**Data Cleaning**

**A Case Study**

# Objectives

* Given a dataset, explore how data cleansing can be performed
* Input: a set of raw data
* Output: a set of “cleaner” data

**Note**: The “cleaner” data means that all data in a column match the metadata description of that column

# Data Cleansing

**Aim**: detect, correct errors and inconsistencies from

data in order to ensure data quality

**General Approach**:

* + Identification of error types
  + Selection of Algorithms
  + Selection of Methods/Approaches
  + Correction of errors

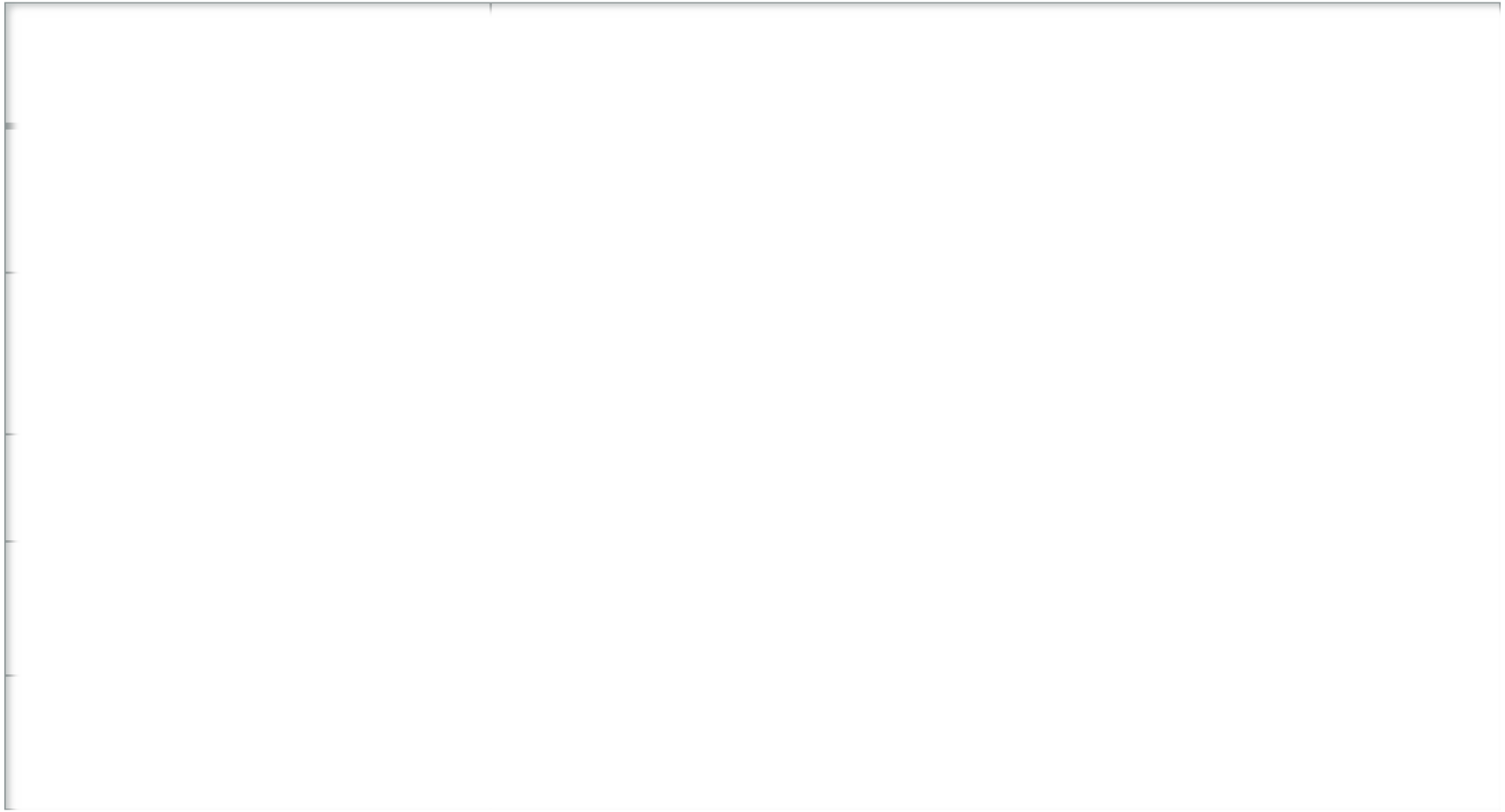
# A Case Study

* Titanic data
* 5 Attributes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Age** | **Gender** | **Survived** | **Family member** |
|  |  |  |  |  |

