King Saud University

College of Computer and Information Sciences

Information Technology department

Data Mining

Course Project

Movie

Project Report

Group members:

Name
Nora almisfer
Latifa albekery
Layan aldhuwayhi
Reem Al- taleb

4/10/2020

1 Problem

We want to know the relationship between the amount of money we put in a movie and the number of ratings on this movie.

Movies are one of the most important means of entertainment and learning in the the same time, and Movies are divided into several sections, there are Documentary Movies for People interested in science, there are action movies for people who love adventuress, and animation movies for children , and so on.

We expect that if we put in additional money, we will be able to bring in professional actors and make professional scenes and in doing so we will produce professional movies resulting in high ratings.

2 Data Mining Task

The data mining is clustring, the input of attribute is numeric, which the attribute are Ratings and Gross. There are two Group

3 Data

We took this data from the machine learning repository site: http://archive.ics.uci.edu/ml/datasets/CSM+%28Conventional+and+Social+Media+Movies%29+Dataset+2014+and +2015

This data sit has 14 attribute and 187 instances (row)

Name of attribute	Data type	Description
Movie	nominal	the name of movies
Ratings	Integer	the ratings for each movie
Year	Integer	the years that the movie was seen
Genre	Integer	the number of genre for movie
Gross	Integer	The number of gross money from the movie
Budget	Integer	the cost of the budget for each movie
Screens	Integer	the number of screens for each movie

Sequel	Integer	number of sequel
Sentiment	Integer	The sentiment of people about the movie
Views	Integer	the number of views for each movie
Likes	Integer	the number of likes for each movie
Dislikes	Integer	the number of dislikes for each movie
Comments	Integer	the number of comment for each movie
Aggregate.Followers	Integer	the number of followers for each movie

Information about Dataset :

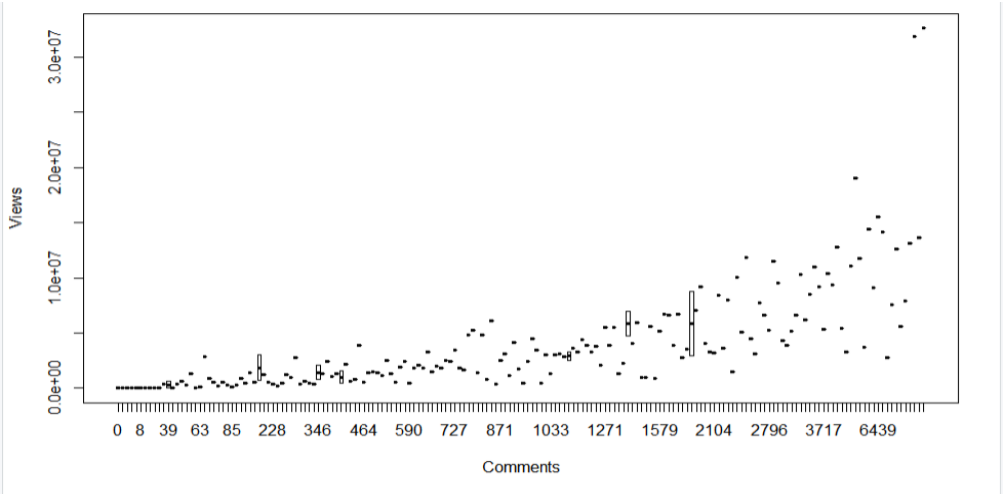
Source: Mehreen Ahmed , Department of Computer Software Engineering , National University of Sciences and Technology (NUST), Islamabad, Pakistan , mahreenmcs '@' gmail.com

Number of Instances: 187, Number of Attributes: 14

Data Set Characteristics: Multivariate, Data type: integer, Missing Values: there is a missing value.

Boxplot in R :

are a measure of how well distributed is the data in a data set. It divides the data set into three quartiles . This graph represents the minimum, maximum, median, first quartile and

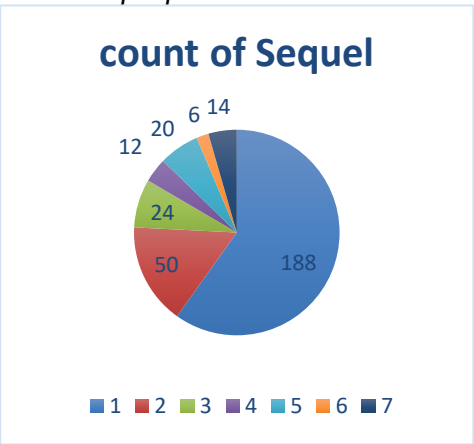


third quartile in the data set

The Boxplot explain the relation between tow attributes Views and Comment and the relation is that comment decrease when the views is decrease.

Pie chart in R:

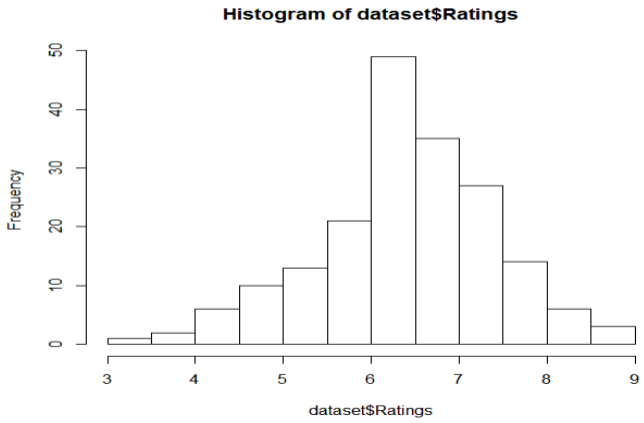
A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion



The chart show how many movies that have the same sequel.

Histogram in R:

A histogram is an accurate representation of the distribution of numerical data.



