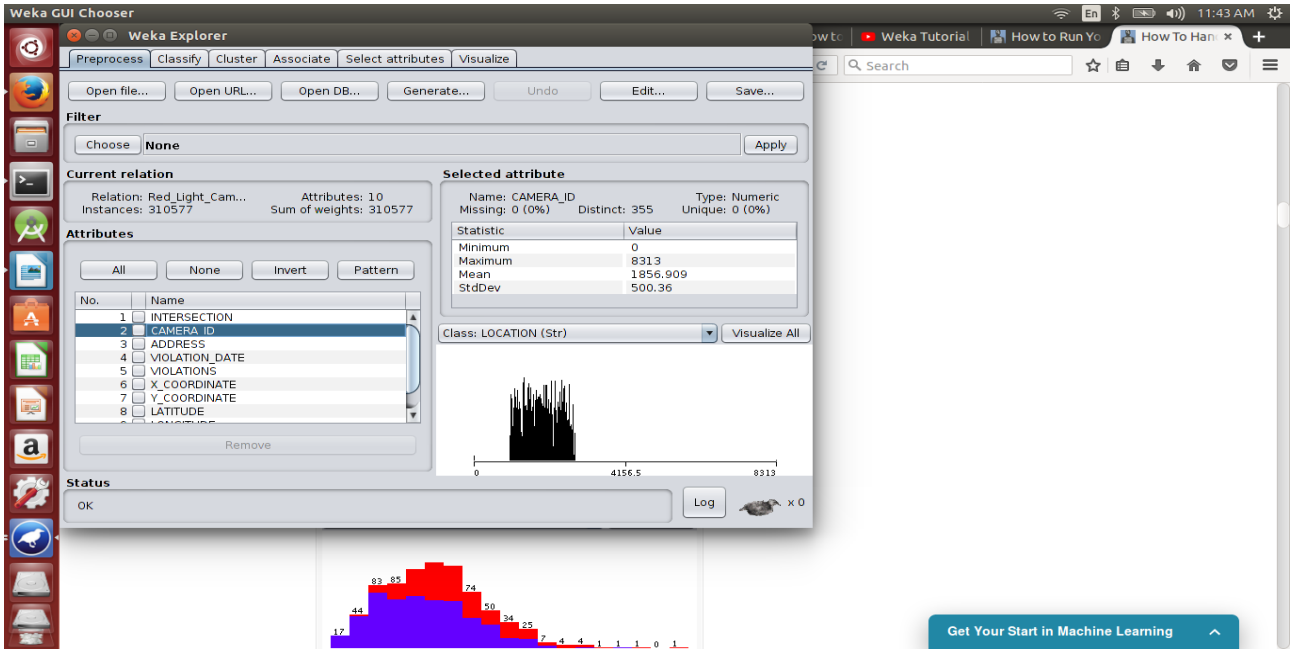
DM ASSIGNMENT

Question-2:

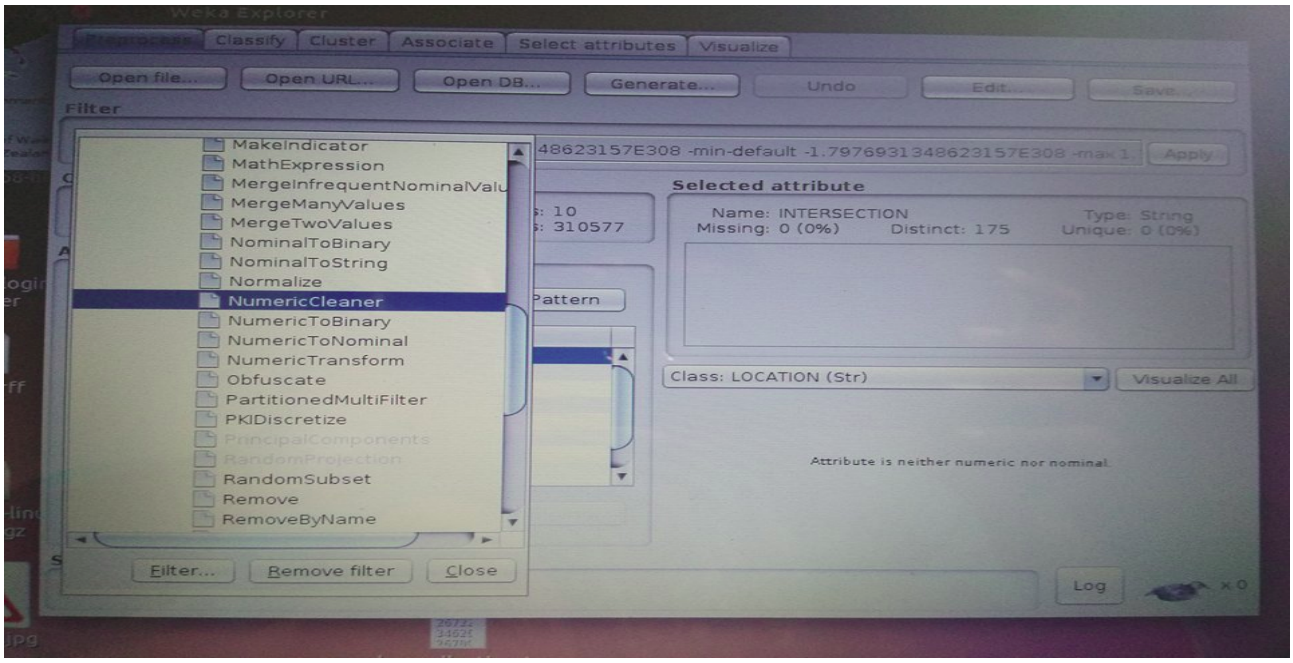
The dataset used is same as Question1.

1)Cleaning Non-numeric Camera Id's :

Open the Weka Explorer. Load the dataset.



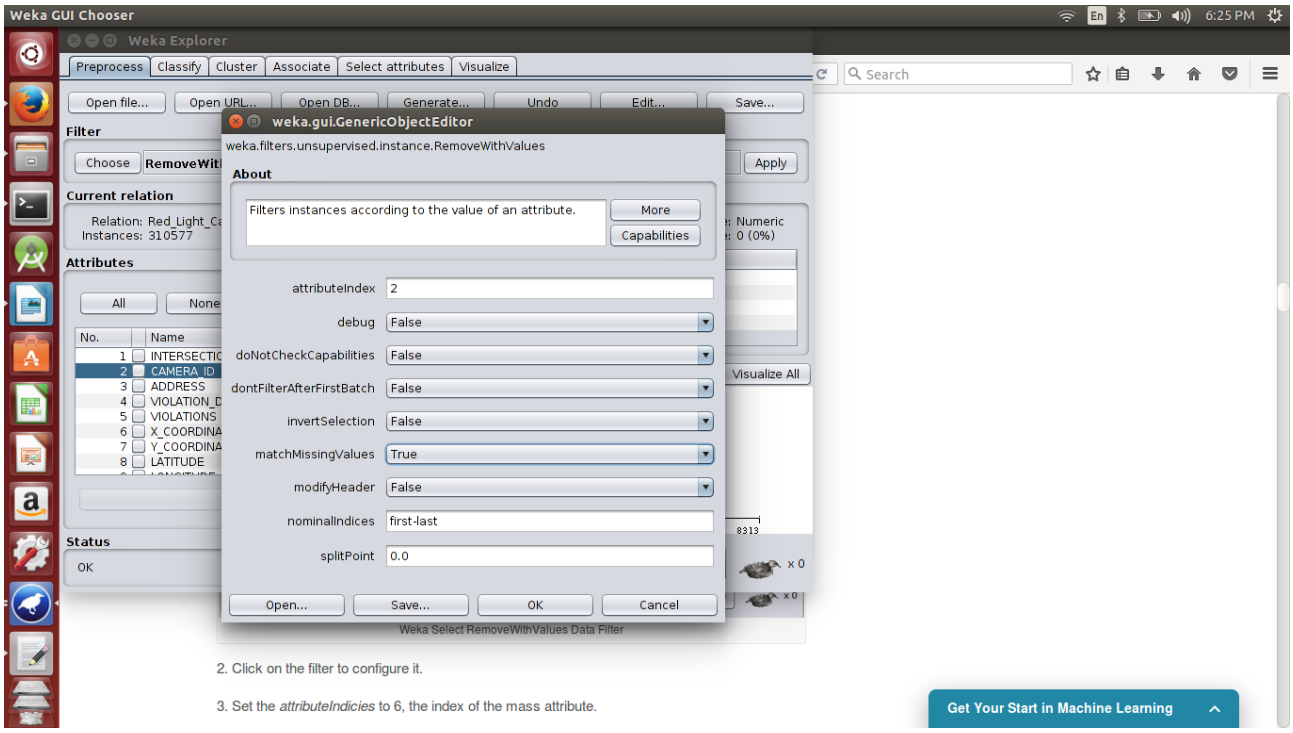
Click the "**Choose**" button for the **Filter and select NumericalCleaner**, it us under **unsupervized attribute.**



