Analytathon2_Report: Classification of Astronomical Light curves

Yogesh Bore

==========

1. Proposed Challenge: Classification of Astronomical Light curves

2. Exploratory Data Analysis:

Data Observations: There are 645 text files given for each objects, which contains light curve measurements of 5 years. Object distance csv file is given which contains uuid and distance of the light curves from earth. UUID is unique and identical to the each object file name. Which is used to merge these to files together.

Data Distribution: Extracted only required 5 variables MJD (Date), uJy-flux (brightness), duJy (error), F filter color and chi/n quality of flux measurement. Plot the box plots for every variable to check the data distribution. In the given fig (a) box plots given for uJy and duJy variable for one light curve. By using the graphs we can say that data is not normally distributed. Multiple outliers are present. Data is not symmetric and for duJy data is rightly skewed. All data is concentrated at the center.

Boxplots of uJy and duJy for single light curve:

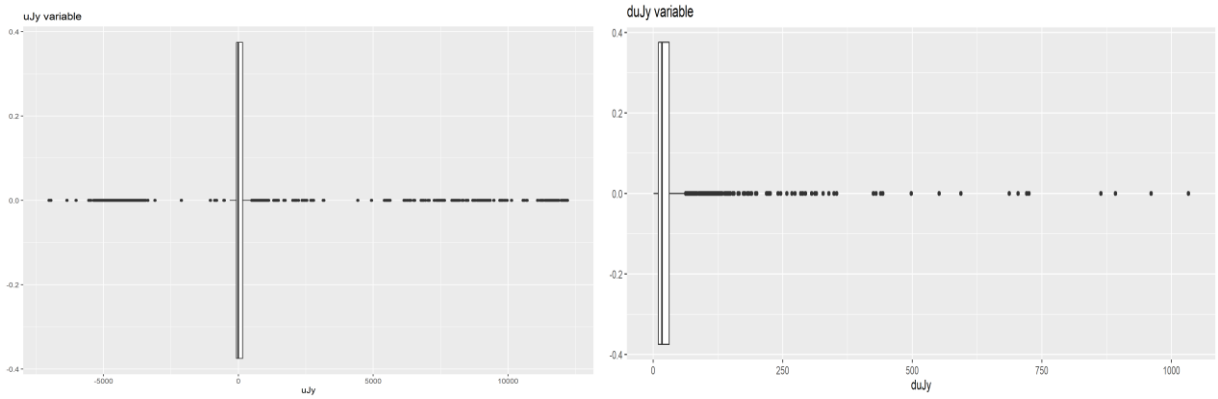


Fig. (a) Box plots for uJy & duJy

Relationship Between variables: To check this we used correlation chart fig (a2) we get histogram for all 4 numeric variables. In the diagonal which show data is not normally distributed. Below diagonal scatter plots given and we can observe that there is no relationship present between all 4 variables. Above diagonal we get the correlation variable and the significant levels for each.

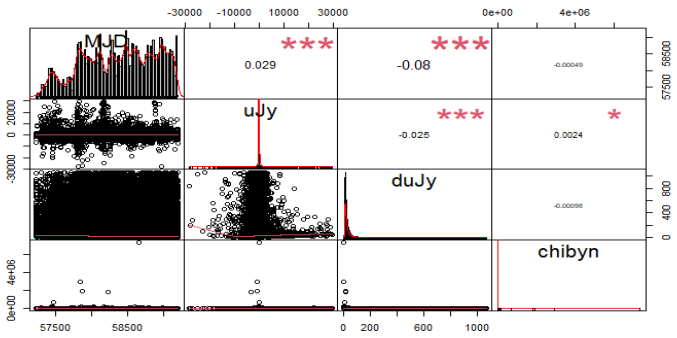


Fig. (a2) Correlation Chart

Correlation Matrix: There is no relationship found in all 4 variables before data cleaning. There are multiple outliers present.

