DataScanR

Quick Guide

Ilona Szczot, Jyotirmoy Das

1 Centrum för social och affektiv neurovetenskap, Linköping University,

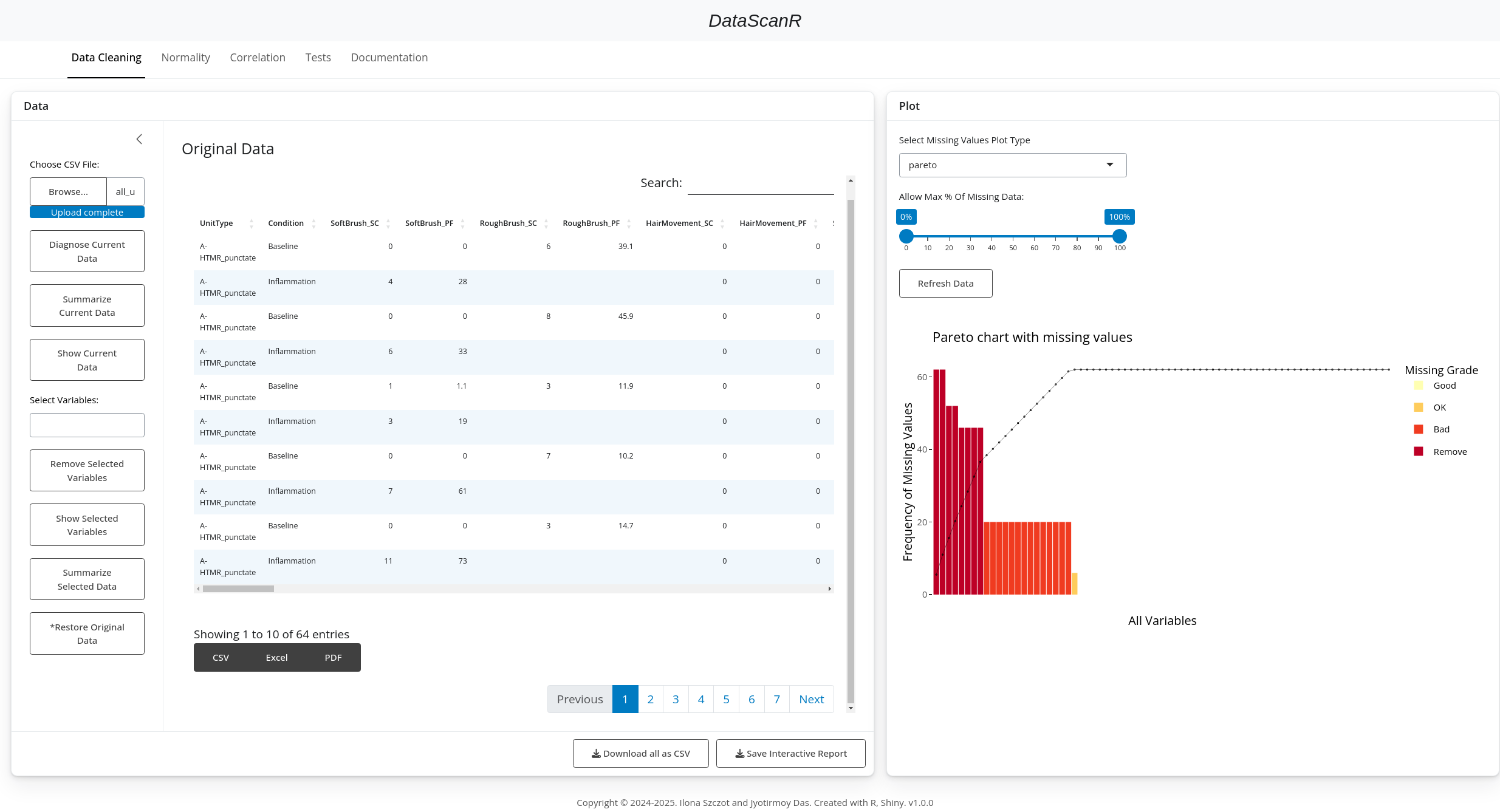
Linköping, Sweden, 2 Core Facility, Faculty of Medicine and Health Sciences, Linköping University, Linköping, Sweden and Clinical Genomics Linköping, Science for Life Laboratory, Sweden

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).

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 basic statistical summary of the data.



Upload Data

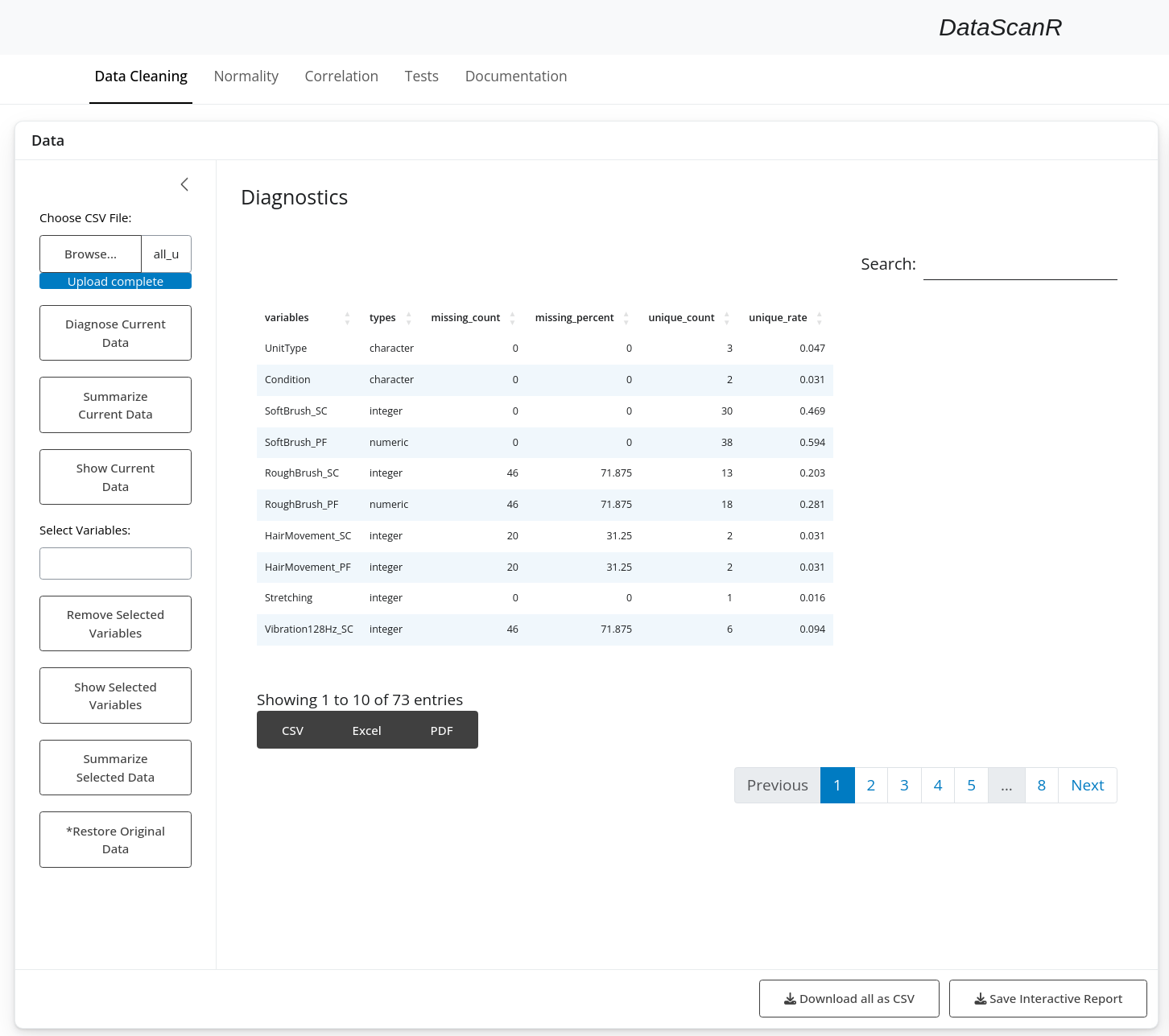
Click on the “Browse...” button to select your data.csv file. This will allow you to see the data in a form of a table. Use “Previous” and “Next” buttons, to see all pages of your data. You can easily export to csv, excel or pdf each displayed table or export all of the pre-processed data to csv file.

