Open Refine Tutorial 3 of 3:
Sorting, Column Splits, Dates, and Geocoding

In this tutorial on Open Refine we'll learn how to reorder and sort columns, create new columns from information in existing cells, manipulate dates, and create geo-coordinates from addresses.

Getting Started

First you need to download a dataset to work with. This tutorial uses open data on festivals from the City of Los Angeles' Department of Cultural Affairs, found here [https://data.lacity.org/A-Livable-and-Sustainable-City/Events-from-LA-Festival-Guide-2014/acy8-72w9].

On the webpage, click the blue 'export' box and export the file as a 'CSV for Excel' to your computer.

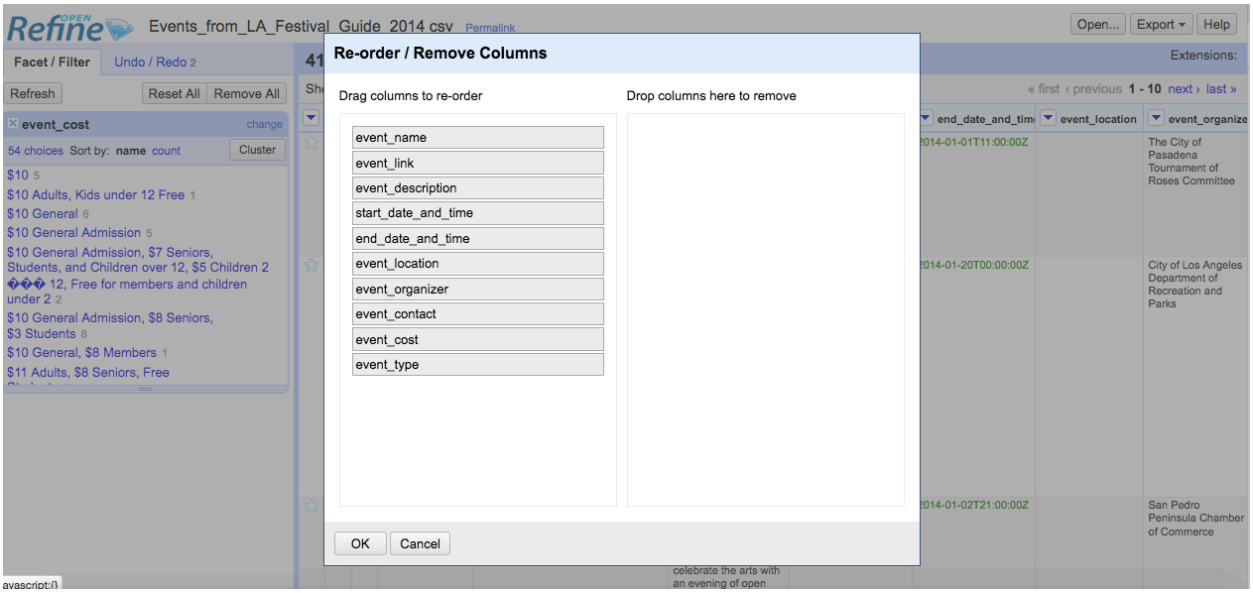
In Open Refine create a new project and click 'browse' to upload your file. Click 'Next'. On the following screen, check the box "commas (CSV)". *Make sure that the box, "Parse cell text into numbers, dates, etc", is unchecked*. Then click 'create project' on the following screen.

Note: For instructions on downloading Open Refine, see Tutorial 1.

Reordering Columns

You may want to reorder or even remove columns in Open Refine. To perform these actions, click the drop down menu at the top of the 'All' column, and choose 'Edit columns' → 'Re-order / remove columns …'

You'll find a list of all columns in the spreadsheet, and you can drag and drop these in the order you desire, or remove columns completely. Go ahead and resort the columns so that 'Event Cost' is next to the 'Event Location' column, or any combination you choose.