As Camera_ID is expected to be unique we remove nonnumeric and blank id's.Set *minThreshold* to 0.1E-8 (close to zero), which is the minimum value allowed for the attribute. Set *minDefault* to NaN, which is unknown

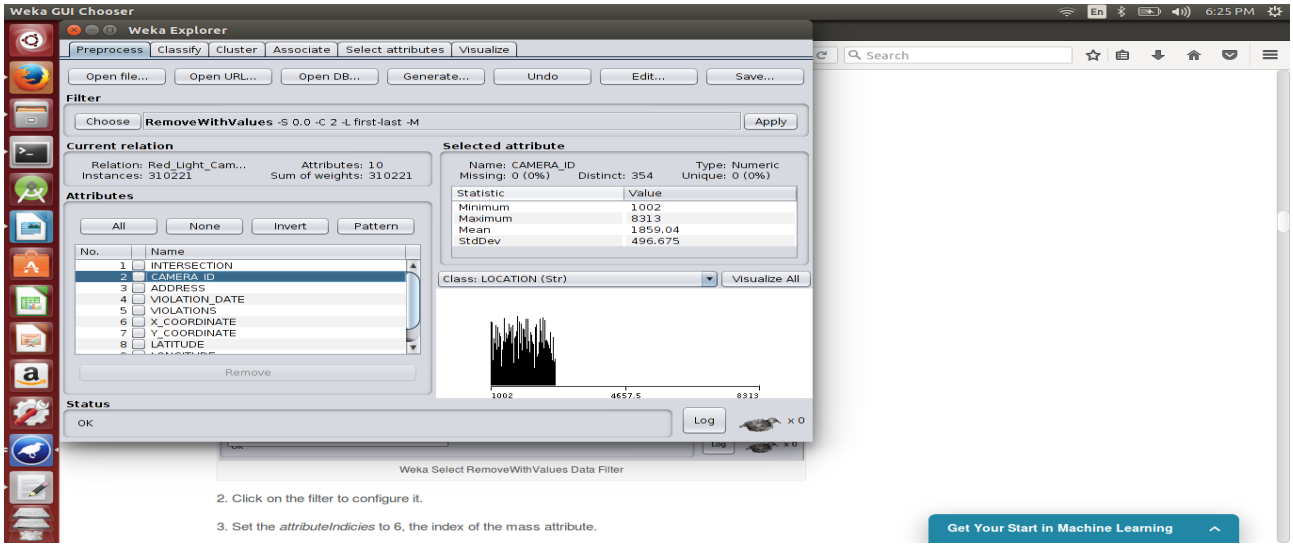
and will replace values below the threshold.So,this marks the missing values.Click the “Choose” button for the Filter and select RemoveWithValues, it's under unsupervized.instance.RemoveWithValues.

Click on the filter to configure it.Set the *attributeIndicies* to 2, the index of the Camera_id attribute.Set *matchMissingValues* to “True”.



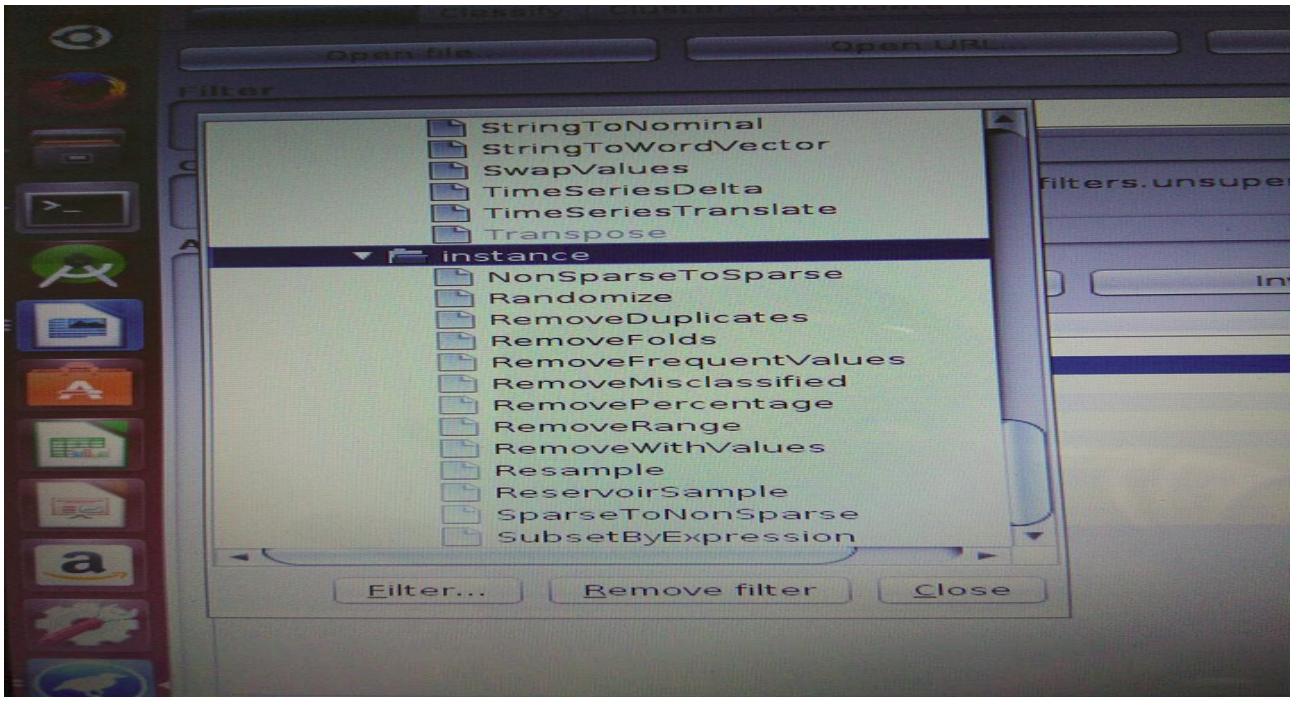
Click the “OK” button to use the configuration for the filter.Click the “Apply” button to apply the filter.

Now the unwanted rows got deleted.



