# A Case Study: Data Cleaning – Google-Refine/OpenRefine

## Objectives

**Input:** a set of raw data

**Output**: a set of “cleaner” data.

**A Case Study**: Processes will be described to show the methods of cleaning data by using **Google Refine/OpenRefine**. Dirty data types considered include: missing values, use of wrong data type (numeric/strings), meaningless values, set violation, erroneous entry.

Google Refine supports the following data formats:

* + TSV, CSV, or values separated by a custom separator you specify
  + Excel (.xls, xlsx)
  + XML, RDF as XML
  + JSON
  + Google Spreadsheets
  + RDF N3 triples

It is suggested that a header row need to be inserted into the file to indicate the meaning of each column, since by default the first row will be treated as the header row: ***Parse next 1 row as column headers*** (see figure 2).

**Dataset:** Titanic Data (with errors)

|  |  |
| --- | --- |
| **Attributes** | **Value** |
| **Class** | **Cabin class, 1st, 2nd or 3rd** |
| **Age** | **adult or child** |
| **Gender** | **male or female** |
| **Survived** | **yes or no** |
| **Family member** | **The number of members in the family, 1 – 4** |

## Introduction to OpenRefine

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; extending it with web services; and linking it to databases like [Freebase](http://www.freebase.com/).

Go to website: <http://openrefine.org/index.html>

to watch the three videos: Explore Data, Clean and Transform Data and Reconcile and Match Data.

## Installation of OpenRefine

Download the stable version – Google Refine 2.5 from this website: <http://openrefine.org/download.html>

Follow the instruction from the following website to install and run OpenRefine: <https://github.com/OpenRefine/OpenRefine/wiki/Installation-Instructions>

## A Case Study

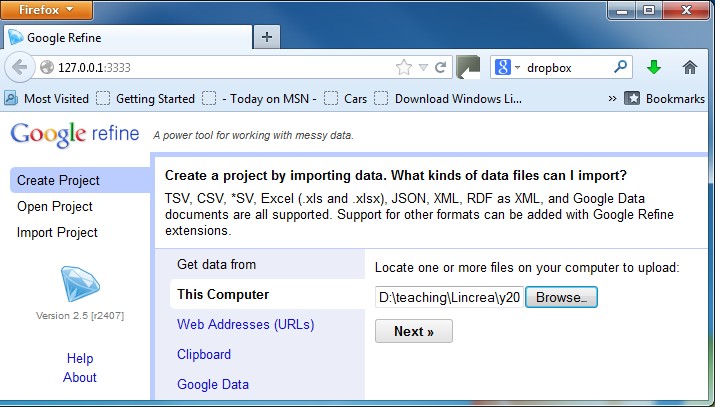
This section shows the procedure of cleaning some of the errors in the Titanic dataset. Step 1: Import data; Click the Browse button to select a file, then click **Next**;

Figure 1 Load a file

Step 2: Name a project: In the Project Name input box, type in a name for the project, e.g., Clean Titanic Data, then click **Create Project**;