*Add screenshot*

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.

This option uses diagnose function from dlookr package.

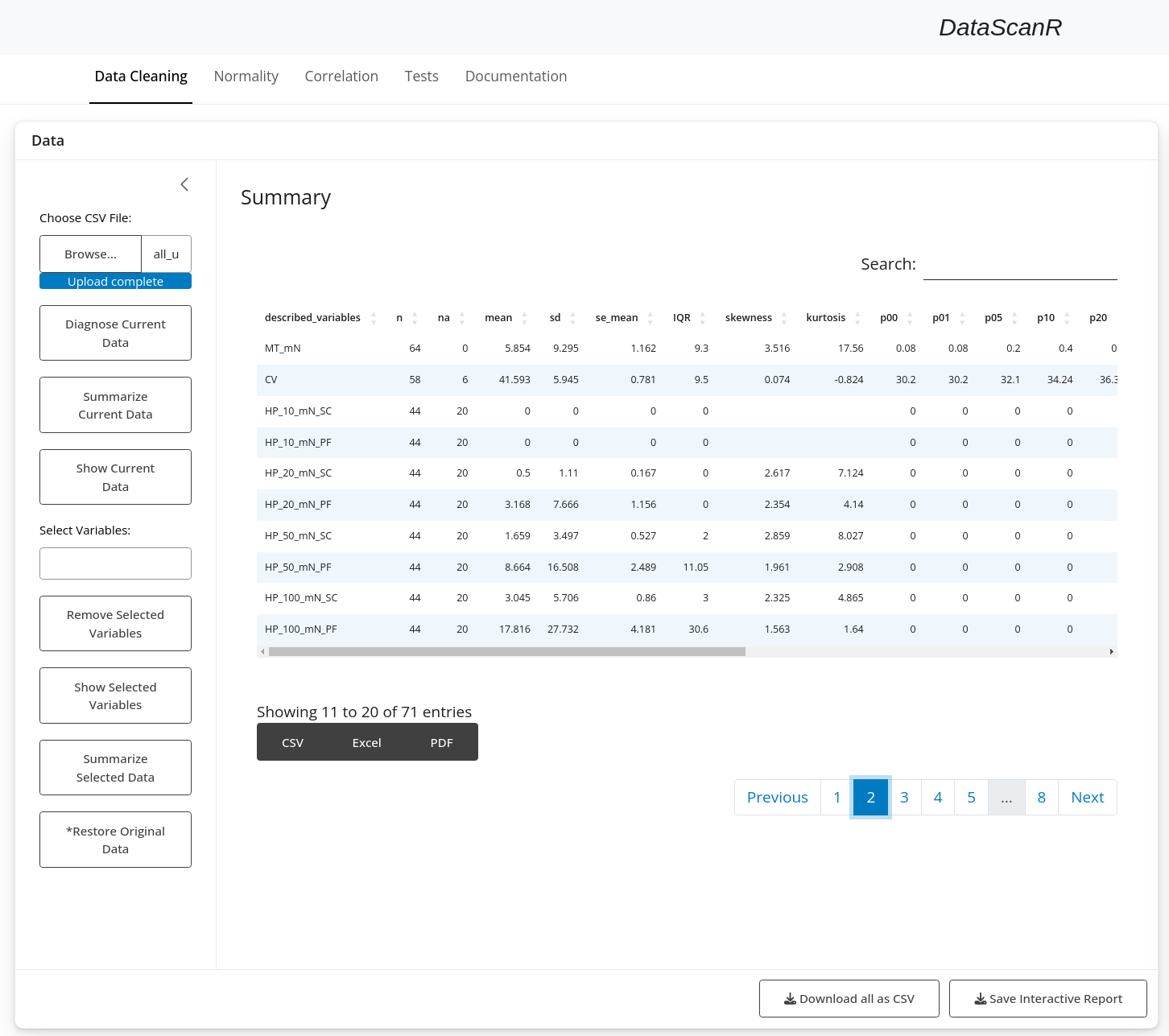


Summarize Current Data

This option calculates descriptive statistics for each variable:

* 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.



Show Current Data

This option displays the current data as a data table.

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.