2)Removing Duplicate Data:

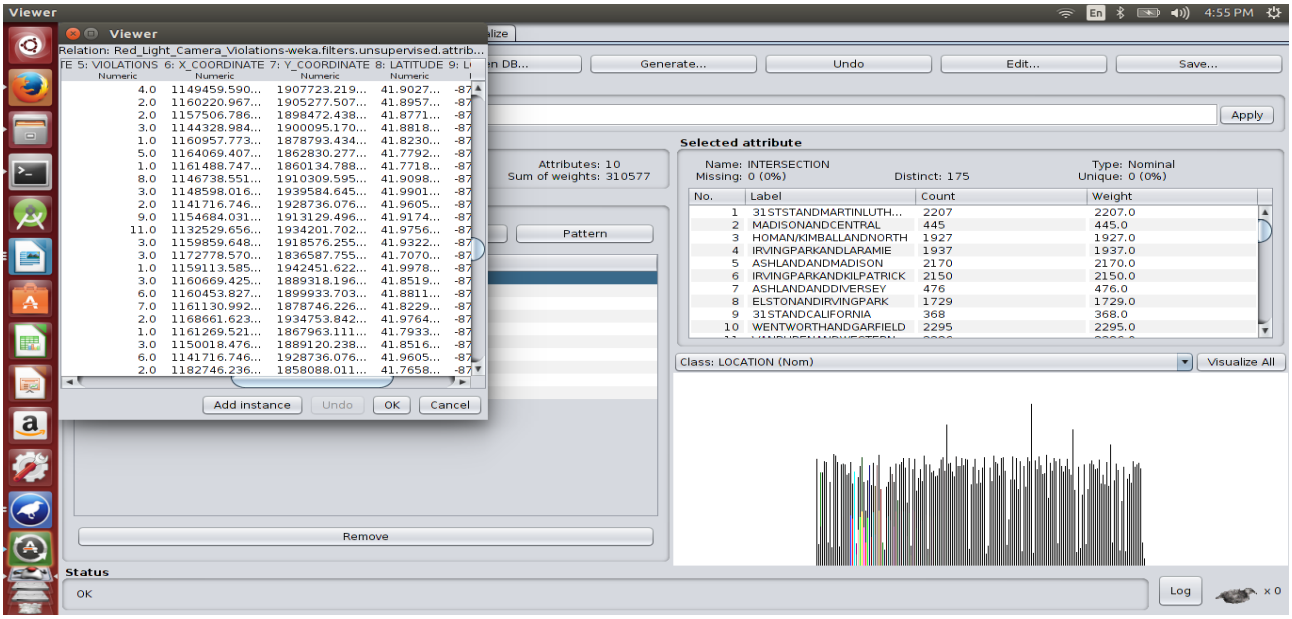
Click the “Choose” button for the Filter and select Remove Duplicate Data, it us under unsupervized.instance.RemoveDuplicate Data.



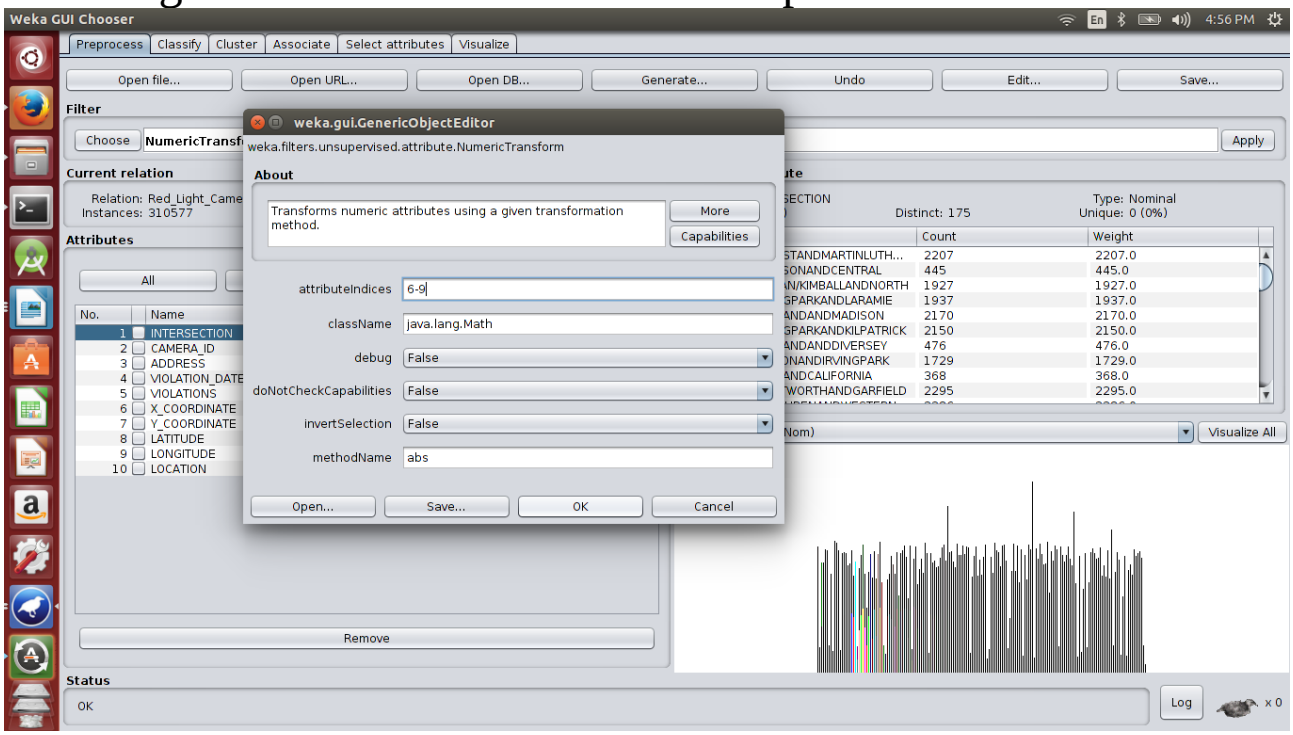
3)Numeric Transform of Attributes:

As the given data consists all numeric attributes and as we filled the misssing values with mean of all other values ,so as the average will be in decimal in case if we need it to convert to absolute value or floor of the value we use numeric transform.We can select this from unsupervised.attribute.numeric transform