Sorting Data

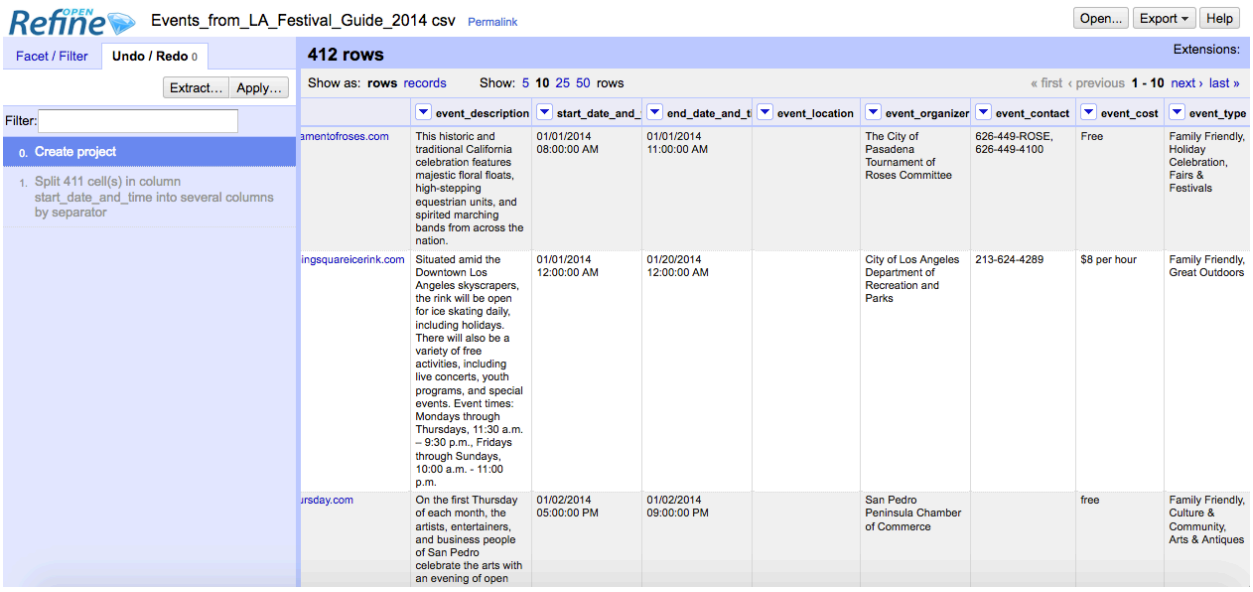
You can also sort data in OpenRefine. Currently this spreadsheet is sorted by date of festival, but say you want to sort by price. Click on the drop down menu for the 'Price' column and choose 'Sort'.

Notice that a new 'Sort' drop down menu is now displayed above the column headings. The 'Sort' drop down menu allows you to reverse the sort order, remove existing sorts, and make sorts permanent. (Unlike Excel 'Sorts', sorts in OpenRefine are temporary). Click on the 'Sort' drop down menu and click 'Reorder rows permanently' to keep the sort you just made.

You can also sort on multiple columns at the same time, so try a few sorting options if you like.

Splitting Columns

Open Refine allows you to make new columns out of data from a single column. To try this function, first make sure your dates are read as text and display like this:



Note: If the dates are read as 'dates', you may need to reupload your data, making sure that the box, "Parse cell text into numbers, dates, etc", is unchecked.

Now go to the drop down menu of the 'start_date_and_time' column to 'Edit column' → 'Split into several columns'. In this case, we want to make a separate column for the festivals start date times.

In the pop up window that appears, make sure 'by separator' is checked, and use a blank space in the text box next to 'Separator' (you won't see anything in the box when you do this). Then you'll want to split into 2 columns. These action tells Open Refine that you want to take all text after the first space in the cells and put this data into a new column. Now click 'OK.'

Now rename the columns now to 'Festival start date' and 'Festival start time.' Go to 'Edit column' → 'Rename this column' on both the old and new columns to perform this action. Go through the same steps now on the 'end_date_and_time' column.

Cleaning and Displaying Dates

Open Refine allows you to display dates in different formats. In this case, you'll use a Transformation command. In the first set of parentheses, you have the current date format. In the second set of parentheses, you have the format you want to convert to:

```
value.toDate('MM/dd/yy').toString('MMMM dd yyyy')
```

There are many different ways to display dates. After the transformation, try playing around with combinations of these commands to see how dates are differently displayed:

