Analython-2 Classification of Astronomical light curves

ABSTRACT

This report aims to represents my analysis by clustering out the supernovae with there light curve shape and understand different patterns.

```
##      dplyr    ggplot2  tidyverse     readxl     plotly      tidyr     readxl
##       TRUE       TRUE       TRUE       TRUE       TRUE       TRUE       TRUE
##   corrplot     naniar       Rcpp      caret    caTools factoextra     visdat
##       TRUE       TRUE       TRUE       TRUE       TRUE       TRUE       TRUE
##       cvms      dplyr    stringr data.table   epitools   dtwclust       roll
##       TRUE       TRUE       TRUE       TRUE       TRUE       TRUE       TRUE
```

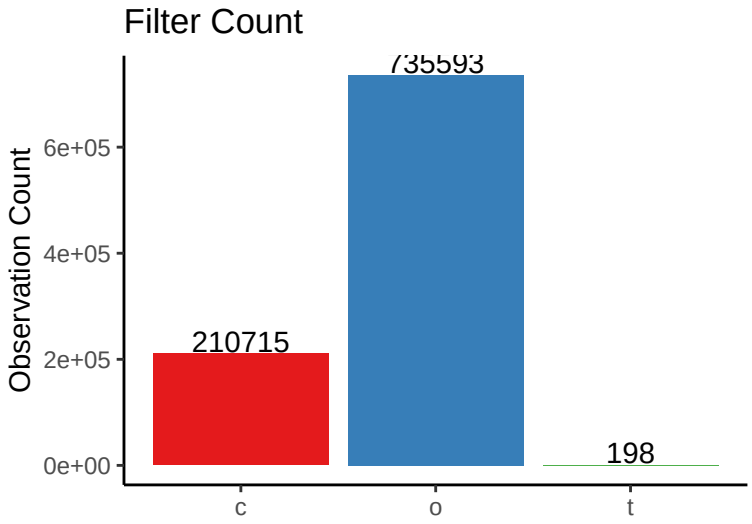
DATA PREPROCESSING

- Extracting useful feature “MDF”,“uJy”,“duJy”,“F”,“chi/N” from the given 645 raw text files Note: MDF = Modfied Julian Date, uJy = Flux (Luminosity), duJy = Error in Flux, F = Telescope used to capture (o/c), chi/N=Quality
- Rename the feature with a valid name “julian_date”, “flux”, “error_flux”, “filter”, “quality”
- Extract the unique Id from the file names and create a new column “series”
- Remove the impact of distance from flux by applying formula and create new column value “flux_intrinsic”

```
##    julian_date flux error_flux filter quality series flux_intrinsic
## 1:    57248.47 -125         92      c    0.94   2776      -15135228
## 2:    57248.49  -84        113      c    1.16   2776      -10170873
## 3:    57248.52 -220        359      c    0.98   2776      -26638002
## 4:    57248.55  125        796      c    1.10   2776       15135228
## 5:    57313.37  -47         19      c    0.87   2776       -5690846
## 6:    57313.40  -14         20      c    0.94   2776       -1695146
```

FEATURE ENGINEERING

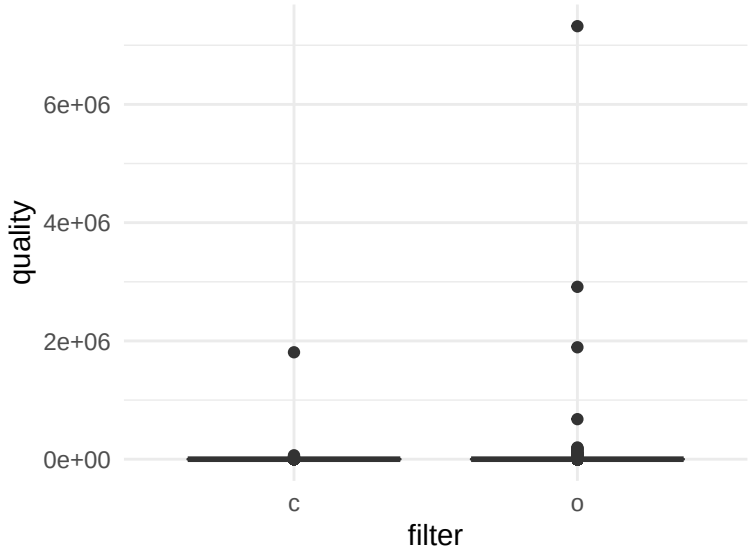
- Apart from “o” and “c” there are around 198 entries with filter name “t”, This is irrelevant and hence to be removed