The Histogram show the rating for the movies between 2014-20115

4 Data preprocessing

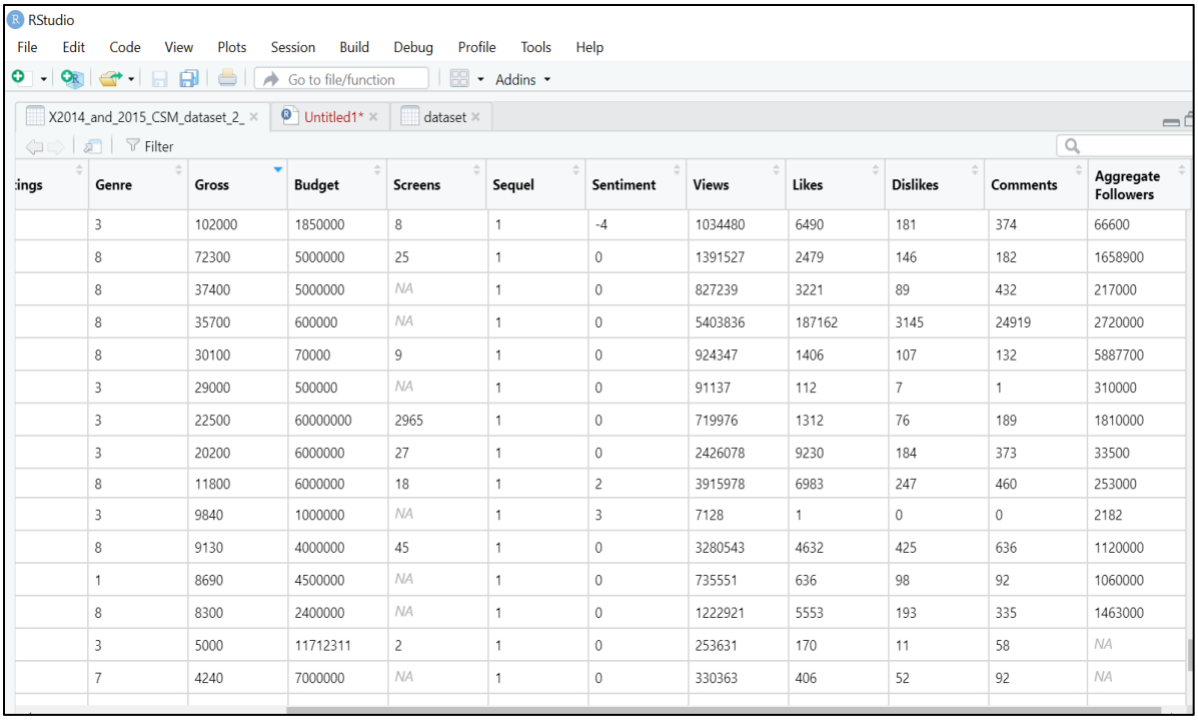
We did choose to do data cleaning to Handle the Missing Values by Filling in the missing value, and we did choose do to Handle Noisy Data that have some random error or variance in a measured variable, and we did some Data Transformation Strategies: Normalization so attribute data are scaled to fall within a smaller, specified range.

Cleaning:

Using the central tendency for the attributes [Screens, Aggregate.Followers, Budget, Dislikes, Comments, Views, Year, Ratings, Genre, Gross, Sequel, Sentiment, Likes] because they all have missing value.

Before:

```
> sum(is.na(dataset))
[1] 59
>
```



ings	Genre	Gross	Budget	Screens	Sequel	Sentiment	Views	Likes	Dislikes	Comments	Aggregate Followers
	3	102000	1850000	8	1	-4	1034480	6490	181	374	66600
	8	72300	5000000	25	1	0	1391527	2479	146	182	1658900
	8	37400	5000000	NA	1	0	827239	3221	89	432	217000
	8	35700	600000	NA	1	0	5403836	187162	3145	24919	2720000
	8	30100	70000	9	1	0	924347	1406	107	132	5887700
	3	29000	500000	NA	1	0	91137	112	7	1	310000
	3	22500	60000000	2965	1	0	719976	1312	76	189	1810000
	3	20200	6000000	27	1	0	2426078	9230	184	373	33500
	8	11800	6000000	18	1	2	3915978	6983	247	460	253000
	3	9840	1000000	NA	1	3	7128	1	0	0	2182
	8	9130	4000000	45	1	0	3280543	4632	425	636	1120000
	1	8690	4500000	NA	1	0	735551	636	98	92	1060000
	8	8300	2400000	NA	1	0	1222921	5553	193	335	1463000
	3	5000	11712311	2	1	0	253631	170	11	58	NA
	7	4240	7000000	NA	1	0	330363	406	52	92	NA

