

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)

Table of Contents

[2 Introduction :](#_Toc89165719) 3

[3 Literature review :](#_Toc89165722)  4

[4 Data Detail :](#_Toc89165723)  4

[5 Data Mining / Analytics / Visualization :](#_Toc89165724)  5-14

[6 Result and Conclusion :](#_Toc89165725) 15

# 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.

# 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

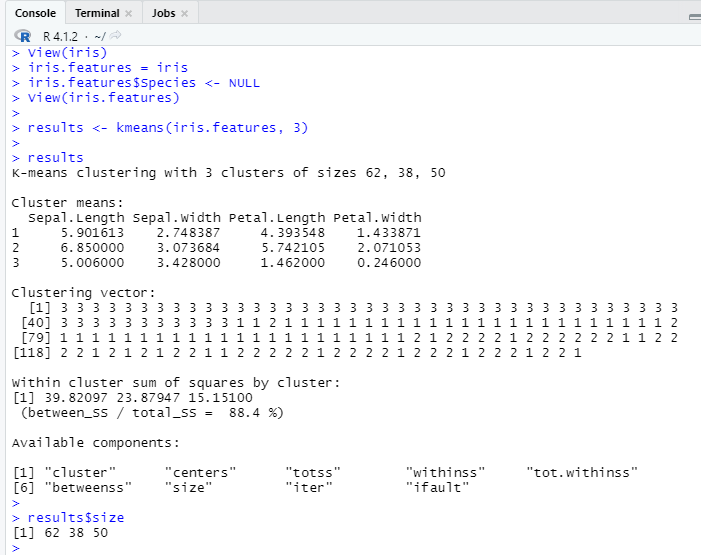
i.e length and width of the sepals and petals and based on the combination of these four features

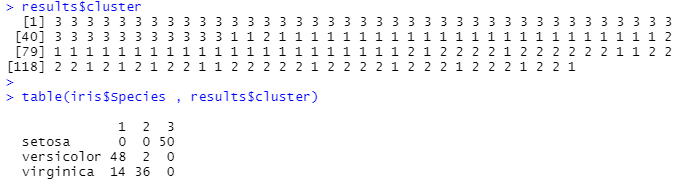
Performing K-Means Clustering on Dataset :

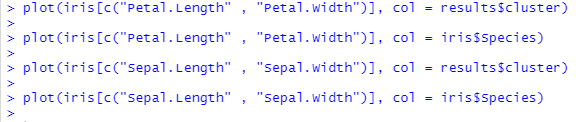
Using K-Means Clustering algorithm on the dataset which includes 150 persons and 4 variables or attributes .

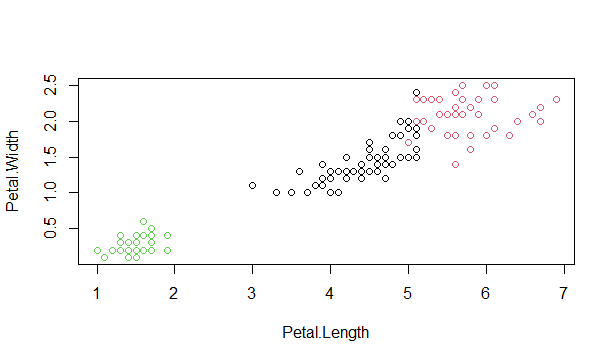
# 5 DATAMINING / ANALYTICS / VISUALIZATION :