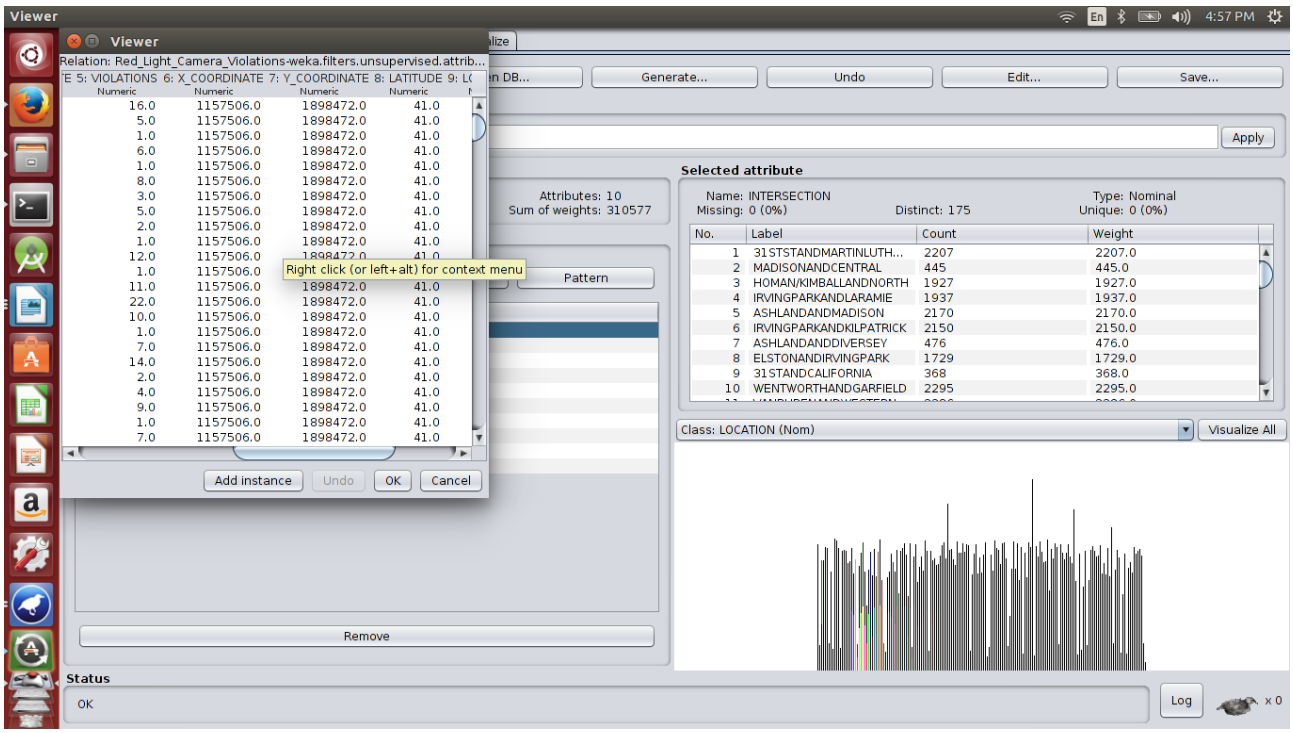
Before numeric transform:



Selecting Numeric Transform from unsupervised Attribute:



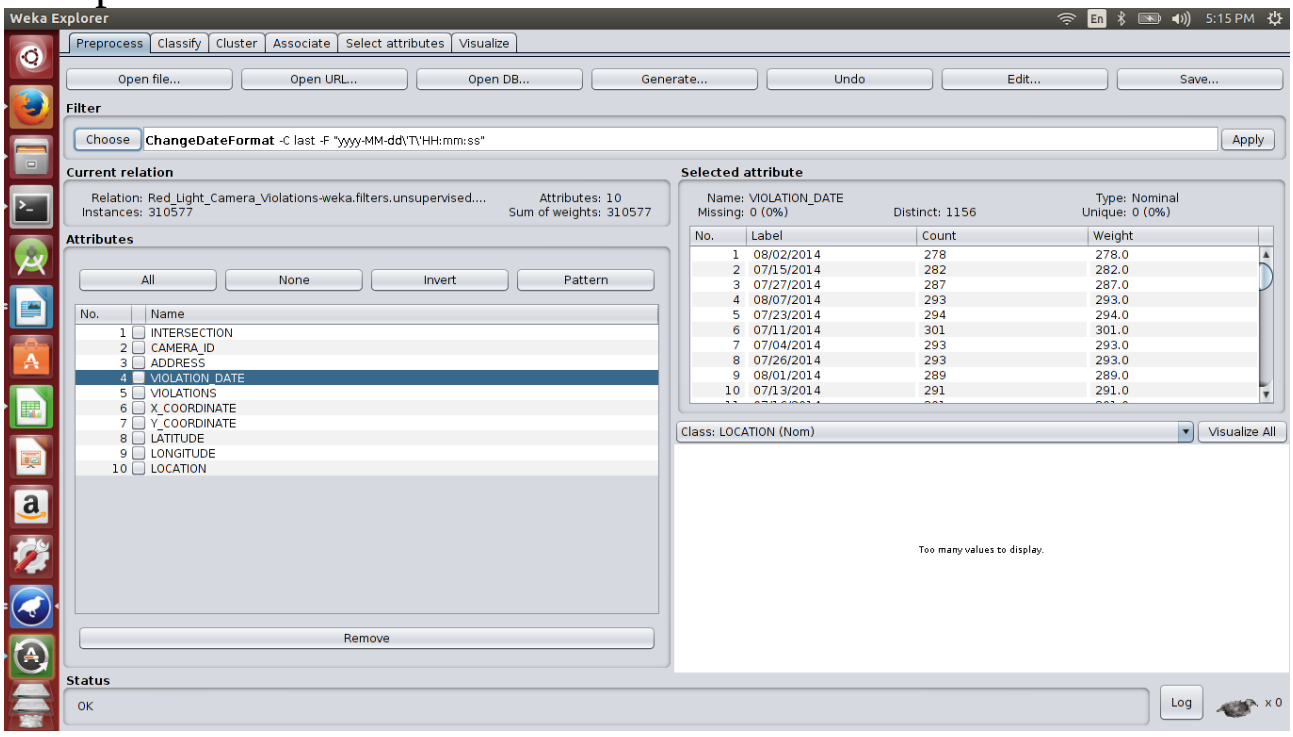
Now after selecting method name as floor the data looks like:



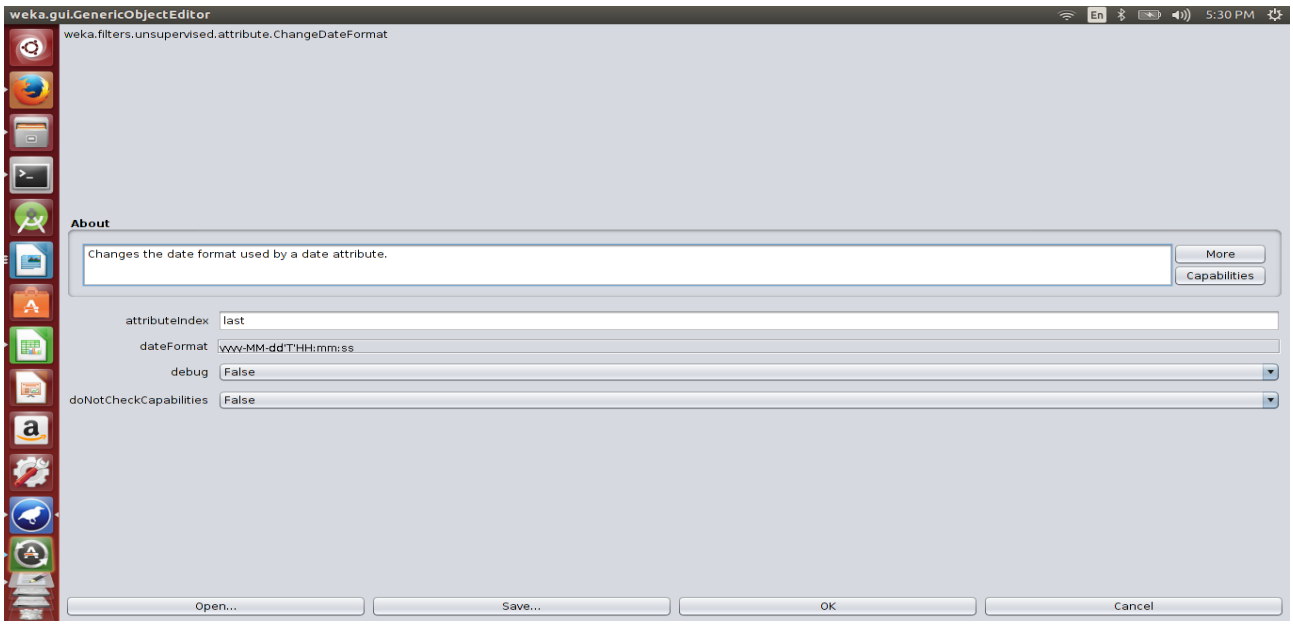
4)Changing Date format:

IN our data we have an attribute violation data which is of the form MM/dd/yyyy.The dataset is clean regarding dates.

