Group Project

K-Means Clustering and PCA

Department: UBIT

Course: Data Warehousing & Data Mining (CS-626)

Submitted To : Sir Dr Tehseen Ahmed Jilani

Batch: MCS Final 2021 (4th Semester)

GROUP MEMBERS:

1. Muhammad Rizwan (P19101039)
2. Khuda Bakhsh (P19101030)
3. Muhammad Humayun (P1810028)
4. Asadullah Siddiqui (P1510010)

1 TABLE OF CONTENTS

2 INTRODUCTION :

Clustering :

- Set of methodologies for automatic classification of samples into a number of groups using a measure of association, so that the samples in one group are similar and samples belonging to different groups are not similar.

- Samples for clustering are represented as a vector of measurements, or more formally, as a point in a multidimensional space.

- Clustering is a very difficult problem because data can reveal clusters with different shapes and sizes in an n-dimensional data space.

K-Means Clustering :

- K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.

- K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

The k-means clustering algorithm mainly performs two tasks :

- Determines the best value for K center points or centroids by an iterative process .

- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

3 LITERATURE REVIEW :

In data mining, clustering is a technique in which the set of objects are assigned to a group called clusters. Clustering is the most essential part of data mining. K-means clustering is the basic clustering technique and is most widely used algorithm. It is also known as nearest neighbor searching. It simply clusters the datasets into given number of clusters. Numerous efforts have been made to improve the performance of the K-means clustering algorithm. In this paper we have been briefed in the form of a review the work carried out by the different researchers using K-means clustering. We have discussed the limitations and applications of the K-means clustering algorithm as well. This paper presents a current review about the K means clustering algorithm.

4 DATA DETAIL :

Dataset consist of 150 samples from each of 3 species (Setosa , Virginica , Versicolor)

Four features were measured from each sample
i.e length and width of the sepals and petals and based on the combination of these four features

Performing K-Means Clustering on Dataset :

Using K-Means Clustering algorithm on the dataset which includes 150 persons and 4 variables or attributes .

5 DATAMINING / ANALYTICS / VISUALIZATION :

K-means Working :