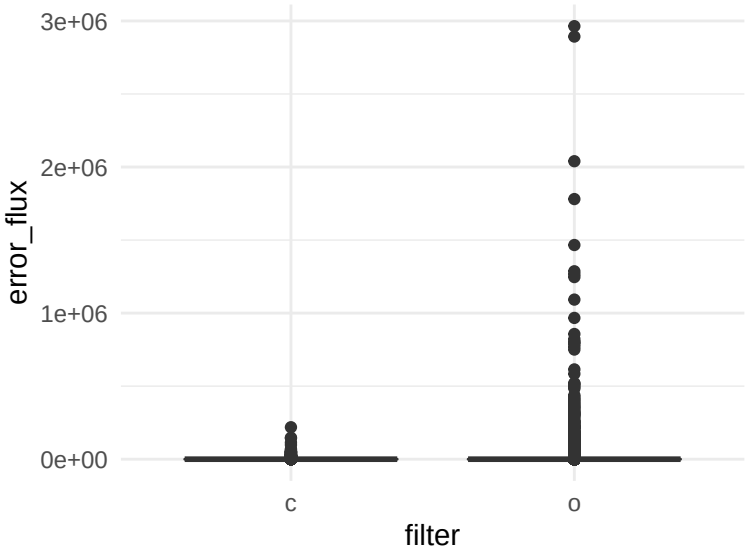


- There are around 78 NA entries in Quality column and these are removed from the data set

```
##                Total_NA Percentage_of_NA
## quality              75             0.01
## julian_date           0             0.00
## flux                  0             0.00
## error_flux            0             0.00
## filter                0             0.00
## series                0             0.00
## flux_intrinsic        0             0.00
```

- There are plenty of outliers seen in the Error and Quality features and these are been removed from dataset





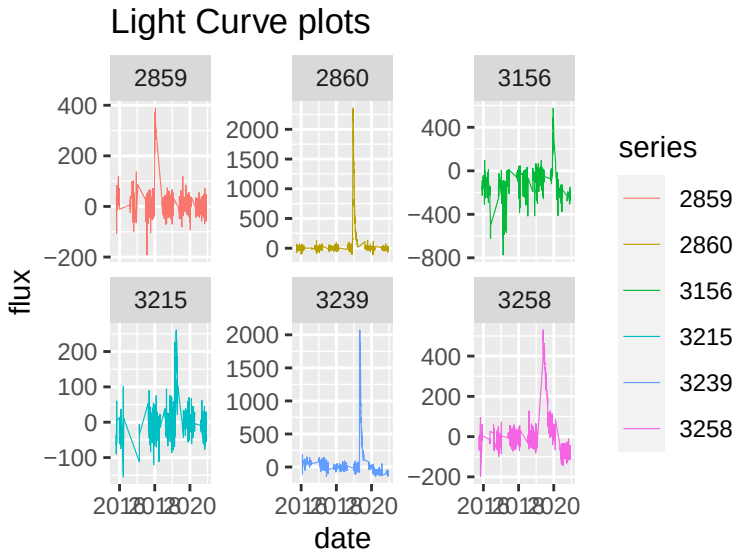
- The Modified julian date is convert Calendar date by creating a new column for better understanding

