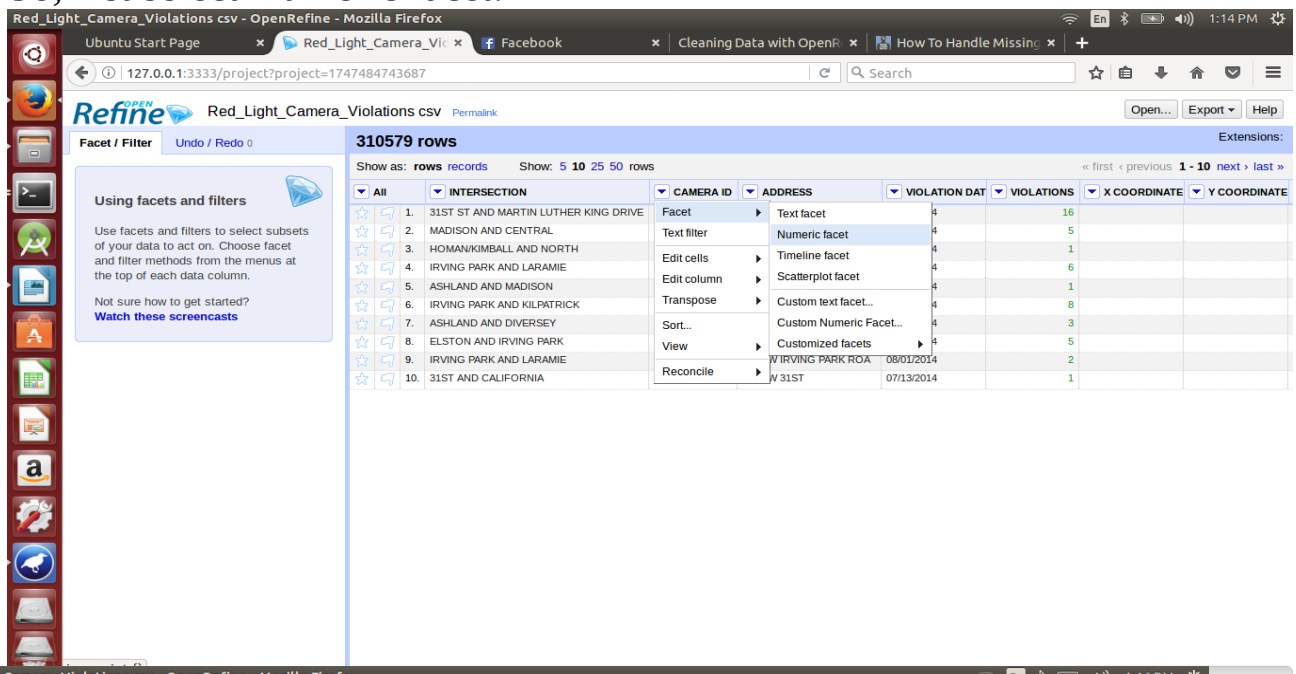
DM ASSIGNMENT

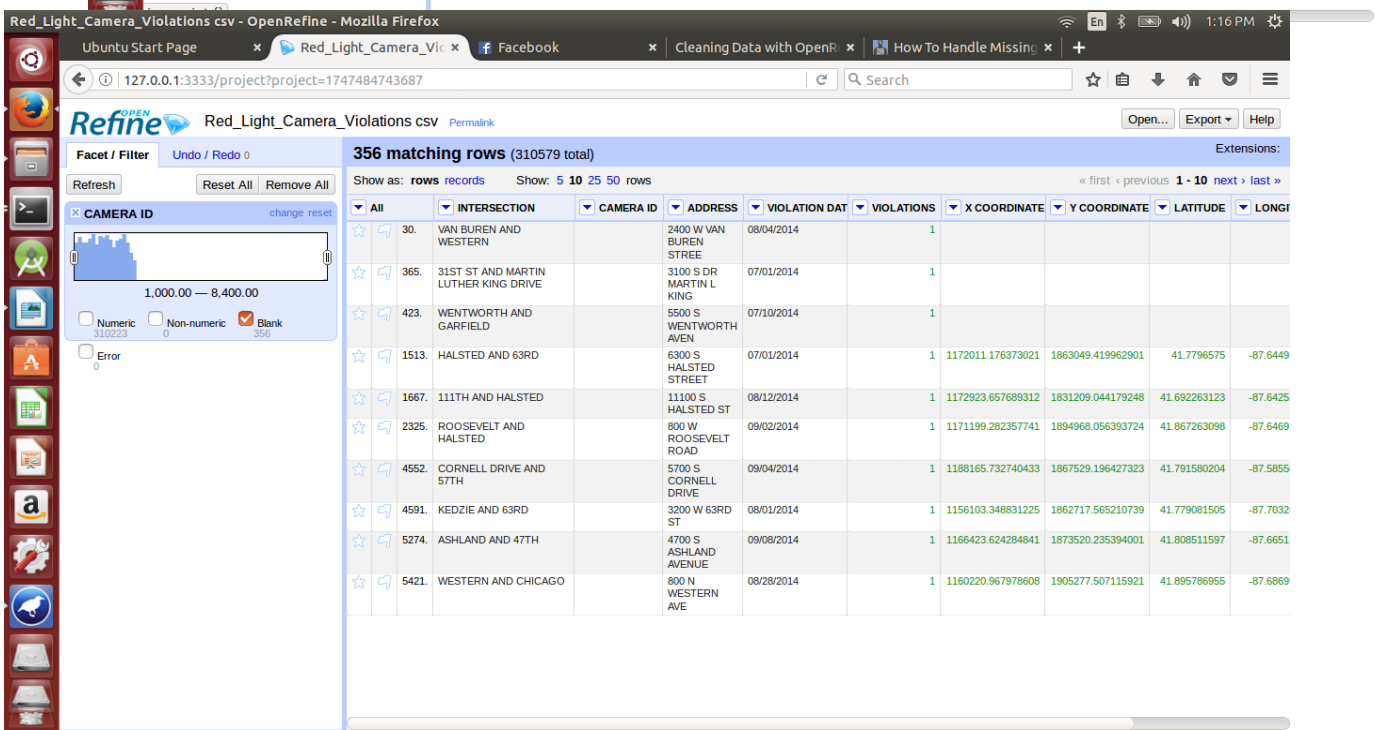
Question-1:

The dataset taken has 10 attributes in total.They are Intersection(Nominal),Camera ID(assumed to be unique,Numeric),Address(Nominal),Violation Date(Nominal),Violations(Numeric),X-coordinate(Numeric),Y-Coordinate(Numeric),Latitude(Numeric),Longitude(Numeric), Class Attribute:Location which is an ordered pair of Latitude and Longitude and this is a Nominal attribute.And the number of Rows are 3,10,579.