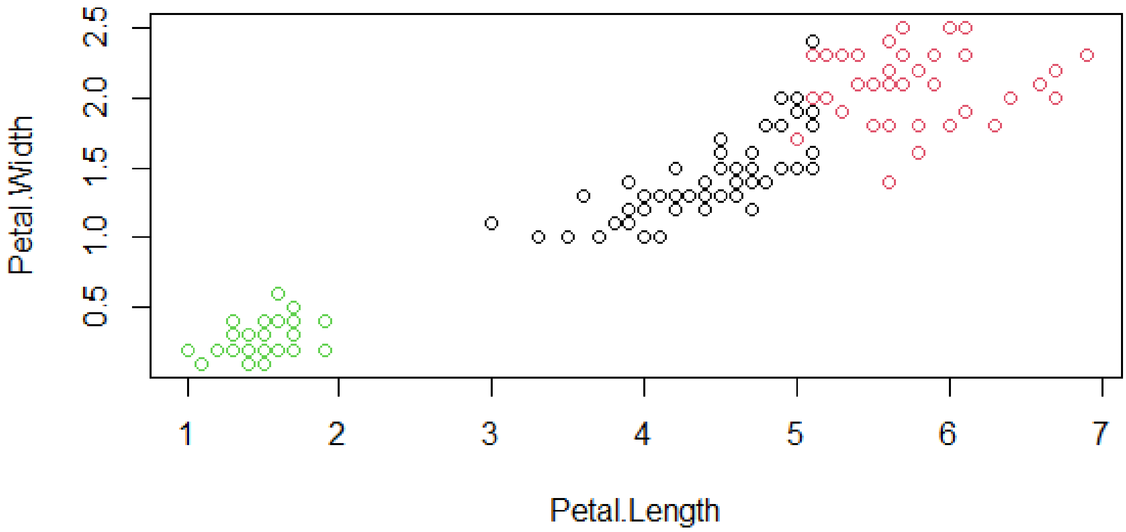
Console Terminal Jobs

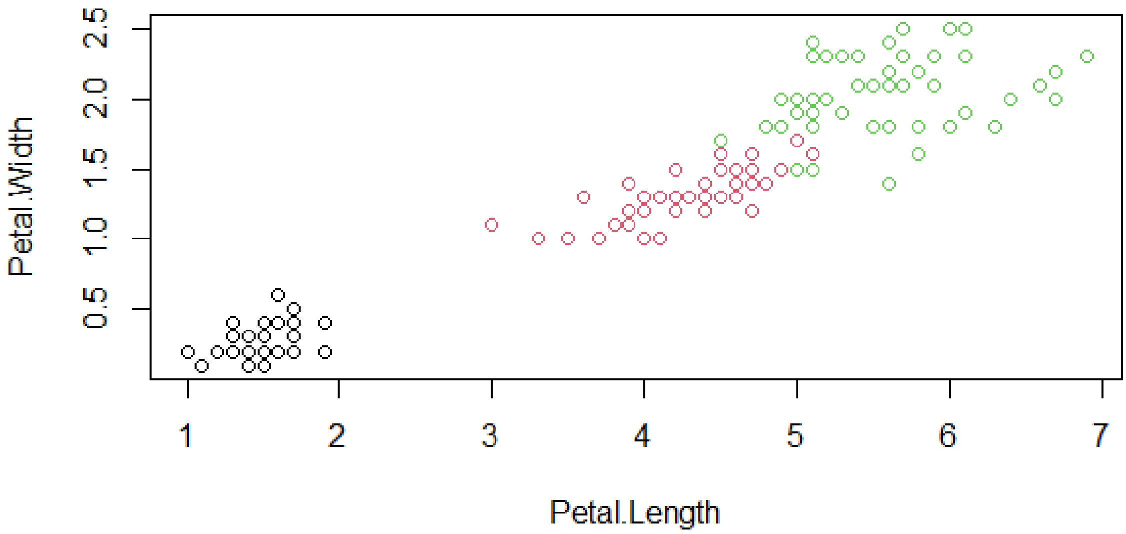
R 4.1.2 · ~/

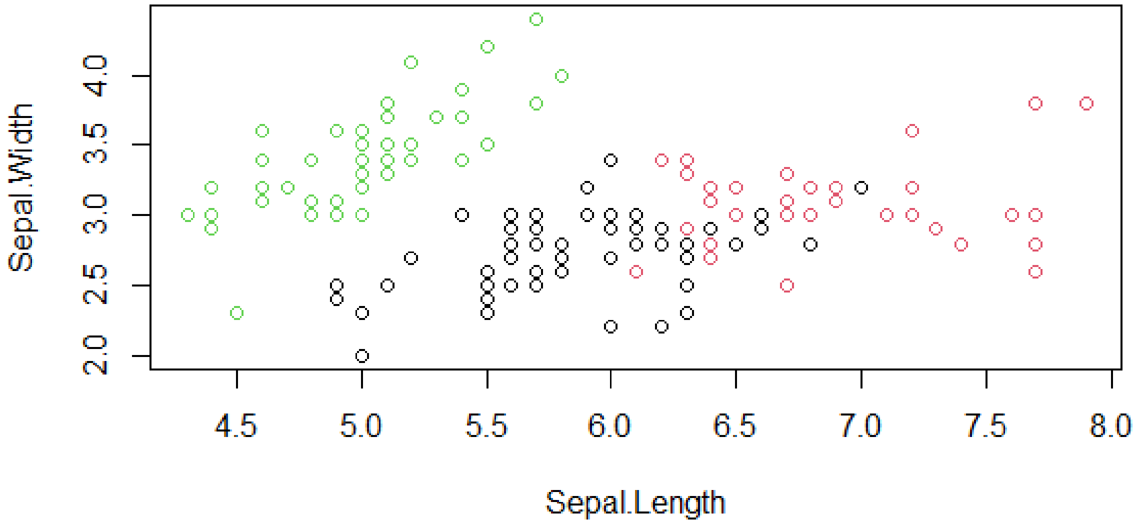
```
> View(iris)
> iris.features = iris
> iris.features$Species <- NULL
> View(iris.features)
>
> results <- kmeans(iris.features, 3)
>
> results
K-means clustering with 3 clusters of sizes 62, 38, 50

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     5.901613    2.748387     4.393548    1.433871
2     6.850000    3.073684     5.742105    2.071053
3     5.006000    3.428000     1.462000    0.246000

Clustering vector:
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [40] 3 3 3 3 3 3 3 3 3 3 3 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2
 [79] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2 2 2 1 1 2 2
[118] 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 1 2 2 2 2 1 2 2 2 1 2 2 2 1 2 2 1

Within cluster sum of squares by cluster:
[1] 39.82097 23.87947 15.15100
 (between_SS / total_SS =  88.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
>
> results$size
[1] 62 38 50
>
```

```
> results$cluster
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [40] 3 3 3 3 3 3 3 3 3 3 3 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2
 [79] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2 2 2 1 1 2 2
[118] 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 1 2 2 2 2 1 2 2 2 1 2 2 2 1 2 2 1
>
> table(iris$Species , results$cluster)

              1  2  3
  setosa      0  0 50
  versicolor 48  2  0
  virginica  14 36  0
```