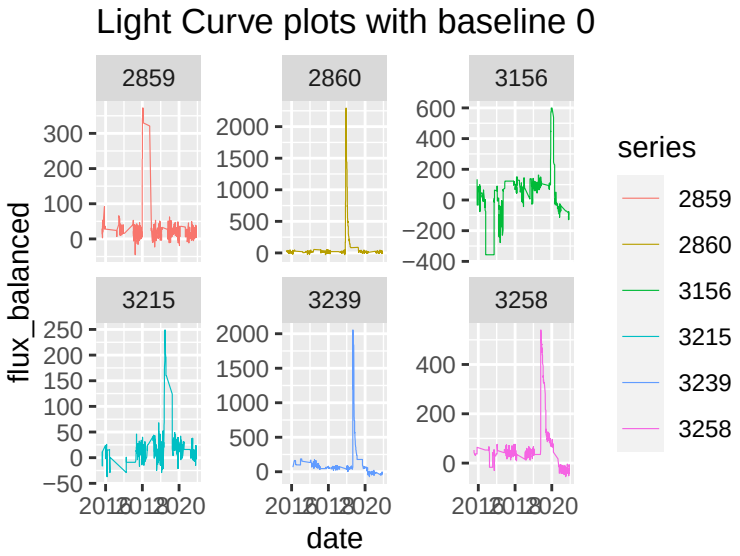
```
##    julian_date flux error_flux filter quality series flux_intrinsic       date
## 1:    57318.26 -102         46      o    0.73   2776      -12350346 2015-10-23
## 2:    57318.28 -161         38      o    1.24   2776      -19494174 2015-10-23
## 3:    57318.30 -128         38      o    1.16   2776      -15498474 2015-10-23
## 4:    57318.32 -118         39      o    0.84   2776      -14287656 2015-10-23
## 5:    57323.37 -169         46      o    1.39   2776      -20462829 2015-10-28
## 6:    57323.40  -67         55      o    0.88   2776       -8112482 2015-10-28
```

- Rather than taking the actual flux value, we did take a rolling median 5 on flux which would help us in getting much better Flux Intensity and these which will help us in spotting different patterns.

```
##          date julian_date flux flux_intrinsic error_flux filter quality series
## 1: 2015-11-08    57334.36  -54       -6538419         18      c    0.99   2776
## 2: 2015-11-08    57334.39  -80       -9686546         17      c    0.91   2776
## 3: 2015-11-08    57334.42  -35       -4237864         19      c    1.02   2776
## 4: 2015-11-08    57334.45  -58       -7022746         22      c    0.85   2776
## 5: 2015-11-15    57341.32  -33       -3995700         18      c    0.70   2776
## 6: 2015-11-15    57341.35  -51       -6175173         19      c    0.69   2776
##    rolmedian_1 rolmedian_2 rolmedian_3 rolmedian_4 rolmedian_5
## 1:         -54       -31.5         -54       -34.0         -47
## 2:         -80       -67.0         -54       -55.5         -54
## 3:         -35       -57.5         -54       -44.5         -54
## 4:         -58       -46.5         -58       -56.0         -54
## 5:         -33       -45.5         -35       -46.5         -54
## 6:         -51       -42.0         -51       -43.0         -51
```

- There are few light curve series which has flux value negative. We are verifying the number of point which are below and if the number is greater than threshold then we create a baseline where 75% of the data is present and moving it to 0 and thus we can see there are some high intensity explosions seen.





MACHINE LEARNING (UNSUPERVISED) MODEL

Hierarchial clustering with K shape

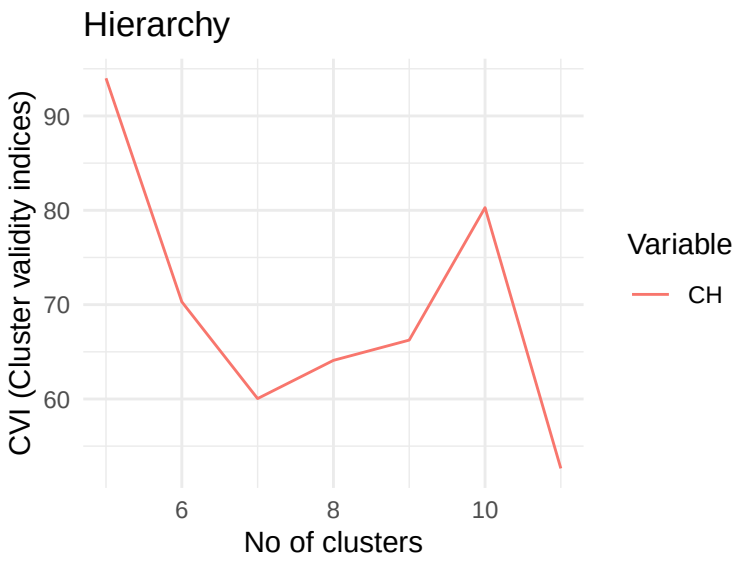
As part of Hierarchical clustering model, rather than taking all the files together and treating as one single time series, we decided to deal with each individual light curve, where every file is a time series object. Every object is represented by its unique id.