Remove Selected Variables

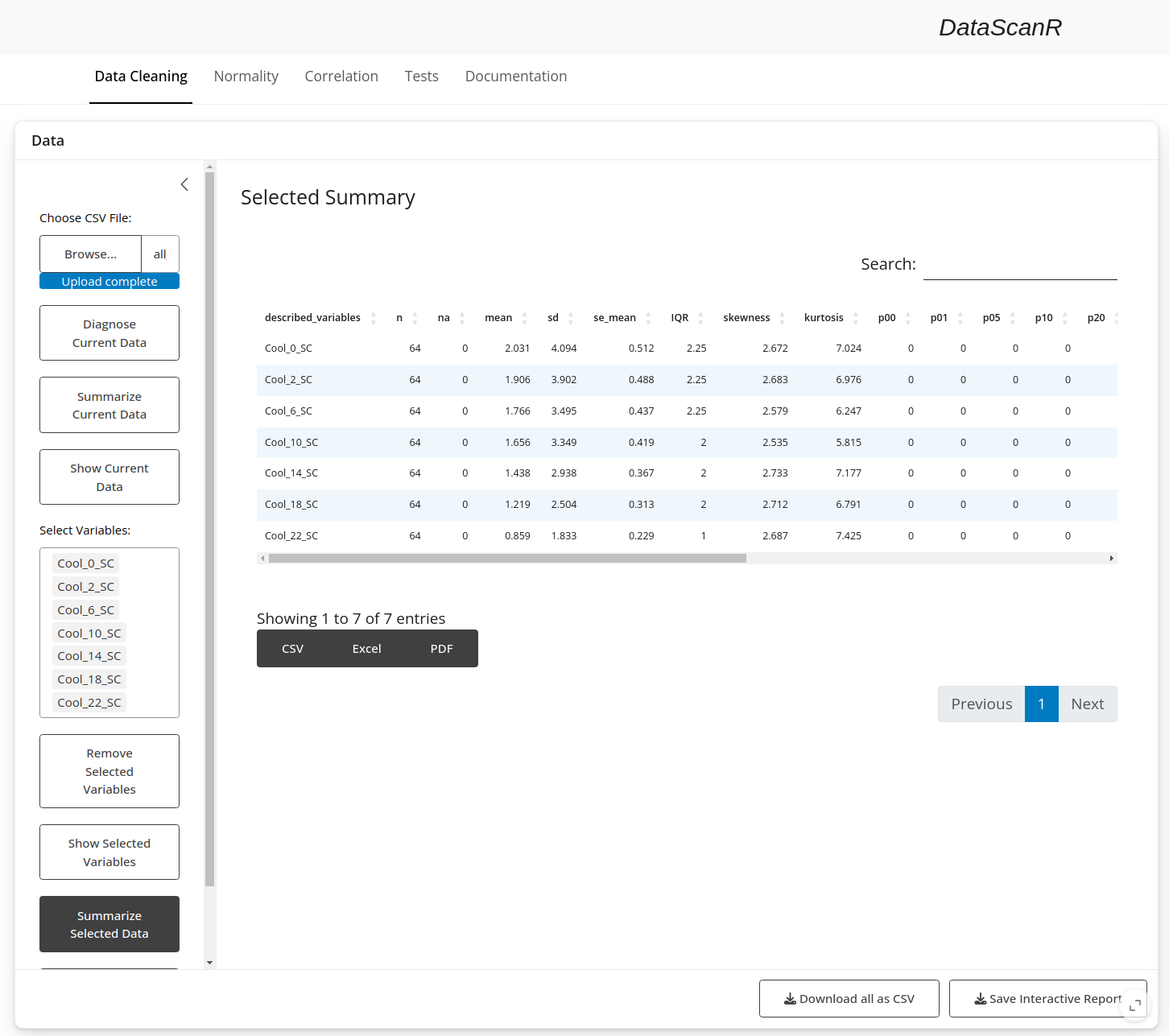
Removes from the dataset all variables that were selected under “Select Variables” section.

Show Selected Variables

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

Summarize Selected Data

Calculates descriptive statistics only for the variables that were selected under “Select Variables” section. 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.



Restore Original Data

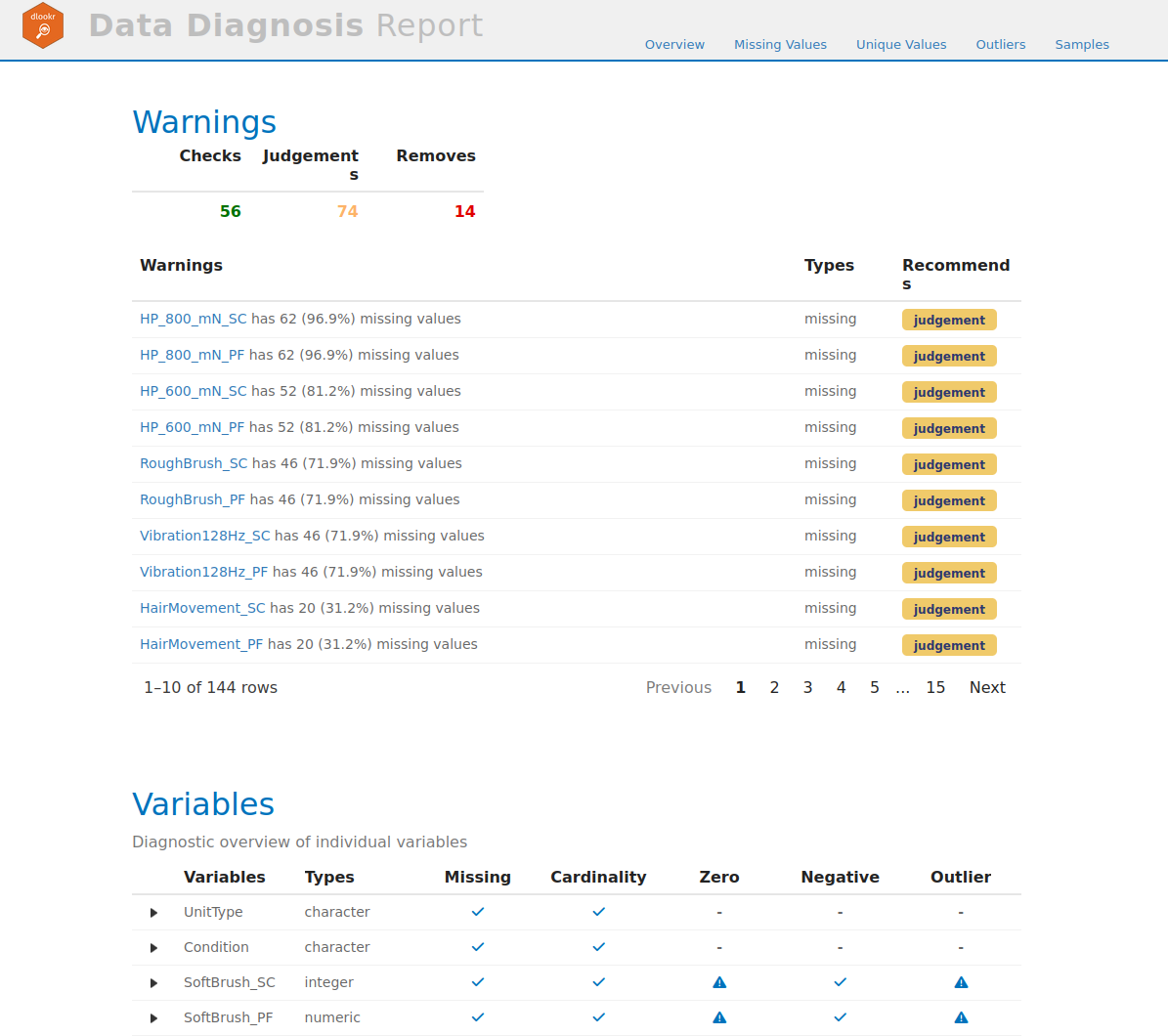
At any time, after removing some variables from the working dataset, the user can “go back” to working on all original data. Note, that the original data file is never modified by the application!