1)Cleaning Non-numeric Camera Id's And Blank Id's:

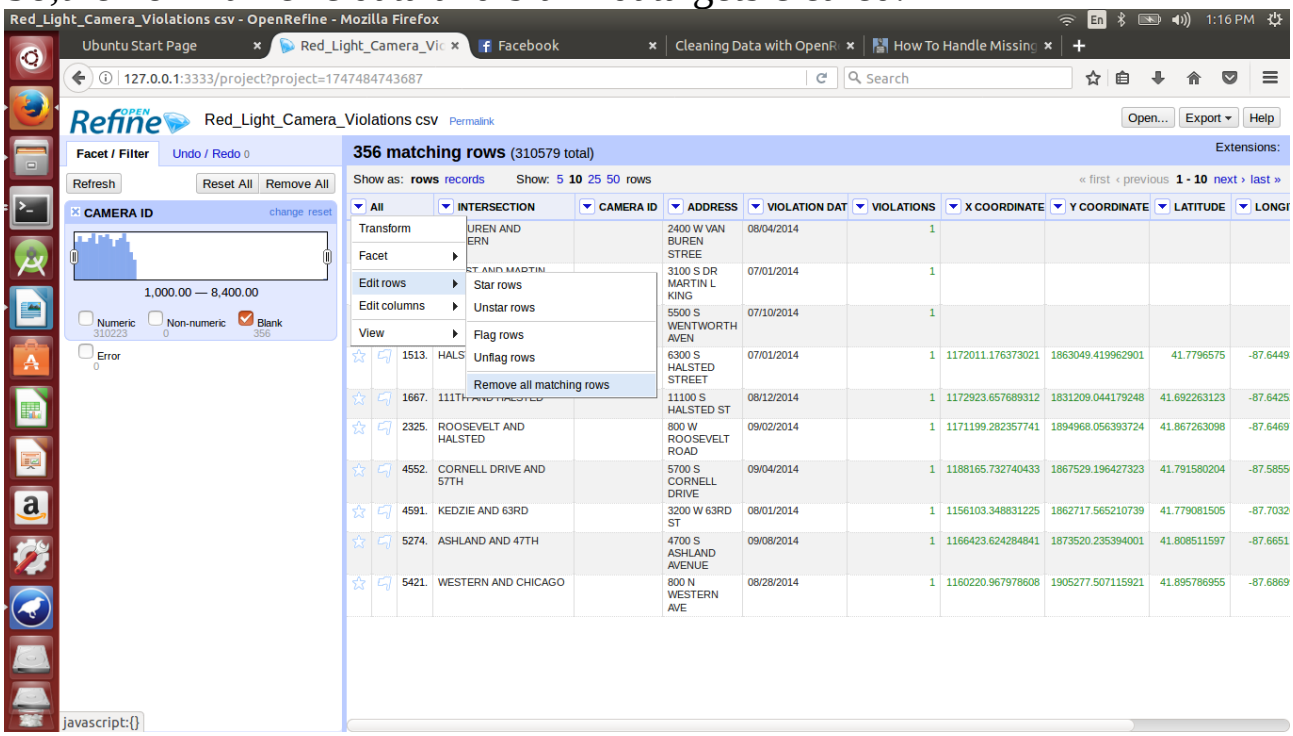
So,first select Numeric facet:





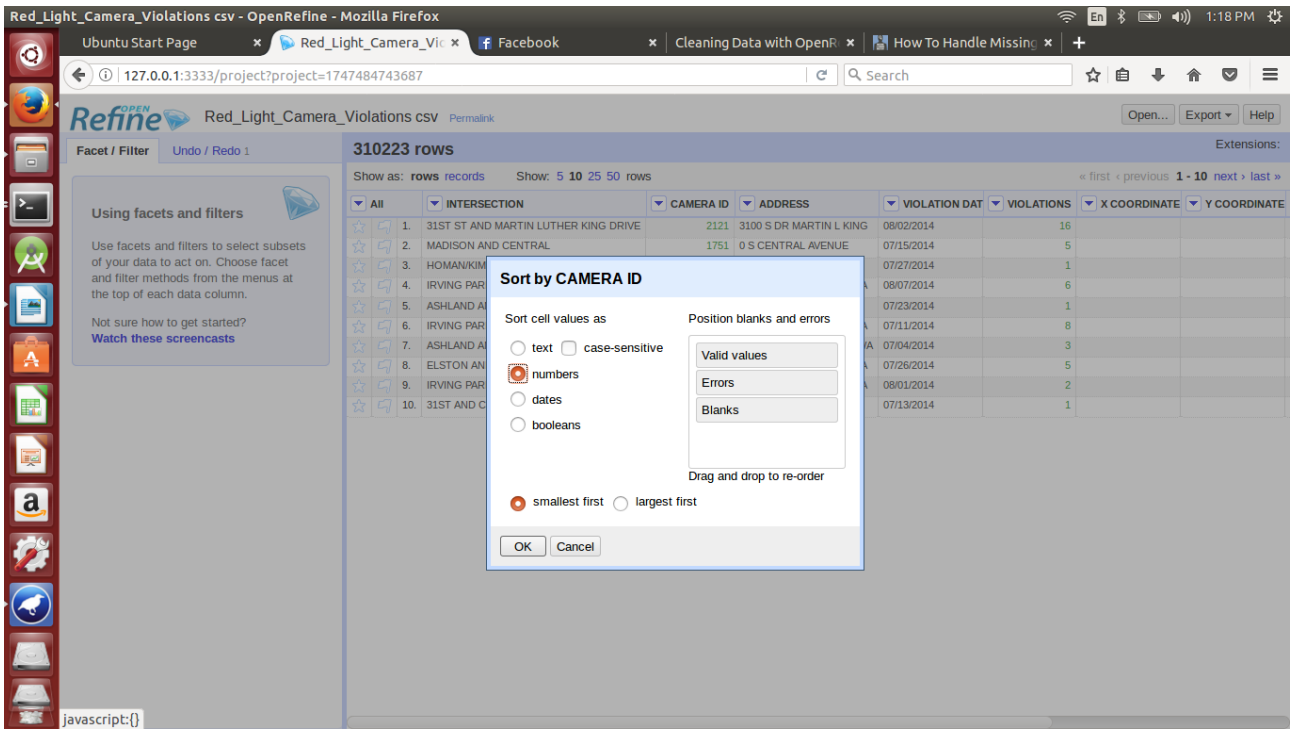
Now,unselect numeric values so we are left with non-numeric and blank values .As Camera Id is expected to be a numeric value,so we delete all these non-numeric values which is cleaning data(Assumption)
Now click on **All--select Edit Rows, and then Remove all matching rows.**

So,the non-numeric data and blank data gets cleared.