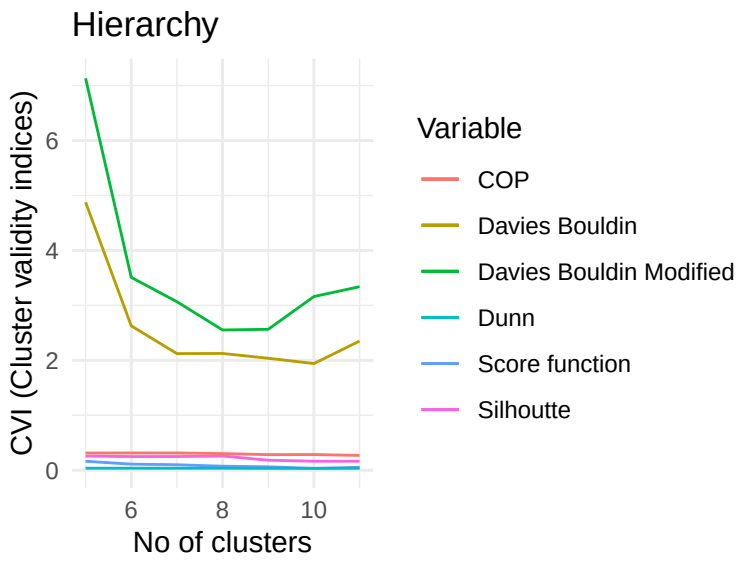
To make the model execution simpler and to verify the patterns captured by different telescopes we decided to separate the data based on the filters.

Time-series shape extraction based on optimal alignments as proposed by Paparrizos and Gravano (2015) for the k-Shape clustering algorithm is the one of the best approaches proven for clustering out time series. It used cross-correlation distance measure to compare different time series.

The model was evaluated on clusters range from 5 to 11 for filter 0 and cluster validity indices (CVIs) metrics is used to spot the right optimal model. Within CVI we have different distance measures between the points and definitely the one having the maximum value will be considered as an optimal model and here we see that a model with 8 clusters is giving us good number.

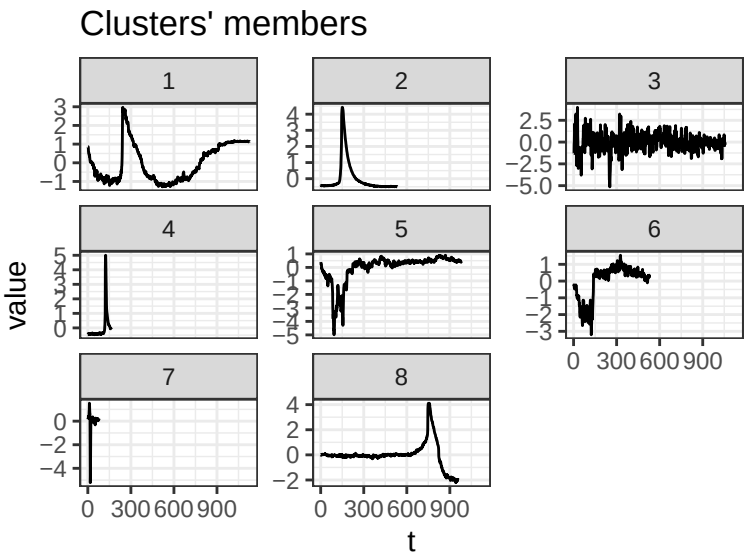
```
##          Sil         SF       CH       DB   DBstar          D       COP Cluster
## V1 0.2571002 0.16331099 93.99667 4.878716 7.134001 0.03645821 0.3132030       5
## V2 0.2486593 0.11106844 70.29681 2.627880 3.510228 0.03648397 0.3143390       6
## V3 0.2493240 0.09992154 60.04122 2.122080 3.067318 0.03649845 0.3137466       7
## V4 0.2583913 0.07301702 64.09330 2.124993 2.552226 0.03669359 0.3026387       8
## V5 0.1829315 0.06249999 66.24121 2.037908 2.564769 0.03397627 0.2837012       9
## V6 0.1631523 0.03309889 80.28204 1.941906 3.160972 0.03425107 0.2849752      10
```





The first set of colorful graphs gives an overview of the series falling under each cluster category, which means all 645 light curves are getting represented here and there other set of graph shows the underlying shape or protype or patterns detected by model and this gives us a clear indication on how some of the light curves don't fit into any class and some are too noisy to tell.

