Rumainum-HW8

Lince Rumainum

The data set used for this homework is about the online news popularity. These data set is based on the news from *masahable* website. It has close to 60 attributes that could help predict how many times certain news is shared. This data set can be found at University of California, Irvine (UCI)'s website: https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity.

There are three types of clustering performed on this data set. **Figure 1** shows the k-mean clustering, **Figure 2** shows the k-medoids clustering, and **Figure 3** shows the hierarchical clustering with the red box showing the groups when we "cut" the data set to create 2 distinct clusters.

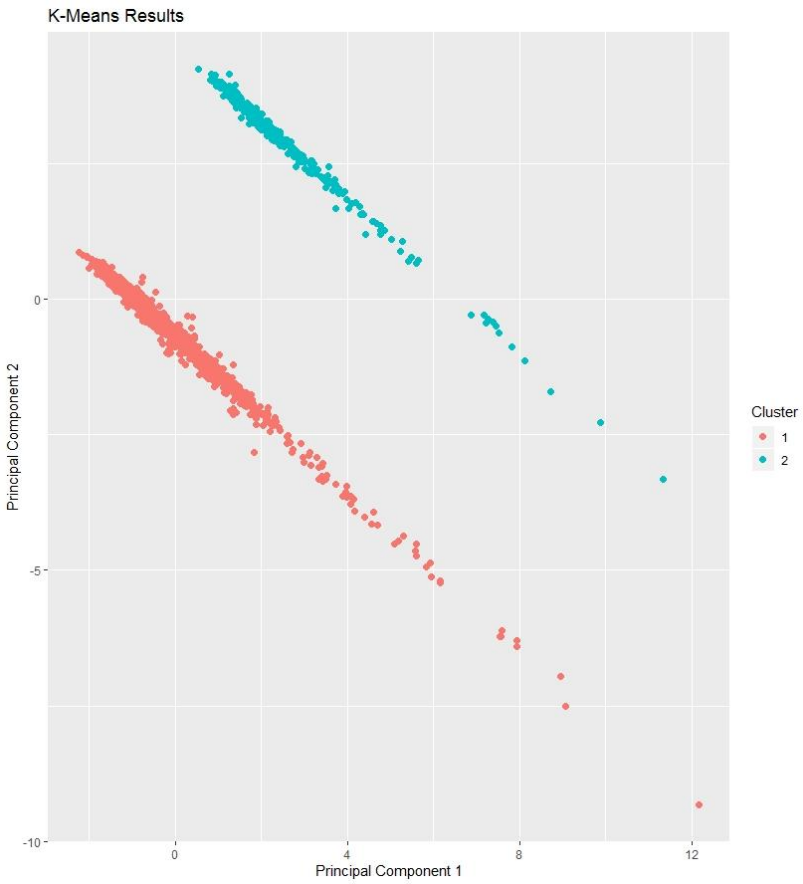

Figure 1 *K-Mean Clustering of the Data Set*

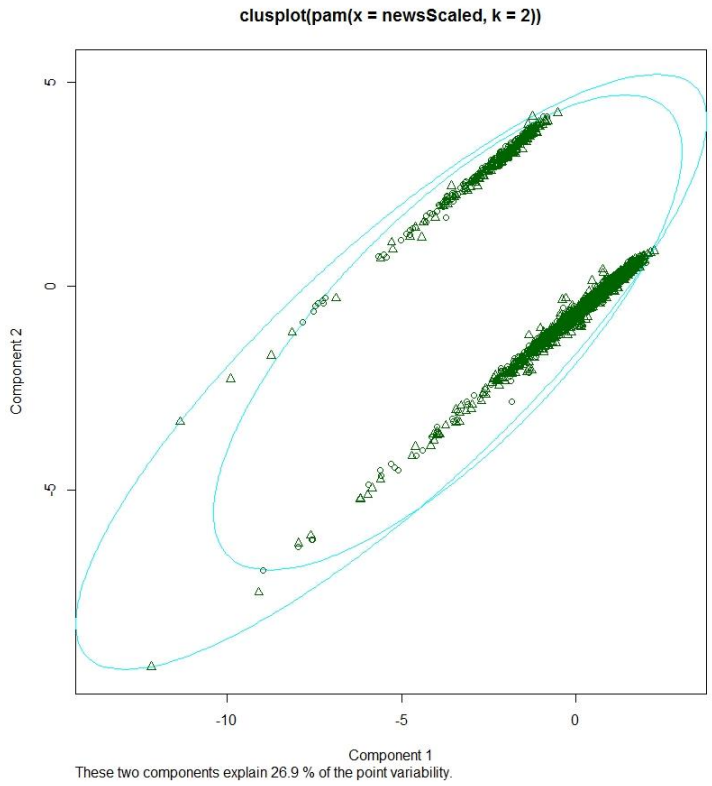

Figure 2 *K-Medoids Clustering on the Data Set*

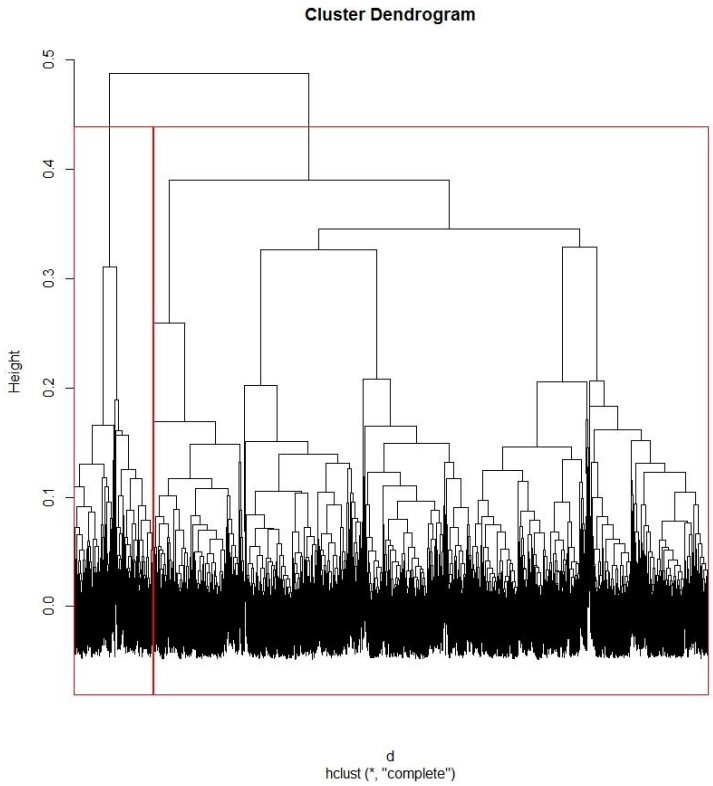

Figure 3 *Hierarchical Clustering on the Data Set*

From **Figure 1** and **Figure 2,** you can see the visualization of the difference of the k-mean and k-medoids clustering and also in **Figure 5** you can see the different values of the centroids and medoids of each attributes for each cluster. On **Figure 4,** it shows the hierarchical clustering from **Figure 3** where the data is "cut" into two clustering data in regular PC2 vs PC1 plot.