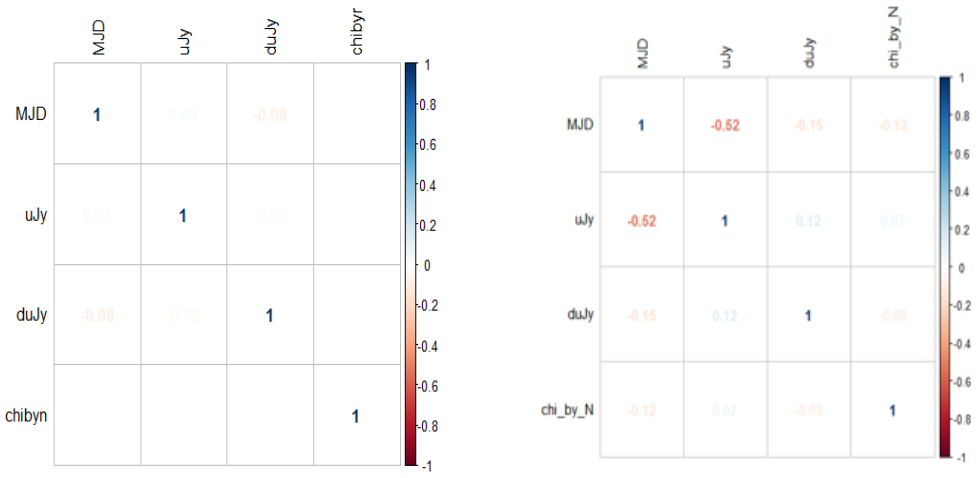


Fig. (a3) Correlation Before and after data cleaning

In the above Fig. (a3) in first matrix we did check the relation between all 4 variables and we found that there is no relationship present. However after cleaning the noisy data and removing outliers we again check the relationship. We found -0.52 moderate negative relationship present between MJD (date) and uJy (brightness) variable.

Data Cleaning:

After filtering out non-useful and missing values: By the exploratory data analysis part we came to know we have to do some data cleaning on the given data. We only consider 5 variables to improve the data retrieval time. Merged the data for all the light curves and introduced a new variable UUID. It is the unique for all the light curves and which we get it in the object distance csv file and which is same as the object file name. We also rename the variables which contains special characters such as MJD and chi/n in order to prevent any run time issues. For star 1000350600112828900.

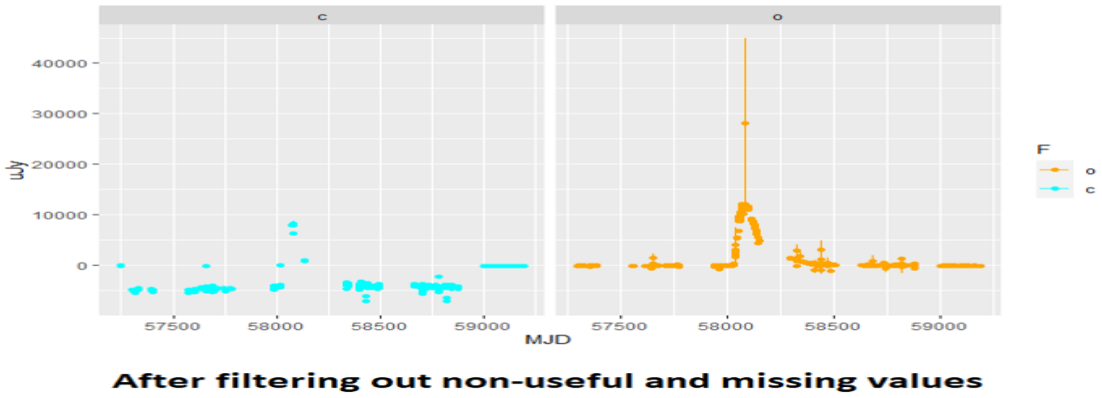


Fig. (b1)

We can see in the fig. (b1) the trend of the light curve after filtering the non-useful and missing values. However we still find some outliers in the data upon checking the messiness patter of the data. We found that there are 75 records with missing chi / n values and 184 records with negative chi/n values. As they as introduced inaccuracies in our analysis we filtered these out as well.