After:

```
> sum(is.na(dataset))
[1] 0
>
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

X2014_and_2015_CSM_dataset_2_ | Untitled1* | dataset

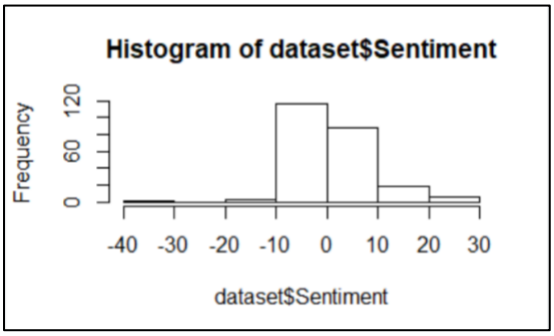
Filter

Year	Ratings	Genre	Gross	Budget	Screens	Sequel	Sentiment	Views	Likes	Dislikes	Comments	Aggregate.Followers
2014	7.5	8	37400	5000000	2209.244	1	0	827239	3221	89	432	217000
2014	5.2	8	35700	600000	2209.244	1	0	5403836	187162	3145	24919	2720000
2014	5.6	8	30100	70000	9.000	1	0	924347	1406	107	132	5887700
2014	4.6	3	29000	500000	2209.244	1	0	91137	112	7	1	310000
2014	6.2	3	22500	60000000	2965.000	1	0	719976	1312	76	189	1810000
2014	5.8	3	20200	6000000	27.000	1	0	2426078	9230	184	373	33500
2014	6.6	8	11800	6000000	18.000	1	2	3915978	6983	247	460	253000
2014	7.0	3	9840	1000000	2209.244	1	3	7128	1	0	0	2182
2014	6.3	8	9130	4000000	45.000	1	0	3280543	4632	425	636	1120000
2014	5.7	1	8690	4500000	2209.244	1	0	735551	636	98	92	1060000
2014	5.7	8	8300	2400000	2209.244	1	0	1222921	5553	193	335	1463000
2014	6.8	3	5000	11712311	2.000	1	0	253631	170	11	58	3038193
2014	4.8	7	4240	7000000	2209.244	1	0	330363	406	52	92	3038193

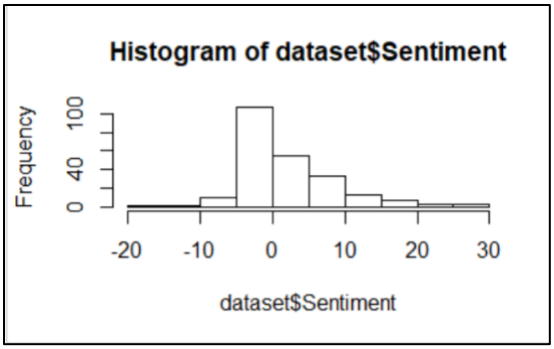
Handle Noisy Data:

by Outlier Analysis on Sentiment attribute.

Before:



After:



Normalization:

by the Min-max normalization on the Screens attribute.

Before:

After:

5 Data Mining Technique

- **Kmeans :**
  Kmeans a classic partitioning method for clustering

  that exist in package fpc .

