Initial Data Analysis Tool

Maciej Zdanowicz

In our project we have decided to create the tool for initial data analysis. The most popular tool like that widely available is the DataExplorer library, which, in our humble opinion, was not a very convenient tool to use nor did it suit our needs.

Our project consists of three main parts:

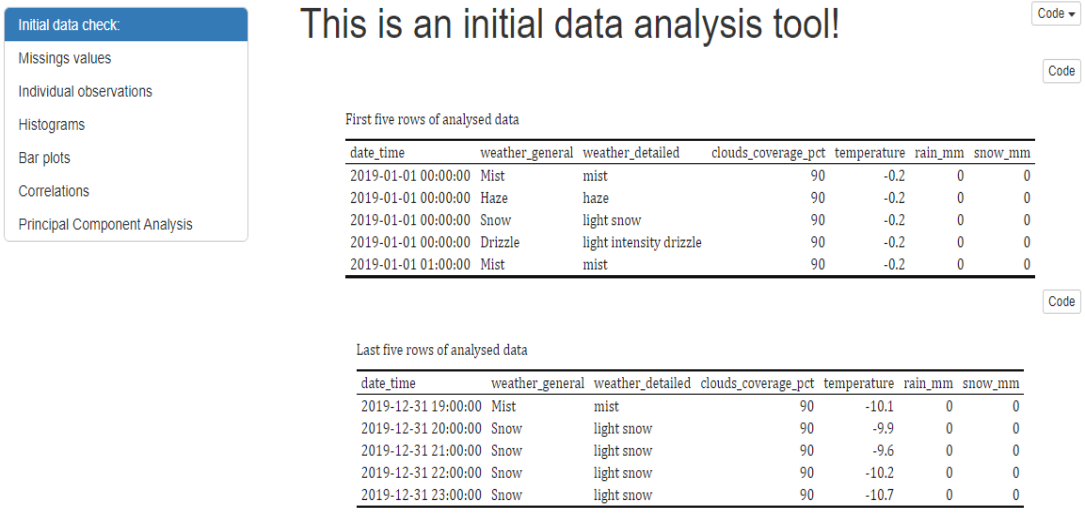
- folder called *functions* contains 6 functions used in our project; functions used for plotting histograms, performing PCA etc.,
- R-markdown file called *template* produces the html output,
- R file called *data_to_md* contains the *raport* function, which takes data provided by the user, renders the markdown template with this data, and creates final report.

How to use our tool?

The first thing needed is the dataset that user wants to analyse. Then, all he or she needs to do is to import the *report* function from the data_to_md file (f.e. by using: *source('data_to_md.R')*) and run it with chosen dataset as the argument (remember to change the directory to file's location!). *test_run* file contains the example of how to use our function.

What information does created output contain?

1. Initial data check, including first and last five rows of data, information about the types of variables, as well as their mean, standard deviation, median etc.



date_time	weather_general	weather_detailed	clouds_coverage_pct	temperature	rain_mm	snow_mm
2019-01-01 00:00:00	Mist	mist	90	-0.2	0	0
2019-01-01 00:00:00	Haze	haze	90	-0.2	0	0
2019-01-01 00:00:00	Snow	light snow	90	-0.2	0	0
2019-01-01 00:00:00	Drizzle	light intensity drizzle	90	-0.2	0	0
2019-01-01 01:00:00	Mist	mist	90	-0.2	0	0

date_time	weather_general	weather_detailed	clouds_coverage_pct	temperature	rain_mm	snow_mm
2019-12-31 19:00:00	Mist	mist	90	-10.1	0	0
2019-12-31 20:00:00	Snow	light snow	90	-9.9	0	0
2019-12-31 21:00:00	Snow	light snow	90	-9.6	0	0
2019-12-31 22:00:00	Snow	light snow	90	-10.2	0	0
2019-12-31 23:00:00	Snow	light snow	90	-10.7	0	0

Initial data check:

Code

x
Number of variables: 7
x
Number of observations: 10591
x
Number of missings: 0
x
Percentage of missings: 0
x
Number of duplicated rows: 9
x
Number of binary variables: 2
x
Number of factors variables: 1
x
Number of continuous variables: 1
x
Number of character variables: 2