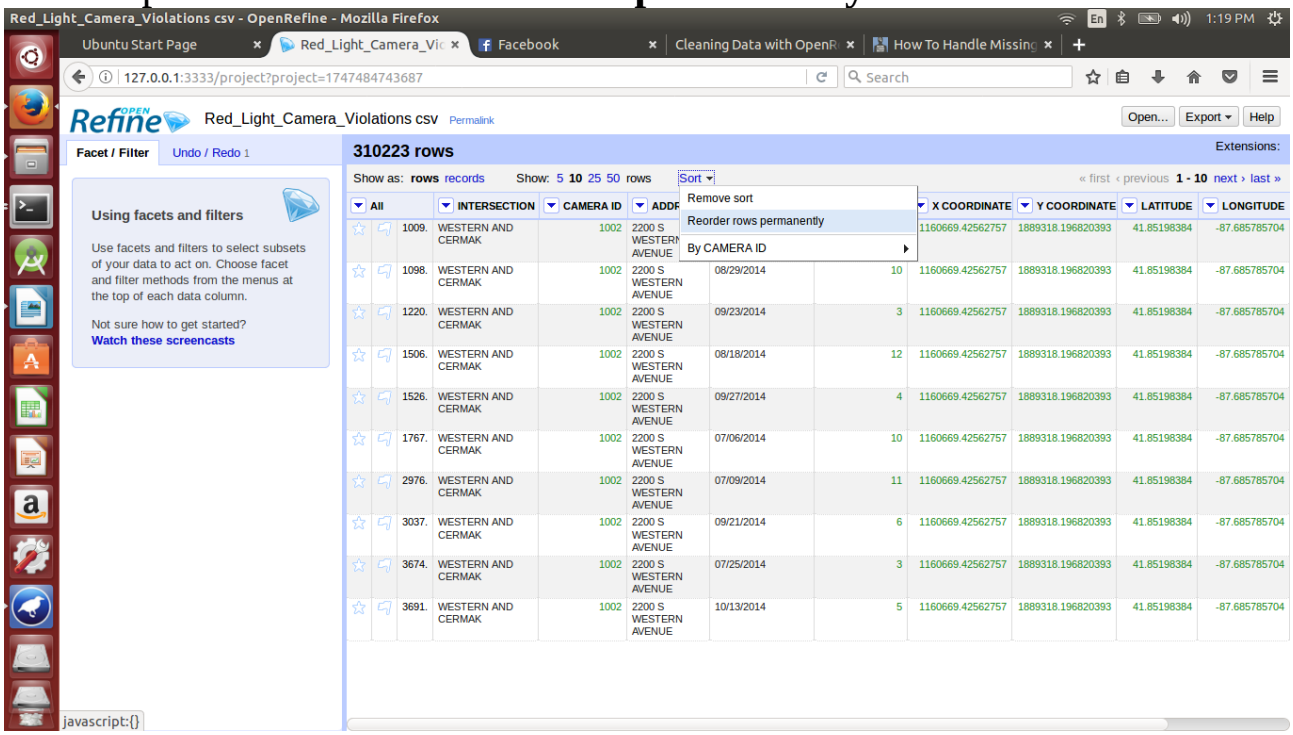


2)Remove Duplicates:

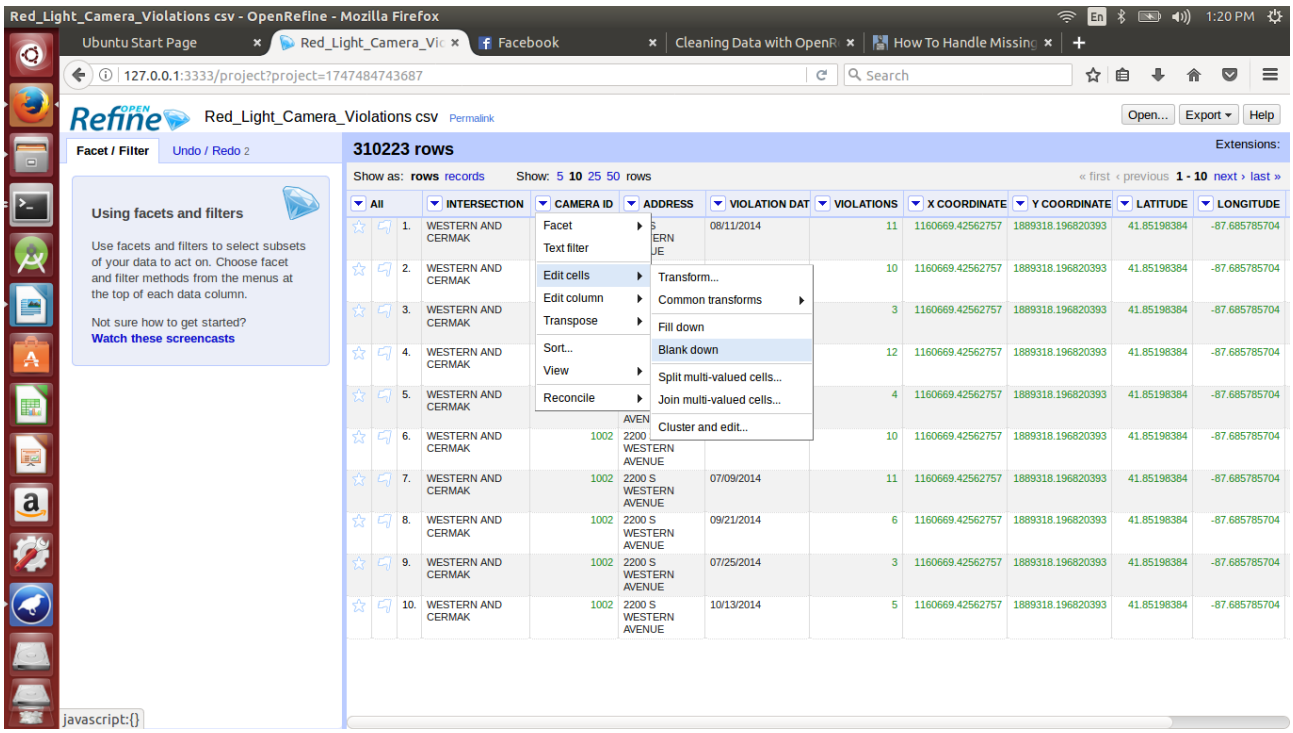
As,in this data set the record id is expected to be unique and so,the dupliactes should be deleted to make data clean.
So,first select **Sort option from dropdown menu of Camera id and then select Numbers radiobutton.**



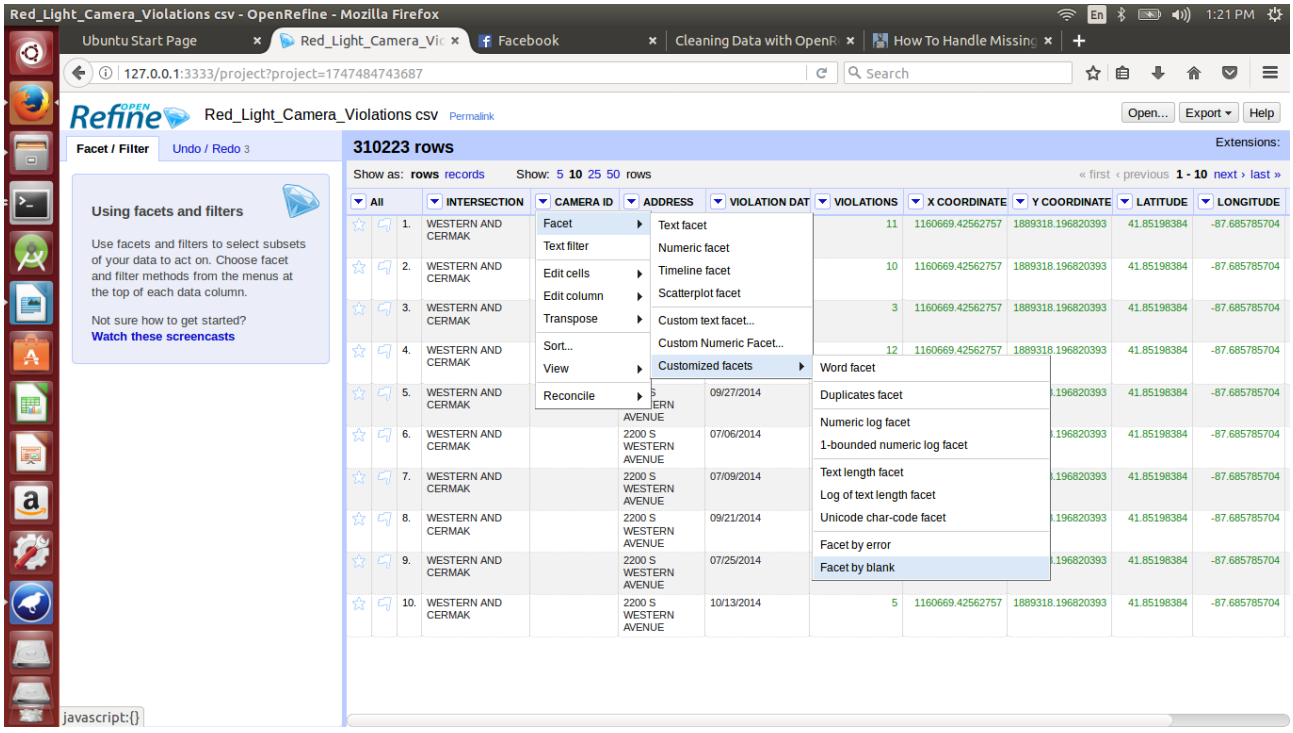
Now,as to prevent variation in order of data later we make these sorted rows permanent.To do this we click the **Sort menu** that has just appeared at the top and choose **Reorder rows permanently.**



