1 Topic

Checking missing and inconsistent values during importation process with Tanagra.

Up to the 1.4.41 version, Tanagra does not handle missing values because it seems interesting to force the students, which are the main users of Tanagra, to think about and to propose the most appropriate solution in relation with the characteristics of their dataset and the goal of their analysis. Thus, Tanagra simply truncates the file to import from the first obstacle. This treatment often disconcerts the users, especially since no error message was sent. They wondered why, whereas the conditions look right, the data were not properly loaded.

With the new version **1.4.42**, the importation of the text file format (tab separator), of the XLS file format (Excel 97-2003), and the data transfer using the add-in for Excel (up to Excel 2010) and LibreOffice 3.5/OpenOffice 3.3, have been modified. Tanagra reads all rows of the base. But it skips the incomplete rows and / or with inconsistencies (e.g. a column contains numeric value whereas this is a discrete attribute). And above all, **an explicit error message counts the number of deleted rows**. Thus, the users are better informed.

This very simplistic approach corresponds to the "listwise deletion" strategy for handling missing values. We know that it is not really a good strategy in most cases[1]. For us, the priority is primarily to alert the users on the problems encountered when reading the data file. They can use this default approach if it seems to be suitable. But they can also treat the missing values outside of Tanagra - with the appropriate approach - before importing again their dataset.

In this tutorial, we show the management of missing data when we send the data from Excel to Tanagra using the add-in Tanagra.xla. Some cells are empty into the Excel data range. This example illustrates the new behavior of Tanagra. We would get the same behavior if we import directly the XLS file or if we imported the corresponding file into the TXT format.

2 Dataset

We use the « ronflement_with_missing_empty.xls »[2] data file. There are 130 instances and 7 attributes. Some rows [14] comprise empty cells. They will be discarded during the data importation.

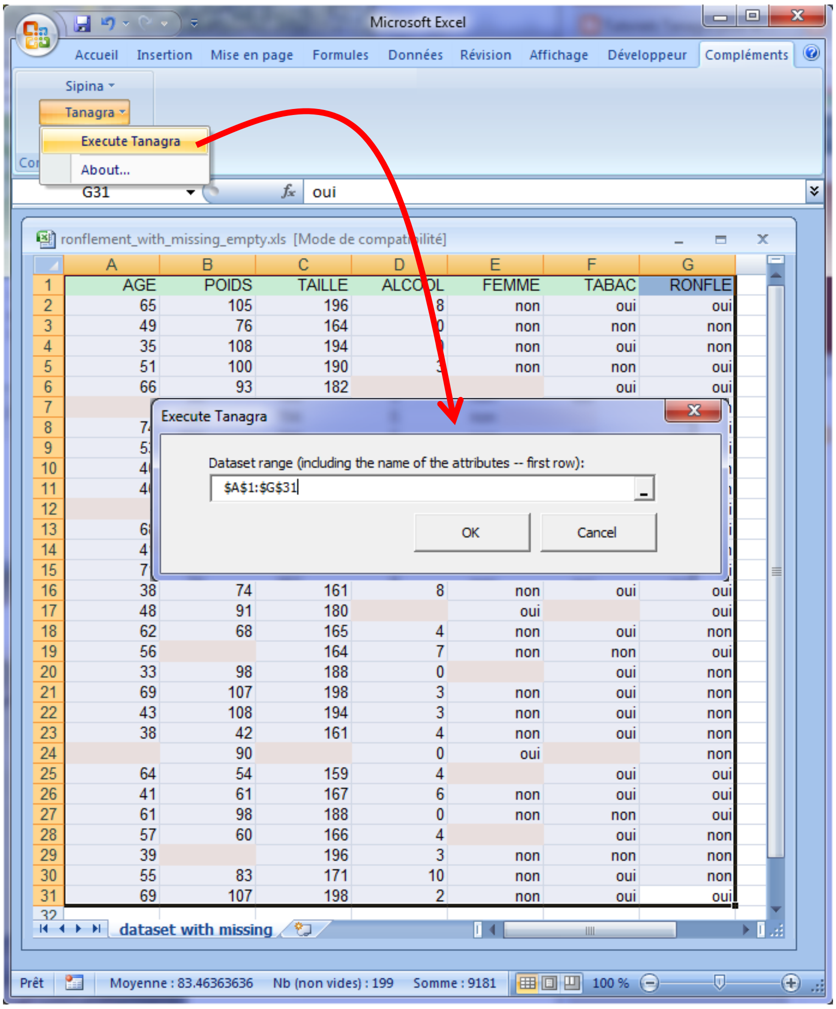
3 Transferring the dataset from Excel to Tanagra

We use the add-in “tanagra.xla” for the importation of the dataset[3]. We load the data file into Excel. The empty cells are colored in pink. Then, we select the data range and we click on the menu COMPLEMENTS / TANAGRA / EXECUTE TANAGRA (probably, you have the ADD-IN menu instead of COMPLEMENTS). We check the cells references. We validate the selection by clicking on the OK button.

