Data Preparation

For each attribute get the frequency of each value;

We know that in the current dataset (titanic), there are 400 rows. The problem with this dataset is that few records presents corrupted data. An example is the row 121, that has the string ‘male’, in the id of the class.

Click on column class -> Facet -> Text facets.

[1st](javascript:%7b%7d) 120

[2nd](javascript:%7b%7d) 152

[3rd](javascript:%7b%7d) 118

[male](javascript:%7b%7d) 1

[(blank)](javascript:%7b%7d) 2

Analyzing the dataset, we clearly see that the original Columns Gender and Age has been swapped. That’s why the column ‘Gender’ was initially containing the values: ‘Adult’ and ‘Child’;

Frequency in base to the Age columns:

[adult](javascript:%7b%7d)287

[child](javascript:%7b%7d)113

For the Gender, just 289 records out of 400 contains a valuable and clear data. In fact, the other 17 records are null, or contains numerical value like 1.0, 2.0 or a text value like:’ Male’;

In according a the dataset, there were in the titanic:

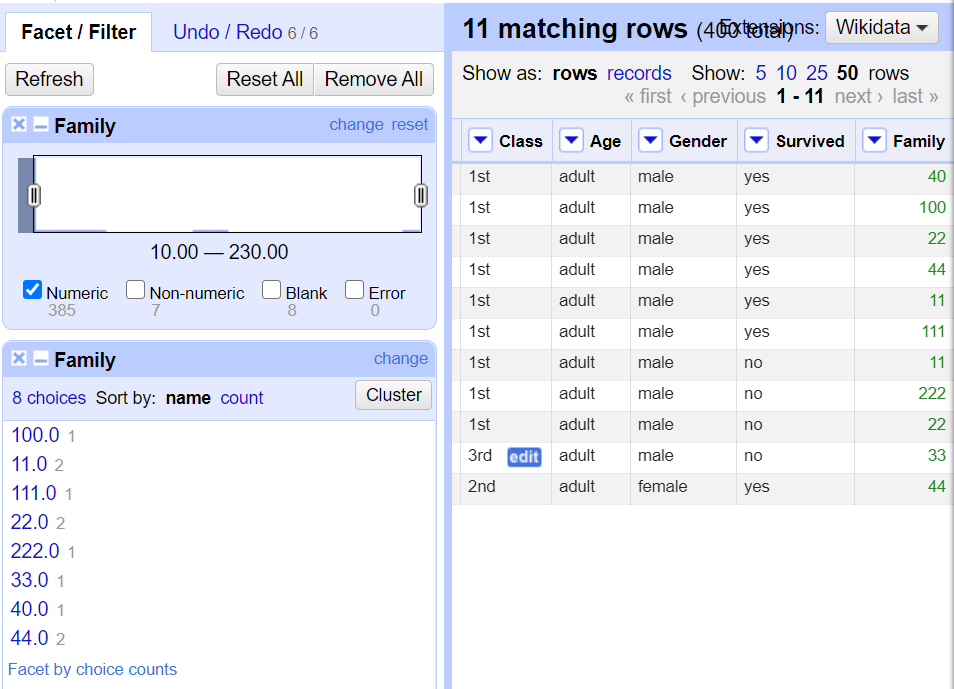
* Male: 249
* Female: 140

For the ‘survived’ columns, 392 records out of 400 contains a clear value. In According to the dataset we had:

* Survived: 213
* NOT SURVIDED: 178
* Unknown (null cell): 1

8 cells In the survived column contains a value: 2.0 (not defined);

In is not clear what the Family columns want to represent. It should be the number of members travelled in a family. It looks like there are different outliers in the family Column, because 374 out of 400 where travelling with a number of family members between 0 and 10.00, but in the dataset there are 11 records where the number of family members has a value between 11 and 222.



From OpenReview, How do I export is as ARFF by the user Generated Method?

Difference between GoogleOpenRefine and Weka.

Transform all values in cleaned titanic into nomical values

* Open Weka: load the cvs/excel file/click the botton choose(Filter) -> unsupervised -> NumericToNomianl -> Apply;

Transform all values in cleaned titanic dataset into numeric values.

In traffic datast, transform casuality class (CASU\_CLS), spped limit (SP\_LIM) and Age of casuality (AGE)CASUY) FOM NUMERIC TO Nomnal values.

Convert ALL the dataset at the end to ARFF format by the system generated method, lead it to Weka and examine the dataset.

For the following 3 exercises, Python is needed within OpenRefine.

10. Transform all values in the cleaned Titanic Revised dataset into nominal values

11. Transform all values in the cleaned Titanic Revised dataset into numeric values

12. On the cleaned Traffic dataset, transform values on casualty class (CASU\_CLS),

speed limit (SP\_LIM) and Age of Casualty (AGE\_CASU) from numeric to

nominal values. (Note: values might need to be categorised).

13. Convert the datasets you cleaned/transformed in questions 10, 11, 12 and 13 to

ARFF format by the system generated method, load it to Weka and examine the

dataset.

What is the metadata of a dataset/database.