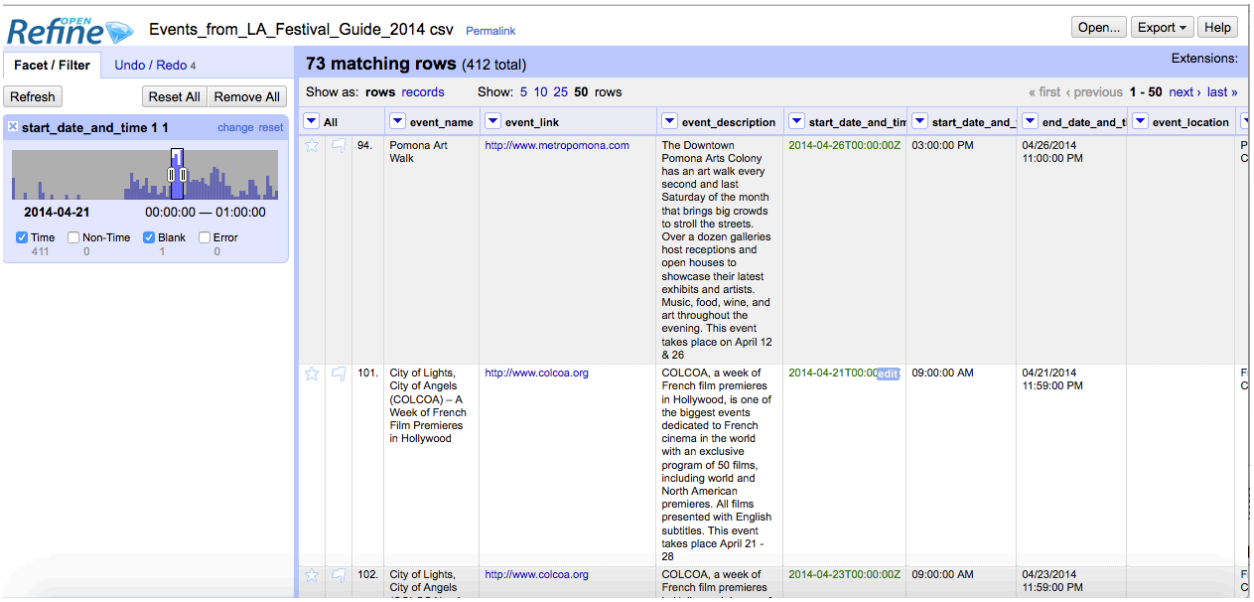
```
value.toDate('MMMM dd yy').toString('MM dd yy')
```

```
value.toDate('MMMM dd yy').toString('MMM dd yy')
value.toDate('MMMM dd yy').toString('MM-dd-yy')
```

Timeline Faceting

We can also tell Open Refine to read this column as a date. Go to 'Edit cells' → 'Common Transformations' → 'To date'. This conversion undoes the formatting from the previous exercise. However, with the date format we can now perform a timeline facet on the column. This action allows some quick analysis, such as seeing which months attract the most festivals.

To create the timeline facet, go to the 'Festival start date' column, click 'Facet' → 'Timeline facet.' Now you can toggle the histogram sliders so that they bracket the months with the most festivals:

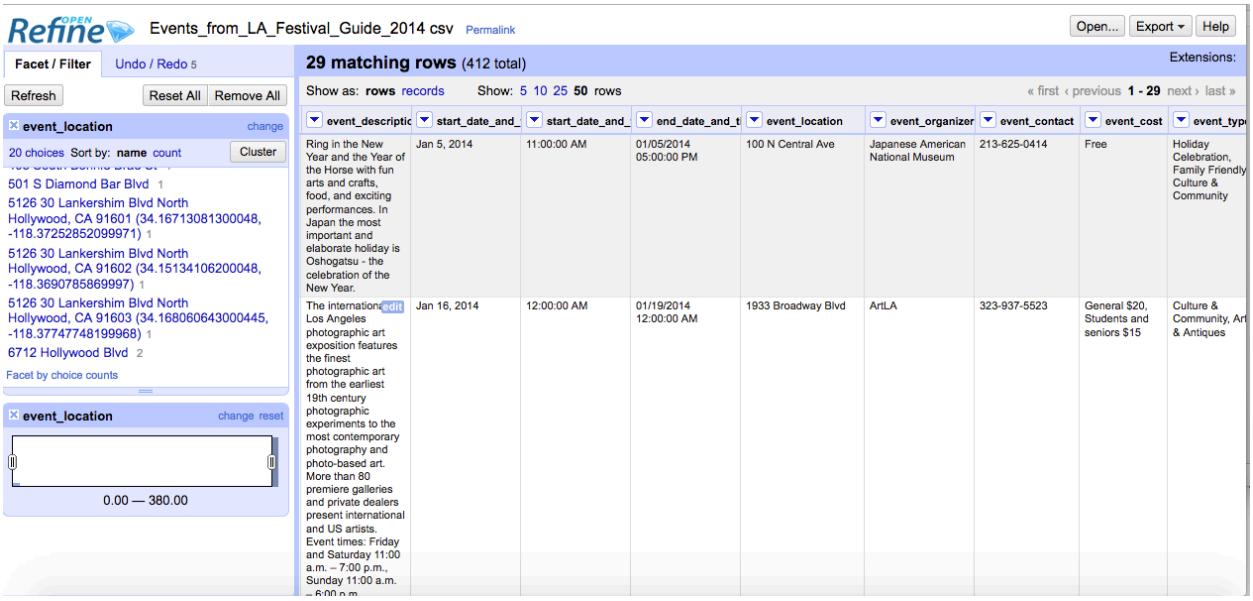


You can see that April, May, June, and July are the months that attract the most festivals in Los Angeles.