Also our focus of our analysis was that of the clustering of orange and cyan filter data. We filter out the values of color T.

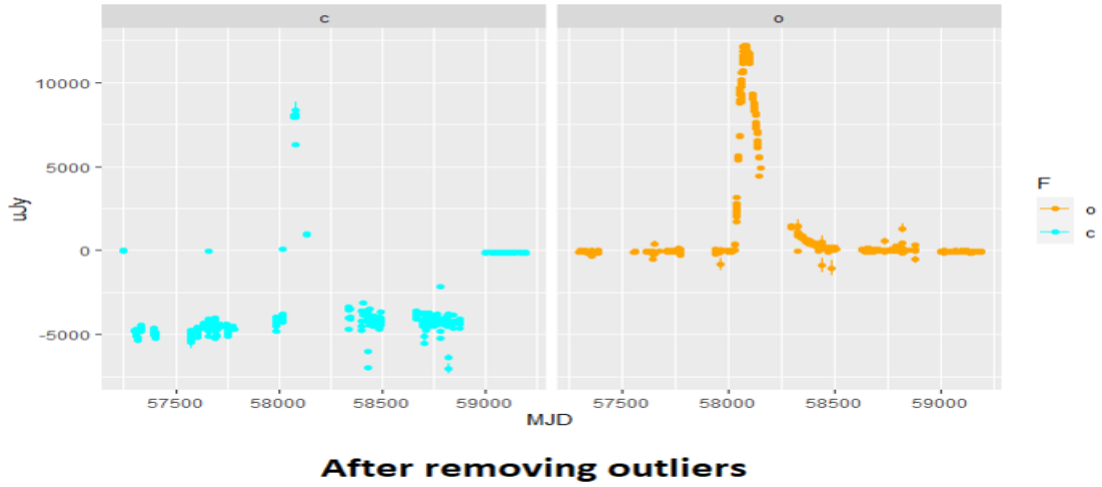


Fig. (b2)

After removing outliers: For removing the obvious outliers, we removed the value outside of two standard deviation of median flux. However as we knew that the flux data can be noisy, we further filtered out values our side of 1 standard deviation of the median flus error since lower error margins would give us better and accurate results. In the graph (After removing outliers) we can see trend of the light curve after the removal of outlier. The graph for cyan filter we can see a constant negative trend of data points. This must have happened as the baseline was adjusted after the introduction of a new target image in order to account for these negative values we calculated a rolling median over a span of 15 days and added this as a constant to the negative values for account them.