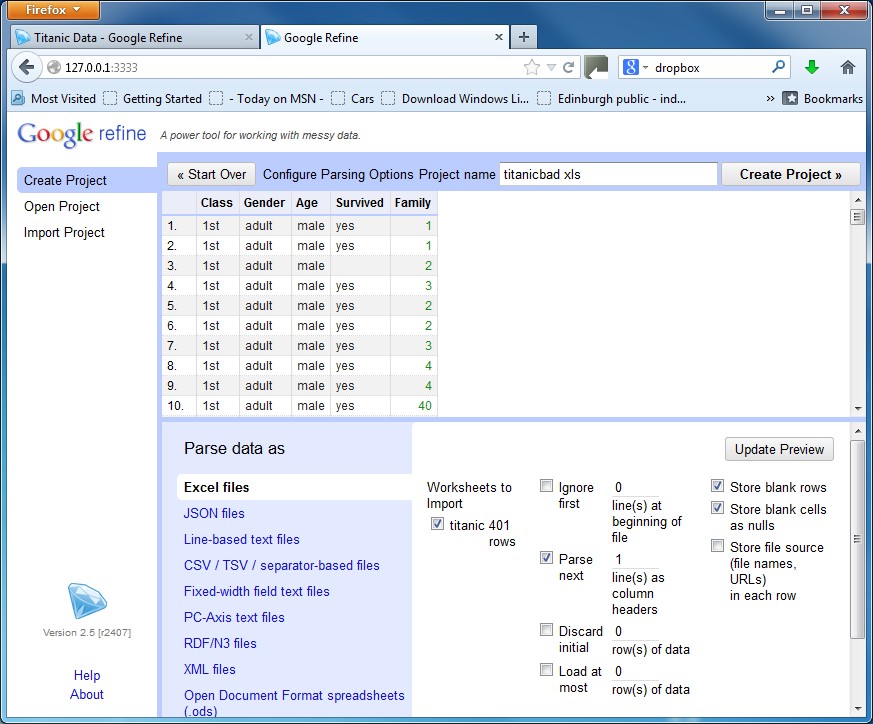


Figure 2 Create a project with options selected

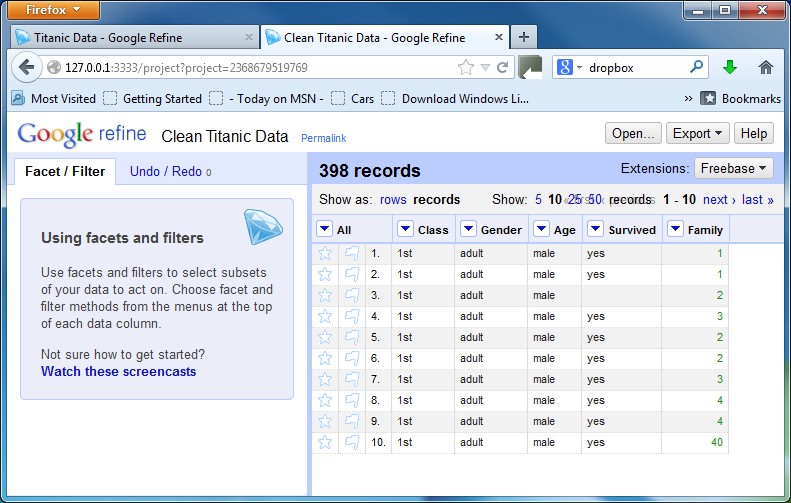


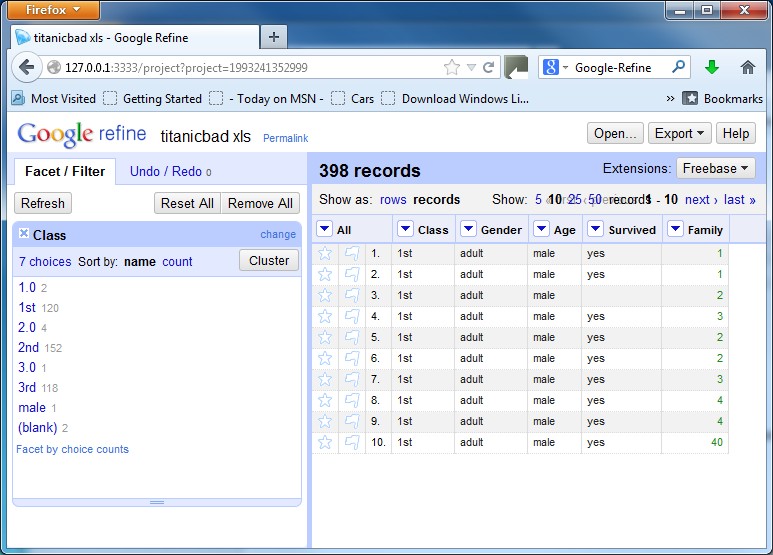
Figure 3 A project is named and created

Step 3: Identify dirty data: Use Facet to check each column; The process can start from any column.

1. **Column 1**. Since the value for this column is a string, the text facet is used: from the drop-down menu in front of “Class”, select Facet Text facet

The result is “facet by choice account”, which summarises the cells in that column to give you a big picture on that column, and allows you to filter some subset of rows for which their cells in that column satisfy some constraint. A choice means a unique value in that column. The count is the number of occurrence of a choice. For example, value 1st is one of the choices and occurs 120 times (the count is

120) in that column.

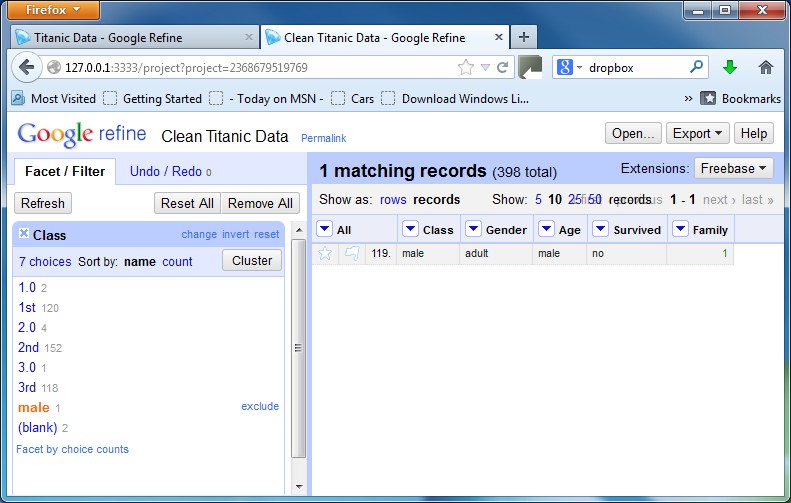


There is a panel with two options, Facet/Filter and Undo/Redo, at the left. When the option Facet/Filter is on, it shows the result of the text facet. The value of the choices is ordered in the alphabetical order with the number of occurrence of the choice next to it. It shows that among the 398 records, there are 7 different choices in the dataset. It also shows the number of occurrence of each value. For example,

value 1st is one of the choices and occurs 120 times in this column.