* *400 rows*

# Attributes



|  |  |
| --- | --- |
| **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** |

**Dirty Data Types**

* missing values
* use of wrong data type (numeric/strings)
* meaningless values
* set violation
* erroneous entry
* duplicates

# Tool – OpenRefine

OpenRefine supports the following data formats:

* TSV, CSV, or values separated by a custom

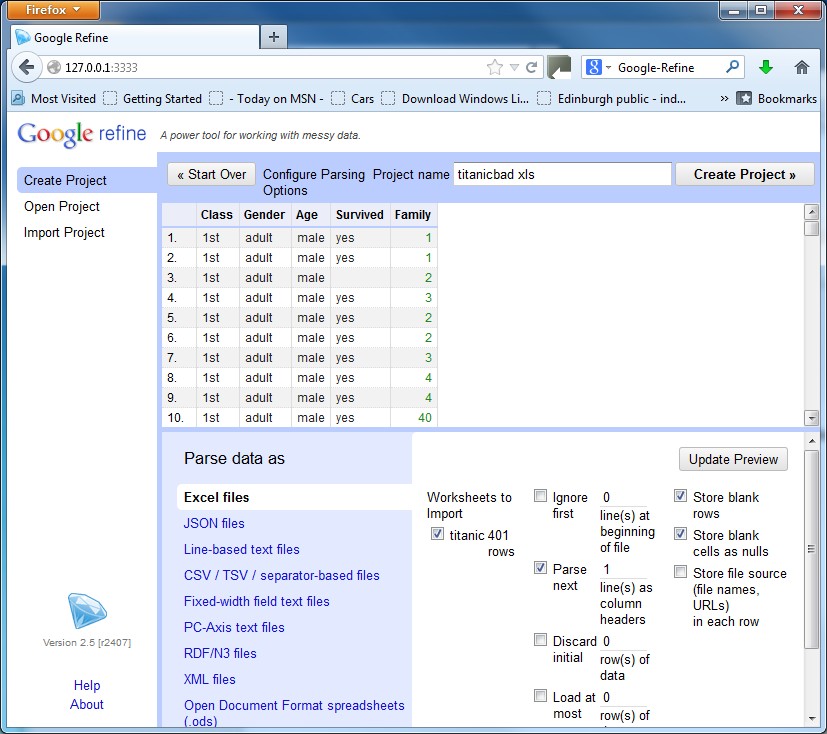
separator you specify

* Excel (.xls, xlsx)
* XML, RDF as XML
* JSON
* Google Spreadsheets
* RDF N3 triples

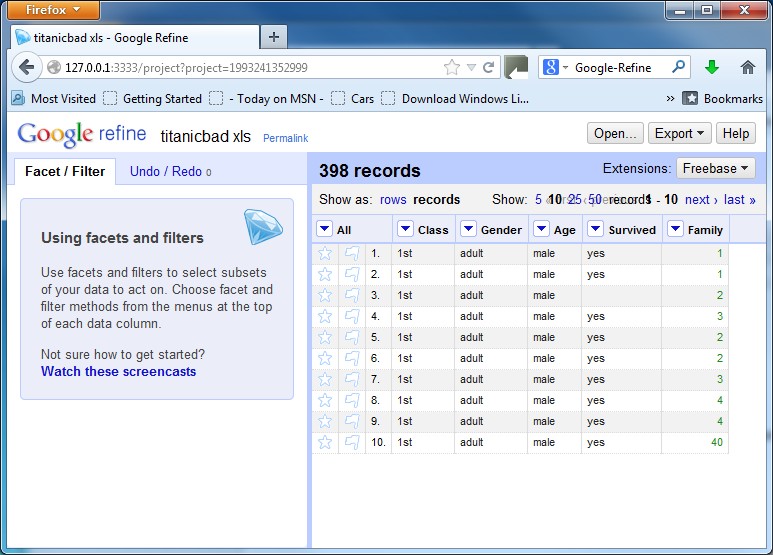
# Prepare the data

* It is suggested that a header row need to be inserted into the file to indicate the meaning of each column
* by default the first row will be treated as the header row: Parse next 1 row as column headers