EDA and Data Cleaning Results

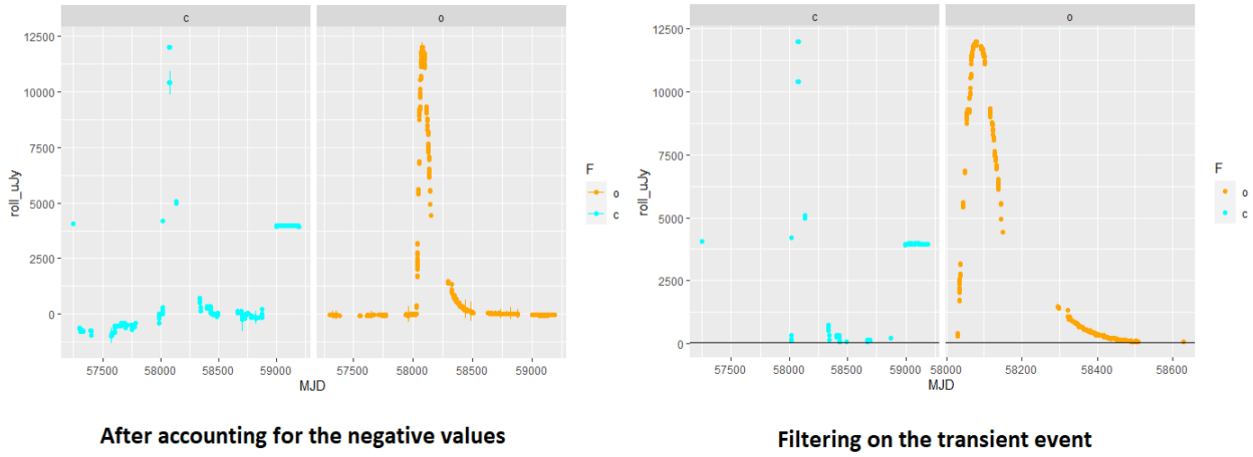


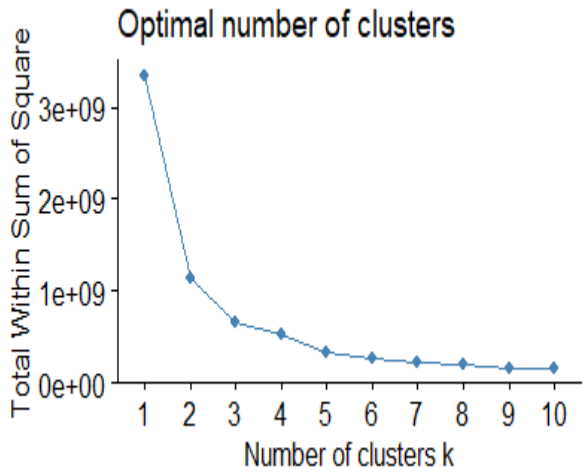
Fig. (b3)

After account the negative values they were better centered close to '0' and we also manage to preserve the trend of the light curve. The main target is the clustering based on the shape of the transient event we applied a filter on the rolling median to retain the data points greater than 60 microJanskys. This is our data ready for the clustering part.

3. Cluster Analysis:

Clustering: We get the number of classes of the light curves for the given data. Our main goal is to segregate groups with similar traits and assign them into clusters. According to the majority rule, the best number of clusters.

We set an arbitrary value of 60 microJanskys by inspecting multiple graphs as suitable point for most of the data. To isolate the light curve from rest of the relatively inactive points from the data without losing too much of the peak. Which is shown in the graph for light curve 105.




```
 1  2
74 494

 1  2  3
74 257 237

 1  2  3  4
37 39 256 236
```

```
  generated_uid peak_height totaltime
1      263.7297    205.4730 942.18828
2      332.5931    178.3441  83.25902
  generated_uid peak_height totaltime
1      263.7297    205.4730 942.18828
2      482.5642    175.4864  82.73893
3      169.9662    181.4430  83.82300
  generated_uid peak_height  totaltime
1      140.5946    234.0541 1101.21779
2      386.0000    178.5128  768.31385
3      482.6328    175.3047   81.07143
4      169.5212    181.3856   82.15383
```

Fig. (c) Cluster Numbers

The table shows the output from each of the popular cluster. The left side table shows total number in each cluster. The second table shows the centers of these clusters. Observing the values we get that 4 clusters best represented the data. Set flux to be above 60, works for majority K-means clustering GSS Graph Plot is for star 105 Extremely High peak flux (Star 195 shown)