[1] http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html

[2] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/ronflement_with_missing_empty.zip

[3] http://data-mining-tutorials.blogspot.com/2010/08/tanagra-add-in-for-office-2007-and.html (for Excel 2007, 2010); http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html (for the previous versions).



	A	B	C	D	E	F	G
1	AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
2	65	105	196	8	non	oui	oui
3	49	76	164	0	non	non	non
4	35	108	194	[illegible]	non	oui	non
5	51	100	190	3	non	non	oui
6	66	93	182			oui	oui
16	38	74	161	8	non	oui	oui
17	48	91	180		oui		oui
18	62	68	165	4	non	oui	non
19	56		164	7	non	non	oui
20	33	98	188	0		oui	non
21	69	107	198	3	non	oui	non
22	43	108	194	3	non	oui	non
23	38	42	161	4	non	oui	non
24		90		0	oui		non
25	64	54	159	4		oui	oui
26	41	61	167	6	non	oui	oui
27	61	98	188	0	non	non	oui
28	57	60	166	4		oui	non
29	39		196	3	non	non	non
30	55	83	171	10	non	oui	non
31	69	107	198	2	non	oui	oui

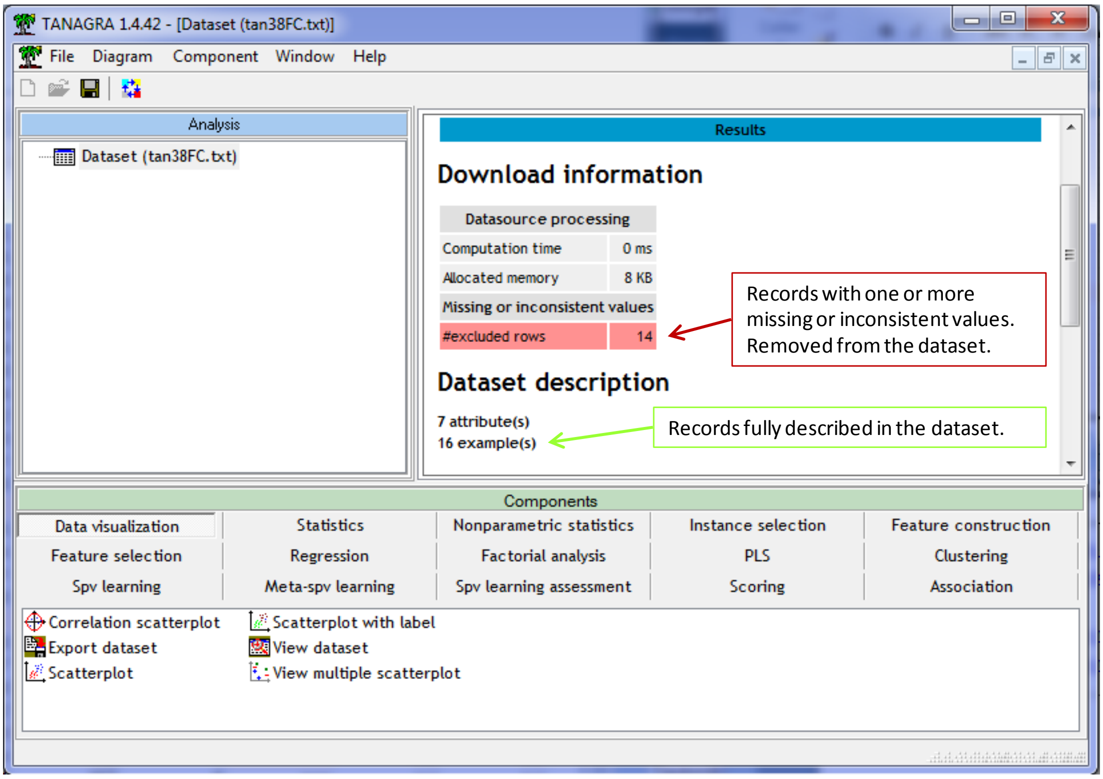
Execute Tanagra

Dataset range (including the name of the attributes -- first row): $A$1:$G$31

OK Cancel

Moyenne : 83.46363636 Nb (non vides) : 199 Somme : 9181

Tanagra is automatically launched. We observe that 16 instances are available, 14 are skipped.