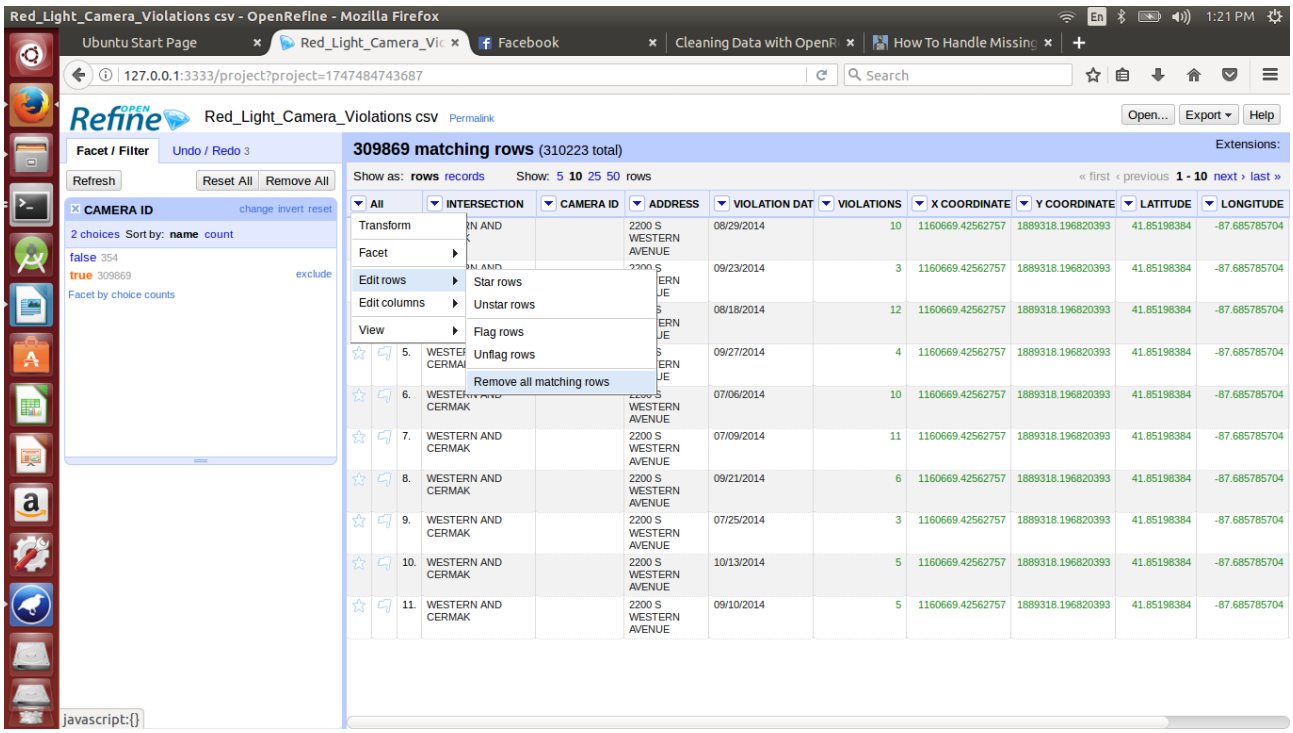
Now,**blank the Camera ID** of rows that have the same CameraID as the row above them, marking them duplicates**.** To do this, click on the Record ID choose **Edit cells -Blank down.**



Now,to sepearte these blank cells from other select **Facet- Customized facets-Facet by blank.**