We can change date format by choosing ChangeDateFormat from unsupevised folder.



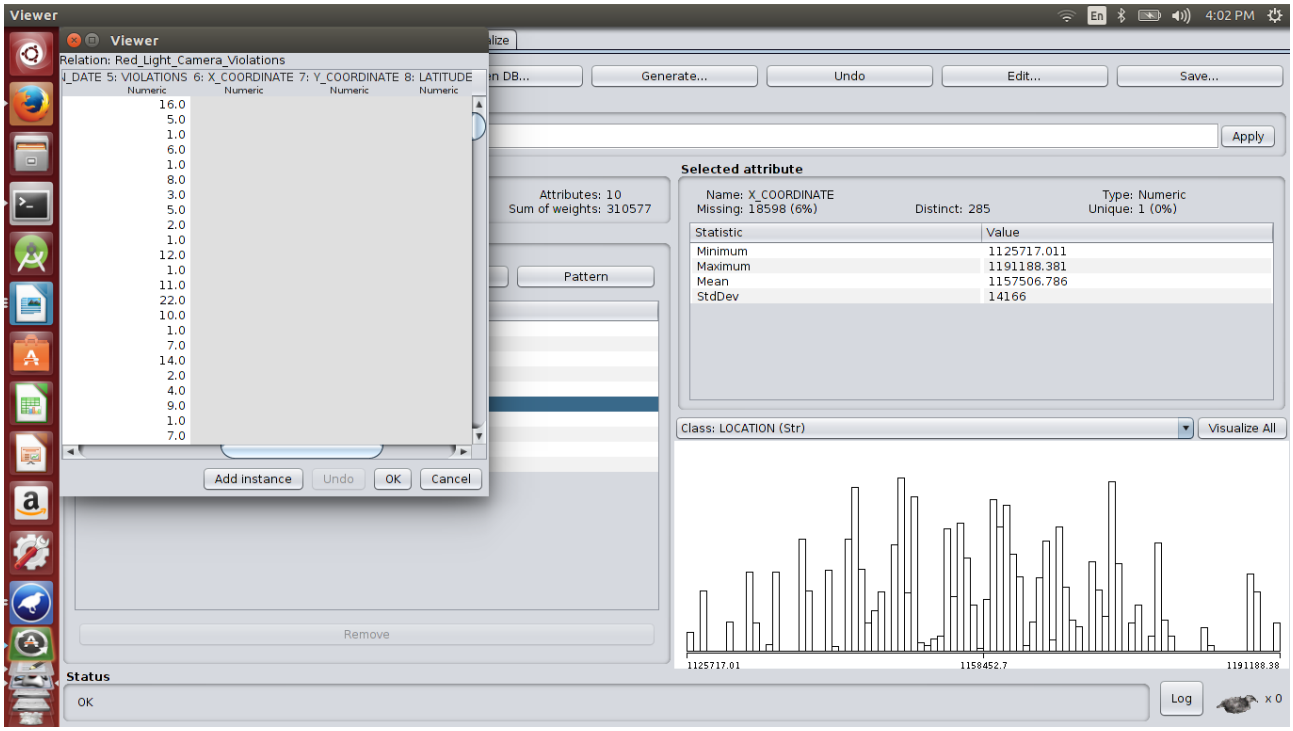
Changing date format:



5)FILLING MISSING VALUES:

Data Set initially with missing values:

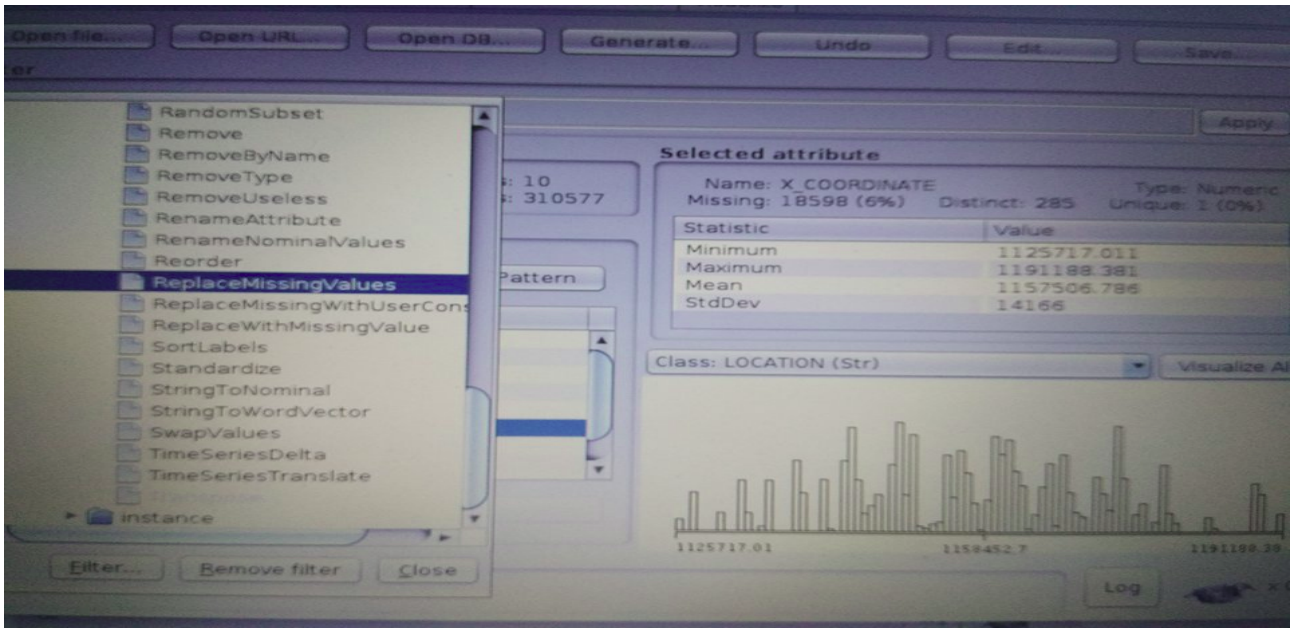
Let us fill missing values for X-coordinate .So,We select X-coordinate from preprocess tab and proceed.



ReplaceMissingValues filter.Click the “Choose” button for the Filter and select ReplaceMissingValues, it us under unsupervized.attribute.ReplaceMissingValues.