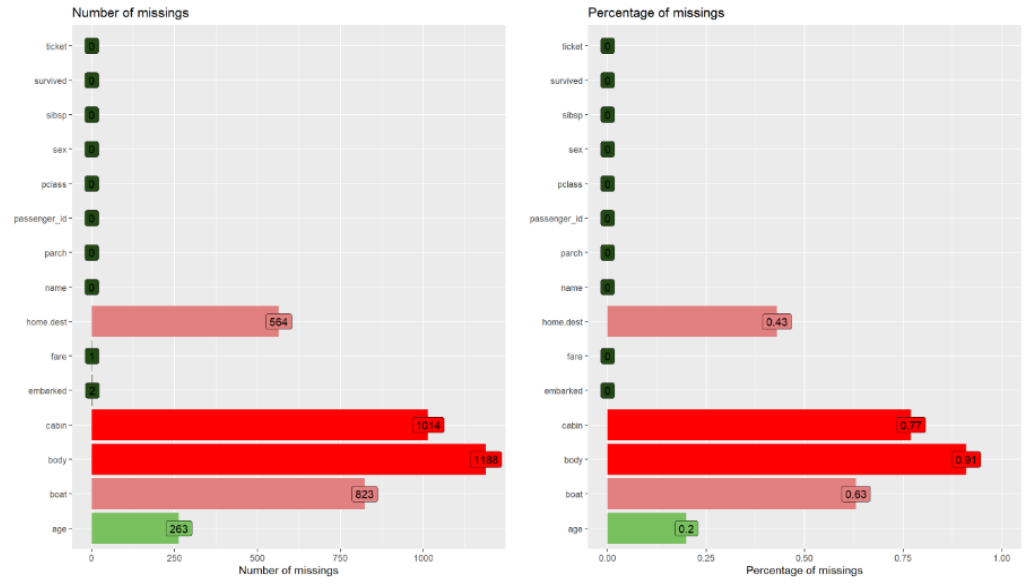
Code

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
date_time	1	10591	NaN	NA	NA	NaN	NA	Inf	-Inf	-Inf	NA	NA	NA
weather_general*	2	10591	3.54	2.71	2.0	3.19	1.48	1	10.0	9.0	0.820	-0.662	0.026
weather_detailed*	3	10591	15.79	7.87	17.0	16.53	10.38	1	32.0	31.0	-0.561	-0.863	0.076
clouds_coverage_pct	4	10591	49.76	39.55	75.0	50.83	22.24	0	92.0	92.0	-0.212	-1.769	0.384
temperature	5	10591	8.24	11.69	9.2	8.78	12.90	-27	33.9	60.9	-0.384	-0.391	0.114
rain_mm	6	10591	0.00	0.00	0.0	0.00	0.00	0	0.0	0.0	NaN	NaN	0.000
snow_mm	7	10591	0.00	0.00	0.0	0.00	0.00	0	0.0	0.0	NaN	NaN	0.000

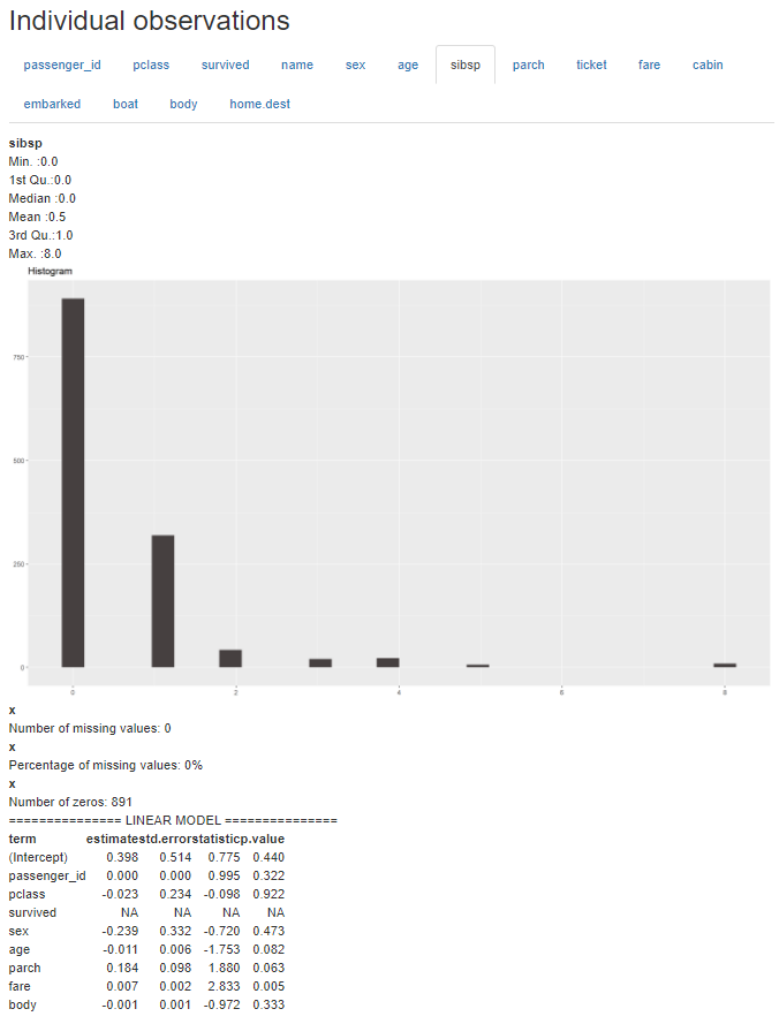
2. Missing values, two plots displaying number and percentage of missing values for each of the variables. Plots change their colours according to following rule:
   - if the proportion of missing values is lower than 10%, the bar is dark green,
   - if the proportion of missing values is between 10% and 35%, the bar is light green,
   - if the proportion of missing values is between 35% and 65%, the bar is light pink,
   - and if the proportion of missing values is higher than 65%, the bar is red.