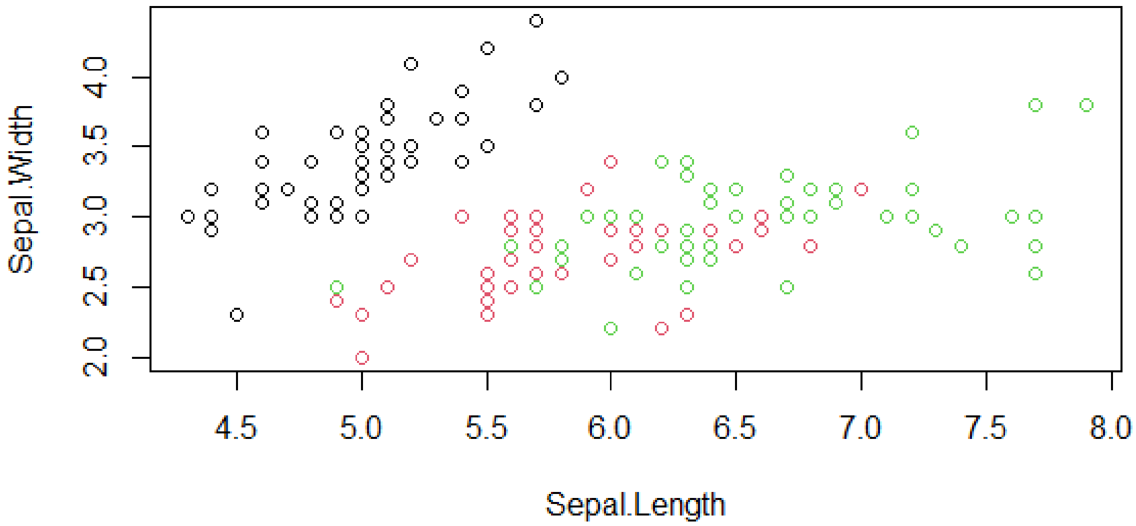
```
> plot(iris[c("Petal.Length" , "Petal.Width")], col = results$cluster)
>
> plot(iris[c("Petal.Length" , "Petal.Width")], col = iris$Species)
>
> plot(iris[c("Sepal.Length" , "Sepal.Width")], col = results$cluster)
>
> plot(iris[c("Sepal.Length" , "Sepal.Width")], col = iris$Species)
>
```


Petal.Width
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
6
7
Petal.Length