Geocoding

Finally, here are the steps to take to create geocode values of text addresses. Geocodes are very powerful to have in your dataset, since they will allow you to map locations into GIS software. These functions use Google Maps behind the scenes to figure out your most likely geographic location.

To begin, we need to isolate all the cells that have addresses (you'll see most are blank). Go to the column 'Event location'. In the drop down menu click 'Facet' → 'Facet by text.' Scroll down to the bottom of your list. You'll see that you have 383 blanks. To facet the blank cells out, click 'Facet by choice counts' at the very bottom of the facet list. This shows a histogram; you can now toggle the sliders so that they exclude the tallest column, which represents the blanks.

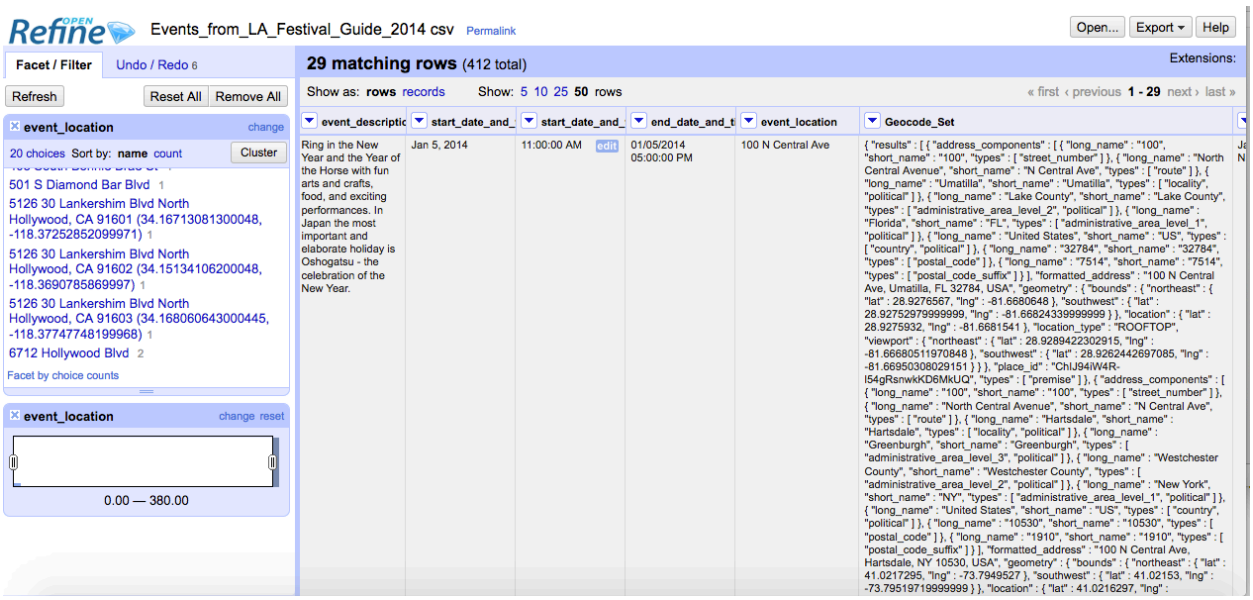


You should now have 29 rows left, all with location information. This amount is a good size for the next function, which can take some time when working with larger datasets.

Now go to 'Edit column' → 'Add column by fetching URLs. This action will create a new column with cell values that come from the internet, in this case Google maps. Enter this command into the text box:

```
"http://maps.google.com/maps/api/geocode/json?sensor=false&address=" +
escape(value, "url")
```

You'll need to name the new column and set the throttle delay to around 500 milliseconds. The new column should look like this:



To convert this into a more readable format, you need to click on the new column, and then select 'Edit column' → 'Add column based on this column.' Create a new name for the column, such as 'Geocoordinates', and enter in the expression below:

```
with(value.parseJson().results[0].geometry.location, pair, pair.lat +", " +
pair.lng)
```

You now have a column containing the geocoordinates.

Credits:
Written by Morgan Currie, September 2016.