In the panel, dirty data can be identified, e.g., missing data ((blank) 2), erroneous data (male 1).

Click the (male) choice in the panel to edit these rows with (male) value:



There is 1 matching record appearing in the data panel.

Looking through the Facet/Filter panel, the following erroneous values can be identified:

1.0

2.0

3.0

male

It is clear that these are entry errors. The correction should be:

1.0  1st

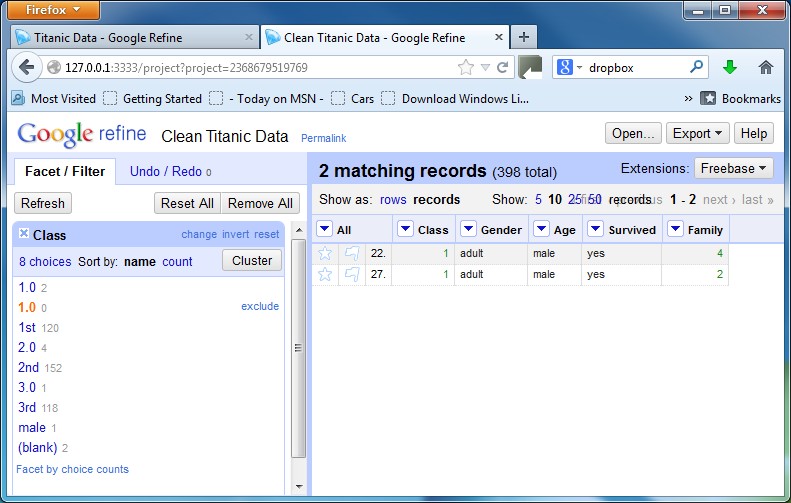
2.0  2nd

3.0 3rd

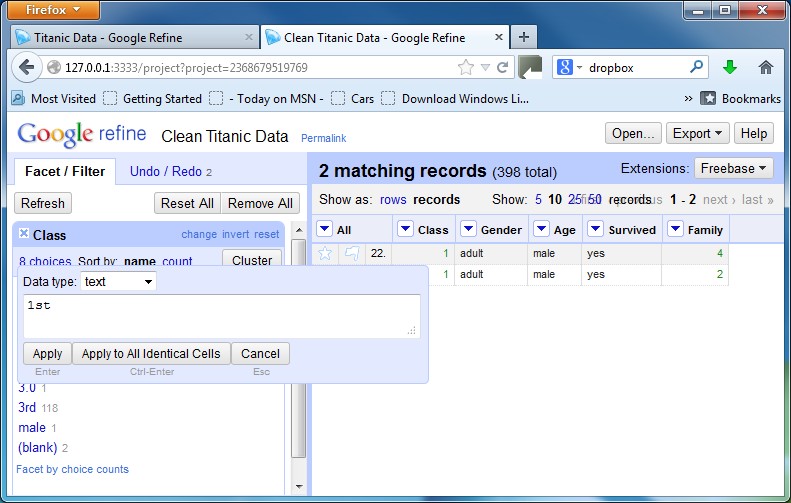
Male  2nd (a guess)