  Is type of unsupervised learning, which is used when you have unlabeled data.

- **hclust** : in clustering that exist in package factoextra,
  Hierarchical cluster analysis on a set of dissimilarities and methods for analyzing it

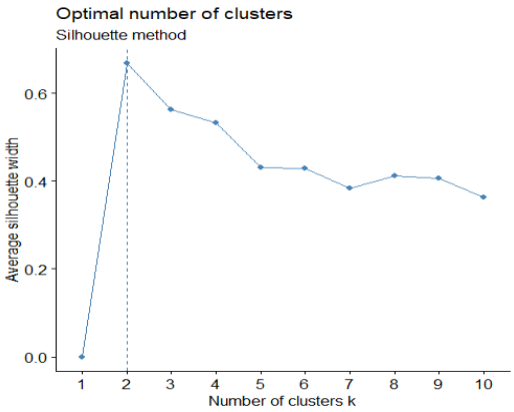
- **fviz_nbclust :** exist in package factoextra Dertemines and visualize the optimal number of clusters using different methods

- **silhouette :** exist in package cluster . method is used to measure the clustering quality and determine the optimal number of clusters

- **NbClust :** exist in package NbClust . directly return the optimal number of clustering based on the frequency distribution histogram

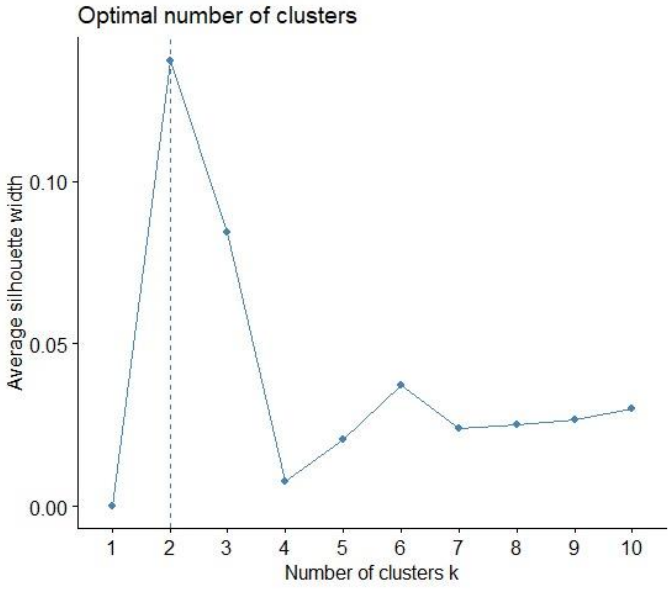
6 Evaluation and Comparison

- **If your mining task is Clustering, compare between at <u>least two algorithms</u> (such as: K-means and k-medoids) considering the following criteria:**
    - <u>optimal number of clusters using the</u> NbClust() method in K-means



plots shows that K=2 is the optimal number of clusters in K-means algorithm with 0.65 average silhouette width.

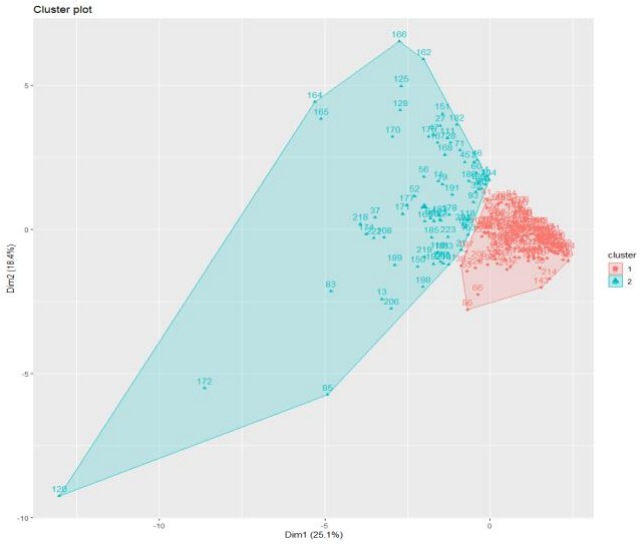
- <u>optimal number of clusters using the</u> NbClust() method in K- medoids



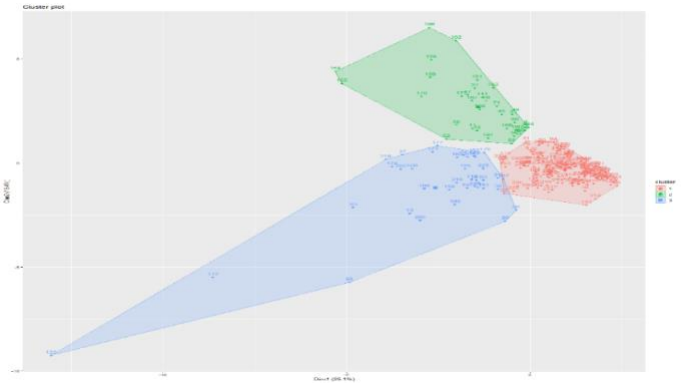
plots shows that K=2 is the optimal number of clusters in K- medoids algorithm with 0.15 average silhouette width.

Algorithm	K-means			k-medoids		
Number of cluster (k)	K=2 **(optimal)**	K=3	K=4	K=2	K=3	K=4
Average silhouette width	0.65	0.6	0.58	0.15	0.9	0.02