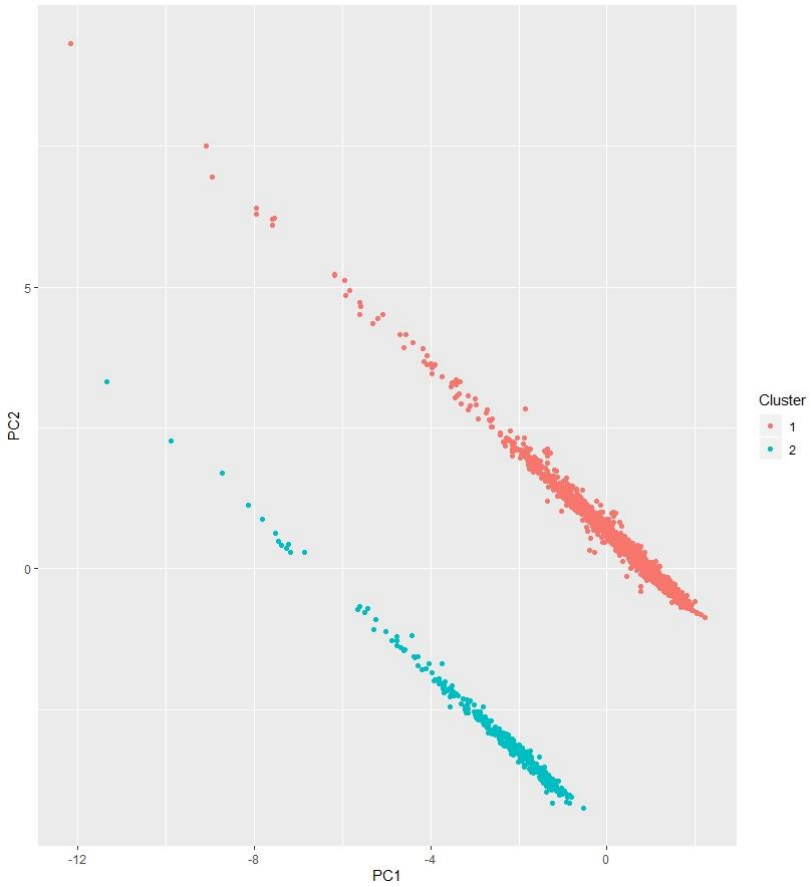

Figure 4 *The Q-Plot of the Hierarchical Clustering on the Data Set*

```
> # compare centroids and medoids
> round(kclus$centers, 3)   # centroids from KM
  n_tokens_title n_tokens_content num_hrefs num_self_hrefs num_imgs num_videos num_keywords weekday_is_monday weekday_is_tuesday weekday_is_wednesday weekday_is_thursday
1          0.001          -0.014    -0.026         -0.027   -0.025      0.005       -0.020            0.065             0.068               0.068              0.069
2         -0.003           0.101     0.182          0.188    0.171     -0.031        0.144           -0.457            -0.476              -0.477             -0.481
  weekday_is_friday weekday_is_saturday weekday_is_sunday is_weekend shares
1             0.058             -0.254           -0.262     -0.378 -0.011
2            -0.410              1.783            1.833      2.647  0.076
> pclus$medoids  # medoids of PAM
     n_tokens_title n_tokens_content num_hrefs num_self_hrefs num_imgs num_videos num_keywords weekday_is_monday weekday_is_tuesday weekday_is_wednesday weekday_is_thursday
[1,]          0.241          -0.223    -0.264         -0.322   -0.191     -0.049        0.416           -0.457             2.101              -0.477             -0.481
[2,]         -0.226           0.031    -0.090         -0.322   -0.313     -0.049       -0.644           -0.457            -0.476               2.094             -0.481
     weekday_is_friday weekday_is_saturday weekday_is_sunday is_weekend shares
[1,]             -0.41             -0.254           -0.262     -0.378 -0.223
[2,]             -0.41             -0.254           -0.262     -0.378 -0.118
```

Figure 5 *Centroids from K-Mean and Medoids of PAM Clustering Methods*

Figure 6 shows the difference of the centroids and medois of K-Mean and PAM, respectively on two different plot, which are shares Vs. weekday_is_tuesday and shares Vs. weekday_is_wednesday.