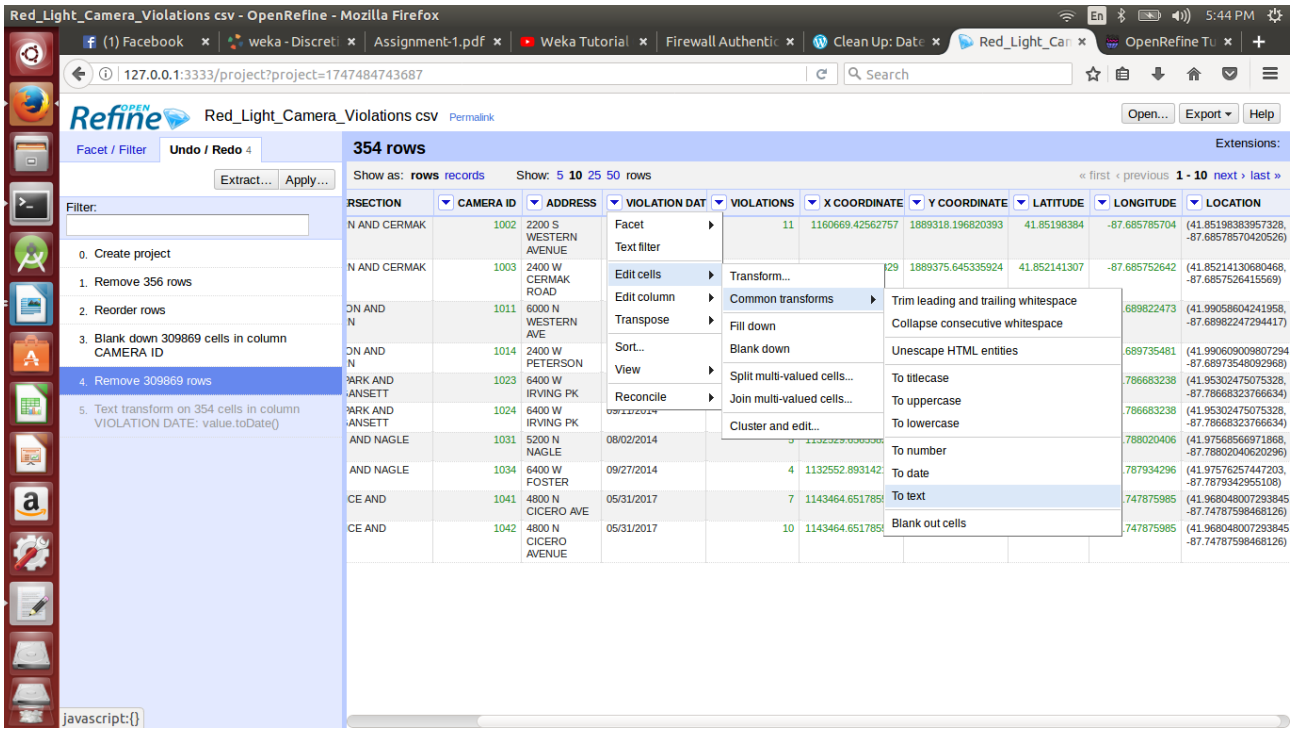


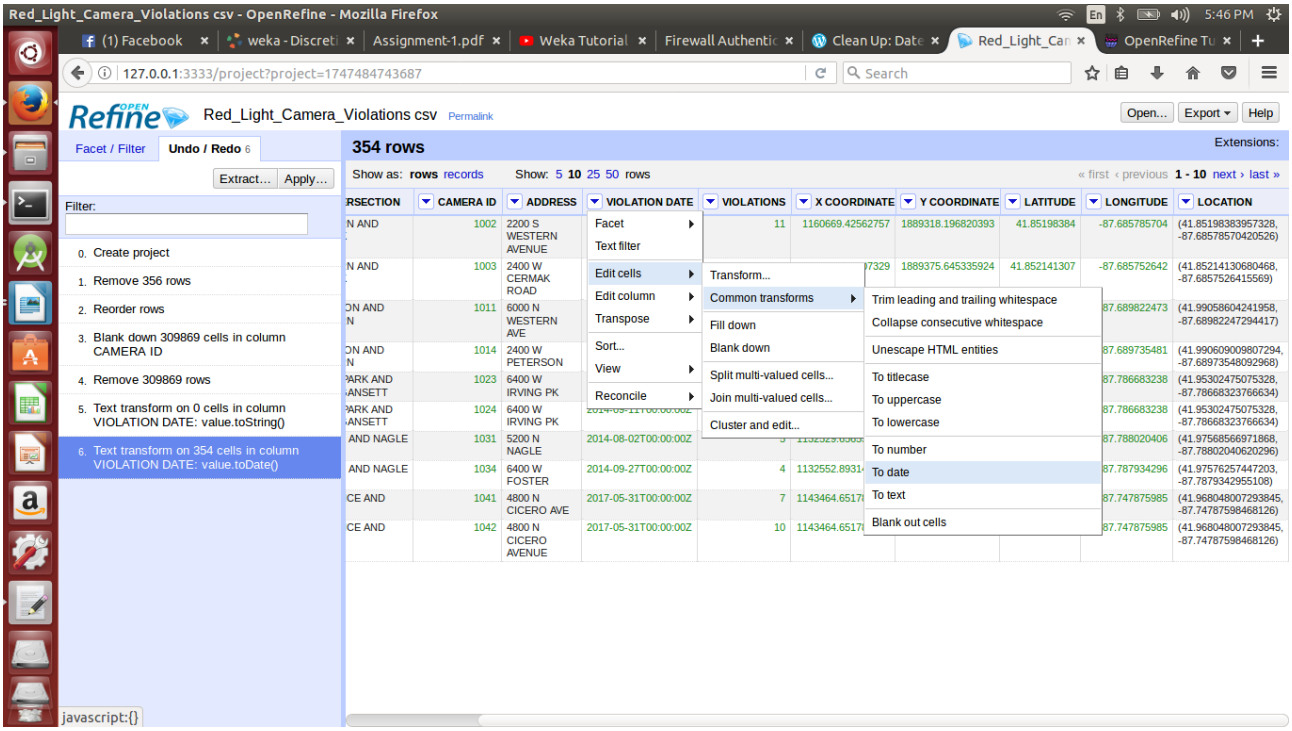
Then select true (i.e blank rows).Now to delete these rows we select **All-Edit rows-Remove all matching rows.**



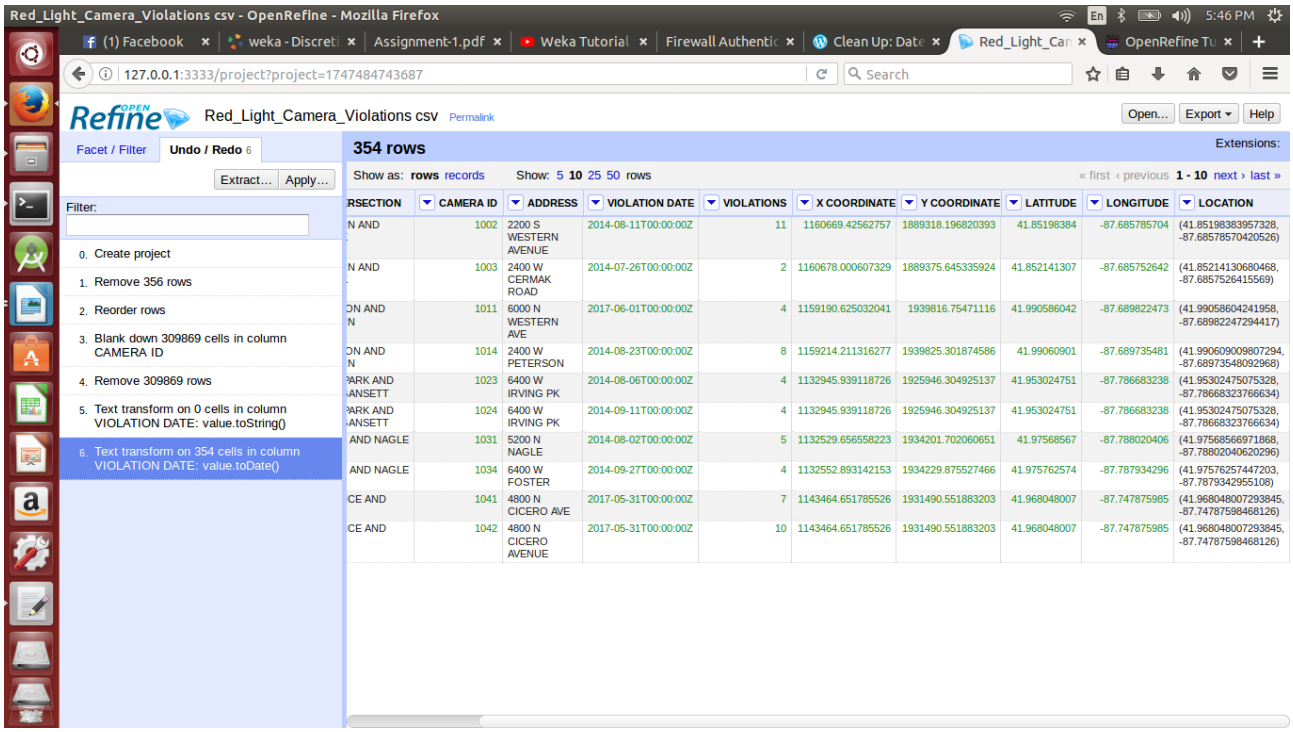
3)Extracting and Cleaning values for dates:

First we want to convert everything to text - Edit cells -> Common transformations -> To text, and then you need to Edit cells -> Common transformations -> To date. If you did not convert all the values to text first, then you may find that some of the years are represented as numbers, and have not been converted. The initial format of violation date is MM/dd/yyyy





Now the format changed to:
To further clean up the dates, we need to use Facet -> Timeline facet and select only "Non-Time" values.But as the dataset set taken hasno nontime values we need not clean further.

