Data cleaning

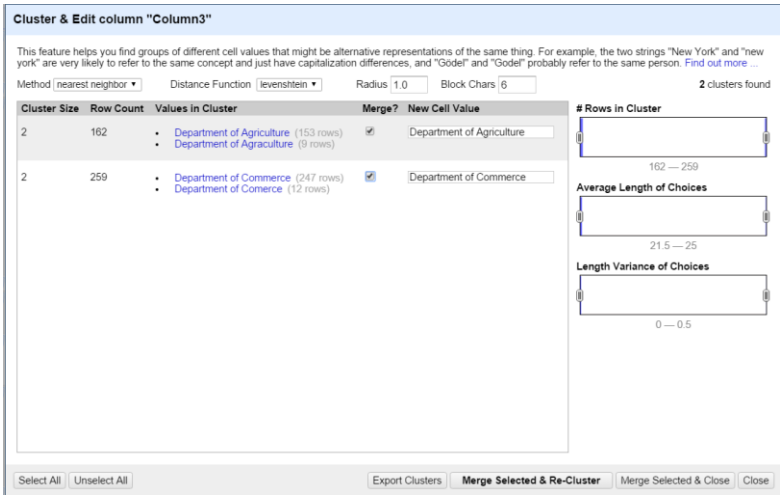
Name: Xinyu Hu
Student ID: 29350379
Email: xh2n17@soton.ac.uk

Overview of the dataset

This dataset is about the US government spending on the project. However, it contains so many errors. In order to make this dataset available for further use, we have to detect all the errors and find out solutions to fix them. Here are some errors I find:

1. *Multiple representation problem in 'Department Name'*

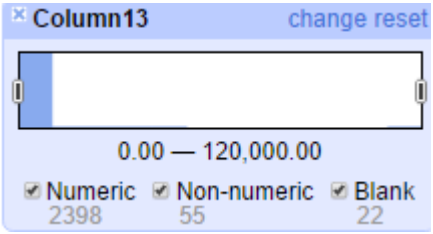
When we go through the first few rows in the 'department name' column, we can discover there are some abbreviations like 'DoA', the 'Agriculture Department' or the 'Department of Agriculture', which all represent the same meaning and should be combined as one.
Solution:
We first apply the 'trim leading and trailing whitespace' method this column, and then access the 'cluster' bottom in the facet:



Finally, edit the abbreviations to the full name in the facet.

2. *Redundant information and mixed use of numerical scales in 'Lifecycle cost'*

As we can see, there are some redundant information like the '($m)' appearing so many times. Since we default all the numbers in the column are in the units of '$million', we have to accurately point out each of them, and delete the '($m)' by using the tool.
Solution: as we apply the numeric facet to this column, we find out that there is no numeric cell, so we first apply ().toNumber action.



Then in order to get rid of the data with'($m)', we use the value.replace() function to deal with.

3. Mixed formats in the 'Completion date'

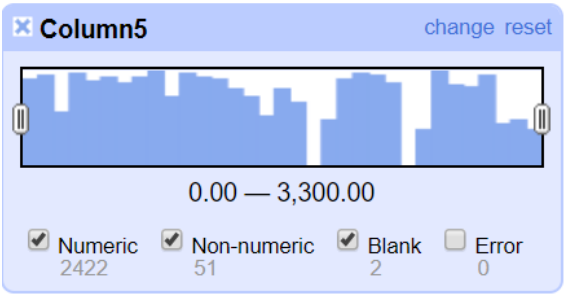
The mixed format like '31/03/2012' and '2012-30-09' are quite obvious in the 'Completion date' column. To deal with this problem, we just transfer the one with '-' to the one with'/', as the slash one is more common to see.

Solution: use the method toString(toDate(value),"dd/MM/yyyy") to make the transformation.

4. Mixed formats in the project ID

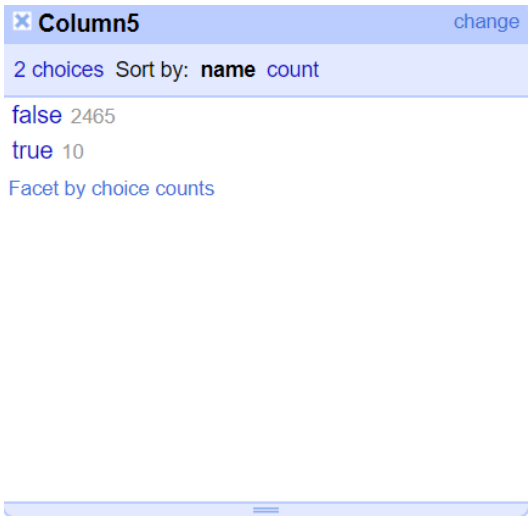
When we check the inconsistency in the 'project ID' column, it's easy to there are different kinds of formats like'1,923' and '1025'. To uniform the data in this column, we have to remove the comma.

Solution: we first check the numeric facet of this part, realizing that there are 50 non-numeric rows. Therefore we use the value.replace() function to cope with it.



5. Duplicated formation in the project ID

Noticing that each ID may be unique, we have to exclude all duplicated ones. To complete this step, we first check how many rows are identical. And then use the 'Blank down' method to cope with it.



6. Summation records

For the summation records in the dataset, we just delete them directly.

Solution: to remove them, we first choose the text facet in the second column, and then 'star' them all and choose the 'remove all matching rows'.