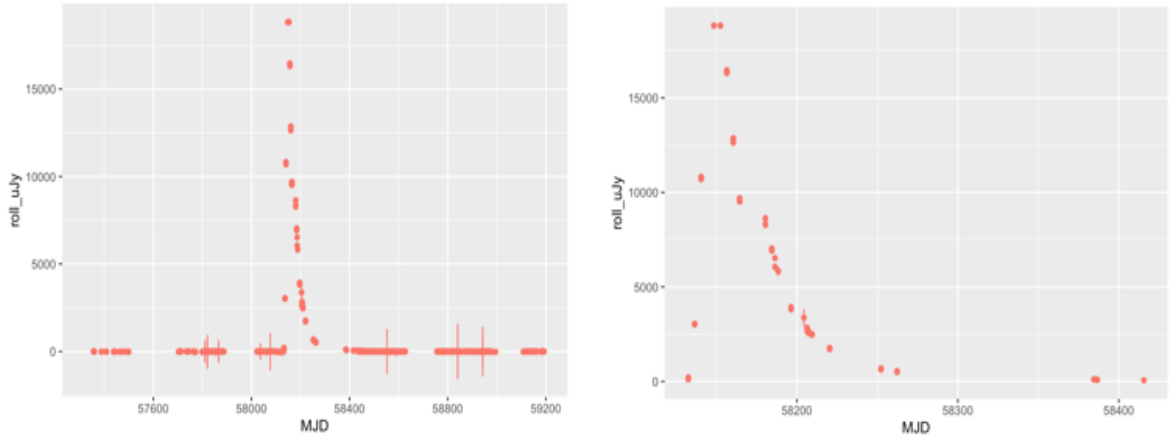


Fig. (c1) Cluster1: Extremely High peak flux

Cluster 1 contains only about 4 light curves. These all have extremely high peak flux. We can observe in the above Fig. (c1) the peak for the orange data for a randomly selected light curve.

Fig. (c2) Cluster2: Extremely long time elapsed due to arbitrary flux cut off

Cluster 2 conatns 2 light curves, Cluster possibly contains peaks that would possibly work better in other categories. This is shown for the orange data for a randomly selected light curve.

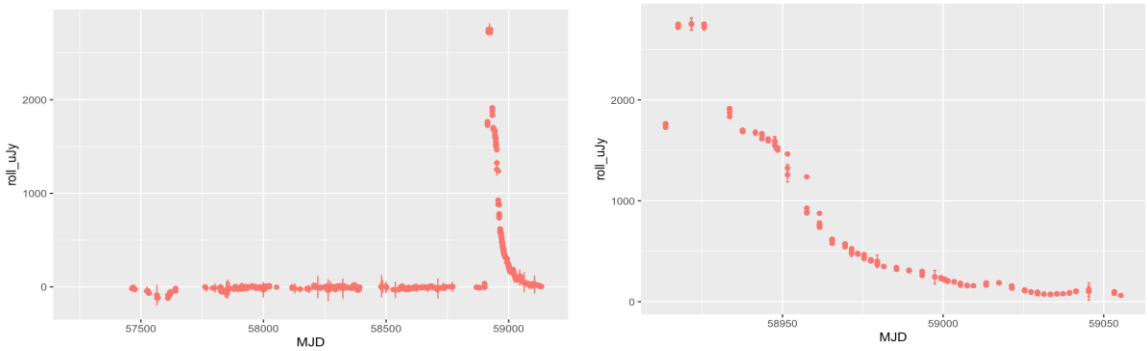


Fig. (c3) Cluster3: High peak, long tail

Cluster 3 shows relatively high peaks and long tail these are possibly type one supernova. Above curve is for orange data. However we got same observations for cyan data for a single light curve.