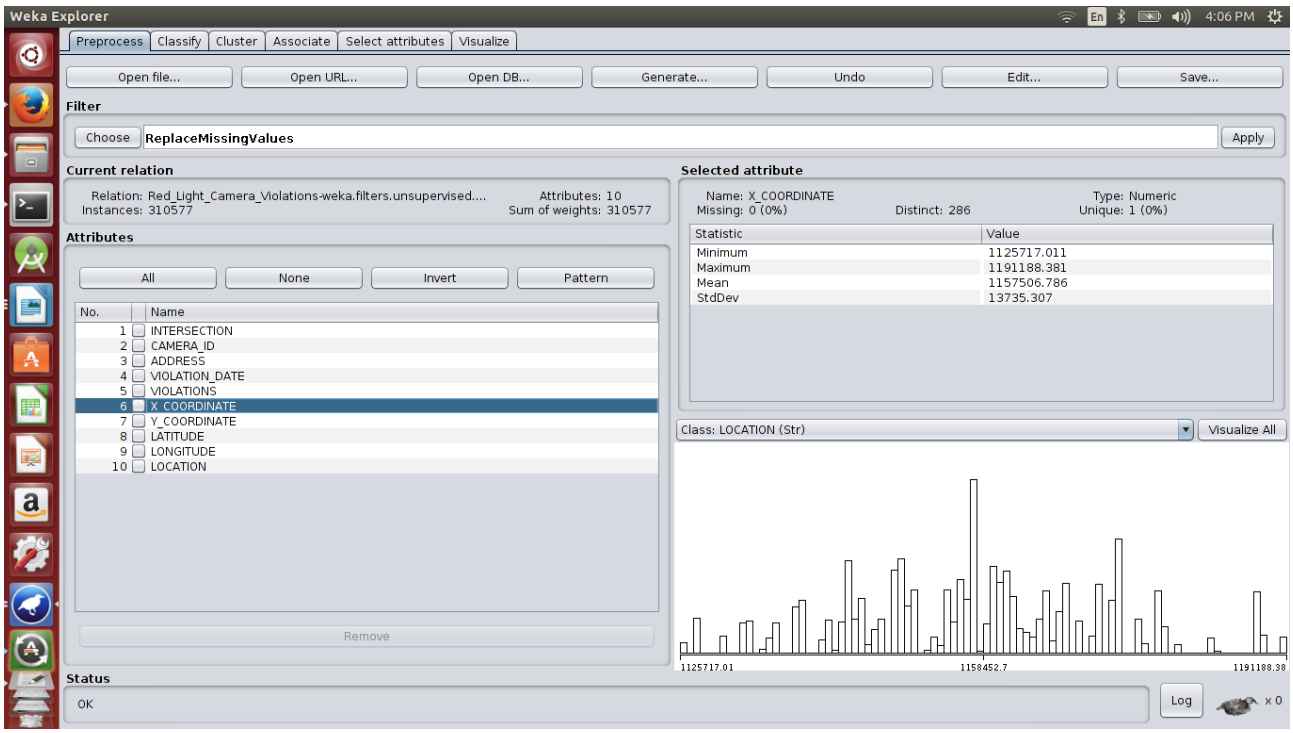
So,the missing values got replaced with mean.

Also we can choose the indices of the attributes of whose missing values we wan't to fill.Default it fills all the missing values of all attributes.



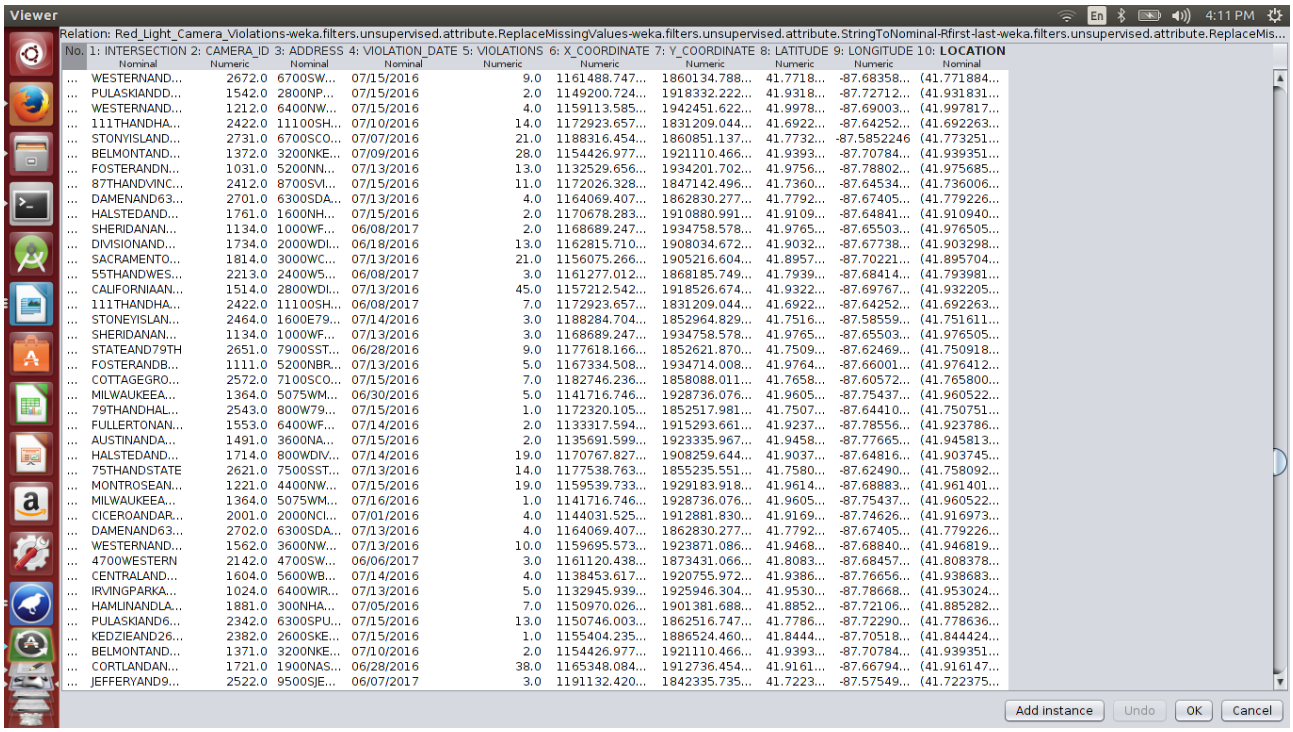
After clicking on Apply button:

We observe there are no more missing values.



Now the dataset looks like this:

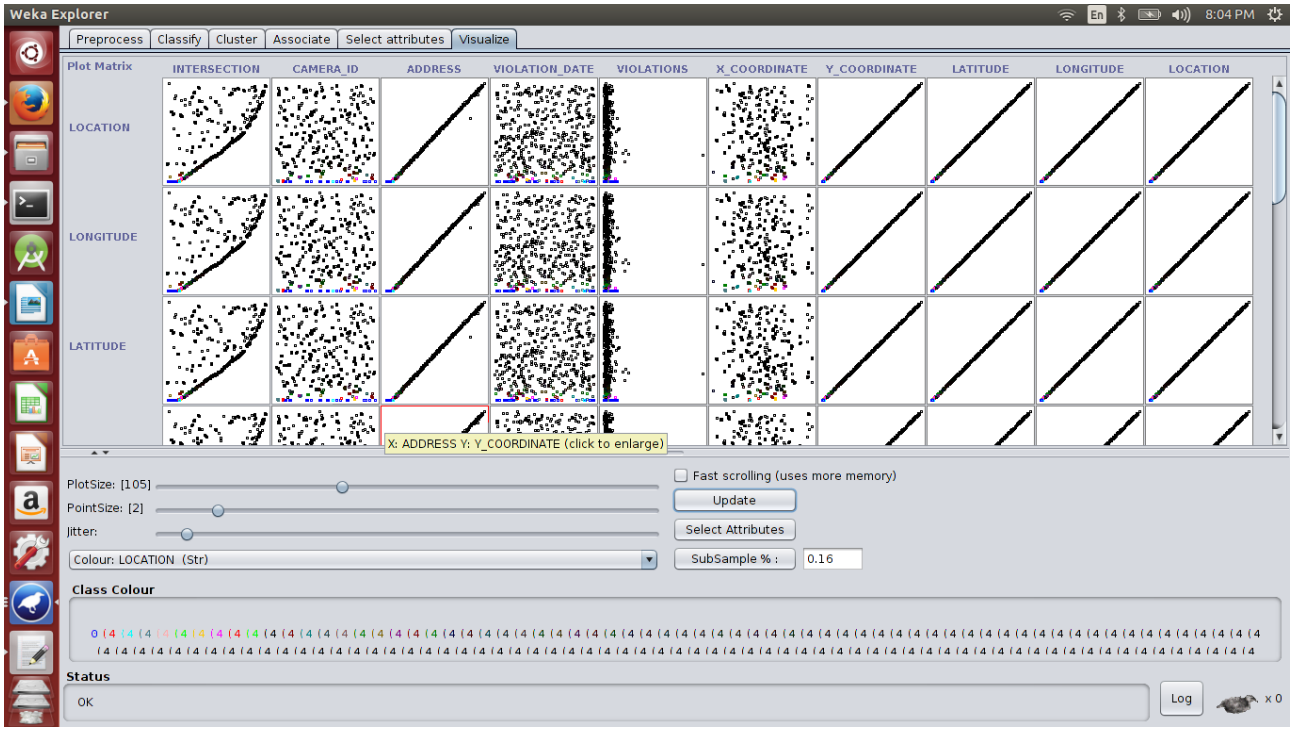
All the missing values got replaced with mean of values.