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.

Save interactive Report

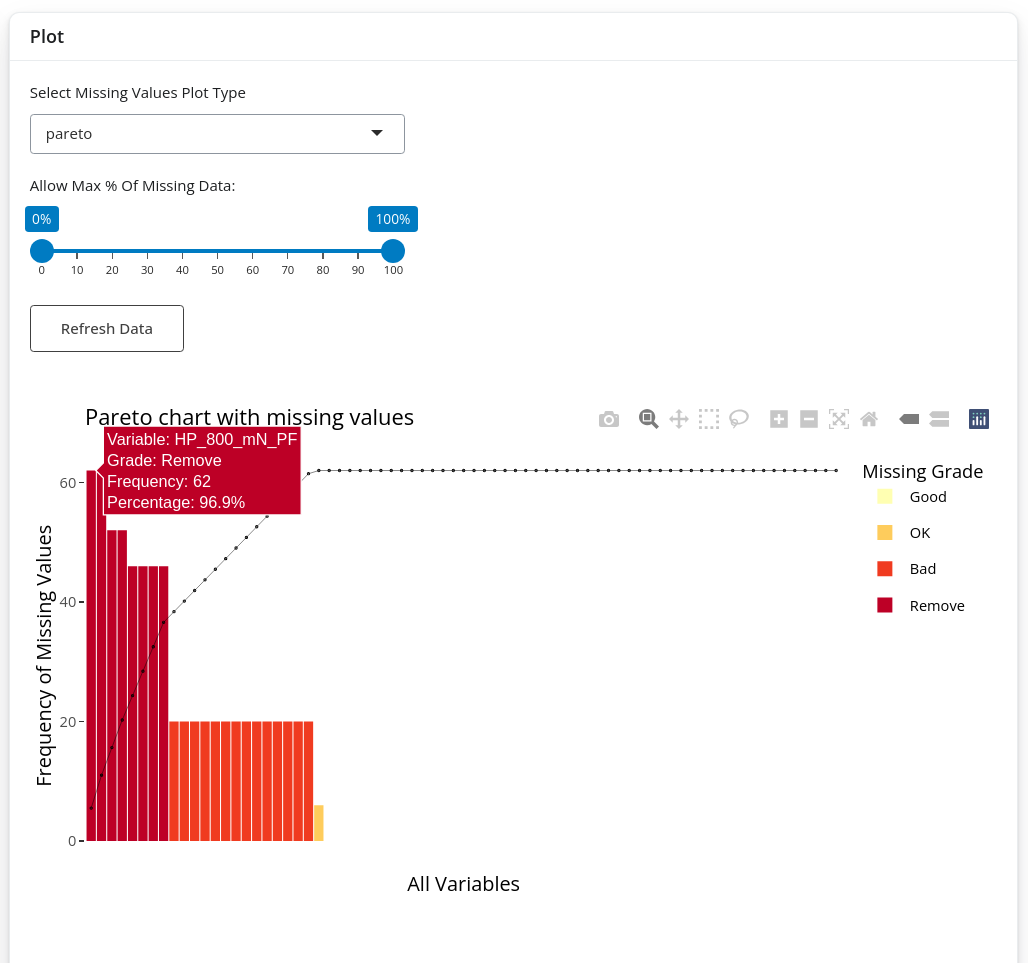
When the data table is displayed, it will generate html report with data diagnostics. That allows to store and share the information about the data quality without the need of using DataSanR. The shows the general summary of data types, missing, data, outliers and possible duplicates.