Download information

Datasource processing	
Computation time	0 ms
Allocated memory	8 KB
Missing or inconsistent values	
#excluded rows	14

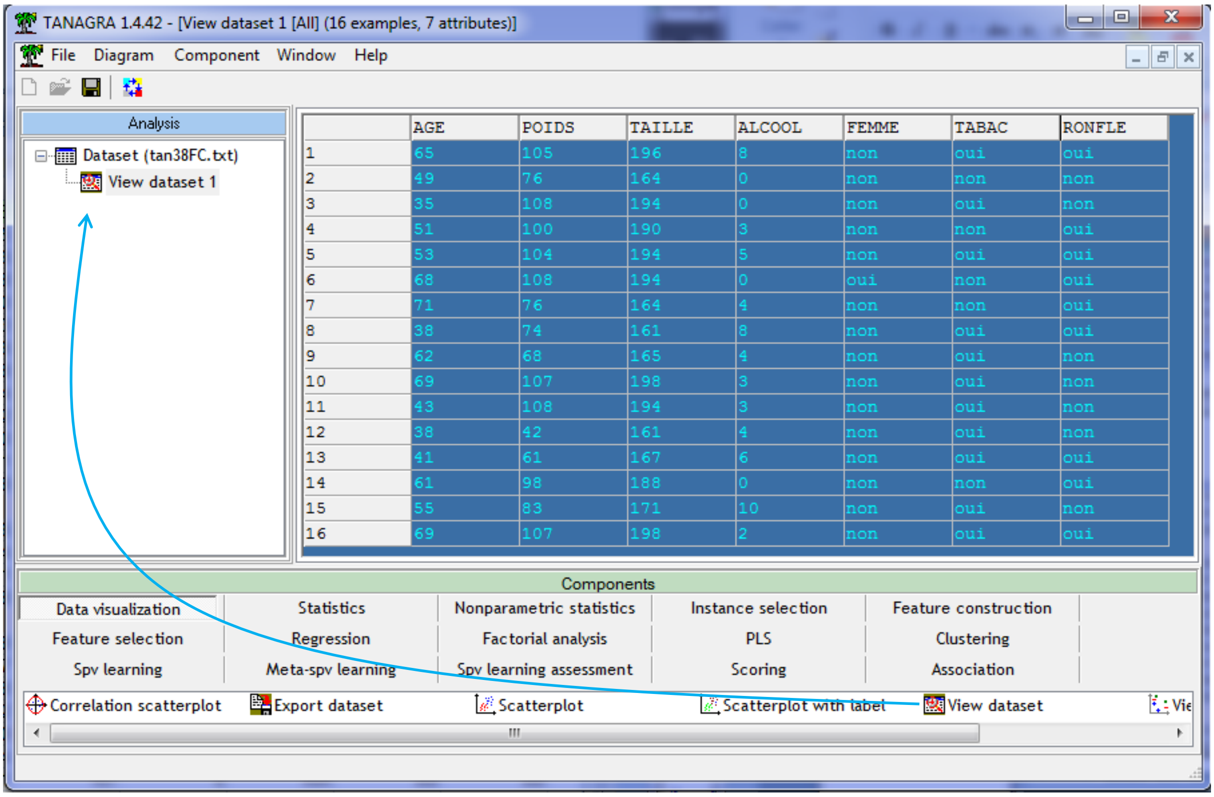
Records with one or more missing or inconsistent values. Removed from the dataset.

Dataset description

7 attribute(s)
16 example(s)

Records fully described in the dataset.

We can visualize the imported instances by using the VIEW DATASET component.



	AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
1	65	105	196	8	non	oui	oui
2	49	76	164	0	non	non	non
3	35	108	194	0	non	oui	non
4	51	100	190	3	non	non	oui
5	53	104	194	5	non	oui	oui
6	68	108	194	0	oui	non	oui
7	71	76	164	4	non	non	oui
8	38	74	161	8	non	oui	oui
9	62	68	165	4	non	oui	non
10	69	107	198	3	non	oui	non
11	43	108	194	3	non	oui	non
12	38	42	161	4	non	oui	non
13	41	61	167	6	non	oui	oui
14	61	98	188	0	non	non	oui
15	55	83	171	10	non	oui	non
16	69	107	198	2	non	oui	oui

4 Conclusion

Compared to the previous versions of Tanagra, the importation time is increased because of these additional checks. We note however that the increase is negligible on most of the datasets. It is only discernible on large databases. To give an order of magnitude, when we import a dataset with 300,000 rows and 122 attributes[4], the importation time which is 9 seconds on Tanagra 1.4.41 becomes 11 seconds on Tanagra 1.4.42. But we note that the importation time remains very low compared with the other data mining tools (on the same dataset: 34 seconds for Knime 2.4.2; 51 seconds for R 2.13.2; 63 seconds for Weka; etc.).

[4] http://data-mining-tutorials.blogspot.com/2012/02/logistic-regression-on-large-dataset.html