Cluster category 1,2,4,8 gives a good pattern to Identify Light curve Cluster category 3,6,7 is too noisy and may be due to external noise factors and not supernova

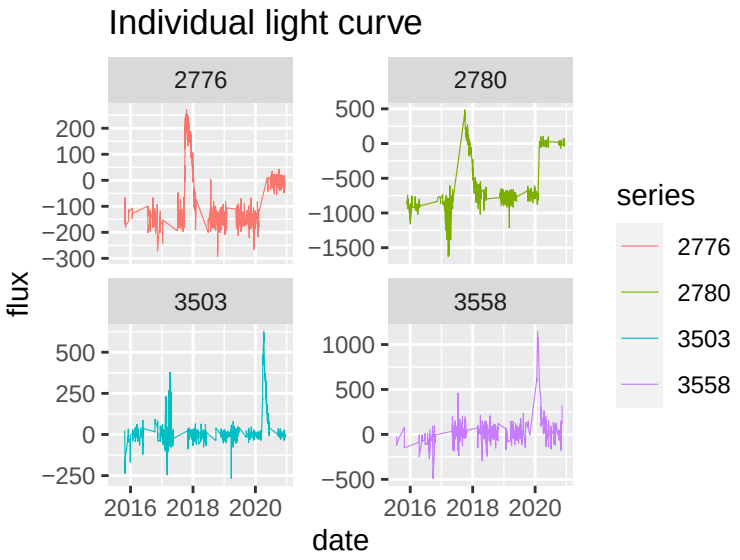


The first table gives the count of light curves falling into each cluster category and second table reference provides a glimpse for each light curve and the cluster bucket its falling into and we also did plot these time series just to make sure it does match with the shapes created by model.

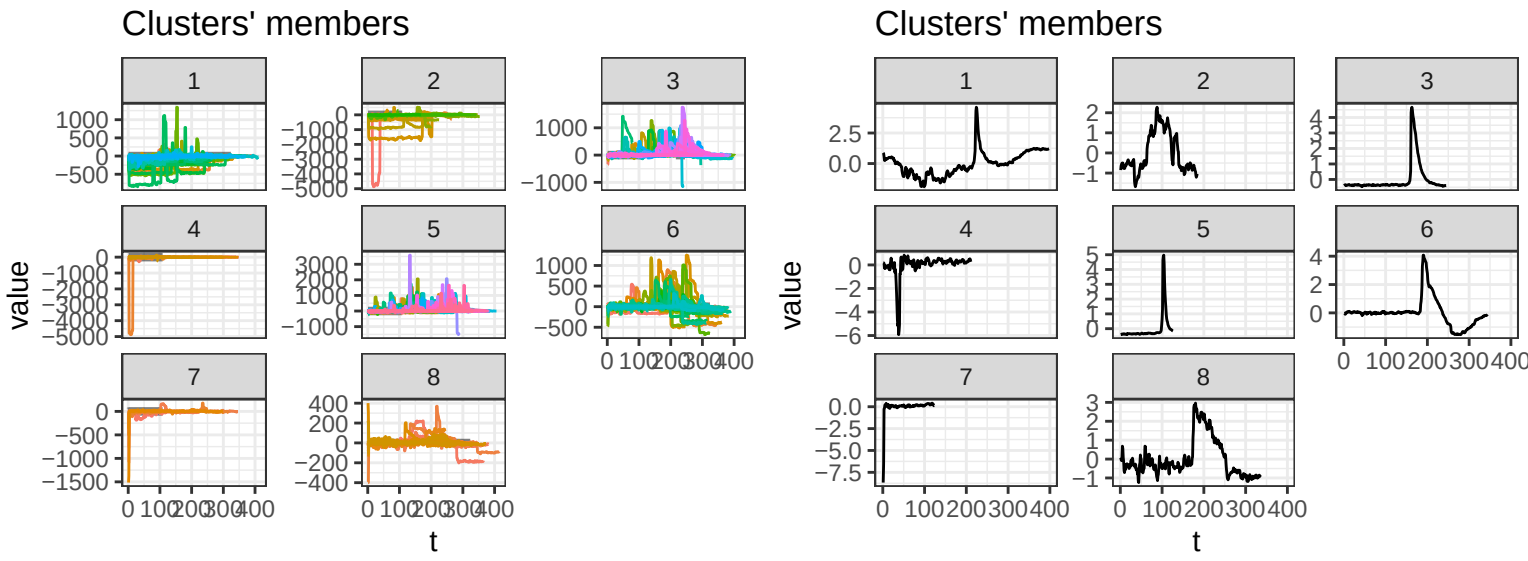
```
##   Cluster Category Count
## 1                1    35
## 2                2   383
## 3                3     2
## 4                4    96
## 5                5    35
## 6                6    22
## 7                7     7
## 8                8    65
```

```
##      Cluster Category
## 2776                1
```

```
## 2777                1
## 2778                2
## 2779                2
## 2780                1
## 2781                1
## 3501                2
## 3502                2
## 3503                2
## 3556                8
## 3557                2
## 3558                4
```