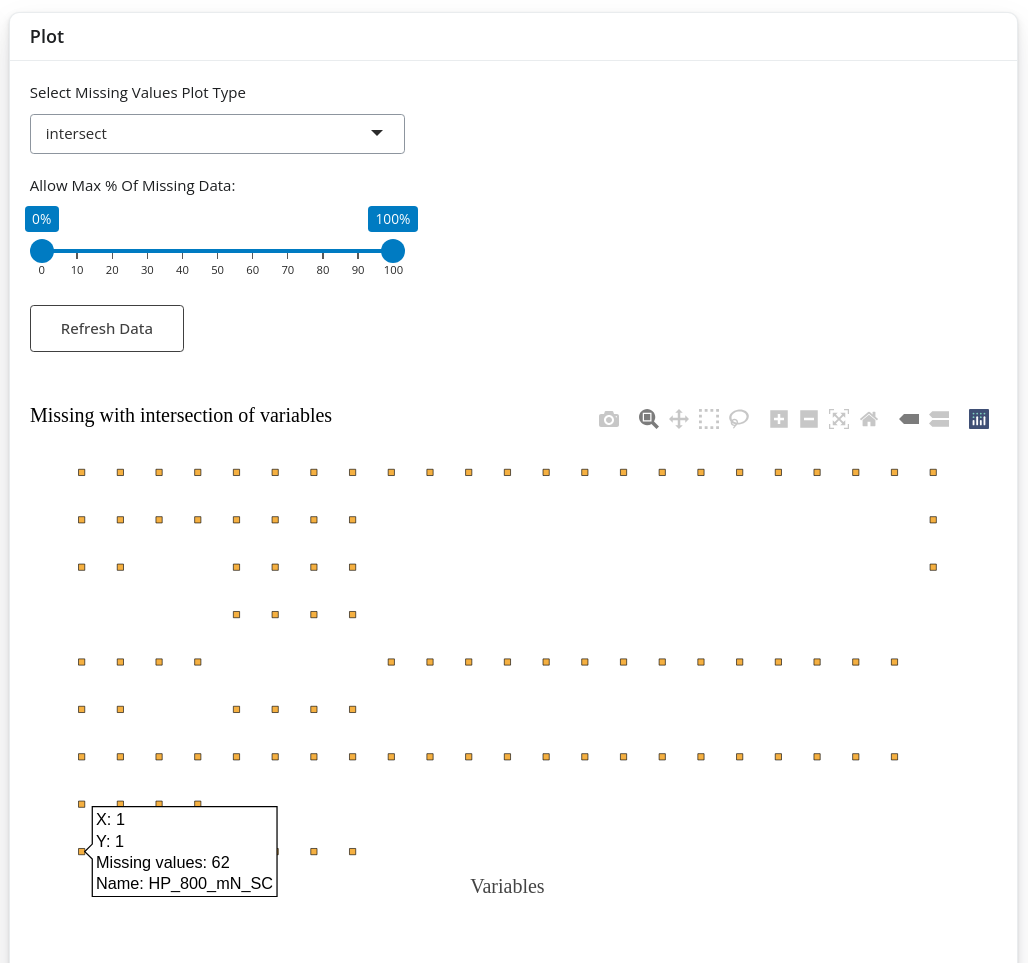


Plot

In this section, one can look at a graphical view of missing data for different variables. 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, 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.



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

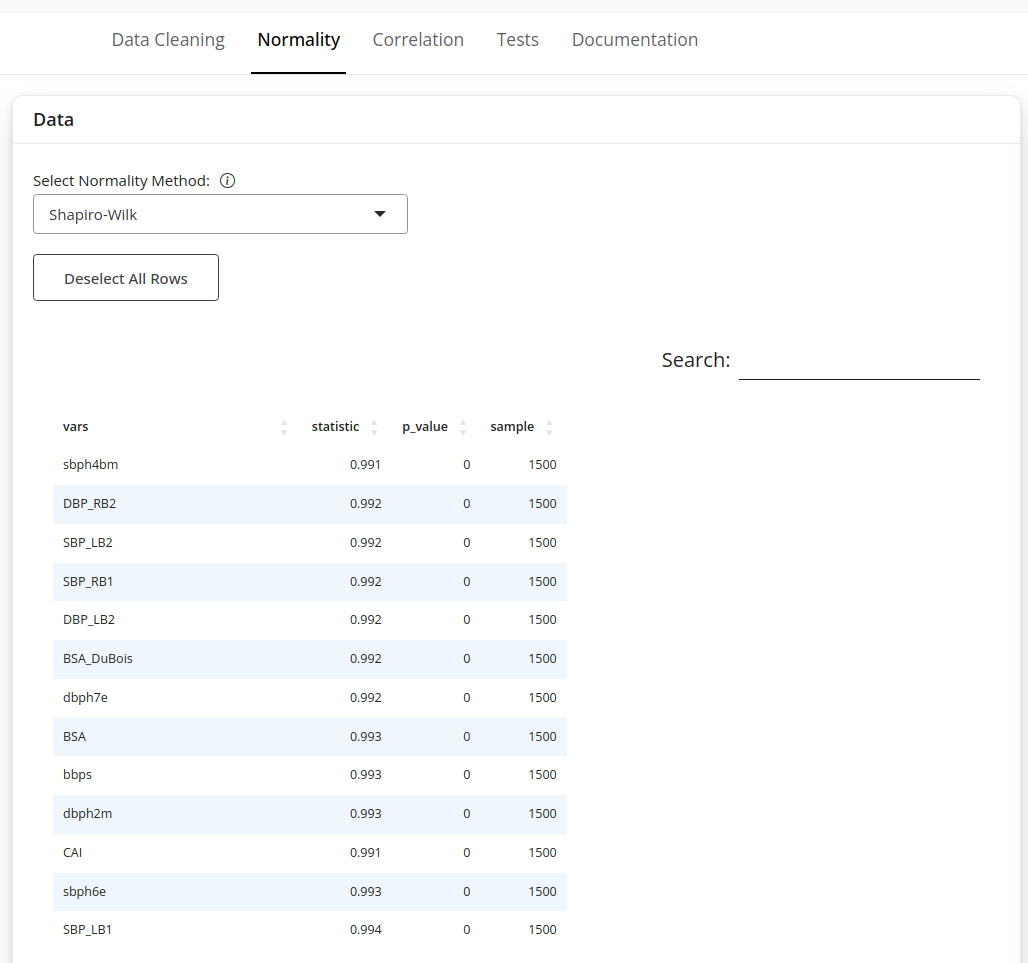
* Shapiro-Wilk: for dataset < 2000 observations.
* Kolmogorov-Smirnov: for dataset > 2000 observations.

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.

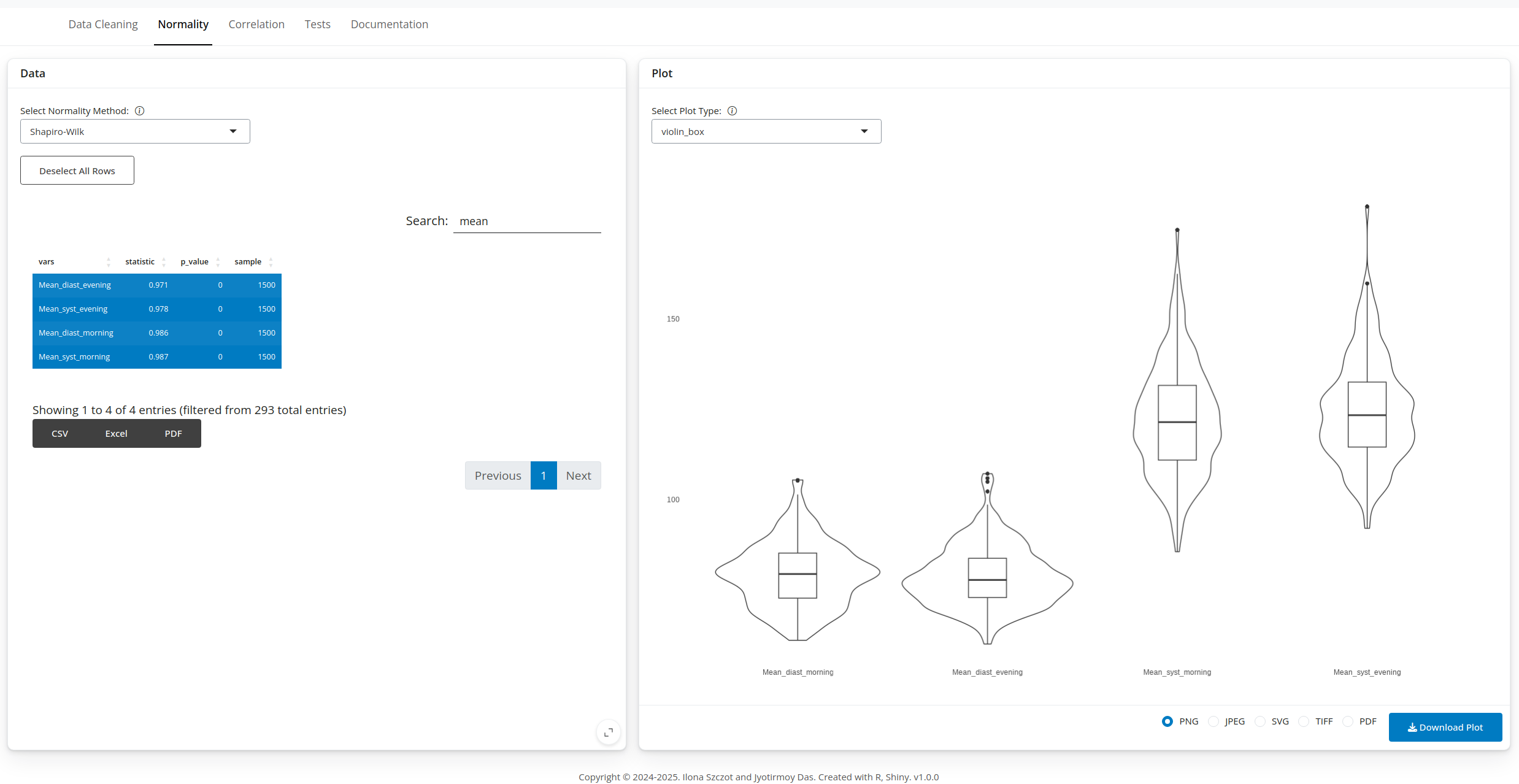
The users can choose the normality method themselves, by changing the method in the drop down menu for “Select Normality Method”.