Petal.Width
0.5
1.0
1.5
2.0
2.5
1
2
3
4
5
6
7
Petal.Length



Sepal.Width
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
Sepal.Length



Sepal.Width
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
Sepal.Length

PCA Working :

R 4.1.2 · ~/

```
> data(iris)
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
>
>
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
>
> mypr <- prcomp(iris[, 5])
Error in colMeans(x, na.rm = TRUE) : 'x' must be numeric
> mypr <- prcomp(iris[2 , 5])
Error in colMeans(x, na.rm = TRUE) : 'x' must be numeric
>
>
> prcomp(~Sepal.Length + Petal.Width, data = iris)
Standard deviations (1, .., p=2):
[1] 1.0734371 0.3382787

Rotation (n x k) = (2 x 2):
                   PC1        PC2
Sepal.Length 0.7419133 -0.6704958
Petal.Width  0.6704958  0.7419133
>
```

Rectangular Snip

R 4.1.2 · ~/

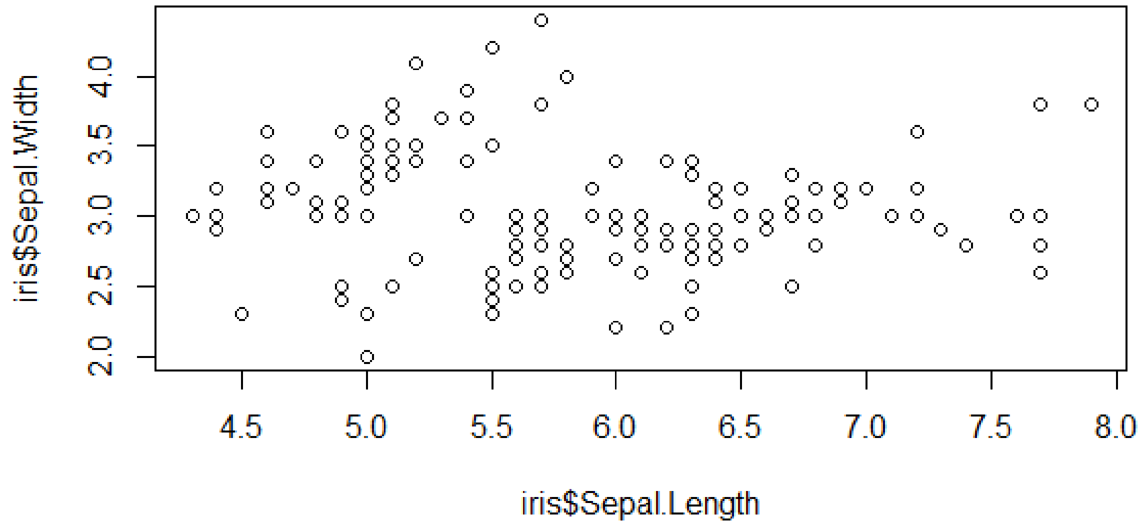
```
> plot(iris$Sepal.Length , iris$Sepal.Width)
>
> plot(scale(iris$Sepal.Length) , scale(iris$Sepal.Width))
>
> summary(mypr)
Importance of components:
                          PC1     PC2    PC3     PC4
Standard deviation     2.0563 0.49262 0.2797 0.15439
Proportion of Variance 0.9246 0.05307 0.0171 0.00521
Cumulative Proportion  0.9246 0.97769 0.9948 1.00000
>
> plot(mypr , type = "l")
>
> biplot(mypr , scale = 0)
>
>
> str(mypr)
List of 5
 $ sdev    : num [1:4] 2.056 0.493 0.28 0.154
 $ rotation: num [1:4, 1:4] 0.3614 -0.0845 0.8567 0.3583 -0.6566 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:4] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width"
  .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
 $ center  : Named num [1:4] 5.84 3.06 3.76 1.2
  ..- attr(*, "names")= chr [1:4] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width"
 $ scale   : logi FALSE
 $ x       : num [1:150, 1:4] -2.68 -2.71 -2.89 -2.75 -2.73 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : NULL
  .. ..$ : chr [1:4] "PC1" "PC2" "PC3" "PC4"
 - attr(*, "class")= chr "prcomp"
>
```