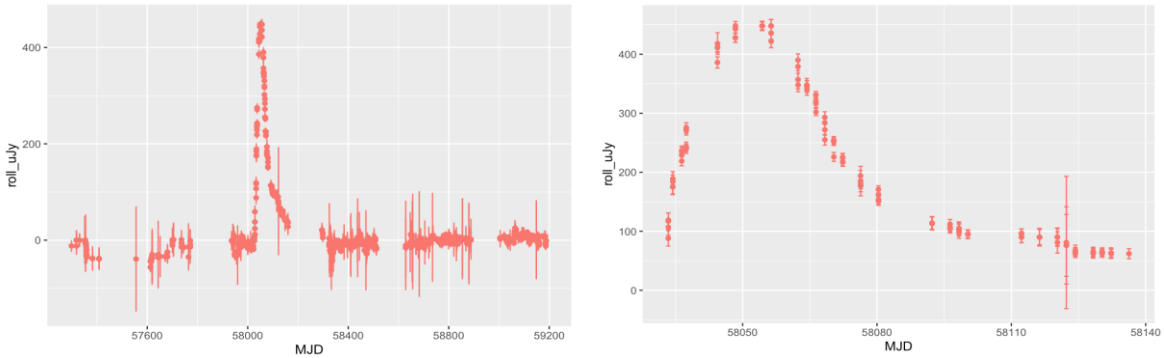


Fig. (c4) Cluster4: Low peak, short tail

Cluster 4 shows light curves with lower peak short tails and these are possibly type 2 supernova. For all 4 cluster we try to plot the graphs for both orange and cyan data. However here we have used only orange data graphs. But we got same nature for the cyan data as well. This graph is plotted for a randomly selected light curve from each cluster. So when we run the same script for other light curve you may get some different results for both orange and cyan data.

4. Possible Further Analysis

Use light curve functions Find better way to isolate peaks, remove later spikes in flux equal to flux in peak. Plot is for light curve 200.

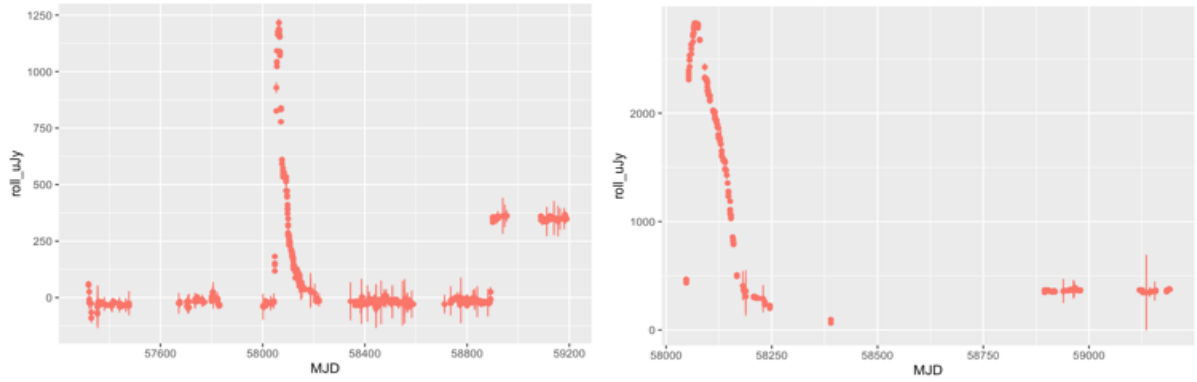


Fig. (c5) Remove the later spikes

Able to remove outliers effectively Able to classify most curves, into type 1 and 2 Unable to identify subclasses of each type Unable to isolate every peak in the different light curves.

GAM (generalized additive model): Time series analysis is a technique to derive a trend across time, which might be used to predict future values. A GAM does this by identifying and summing multiple functions that results in a trend line that best fits the data. GAM analysis before of the rolling median for the flux variable and check the results. Then applied the filter on the flux and set an arbitrary value of 60 microJanskys and tried to apply the gam again and check the result.

GAM Model Data	GAM Model Result R^2
With Whole Flux	59.30%
Flux above 60	76.20%

Table (1) GAM Model Results

In the above table we can see the gam model results after filtering the flux values increase. We tried to apply gam on the whole data first and then apply then filter on the flux we got the data for all the light curves having flux more than or equal to 60 and then same filter data used for gam and observed a big difference in the results it increase by 16.9 %.

5. Conclusion:

We have successfully classify the Astronomical Light curves by using the available 5 years data. There are lot of outliers present in the given data, removed the outliers and noisy data before performing the clustering. Also checked the relationship between the variables by using the box plots and correlation chart. Before data cleaning we didn't find any relationship present between the data. However we checked the same correlation matrix after data cleaning found the moderate negative relationship present between the data. Then applied k-means algorithm on the cleaned data for the classification. We get the number of clusters. Plotted graphs for light curve from each object and observe the trends. Applied generalized additive model on whole flux and compared the result with same model with flux above 60 microJanskys and get the good results.