Missings values

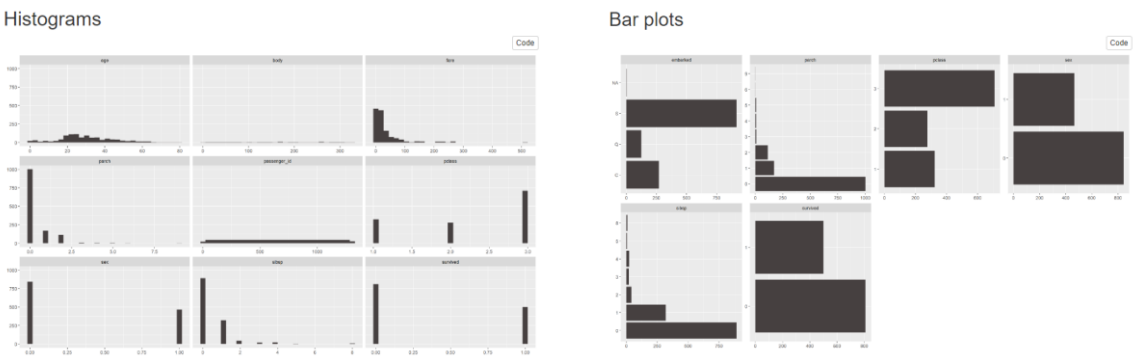
Code



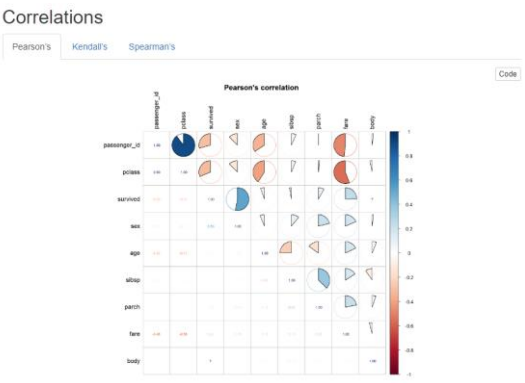
3. Individual observations, including clickable tabsets for each of the variables, containing information such as: minimum value, median and maximum value, histogram or bar plot, percentage of missing values, and the result of linear regression for numerical variables.

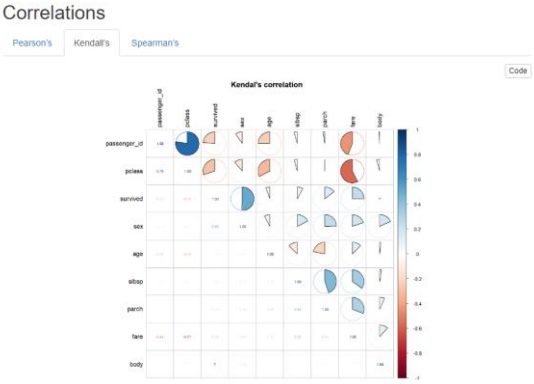


4. Histograms and bar plots, this section contains histograms and bar plots for all variables in the dataset.



5. Correlations, three different correlation matrixes are calculated: Pearson's, Kendall's and Spearman's.





6. Principal Component Analysis (PCA), plots displaying the Total Within Sum of Square and the Average Silhouette width for three different algorithms: PAM, k-means, and clara. These plots are useful in deciding on optimal number of clusters. Additionally, there are two plots representing the similar characteristics as before for the whole dataset that has been transformed into two dimensions. Last two plots present the clusters obtained from the two-dimensional dataset.