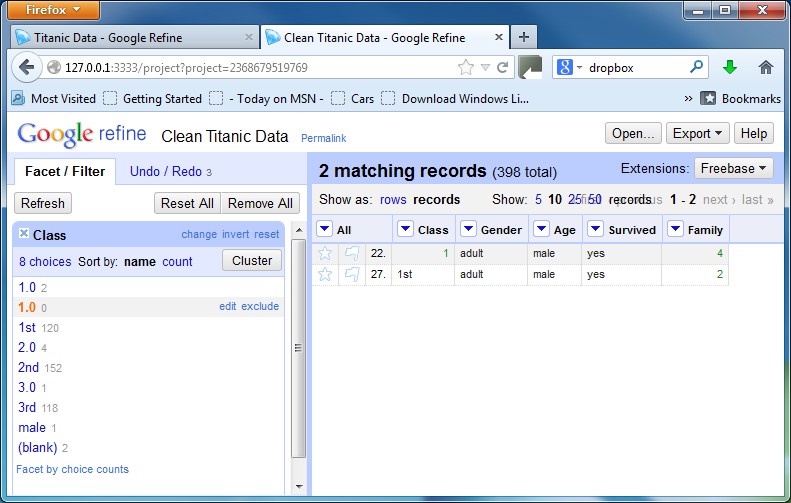
The change to 1.0 is done by

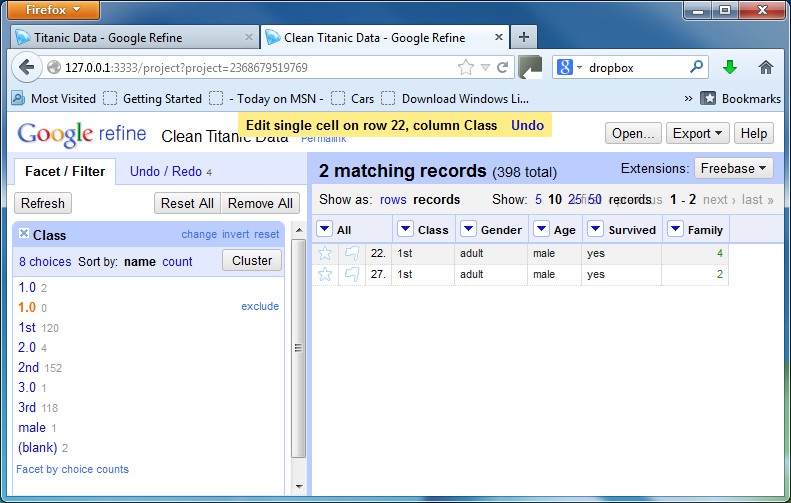
1. Select the row that has 1.0 as a value by clicking 1.0 in the Facet/Filter panel at the left:



1. Edit the value by clicking the cell and change the value to 1st, select the data type as text, then click **Apply**.

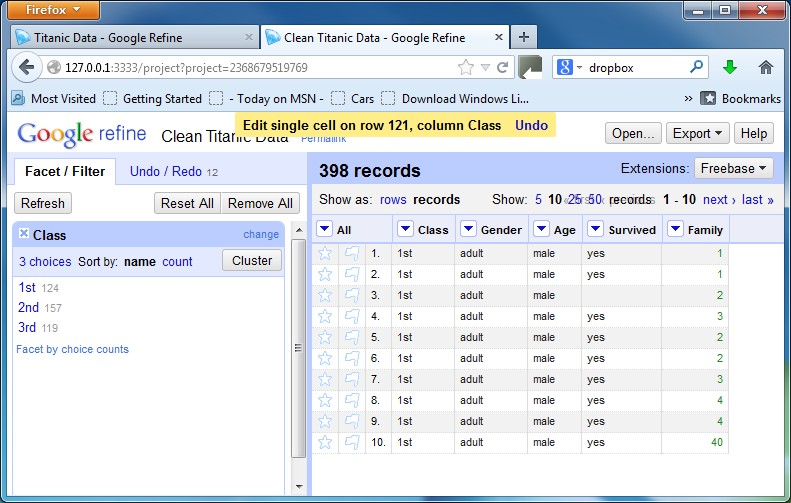




Use the same way to correct the value at row 22.

The change to others can be done similarly.

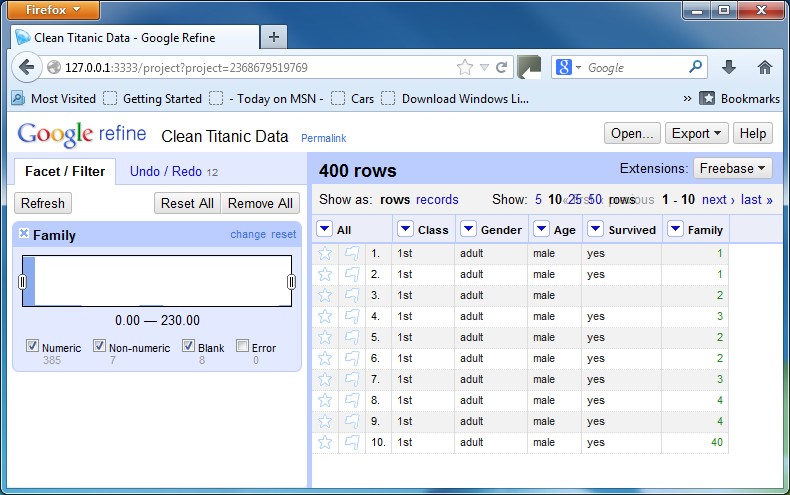
To deal with missing values, expert’s knowledge is required.