# Create a Project



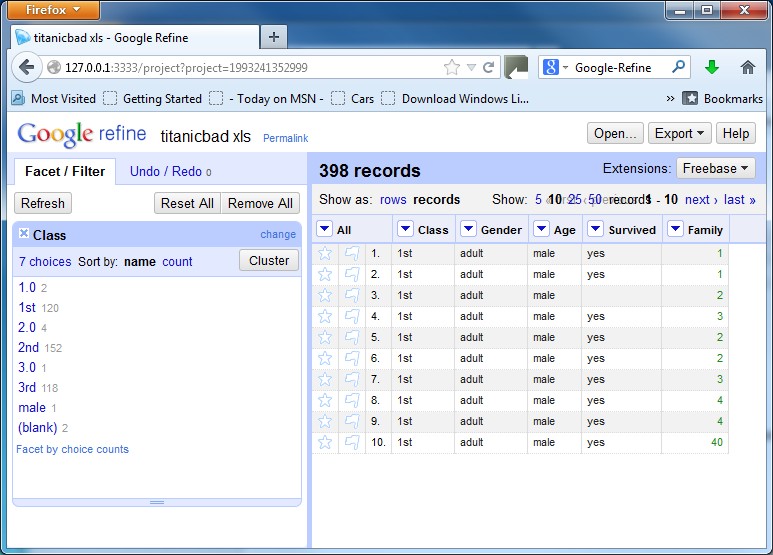
**A Project Created**



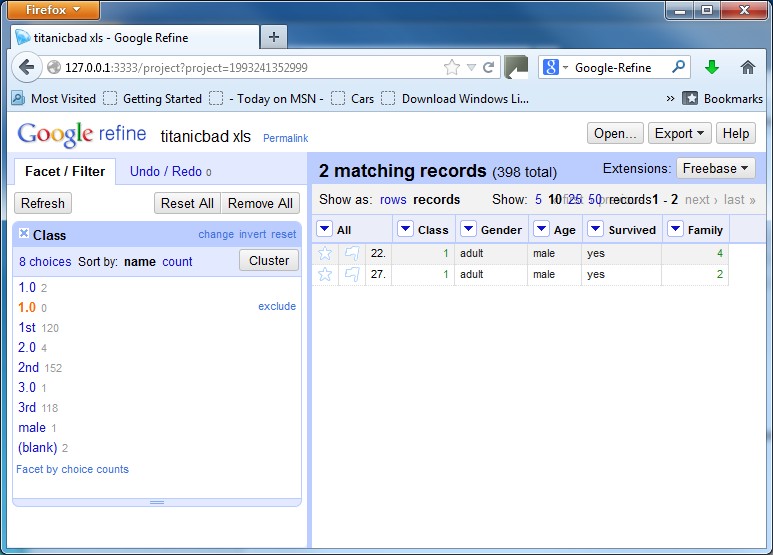
**Identify Dirty Data**

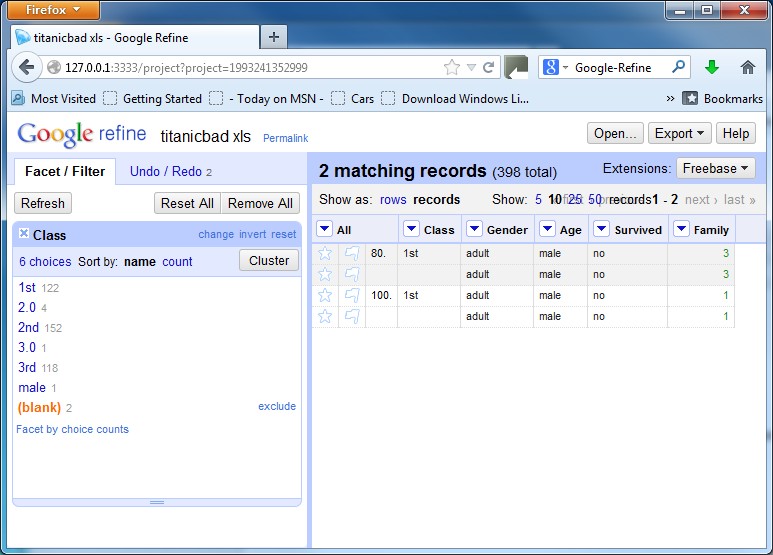
Use Facet to check each column

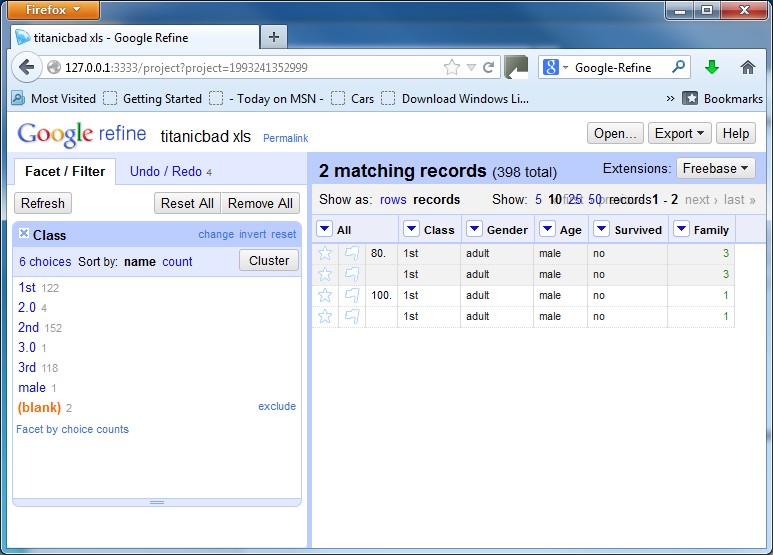
* Text facet
* Numeric facet

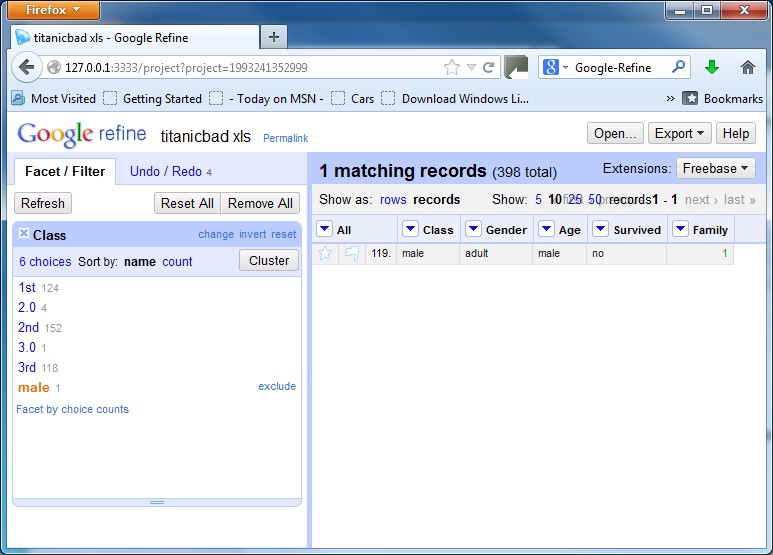