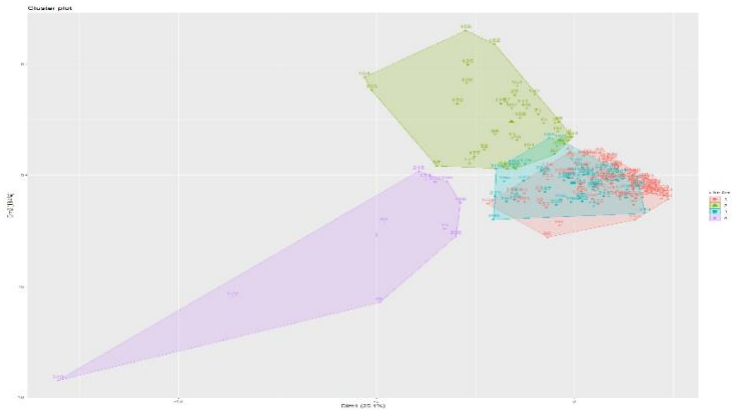
K-means Visualization:



K=2



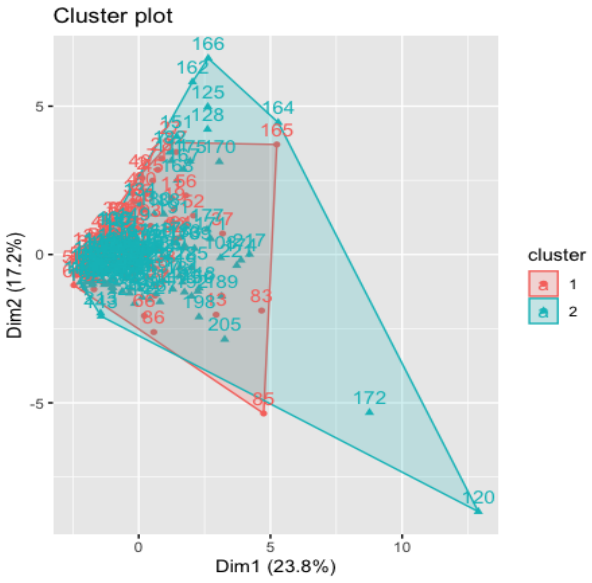
K=3

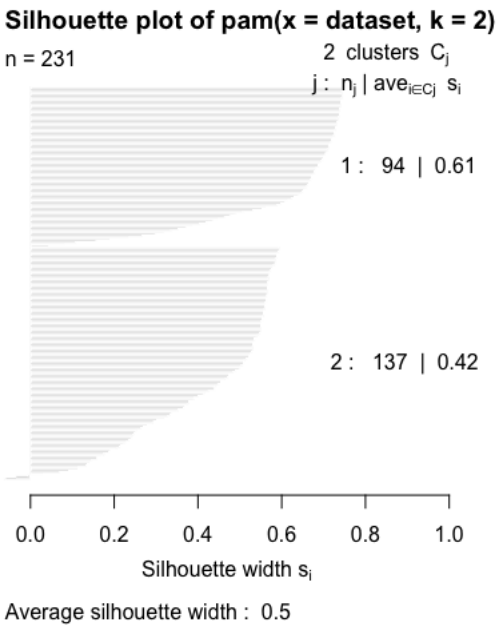


K=4

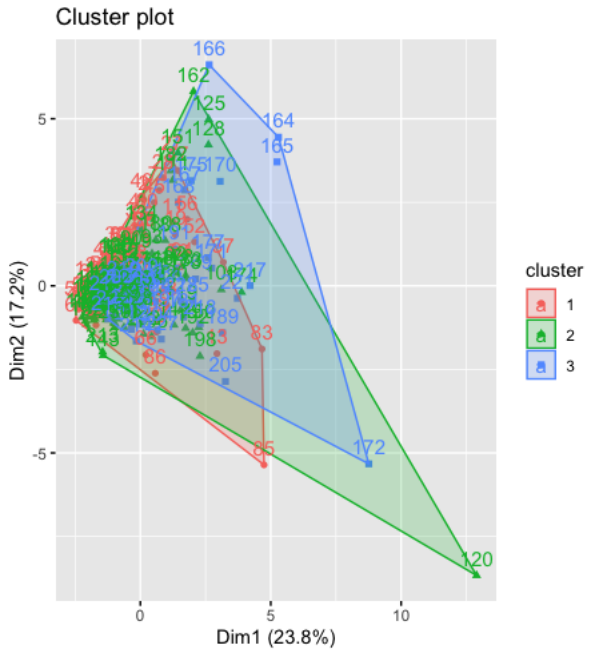
K- medoids Visualization:

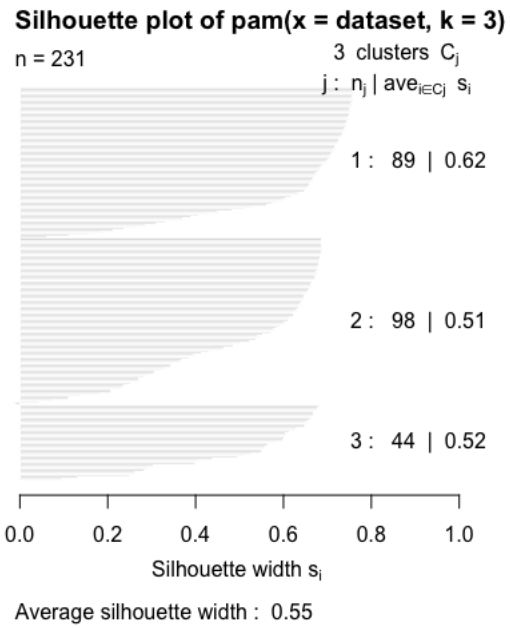
K=2:



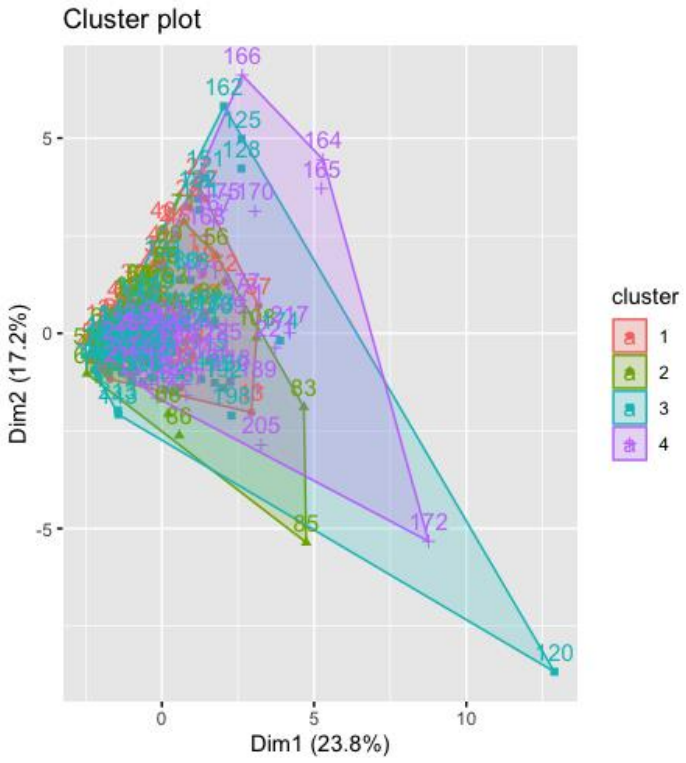

Silhouette plot of pam(x = dataset, k = 2)
n = 231
2 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 94 | 0.61
2 : 137 | 0.42
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.5

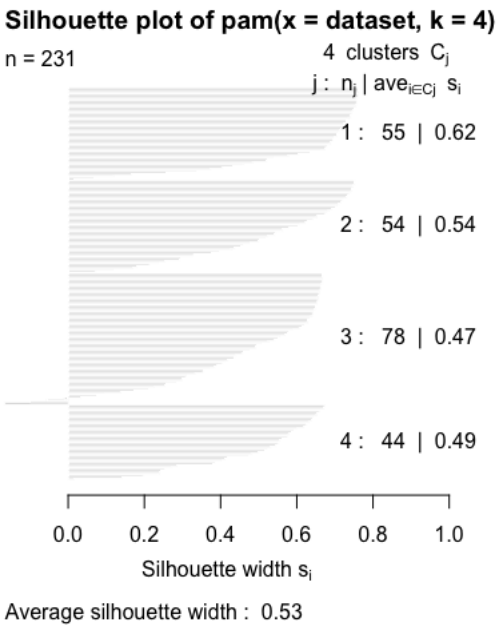
K=3:


Cluster plot
Dim2 (17.2%)
Dim1 (23.8%)
cluster
1
2
3


Silhouette plot of pam(x = dataset, k = 3)
n = 231
3 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 89 | 0.62
2 : 98 | 0.51
3 : 44 | 0.52
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.55

K=4:


Cluster plot
Dim2 (17.2%)
Dim1 (23.8%)
cluster
1
2
3
4



7 Findings

In this section, you should discuss your findings by investigating the following:

- When the best obtained mining results was achieved?
- Decide whether these results are interested or not (Relate the results with the goal of your project).

8 Code

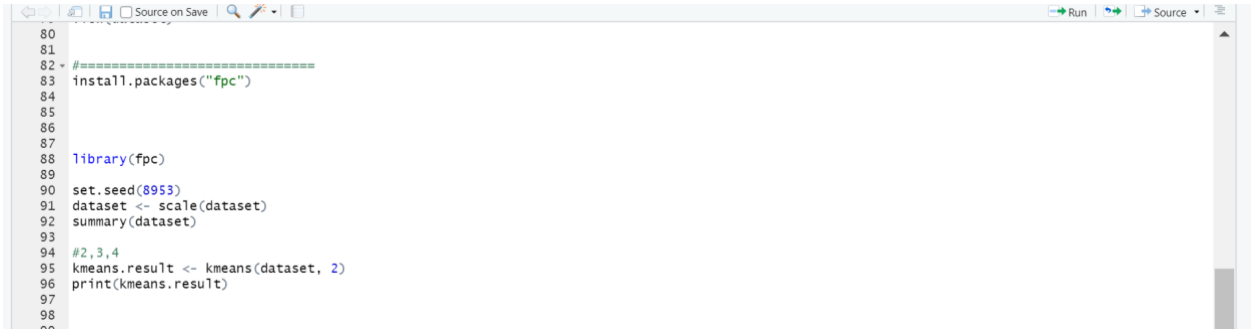
The section contains the R code you write to accomplish the data mining task. Provide all the comments required to understand your code. Divide codes into 3 parts: preprocessing part, data mining task, evaluation part.

1. preprocessing part .

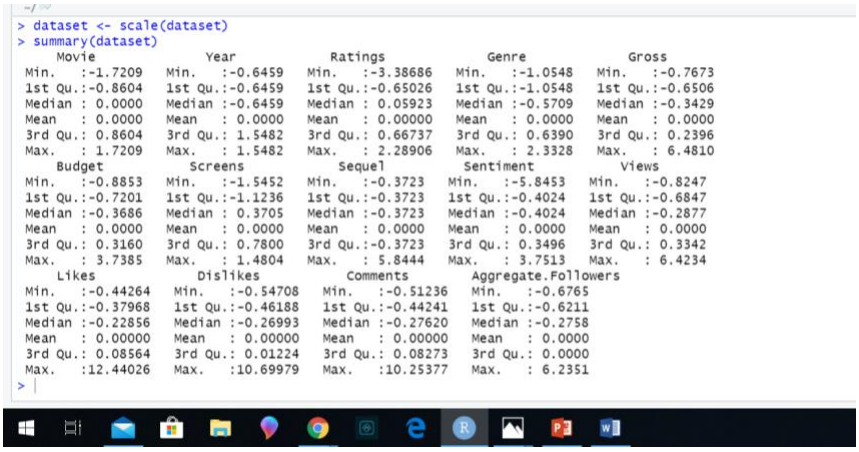
```r
install.packages("outliers")