There are two missing values for Class. Looking through the Facet/Filter panel, it is not clear which value should be added. So we can randomly insert value 1st for these two cells. Then click the (blank) in the panel:

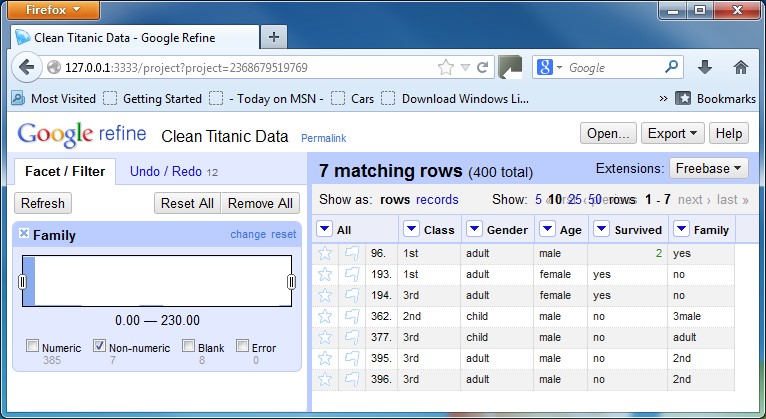
Column 1 is cleaned

## 2. Column Family

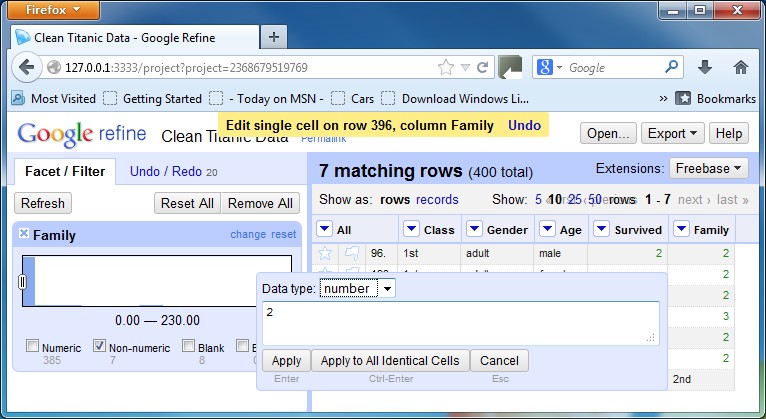
Since the value for this column is numeric, the numeric facet is used first: from the drop-down menu in front of “Family”, select Facet Numeric facet

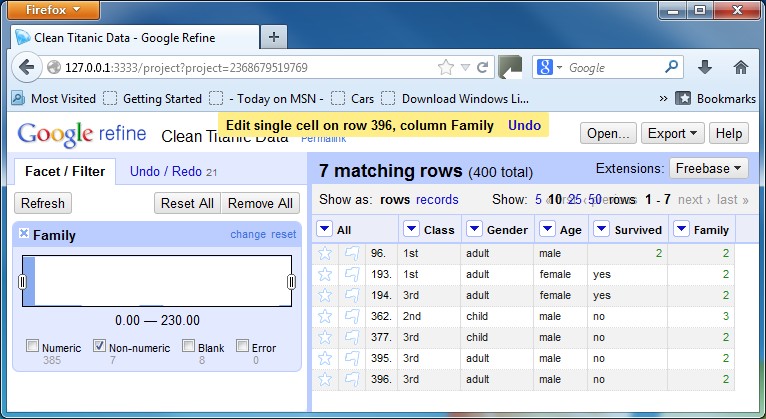


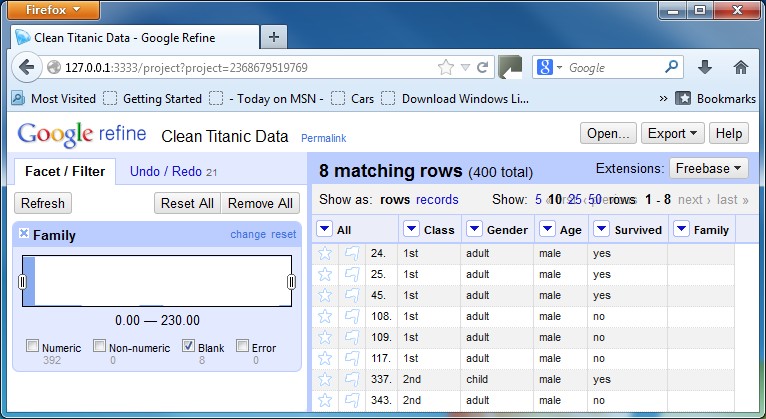
From the Facet/Filter panel at the left, you can see that there are 385 numeric values, ranging from 0.00 to 230, except 7 with a non-numeric value and 8 with blank. Un-tick the box next to the Numeric and blank, you have