Use Text facet

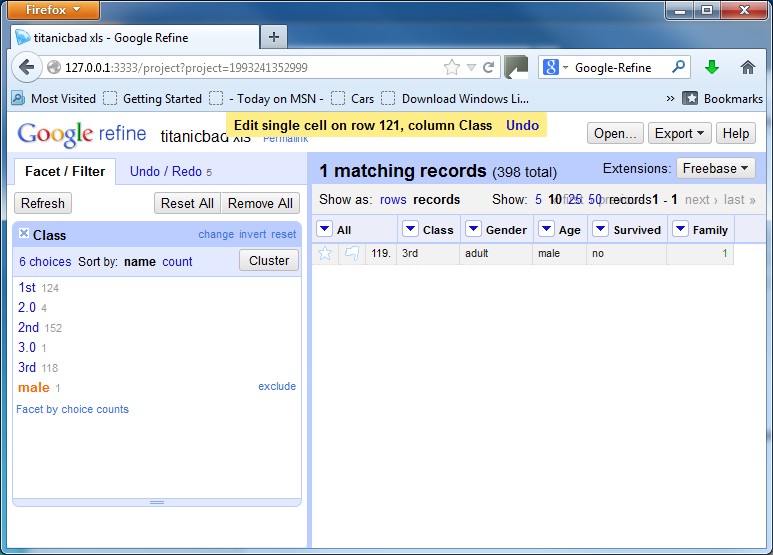
# Identify Dirty Data – Wrong Type





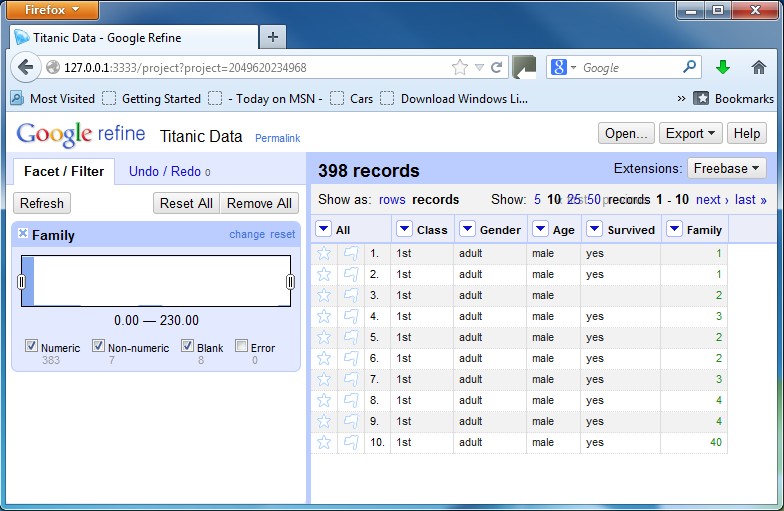




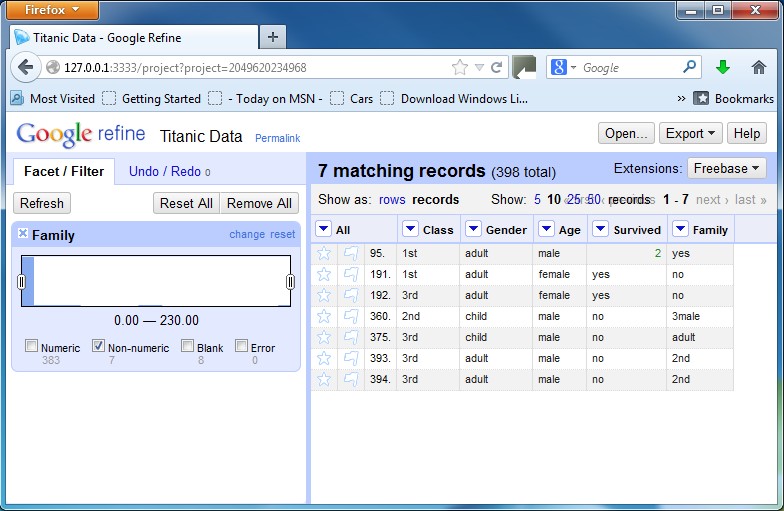


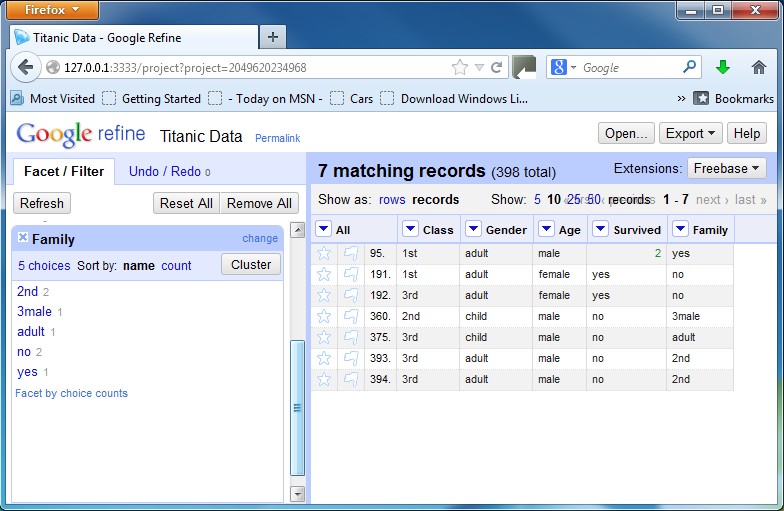
**Identify Dirty Data – Family**

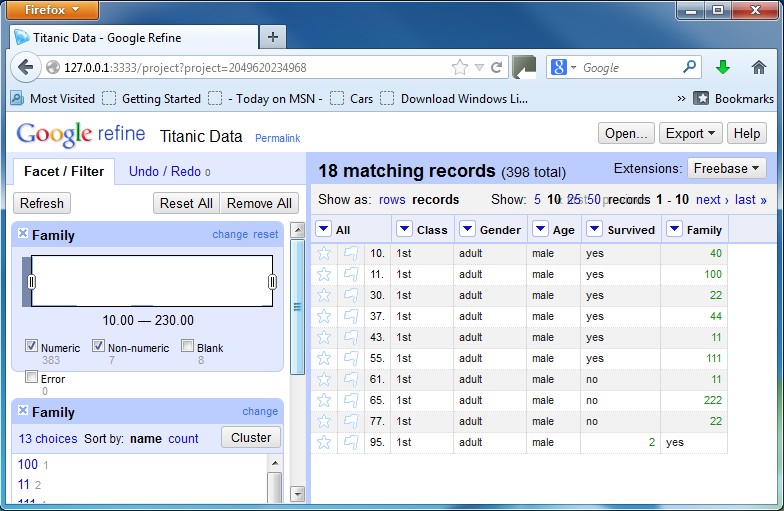