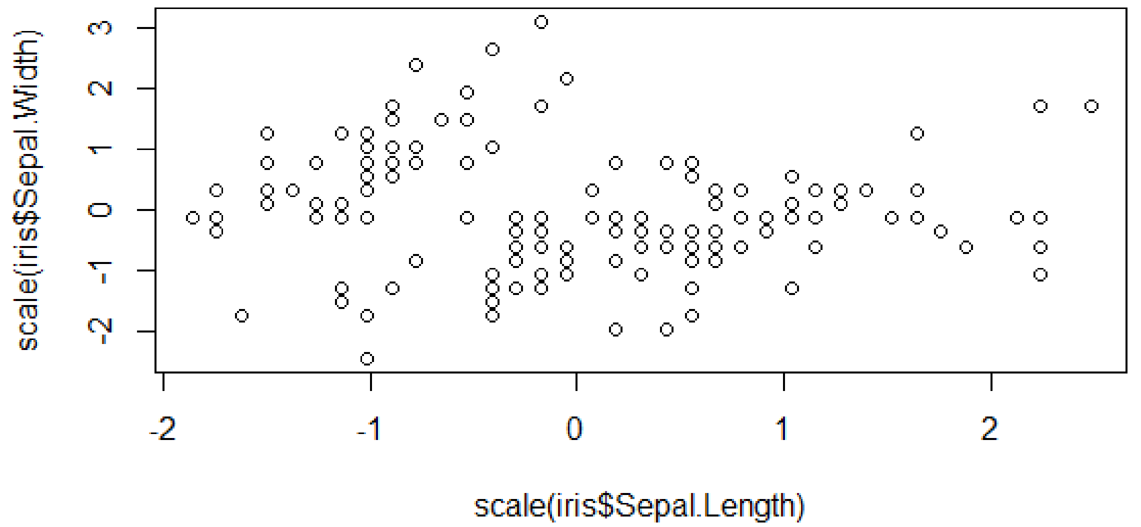
R 4.1.2 · ~/

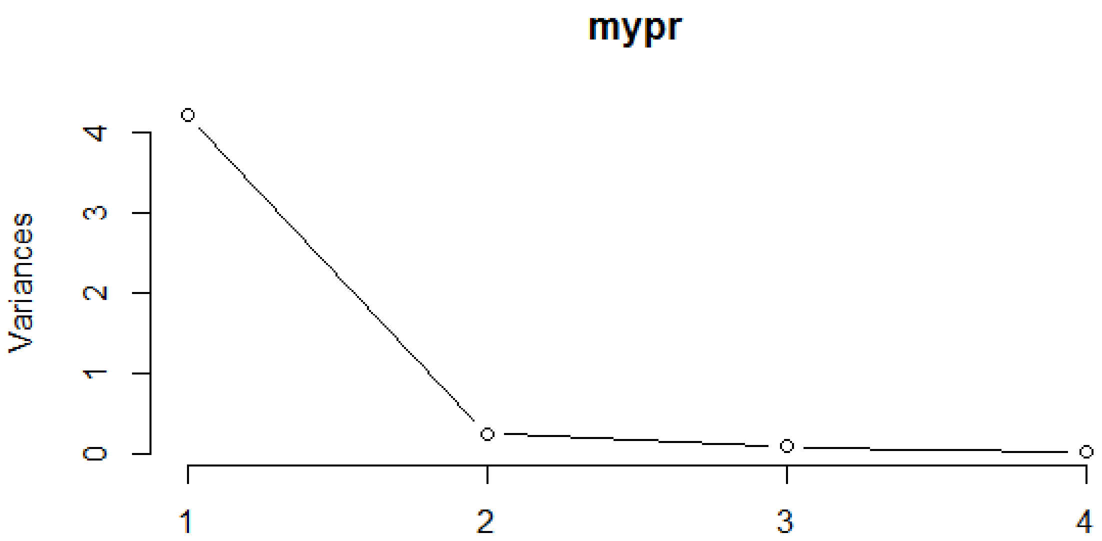
```
>
> mypr$x
             PC1          PC2          PC3           PC4
  [1,] -2.684125626 -0.319397247  0.027914828  0.0022624371
  [2,] -2.714141687  0.177001225  0.210464272  0.0990265503
  [3,] -2.888990569  0.144949426 -0.017900256  0.0199683897
  [4,] -2.745342856  0.318298979 -0.031559374 -0.0755758166
  [5,] -2.728716537 -0.326754513 -0.090079241 -0.0612585926
  [6,] -2.280859633 -0.741330449 -0.168677658 -0.0242008576
  [7,] -2.820537751  0.089461385 -0.257892158 -0.0481431065
  [8,] -2.626144973 -0.163384960  0.021879318 -0.0452978706
  [9,] -2.886382732  0.578311754 -0.020759570 -0.0267447358
 [10,] -2.672755798  0.113774246  0.197632725 -0.0562954013
 [11,] -2.506947091 -0.645068899  0.075318009 -0.0150199245
 [12,] -2.612755231 -0.014729939 -0.102150260 -0.1563792078
 [13,] -2.786109266  0.235112000  0.206844430 -0.0078879115
 [14,] -3.223803744  0.511394587 -0.061299672 -0.0216798118
 [15,] -2.644750390 -1.178764636  0.151627524  0.1592097177
 [16,] -2.386039034 -1.338062330 -0.277776903  0.0065515459
 [17,] -2.623527875 -0.810679514 -0.138183228  0.1677347372
 [18,] -2.648296706 -0.311849145 -0.026668316  0.0776281796
 [19,] -2.199820324 -0.872839039  0.120305523  0.0270518681
 [20,] -2.587986400 -0.513560309 -0.213665172 -0.0662726502
 [21,] -2.310256215 -0.391345936  0.239444043 -0.0150707908
 [22,] -2.543705229 -0.432996063 -0.208457232  0.0410654027
 [23,] -3.215939416 -0.133468070 -0.292396751  0.0044821251
 [24,] -2.302733182 -0.098708855 -0.039123259  0.1483525893
 [25,] -2.355754049  0.037281860 -0.125021083 -0.3003309039
 [26,] -2.506668907  0.146016880  0.253420042  0.0346074722
 [27,] -2.468820073 -0.130951489 -0.094910576  0.0574497158
 [28,] -2.562319906 -0.367718857  0.078494205 -0.0141727423
 [29,] -2.639534715 -0.312039980  0.145908896  0.0657834667
 [30,] -2.631989387  0.196961225 -0.040771079 -0.1239833064
 [31,] -2.587398477  0.204318491  0.077222989 -0.0604622767
 [32,] -2.409932497 -0.410924264  0.145524972  0.2316284917
 [33,] -2.648862334 -0.813363820 -0.225669150 -0.2813723471
 [34,] -2.598736749 -1.093145759 -0.157810813 -0.0953488583
 [35,] -2.636926878  0.121322348  0.143049582  0.0190703413
```