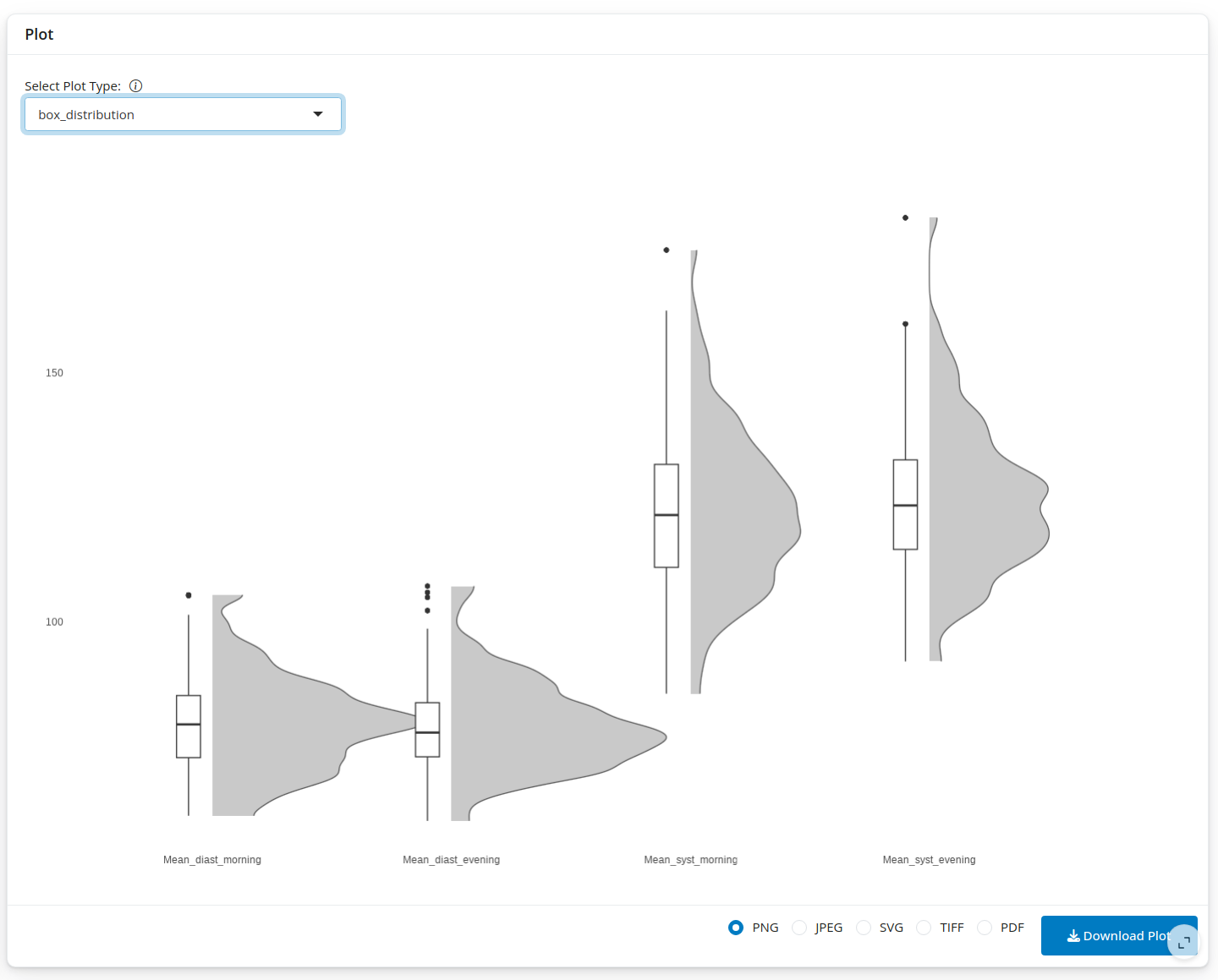


Plot

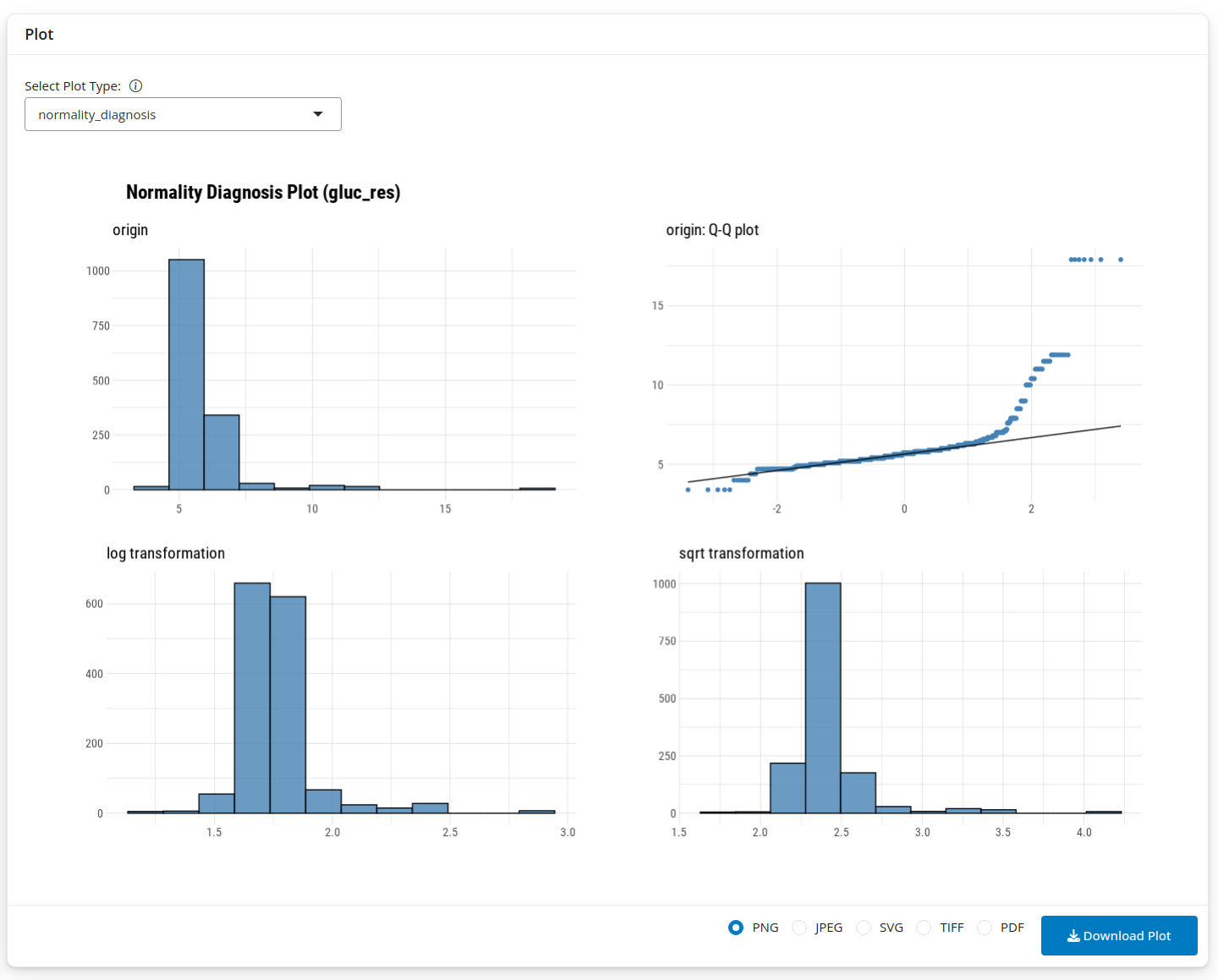
In this part of “Normality” section, one can visualize the distributions of up to 6 variables at a time.

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 most popular