**Numeric facet**

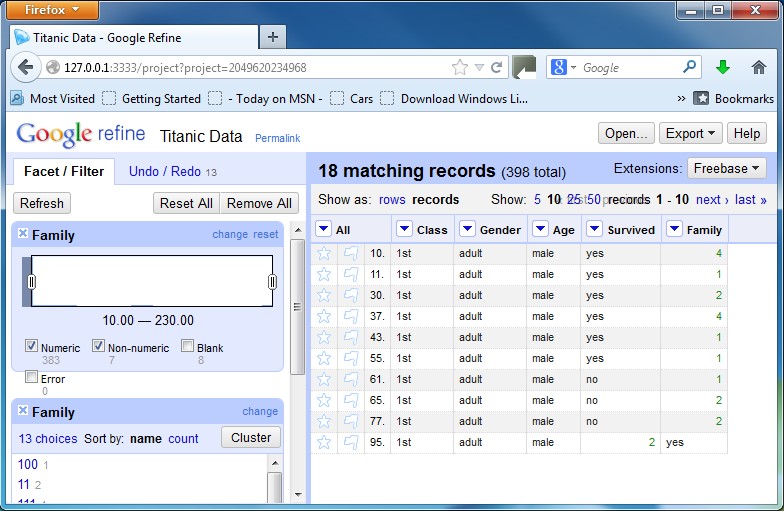


# Identify Dirty Data – non-numeric

**Numeric facet**

**Text facet**

**Finding errors by moving the bars**

**Finding errors by moving the bars -- Correction**

To remove a row, do the following:

* 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 - to merge all the columns into one column

in an Excel file, and then do a check in OpenRefine.

* Step 1: Save the excel data to a CSV 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.

# Export Processed Data

To export the processed data, do the following:

* Click the **Export** button at the top right,
* Select **Excel**, then name, locate and save the cleaned dataset