To be sure about this optimal model with 8 cluster selected we did perform clustering filter c data and yes we did get a good representation of clusters and there pattern. We did get different patterns here because the intensity of explosions captured by both the telescopes are different

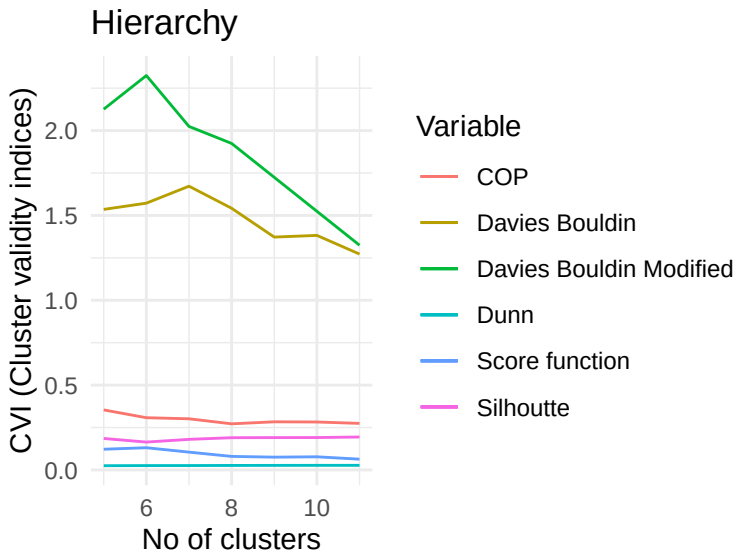


Hierarchial clustering with Dynamic time wrapping

The Metrics for Hierarchical Dynamic time wrapping showed no improvement with the clusters formed.

```
##         Sil         SF        DB    DBstar          D       COP Cluster
```

```
## 1 0.1854152 0.12207209 1.535471 2.126024 0.02483722 0.3536827       5
## 2 0.1641961 0.13086889 1.572307 2.324427 0.02576396 0.3078850       6
## 3 0.1802415 0.10494959 1.672307 2.024427 0.02591147 0.3014194       7
## 4 0.1899439 0.07967956 1.542307 1.924427 0.02647371 0.2718630       8
## 5 0.1905580 0.07542527 1.372307 1.724427 0.02679783 0.2836204       9
## 6 0.1908086 0.07747300 1.382307 1.524427 0.02702483 0.2825687      10
## 7 0.1939641 0.06342924 1.272307 1.324427 0.02709376 0.2743051      11
```



CONCLUSION

With all the metric values observed and cluster patterns plotted we conclude that Shape extraction Hierarchical model did a balanced grouping within the light curves and gave us the better results and that brings us to the end of the presentation

REFERENCES

- https://cran.r-project.org/web/packages/dtwclust/vignettes/dtwclust.pdf
- http://rstudio-pubs-static.s3.amazonaws.com/398402_abe1a0343a4e4e03977de8f3791e96bb.html
- https://rpubs.com/imartinezl/tsclustering
- https://journal.r-project.org/archive/2019/RJ-2019-023/RJ-2019-023.pdf
- https://www.programmersought.com/article/71856995021/
- https://rpubs.com/KaraLynne/382832