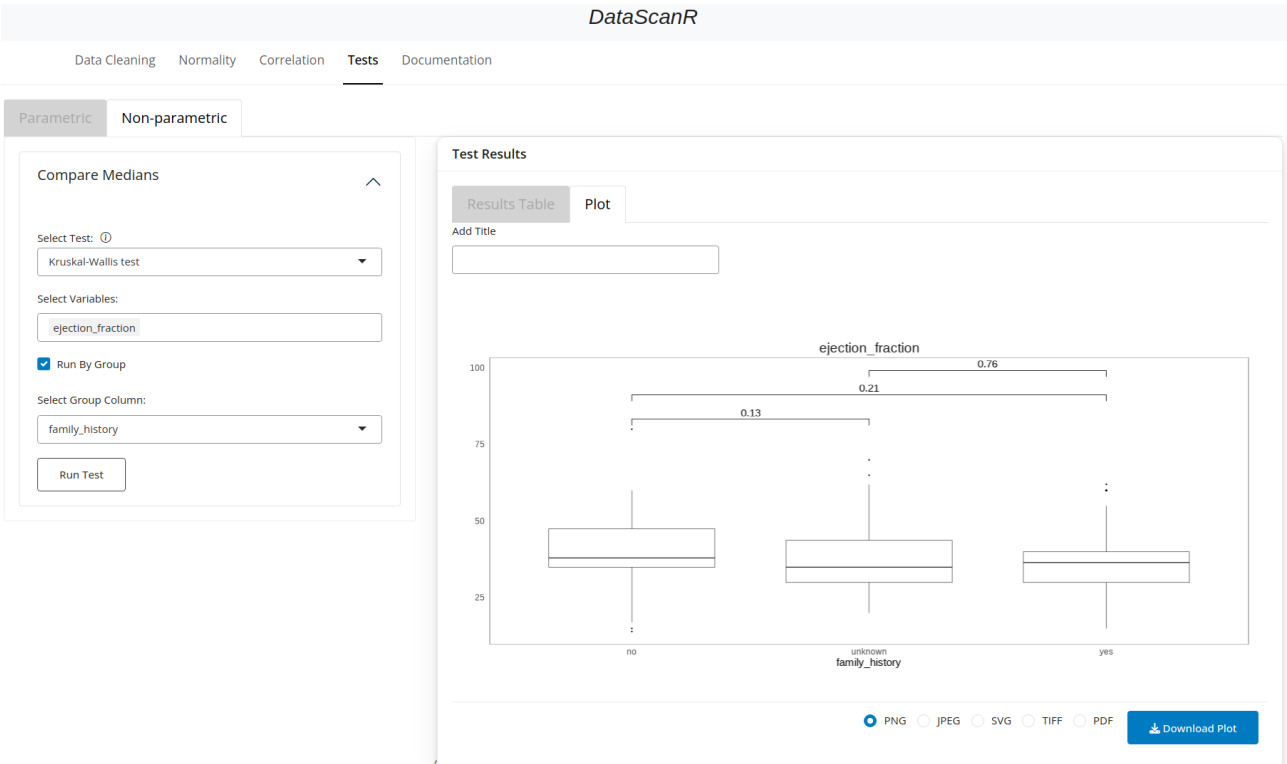


Fig24. Kruskal-Wallis test results.

Test Results				
Results Table				
Kruskal-Wallis test				
Group: family_history				
Search: _____				
vars	p_value	Kruskal_Wallis_chi_squared	parameter	
ejection_fraction	0.27	2.617	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.

6. Example Data

The example data was used to demonstrate DataScanR functionality. Users can download and use that data after clicking on the download link in the bottom of “Data Cleaning” section. (Fig26.) The example data comes from dlookr package. [1] (Ryu, u.d.) and additional variables were simulated by the authors of DataScanR to be able to demonstrate all options of this application.

Current Data

Summarize Current Data

Show Current Data

Select Variables:

Remove Selected Variables

Show Selected Variables

Summarize Selected Data

*Restore Original Data

Download Example Data

Original Data

age	anaemia	cpk_enzyme	diabetes	ejection_fraction	hblood_pressure	pl
75	No	582	No	20	Yes	
55	No	7861	No	38	No	
65	No	146	No	20	No	
50	Yes	111	No	20	No	
65	Yes	160	Yes	20	No	
90	Yes	47	No	40	Yes	
75	Yes	246	No	15	No	
60	Yes	315	Yes	60	No	
65	No	157	No	65	No	
80	Yes	123	No	35	Yes	

Showing 1 to 10 of 299 entries

CSV

Excel

PDF

Previous

1

2

3

4

5

...

30

Next

Download all as CSV

Save Interactive Report

Fig26. Example data available to download.

References

- [1] Ryu C (2024). *dlookr: Tools for Data Diagnosis, Exploration, Transformation*. R package version 0.6.3, <https://CRAN.R-project.org/package=dlookr>.
- [2] Wei, T., & Simko, V. (2017). *R package “corrplot”: Visualization of a Correlation Matrix*. R package version 0.94, <https://github.com/taiyun/corrplot>