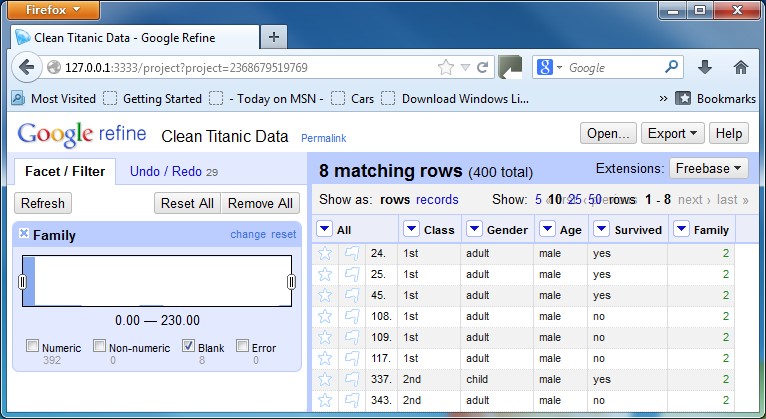
It is clear that the values 3male, 2nd and 2nd should be 3, 2 and 2 respectively. However, the values yes, no and adult should be corrected based on domain knowledge. In this case, we just replace them randomly by 2. After making the changes, don’t forget to change the data type to number, you have



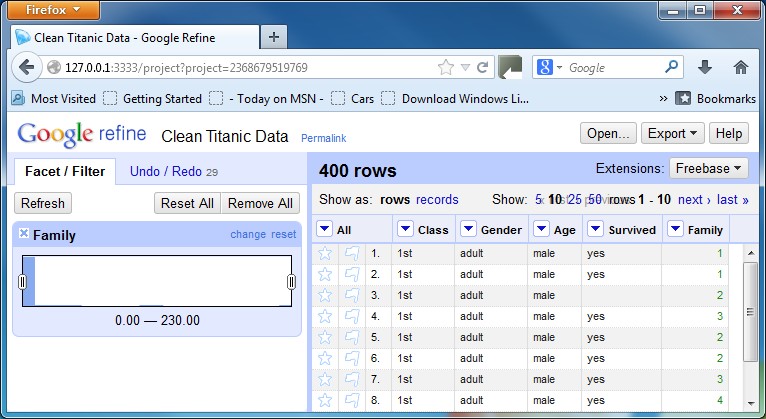


Then tick the box next to Blank,

Again, domain knowledge is needed. In this case, we just replace them randomly by 2.

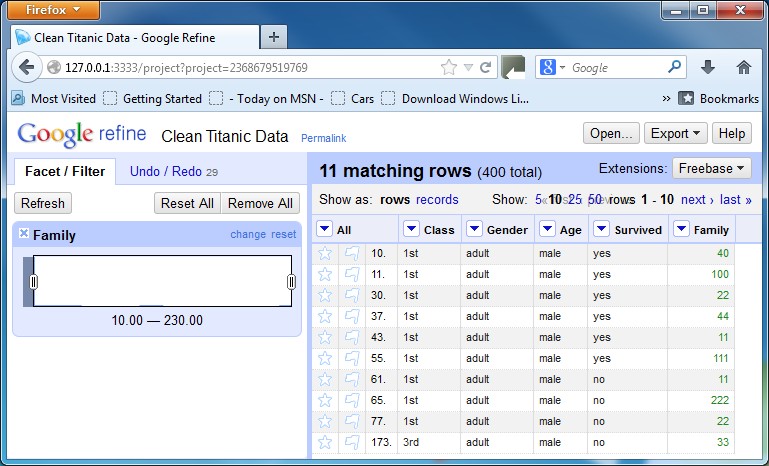


Finally tick the box next to Numeric.



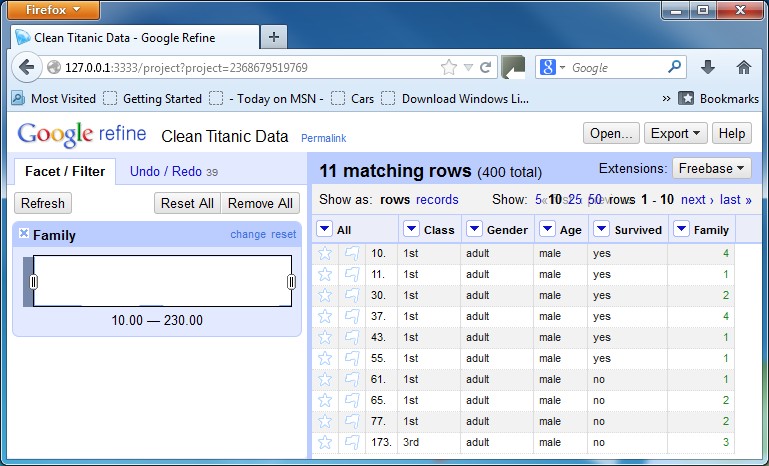
All values are numeric now.

Other than using the Text facet (see above sections), you can also identify outliers by moving the scaling bar on the diagram in the Facet/Filter panel toward left or right. Moving the left bar toward right, you can get the value range from 10 to 2300. There are 11 records within this range, which are outliers.