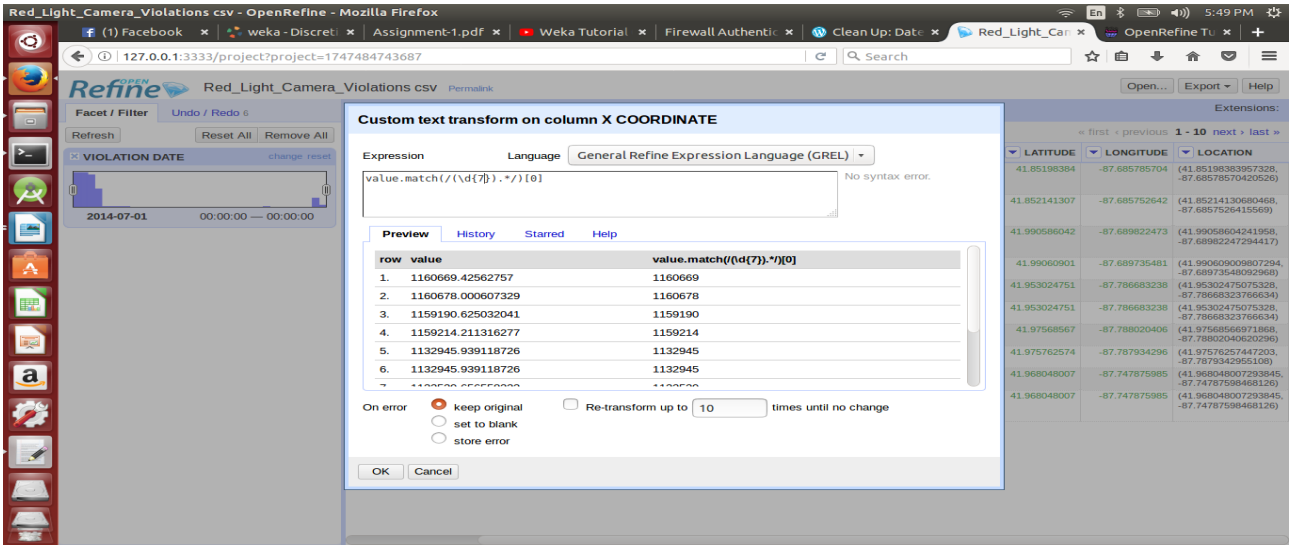


To get only year we can use this grel expression

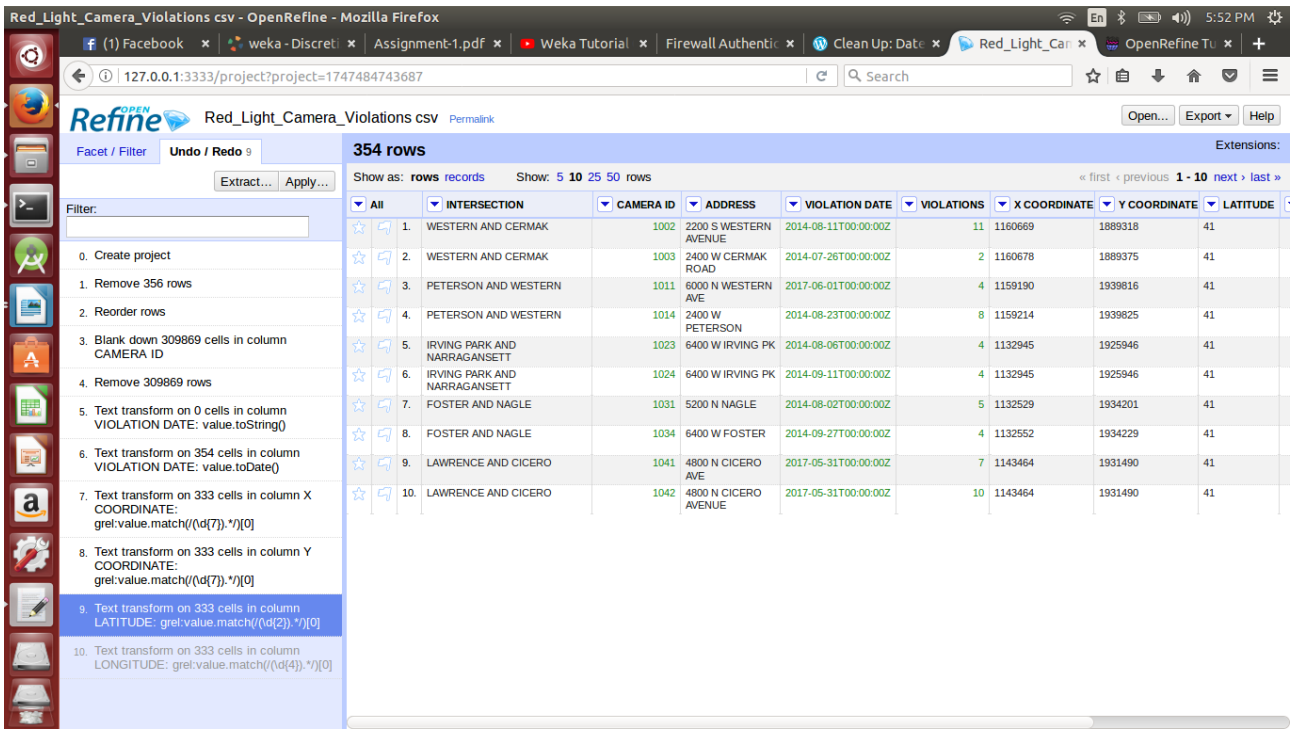
value.match(/.*(\d{4}).*/)[0].

4)Converting decimal values to integers using GREL**(Transformation of Numeric Data):**

Do Edit cells -> Transform, and use the code below. The ".*" means a sequence of zero or more characters (letters, numbers, symbols, etc). The "\d" indicates that we're looking for a digit. The "{7}" shows that we want to match exactly 7 digits. The value.match function returns an array of results, so we use "[0]" to retrieve only the first match.