library(outliers)


dataset = read.csv('/Users/latifaalbkery/Documents/leve6/data mining/project/2014 and 2015 CSM dataset (2).csv')

dim(dataset)
names(dataset)
is.na(dataset)
sum(is.na(dataset))


dataset$Moive <- sapply(dataset$Moive , as.numeric)

dataset$Screens = ifelse(is.na(dataset$Screens),  ave(dataset$Screens,
                         FUN =function(x) mean(x,na.rm=TRUE)), dataset$Screens)

dataset$Aggregate.Followers = ifelse(is.na(dataset$Aggregate.Followers),
                           ave(dataset$Aggregate.Followers, FUN =function(x) mean(x,na.rm=TRUE)), dataset$Aggregate.Followers)

dataset$Budget = ifelse(is.na(dataset$Budget),ave(dataset$Budget,
                              FUN =function(x) mean(x,na.rm=TRUE)), dataset$Budget)


dataset$Dislikes  = ifelse(is.na(dataset$Dislikes ),ave(dataset$Dislikes
                                    , FUN =function(x) mean(x,na.rm=TRUE)), dataset$Dislikes )


dataset$Comments   = ifelse(is.na(dataset$Comments  ),ave(dataset$Comments
                                                          , FUN =function(x) mean(x,na.rm=TRUE)), dataset$Comments  )


dataset$Views   = ifelse(is.na(dataset$Views  ),ave(dataset$Views
                                                    , FUN =function(x) mean(x,na.rm=TRUE)), dataset$Views  )


dataset$Year   = ifelse(is.na(dataset$Year  ),ave(dataset$Year
                                                  , FUN =function(x) mean(x,na.rm=TRUE)), dataset$Year  )


dataset$Ratings   = ifelse(is.na(dataset$Ratings  ),ave(dataset$Ratings
                                                        , FUN =function(x) mean(x,na.rm=TRUE)), dataset$Ratings  )
```

```r
dataset$Genre   = ifelse(is.na(dataset$Genre  ),ave(dataset$Genre
                                                    , FUN =function(x) mean(x,na.rm=TRUE)), dataset$Genre  )


dataset$Gross   = ifelse(is.na(dataset$Gross  ),ave(dataset$Gross
                                                    , FUN =function(x) mean(x,na.rm=TRUE)), dataset$Gross  )


dataset$Sequel   = ifelse(is.na(dataset$Sequel  ),ave(dataset$Sequel
                                                      , FUN =function(x) mean(x,na.rm=TRUE)), dataset$Sequel  )


dataset$Sentiment   = ifelse(is.na(dataset$Sentiment  ),ave(dataset$Sentiment
                                                            , FUN =function(x) mean(x,na.rm=TRUE)), dataset$Sentiment  )


dataset$Likes   = ifelse(is.na(dataset$Likes  ),ave(dataset$Likes
                                                    , FUN =function(x) mean(x,na.rm=TRUE)), dataset$Likes  )


sum(is.na(dataset))

View(dataset)

#==============================

hist(dataset$Sentiment)
OutAge = outlier(dataset$Sentiment, logical =TRUE)
sum(OutAge)
Find_outlier = which(OutAge ==TRUE, arr.ind = TRUE)
dataset= dataset[-Find_outlier,]

hist(dataset$Sentiment)

#==============================


normalize <- function(x) {return ((x - min(x)) / (max(x) - min(x)))}

hist(dataset$Screens)

dataset$Screens<-normalize(dataset$Screens)

hist(dataset$Screens)
```

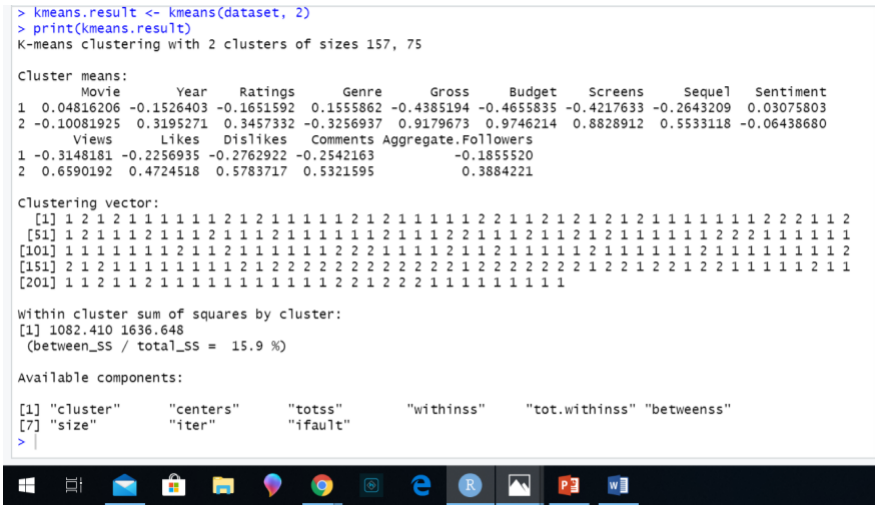
2. data mining task.



```
80
81
82 #============================
83 install.packages("fpc")
84
85
86
87
88 library(fpc)
89
90 set.seed(8953)
91 dataset <- scale(dataset)
92 summary(dataset)
93
94 #2,3,4
95 kmeans.result <- kmeans(dataset, 2)
96 print(kmeans.result)
97
98
```

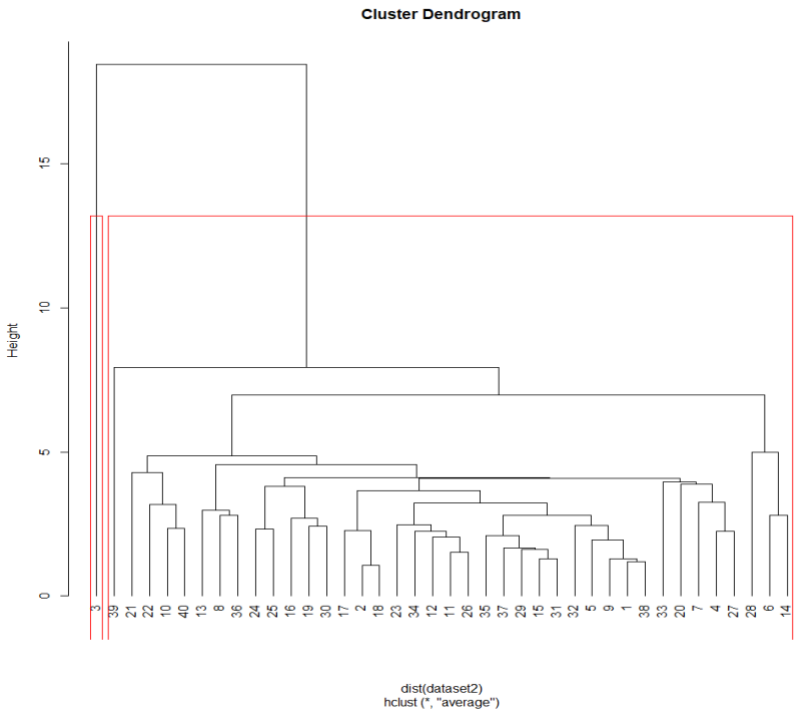