file formats. Here are some of the examples of distribution visualizations:









normality\_diagnosis 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

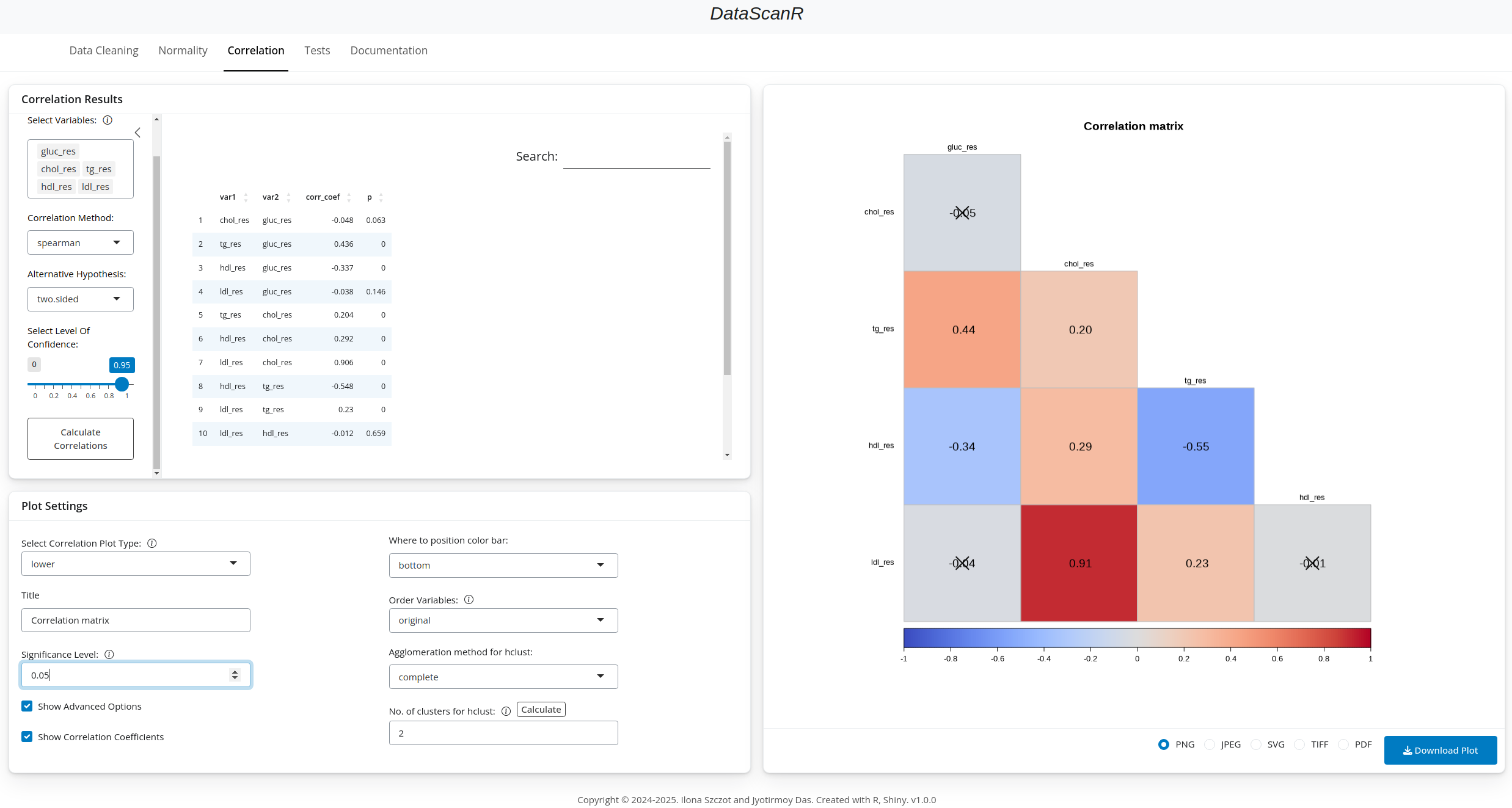
This plot type allows to only visualize one variable at a time.