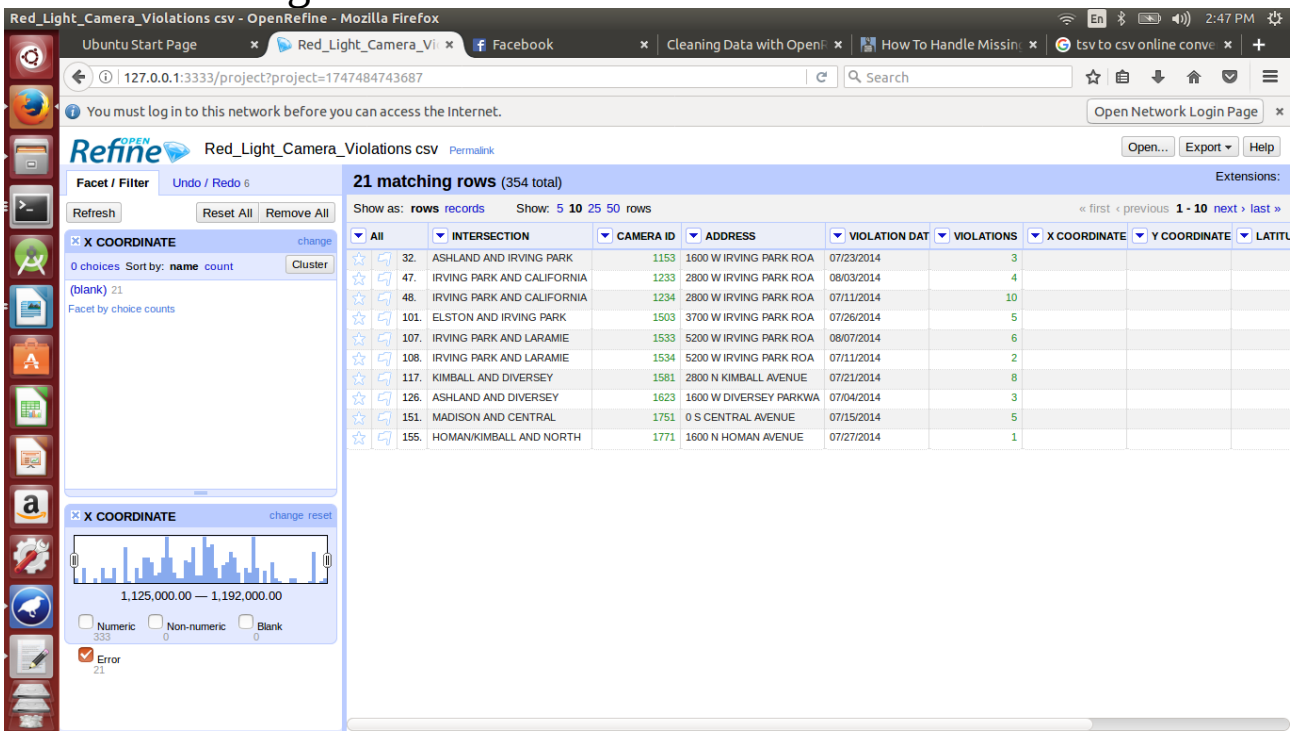
Repeat this for x-coordinate,Y-coordinate,latitude and longitude and this expression convert to their corresponding integer values which helps further. So,now the data set looks like:



So,all the decimal values got replaced with their integer values.

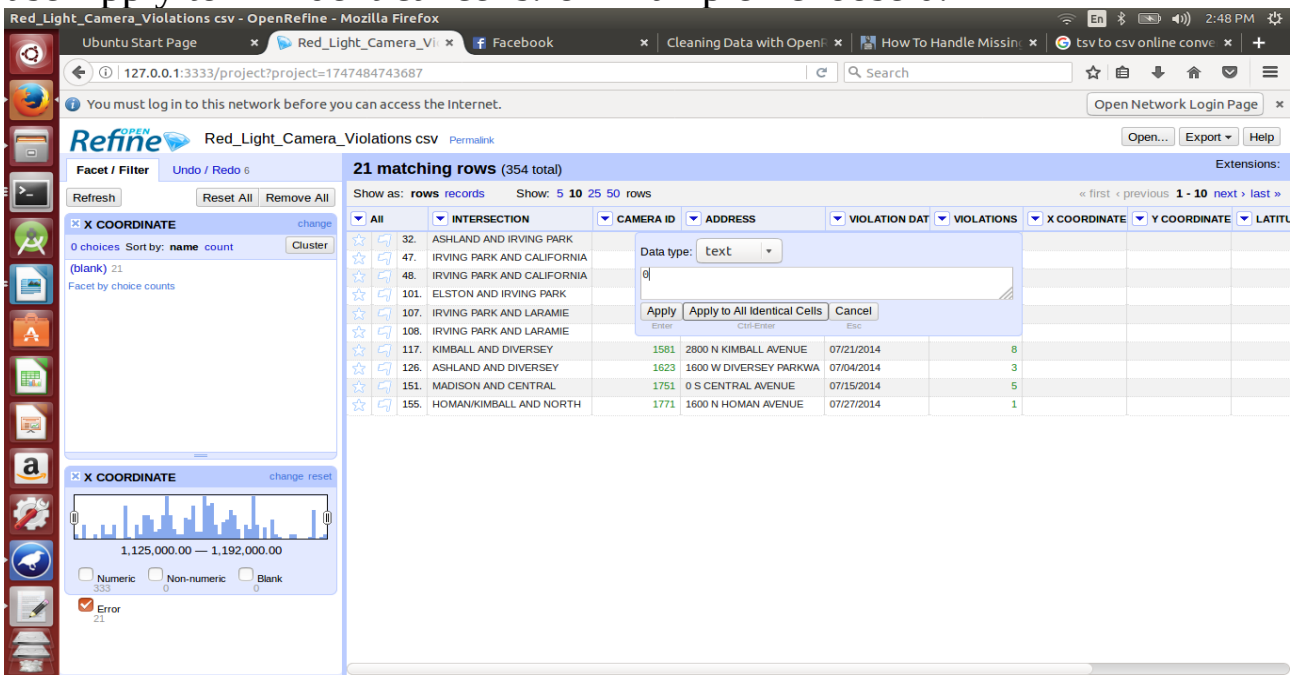
5)MISSING VALUES:

On Selecting Numeric facet from drop-down menu of X-coordinate we get:



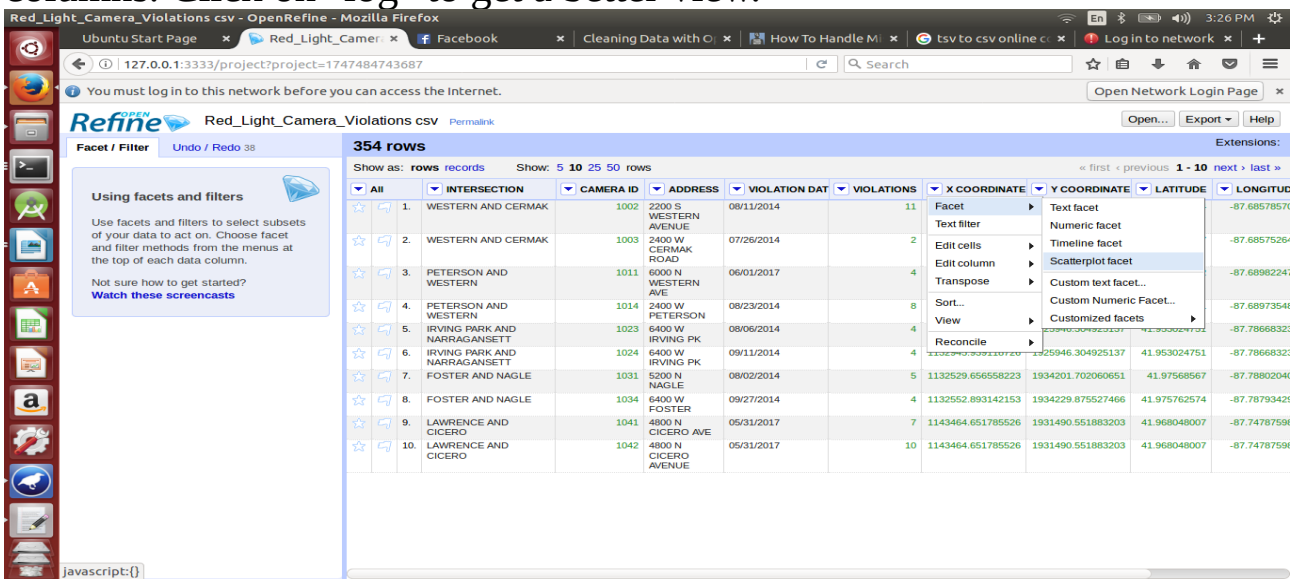
Now on selecting Error tuples,(i.e which have x-coordinate as null)