```
> dataset <- scale(dataset)
> summary(dataset)
     Movie               Year             Ratings             Genre              Gross        
 Min.   :-1.7209   Min.   :-0.6459   Min.   :-3.38686   Min.   :-1.0548   Min.   :-0.7673  
 1st Qu.:-0.8604   1st Qu.:-0.6459   1st Qu.:-0.65026   1st Qu.:-1.0548   1st Qu.:-0.6506  
 Median : 0.0000   Median :-0.6459   Median : 0.05923   Median :-0.5709   Median :-0.3429  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000  
 3rd Qu.: 0.8604   3rd Qu.: 1.5482   3rd Qu.: 0.66737   3rd Qu.: 0.6390   3rd Qu.: 0.2396  
 Max.   : 1.7209   Max.   : 1.5482   Max.   : 2.28906   Max.   : 2.3328   Max.   : 6.4810  
     Budget             Screens            Sequel           Sentiment           Views        
 Min.   :-0.8853   Min.   :-1.5452   Min.   :-0.3723   Min.   :-5.8453   Min.   :-0.8247  
 1st Qu.:-0.7201   1st Qu.:-1.1236   1st Qu.:-0.3723   1st Qu.:-0.4024   1st Qu.:-0.6847  
 Median :-0.3686   Median : 0.3705   Median :-0.3723   Median :-0.4024   Median :-0.2877  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
 3rd Qu.: 0.3160   3rd Qu.: 0.7800   3rd Qu.:-0.3723   3rd Qu.: 0.3496   3rd Qu.: 0.3342  
 Max.   : 3.7385   Max.   : 1.4804   Max.   : 5.8444   Max.   : 3.7513   Max.   : 6.4234  
     Likes             Dislikes          Comments        Aggregate.Followers
 Min.   :-0.44264   Min.   :-0.54708   Min.   :-0.51236   Min.   :-0.6765   
 1st Qu.:-0.37968   1st Qu.:-0.46188   1st Qu.:-0.44241   1st Qu.:-0.6211   
 Median :-0.22856   Median :-0.26993   Median :-0.27620   Median :-0.2758   
 Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000   
 3rd Qu.: 0.08564   3rd Qu.: 0.01224   3rd Qu.: 0.08273   3rd Qu.: 0.0000   
 Max.   :12.44026   Max.   :10.69979   Max.   :10.25377   Max.   : 6.2351   
> 
```

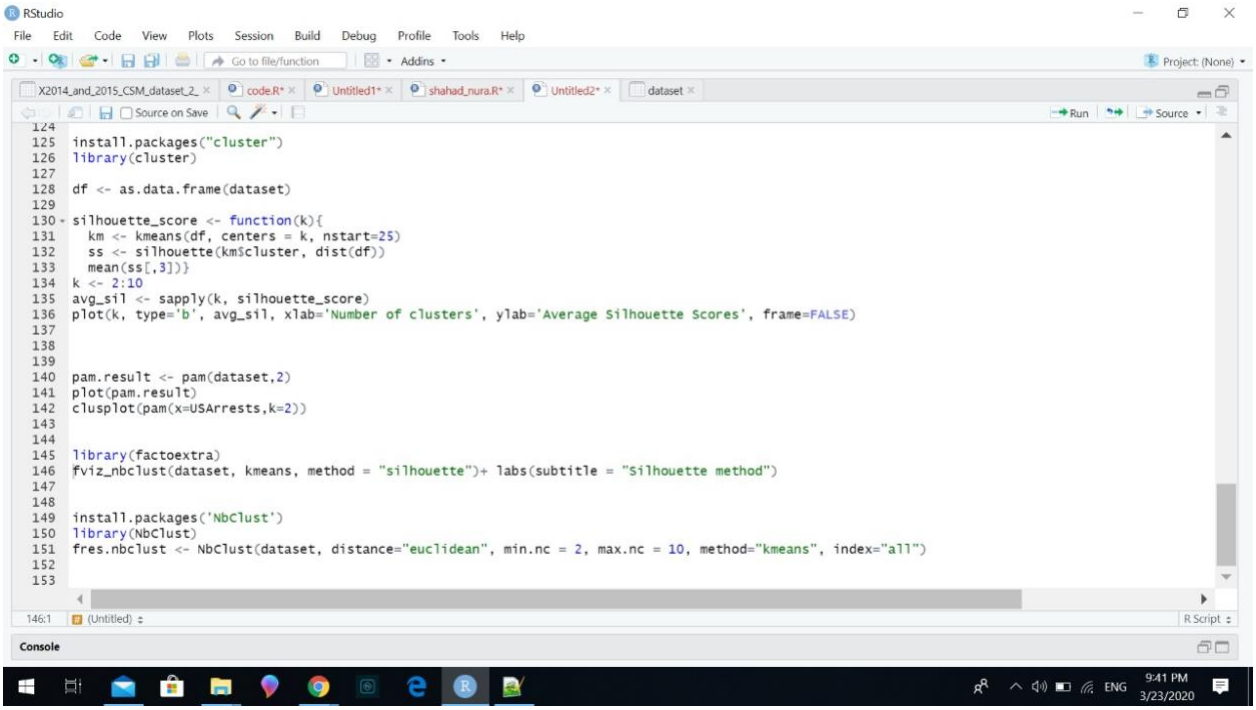


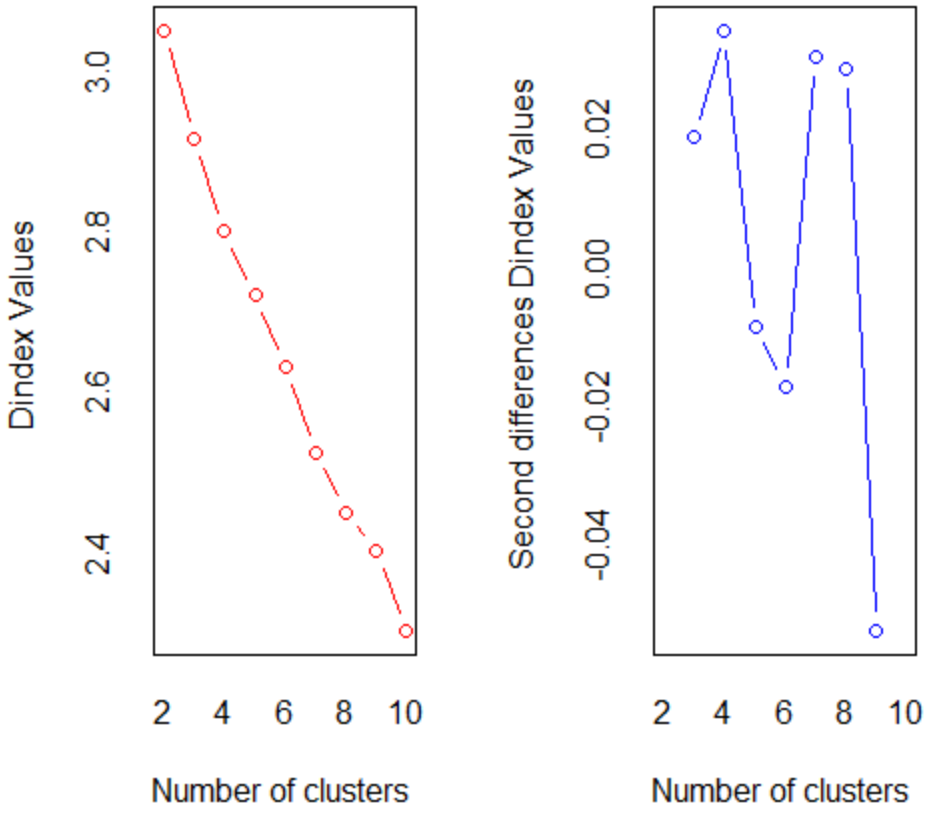
```
> kmeans.result <- kmeans(dataset, 2)
> print(kmeans.result)
K-means clustering with 2 clusters of sizes 157, 75