6)ScatterPlot Matrix:

When attributes are numeric we can create a scatter plot of one attribute against another. This is useful as it can highlight any patterns in the relationship between the attributes, such as positive

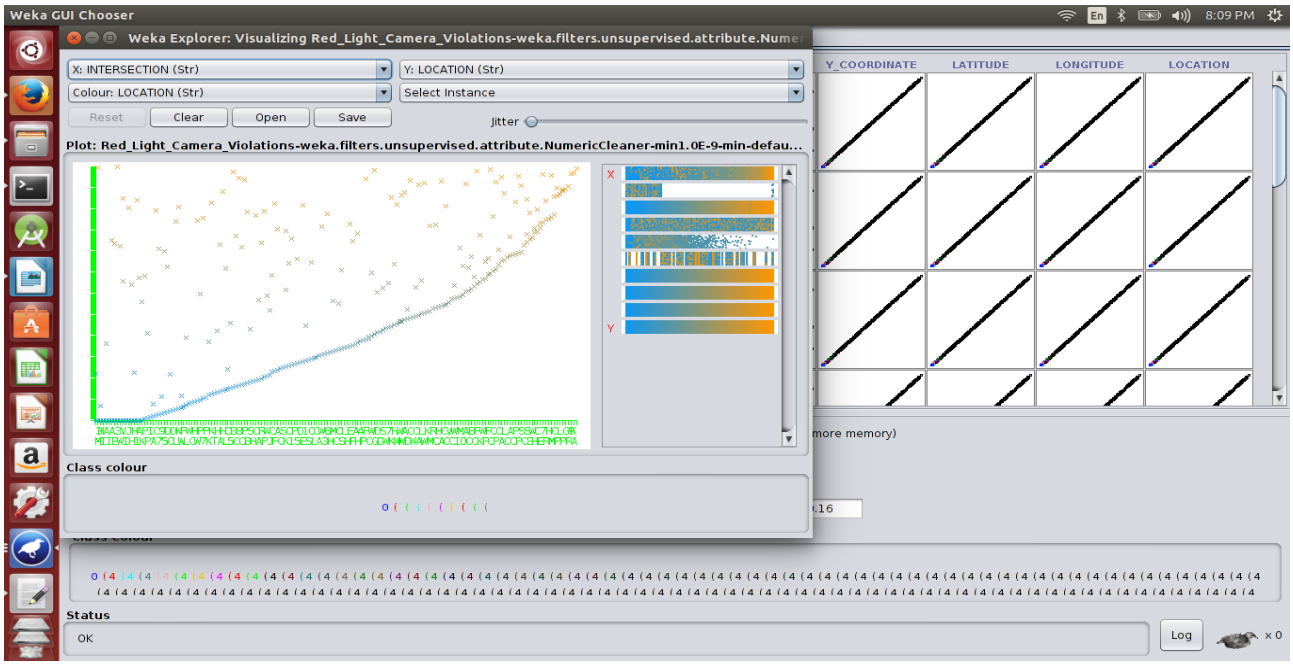
or negative correlations.So,Weka provides us Visualize tab for this purpose.



The dots in the scatter plots are colored by their class value. Clicking on a plot will give you a new window with the plot . all combinations of attributes are plotted in a systematic way.

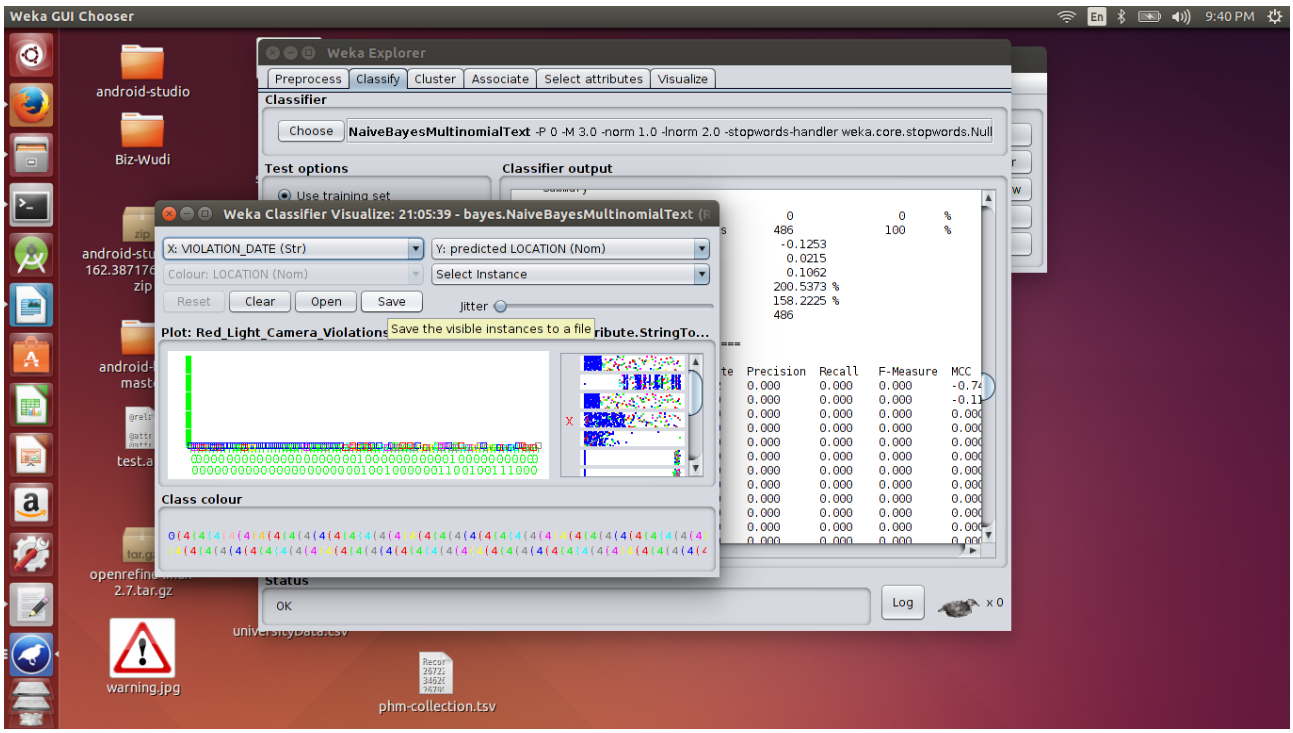
We can also see that each plot appears twice, first in the top left triangle and again in the bottom right triangle with the axes flipped. We can also see a series of plots starting in the bottom left and continuing to the top right where each attribute is plotted against itself.