3. Correlation

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

Correlation Results

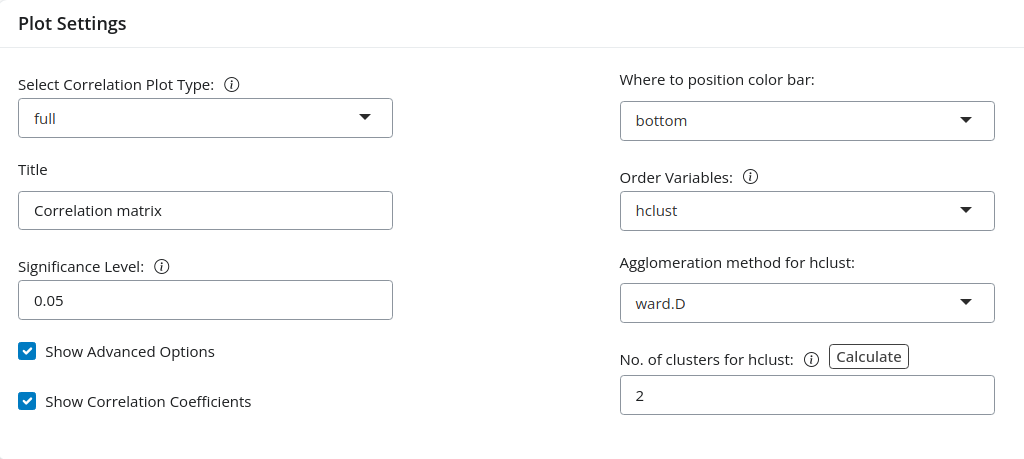
In this part, one can select the numerical variables of interest, the method to use for calculating the correlations (pearson, kendall, spearman), the hypothesis to test (less, greater, two-sided), and the level of confidence. After clicking on “Calculate Correlations” button, the results are displayed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



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.



Advanced options

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.

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

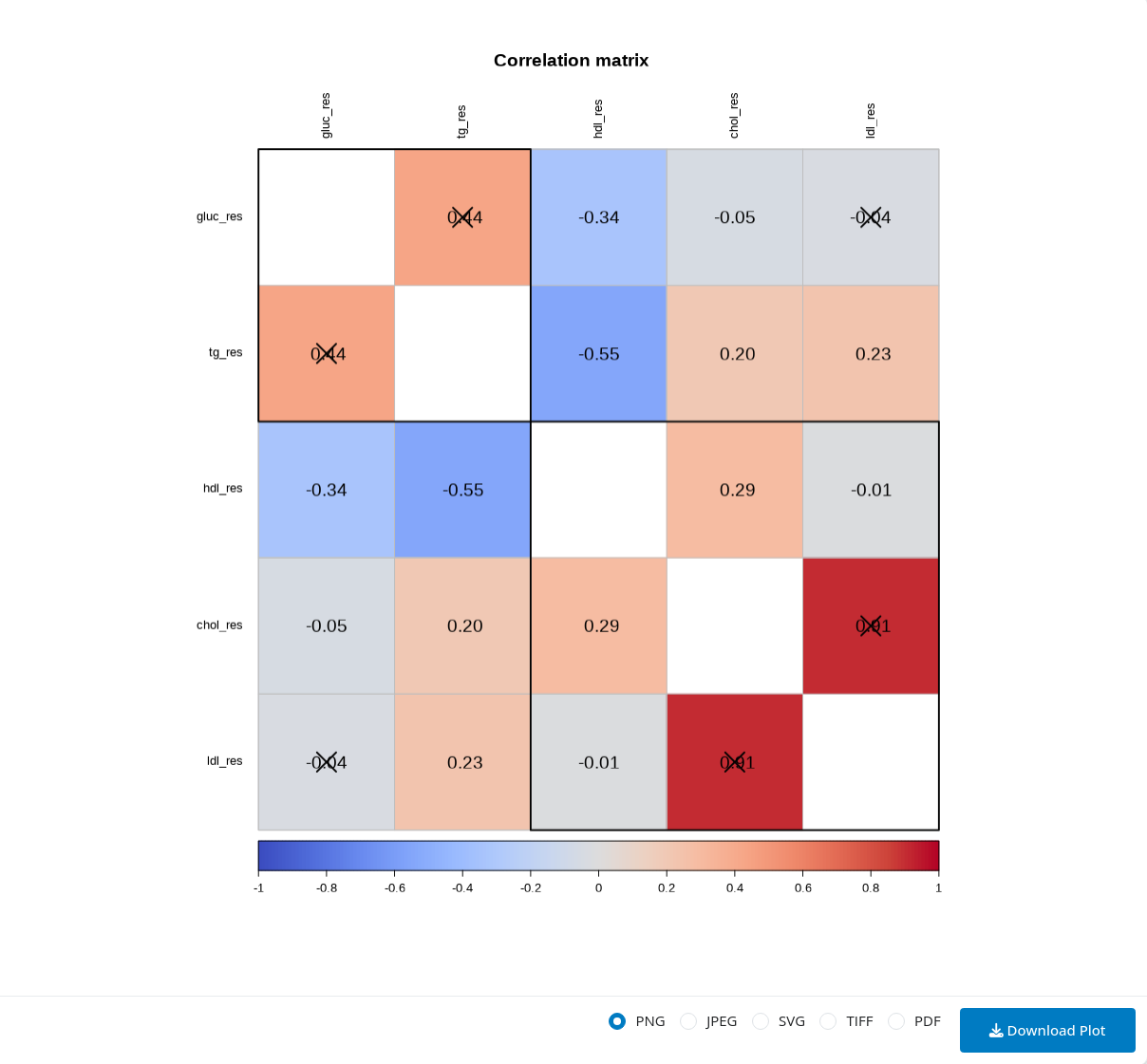
Agglomeration method for hclust

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'.

No. of clusters for hclust

By default, it will show 2 clustering squares on a “full” plot type if hclust order is selected. 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.



4. Tests