Cluster means:
        Movie       Year    Ratings      Genre      Gross     Budget    Screens     Sequel   Sentiment
1  0.04816206 -0.1526403 -0.1651592  0.1555862 -0.4385194 -0.4655835 -0.4217633 -0.2643209  0.03075803
2 -0.10081925  0.3195271  0.3457332 -0.3256937  0.9179673  0.9746214  0.8828912  0.5533118 -0.06438680
       Views      Likes   Dislikes   Comments Aggregate.Followers
1 -0.3148181 -0.2256935 -0.2762922 -0.2542163          -0.1855520
2  0.6590192  0.4724518  0.5783717  0.5321595           0.3884221

Clustering vector:
  [1] 1 2 1 2 1 1 1 1 1 1 2 1 2 1 1 1 1 1 2 1 2 1 1 1 1 1 1 2 1 2 1 2 1 2 1 2 1 1 1 1 1 1 1 2 2 2 1 1 2
 [51] 1 2 1 1 1 2 1 1 1 2 1 1 1 2 1 1 1 1 1 1 2 1 1 1 2 2 1 1 1 2 1 1 2 1 2 1 2 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1
[101] 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1 2 2 2 1 1 1 1 2 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2
[151] 2 1 2 1 1 1 1 1 1 1 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 1 2 2 1 2 2 1 2 2 1 1 1 1 1 2 1 1
[201] 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 2 1 2 2 2 2 1 1 1 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 1082.410 1636.648
 (between_SS / total_SS =  15.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
> 
```

```
install.packages("factoextra")
library(factoextra)



fviz_cluster(kmeans.result, data = dataset)

idx <- sample(1:dim(dataset)[1], 40)
dataset2 <- dataset[idx, ]
hc <- hclust(dist(dataset2), method = "ave")
plot(hc, hang = -1)
rect.hclust(hc, k = 2)
groups <- cutree(hc, k = 2)


kmeans.result2<-kmeans(dataset,2)
fviz_cluster(kmeans.result2, data = dataset)
```



3. evaluation part.


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
X2014_and_2015_CSM_dataset_2_
code.R*
Untitled1*
shahad_nura.R*
Untitled2*
dataset
Source on Save
Run
Source
install.packages("cluster")
library(cluster)
df <- as.data.frame(dataset)
silhouette_score <- function(k){
km <- kmeans(df, centers = k, nstart=25)
ss <- silhouette(km$cluster, dist(df))
mean(ss[,3])}
k <- 2:10
avg_sil <- sapply(k, silhouette_score)
plot(k, type='b', avg_sil, xlab='Number of clusters', ylab='Average Silhouette Scores', frame=FALSE)
pam.result <- pam(dataset,2)
plot(pam.result)
clusplot(pam(x=USArrests,k=2))
library(factoextra)
fviz_nbclust(dataset, kmeans, method = "silhouette")+ labs(subtitle = "Silhouette method")
install.packages('NbClust')
library(NbClust)
fres.nbclust <- NbClust(dataset, distance="euclidean", min.nc = 2, max.nc = 10, method="kmeans", index="all")
146:1
(Untitled)
R Script
Console
ENG
9:41 PM
3/23/2020




```
> fres.nbclust <- NbClust(dataset, distance="euclidean", min.nc = 2, max.nc = 10, method="kmeans", index="al
l")
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 10 proposed 2 as the best number of clusters
* 3 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 5 proposed 5 as the best number of clusters
* 5 proposed 10 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2


*******************************************************************
>
```

9 References

Ahmed M, Jahangir M, Afzal H, Majeed A, Siddiqi I. (2016). UCI Machine Learning Repository: Conventional and Social Media Movies) Dataset. [online] Archive.ics.uci.edu.Available at:http://archive.ics.uci.edu/ml/datasets/CSM+%28Conventional+and+Social+Media+Movies%29+Dataset+2014+and+2015# [Accessed 1 Jun.2016].