Controls at the bottom of the screen. These let us increase the size of the plots, increase the size of the dots and add jitter.

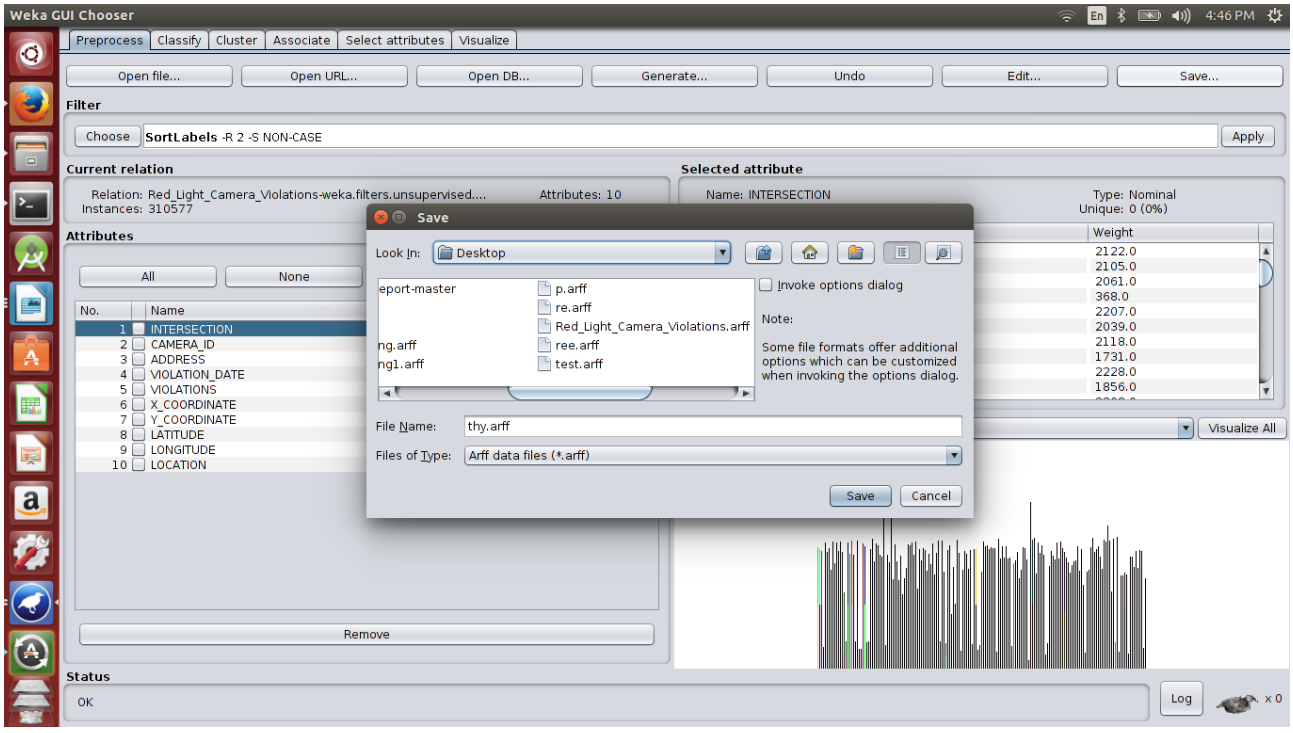


7)**Explore Data:**

Right-click the model, choose "Visualize classifier errors" and Save the data (including the prediction).

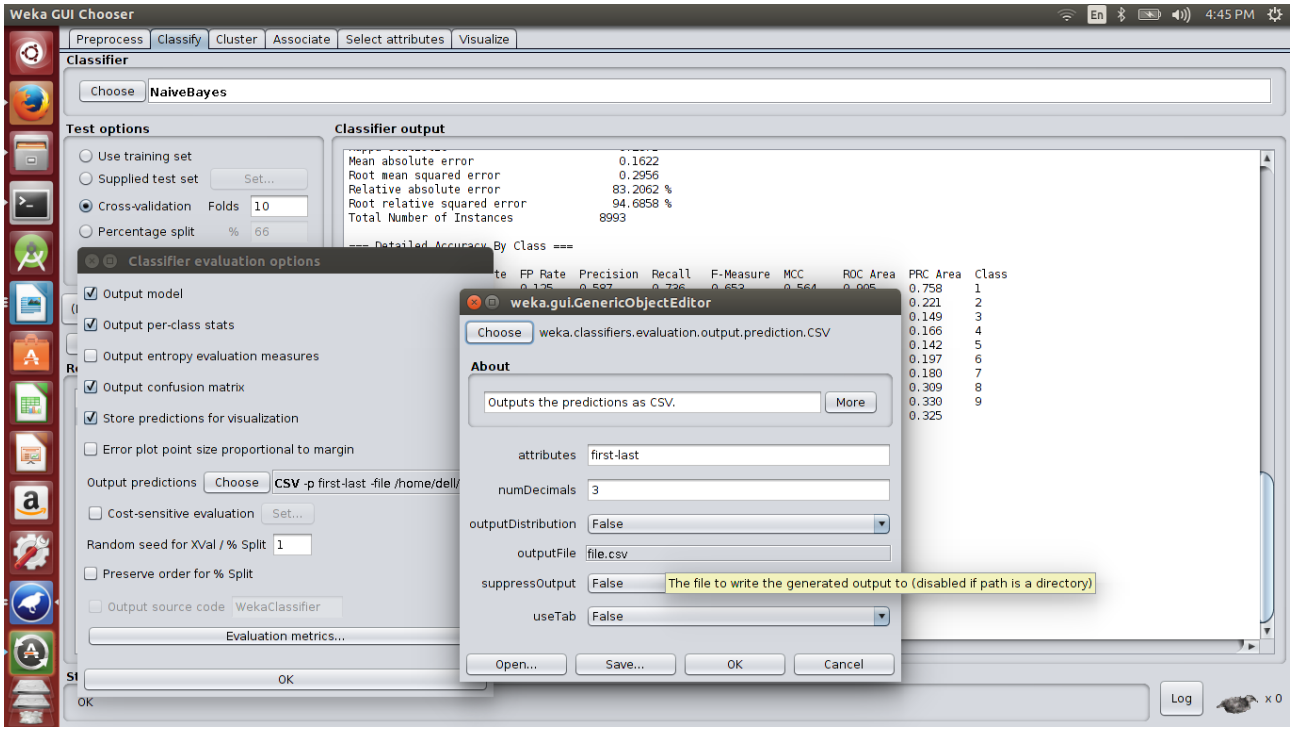


Or in the preprocess tab we have a save option where we can save the data.



And if we want data in csv format back We can do that by

selecting More options button from classify tab and then give the file format that we need and then the output file name.



Observed differences between weka and openrefine:

1)Weka provides many classifiers like naive bayes,Serialized classifier and many unlike open refine

2)Also for filling missing values weka has a direct option unlike openrefine.

3)Also we can change attributes from one type to other(nominal to numeric, String to nomial and many other) .

4)Scatterplot is obtained in both.

5)Both have options for clustering and merging.

6)So,what ever open refine have weka also has,but there are some which are not in openrefine but in weka.