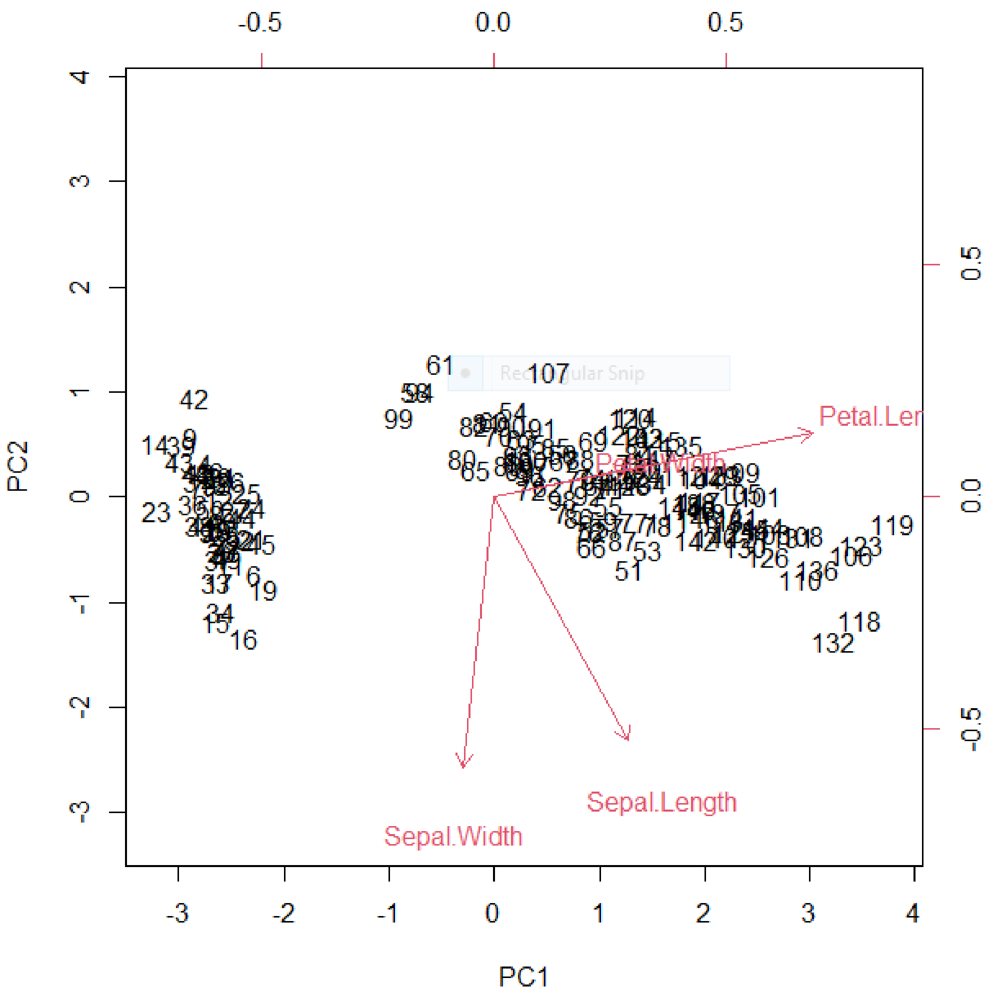
```
> iris2 <- cbind(iris, mypr$x[, 1:2])
> head(iris2)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species       PC1        PC2
1          5.1         3.5          1.4         0.2  setosa -2.684126 -0.3193972
2          4.9         3.0          1.4         0.2  setosa -2.714142  0.1770012
3          4.7         3.2          1.3         0.2  setosa -2.888991  0.1449494
4          4.6         3.1          1.5         0.2  setosa -2.745343  0.3182990
5          5.0         3.6          1.4         0.2  setosa -2.728717 -0.3267545
6          5.4         3.9          1.7         0.4  setosa -2.280860 -0.7413304
>
```

Graphs :


iris$Sepal.Width
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
iris$Sepal.Length


scale(iris$Sepal.Width)
-2
-1
0
1
2
3
-2
-1
0
1
2
scale(iris$Sepal.Length)


mypr
Variances
0
1
2
3
4
1
2
3
4


PC1
PC2
Sepal.Width
Sepal.Length
Petal.Width

4 RESULT AND CONCLUSION :

In this project, we have made a survey on work carried out by different researchers using K-means clustering approach. We also discussed the evolution, limitations and applications of K-means clustering algorithm. It is observed that a lot of improvement has been made to the working of K-means algorithm in the past years. Maximum work carried out on the improvement of efficiency and accuracy of the clusters. This field is always open for improvements. Setting appropriate initial number of clusters is always a challenging task. At the end it is concluded that although there has been made plenty of work on K-means clustering approach, there is a scope for future enhancement.