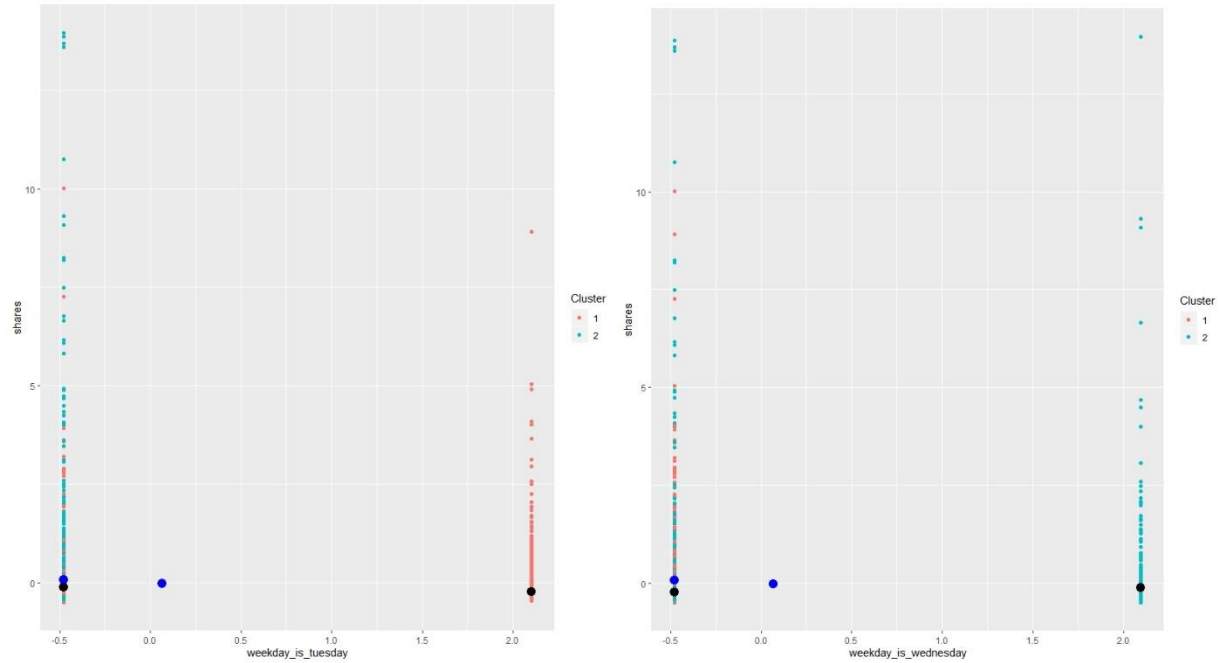

Figure 8 *Plots Showing K-Mean's Centroid (in Blue) and Medoids of PAM (in Black)*

So, from the K-Mean clustering method, **Figure 7** shows that the clustering 1 is news that are published on weekdays while clustering 2 is news that are published on the weekend and most likely with no videos and few keywords. Other than that distinction, the data are almost arbitrary when it comes to the number of tokenized word on the title and content, number of images, videos, and shares.

```
> newsClus <- data.frame(trainData, clust=kclus$cluster, trainData)
> head(newsClus[newsClus$clust==1,c("n_tokens_title","n_tokens_content","num_imgs","num_videos","num_keywords","is_weekend","shares")])
                                                 n_tokens_title n_tokens_content num_imgs num_videos num_keywords is_weekend shares
http://mashable.com/2013/11/13/zoomer-robotic-dog/                       -1.629          -0.741    0.909     -0.313        1.476     -0.378  1.784
http://mashable.com/2013/11/26/britney-jean-spears-album-stream-free-itunes/ -0.694          0.334   -0.436     -0.313       -0.114     -0.378 -0.341
http://mashable.com/2014/02/27/moov-fitness-tracker/                      2.111          -0.612   -0.436      0.479       -1.704     -0.378 -0.341
http://mashable.com/2014/07/15/summer-guide-seattle/                     -0.694          -0.156   -0.436     -0.049        1.476     -0.378 -0.252
http://mashable.com/2013/10/07/little-wrecking-ball/                     -1.629          -0.210   -0.436     -0.313        0.416     -0.378 -0.252
http://mashable.com/2013/05/14/old-spice-ads/                             0.709          -0.625   -0.558      0.215       -1.174     -0.378 -0.282
> head(newsClus[newsClus$clust==2,c("n_tokens_title","n_tokens_content","num_imgs","num_videos","num_keywords","is_weekend","shares")])
                                                n_tokens_title n_tokens_content num_imgs num_videos num_keywords is_weekend shares
http://mashable.com/2013/05/11/social-web-drivers-ed/            0.709          -0.477   -0.558     -0.313       -0.644      2.647 -0.312
http://mashable.com/2013/03/10/sxswi-day-3/                      0.709          -0.642   -0.436     -0.313       -0.114      2.647  0.030
http://mashable.com/2014/10/30/tim-cooks-essay-why-now/         -0.226          -0.326   -0.436     -0.313       -0.114      2.647 -0.104
http://mashable.com/2013/03/17/viral-video-recap-17-2/          -0.226          -0.780   -0.558     -0.313       -0.114      2.647 -0.327
http://mashable.com/2014/05/10/mother-day-gift-finder/           0.709          -0.664   -0.313     -0.313       -1.174      2.647 -0.348
http://mashable.com/2013/02/24/learn-to-use-photoshop-free/     -0.226          -0.341   -0.436     -0.313       -0.114      2.647  0.134
>
```

Figure 7 *Clustering Coefficients for Some of the Attributes from the Data Set*