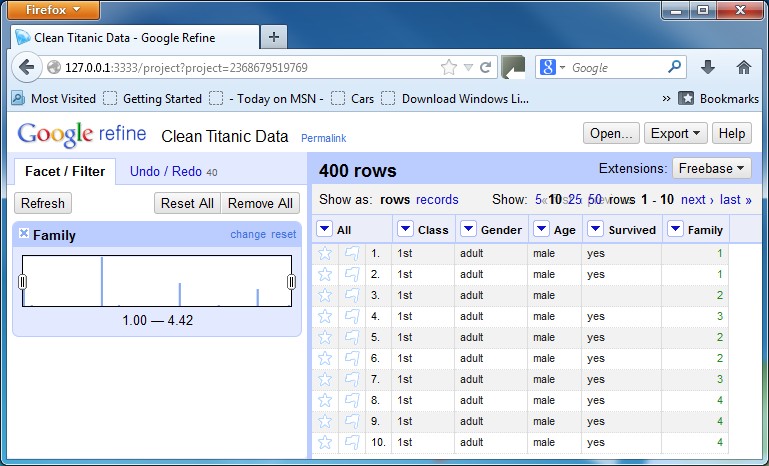


The following change can be made:

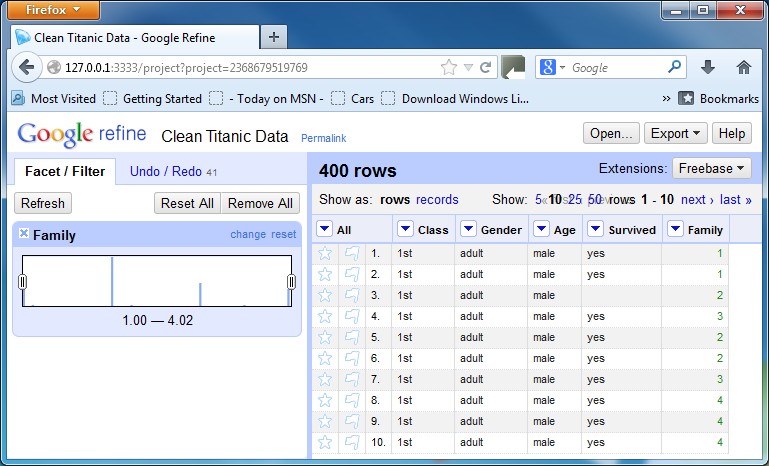
40  4; 100  1; 22 2; 44  4; 11  1; 111  1; 222  2; 33  3.



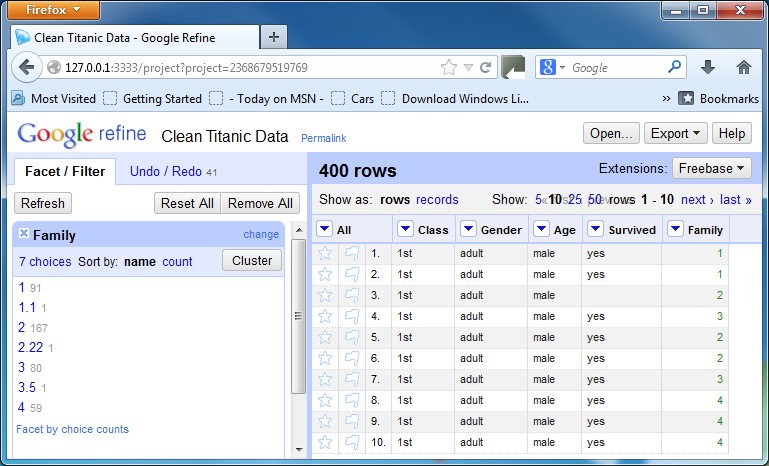
Click Refresh at the top left of the left panel, you have



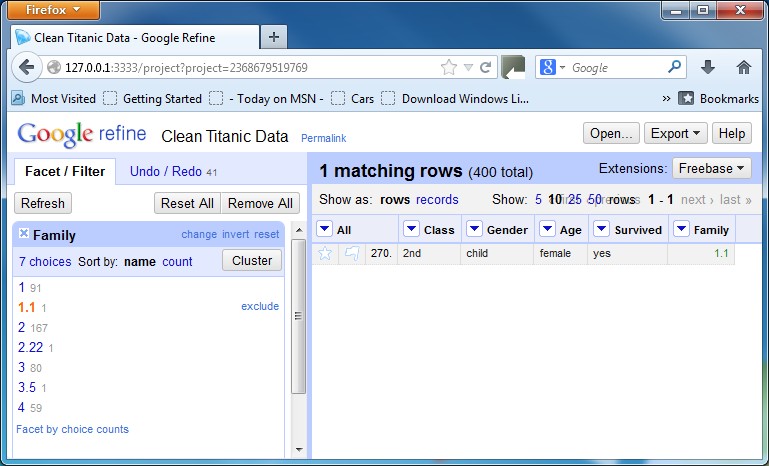
Now the value range is from 1.00 to 4.42. Moving the left bar toward right until 4.08, there is one record with a value of 4.4, which should be 4. Make a change, click Refresh and moving the left bar back:



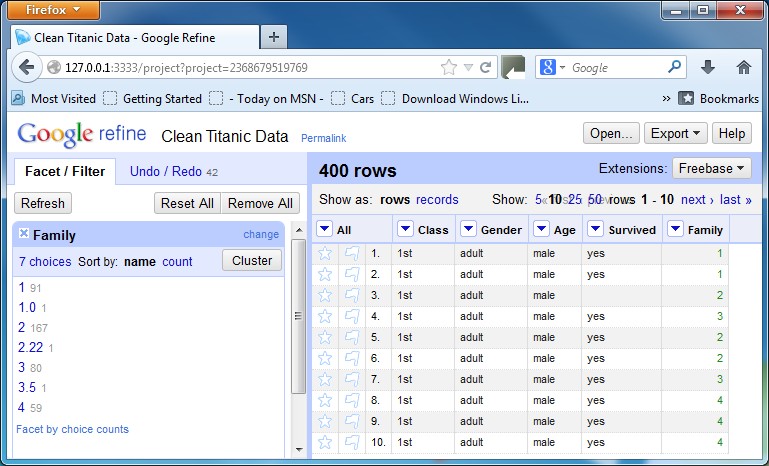
Now the value range becomes from 1.00 to 4.02. We can use Text Facet to check whether there is any more outlier existing. From the drop-down menu in front of “Family”, select Facet Text facet



The values with decimal places should be incorrect. In the left panel, click value 1.1:

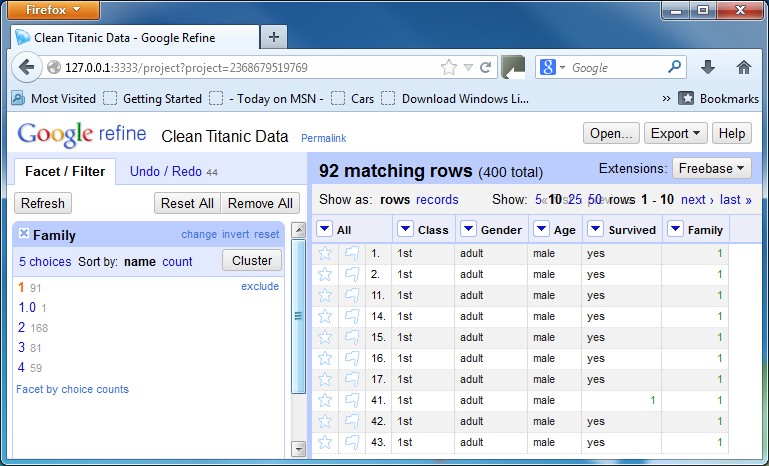


Change value 1.1 to 1, and click 1.1 (in red) again:



**Note:** There is a choice with a value 1, and a choice with a value 1.0. These two should refer to the same choice 1. The 1.0 appears because when you made the change, you selected the data type as numeric, others might be just text. If you change it to text from numeric, the 1.0 should be merged to 1.

Similarly, you can make the following change: 2.22  2; 3.5  3



Columns Gender, Age and Survived can be cleaned in the similar way.

At the end, you can export the processed data for further cleaning at the next stage. To export the processed data, click the **Export** button at the top right, select **Excel**, then name, locate and save the cleaned dataset.

# Other types of errors

1. **Missing ID or Empty rows**. If there are any rows missing ID or empty in your dataset, you might want to remove them from your dataset. You can either remove them from excel or within Google-Refine. To do this in Google-Refine, follow:
   * flag those rows that you want to remove by using the left-most drop-down menu in front of “All” and invoke *Edit rows  Flag rows*.
   * From the left-most drop-down menu in front of "All", select Facet Facet by flag. Then select ***true*** in the Facet/Filter panel to show only the row(s) to remove
   * invoke Edit Rows > Remove all matching rows

## Duplicates

The basic idea of dealing with duplicate data is to merge all the columns into one column in an Excel file, and then do a check using Google Refine.

Step 1: Save the excel data to a CSV (Comma Delimited) file; Step 2: Load this file into a text editor, such as Microsoft Word;

Step 3: Copy the data in the CSV file to a single column in the Excel file; Step 4: Load the Excel file to Google Refine for processing;

Step 5: Use Text Facet to list all choices. If there is any choice that occurs more than once, there is a duplicate.

**Note**: When selecting the Text Facet, if it shows in the Facet/Filter panel that “too many count choices for display”, you might need to low down the default count numbers. By default, it is 4000.

# Reading:

* + Overview -- There are three videos - <http://openrefine.org/index.html>
  + Documentation wiki - <https://github.com/OpenRefine/OpenRefine/wiki>

o Including installation instruction.