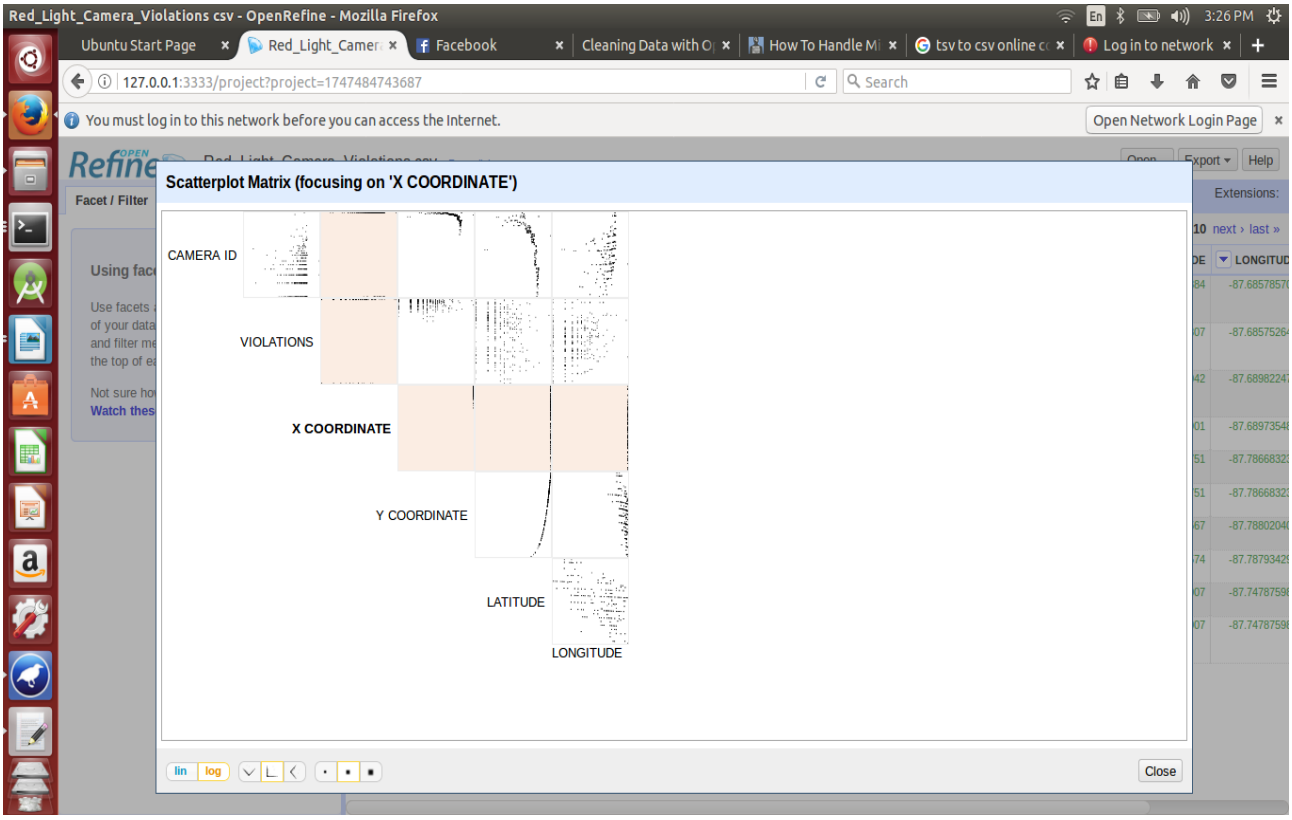
We donot have a direct option in open refine to fill the missing values whereas weka has.we can fill them only by specifying particular value required value on selecting edit write the value and then use Apply to All Identical cells.for Example i choose 0:



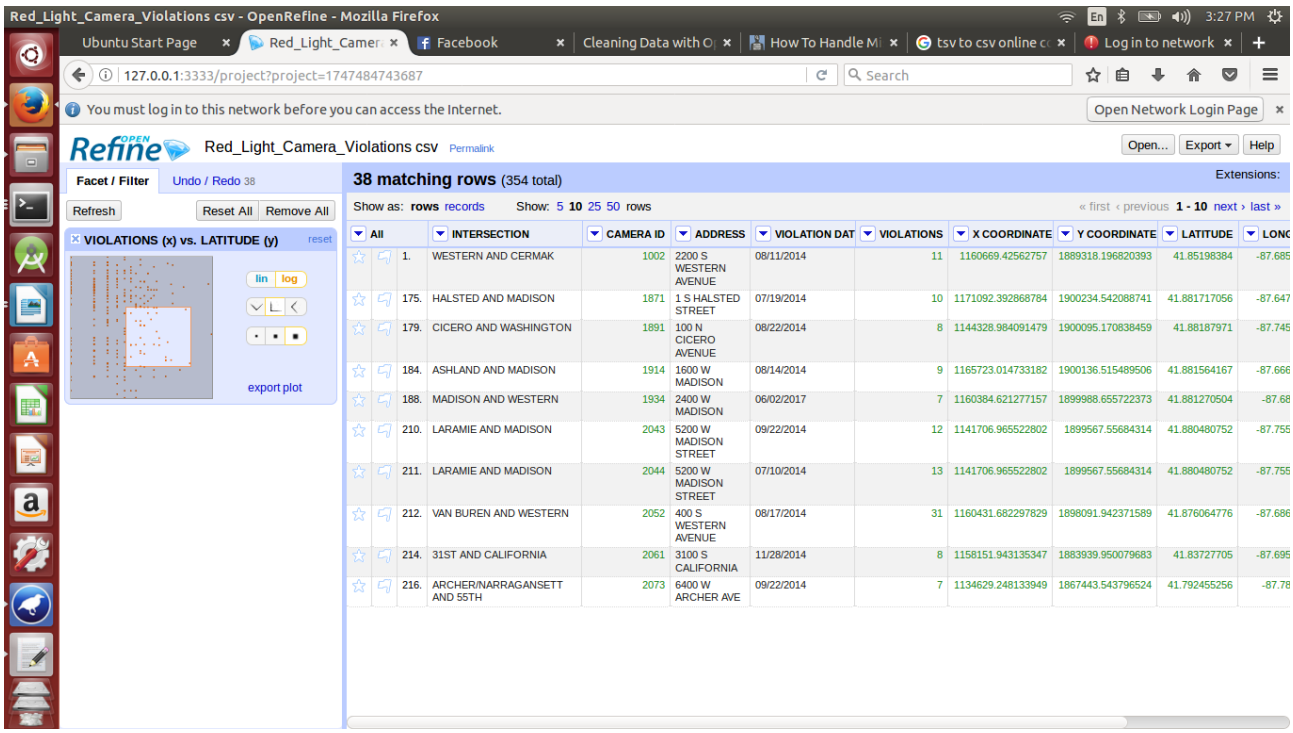
6)ScatterPlot:

Click on the "X COORDINATE" column, Facet -> Scatterplot facetThis shows the relationships between all of the numeric values in each of the columns. Click on "log" to get a better view.





Click on the plot for Violationsvs. X COORDINATE. You can now drag select a portion of the plot, and then see the rows corresponding to that selection.



7)EXPORT DATA:

The data can be exported to formats such as Excel.