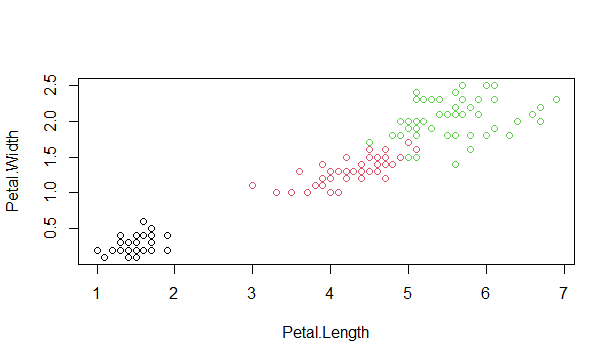
**K-means Working :**

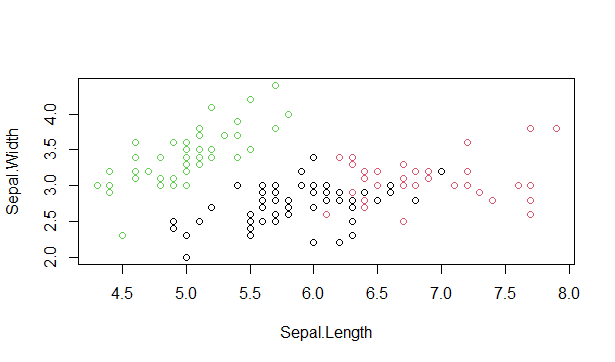


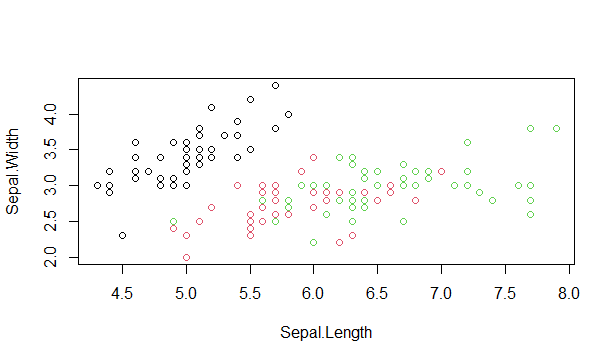




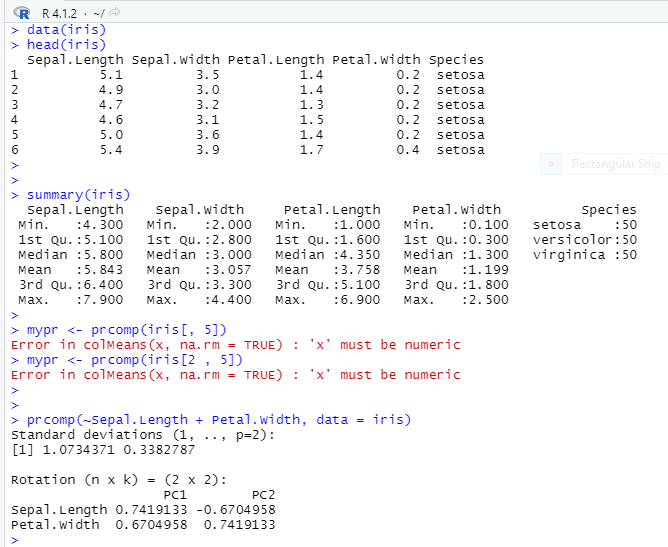


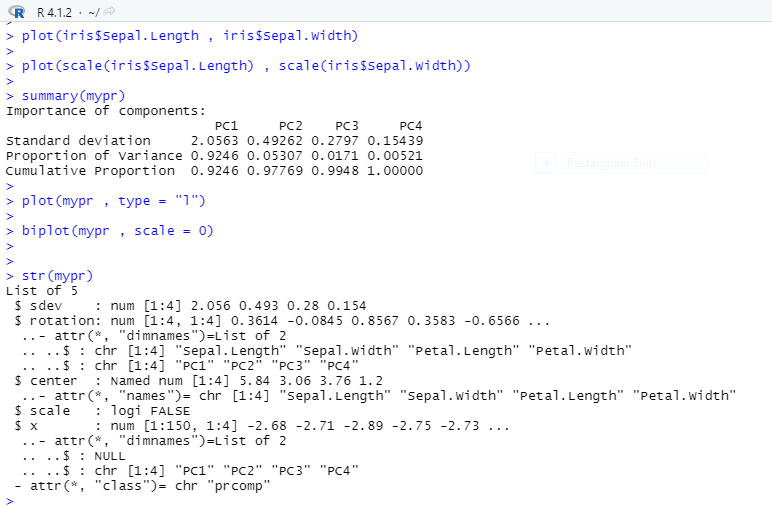


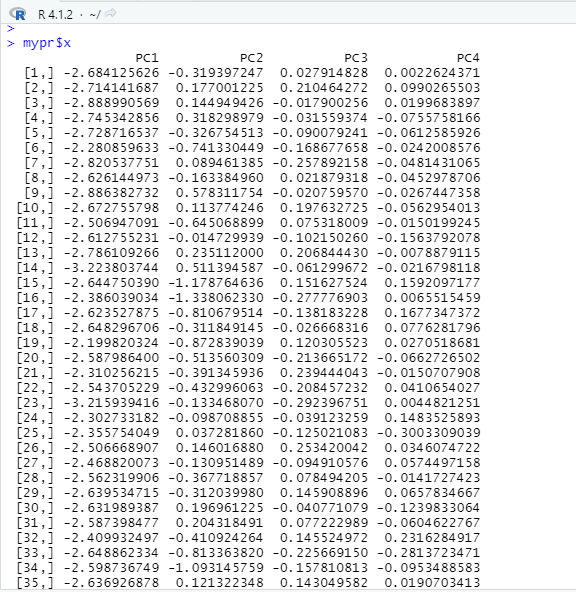


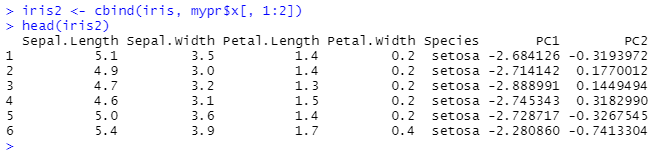


**PCA Working :**

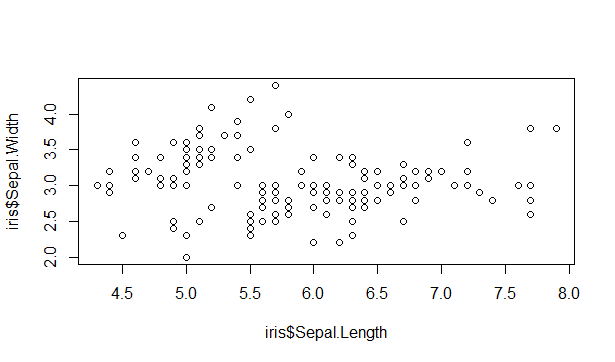
****

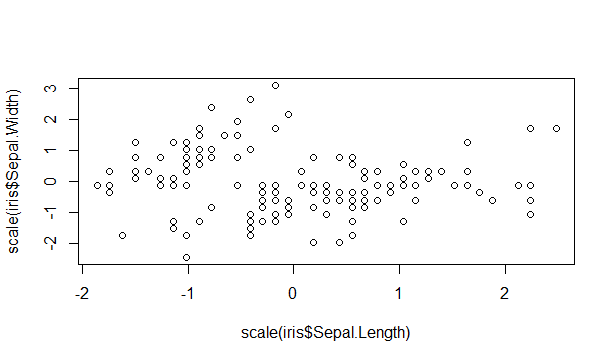
****

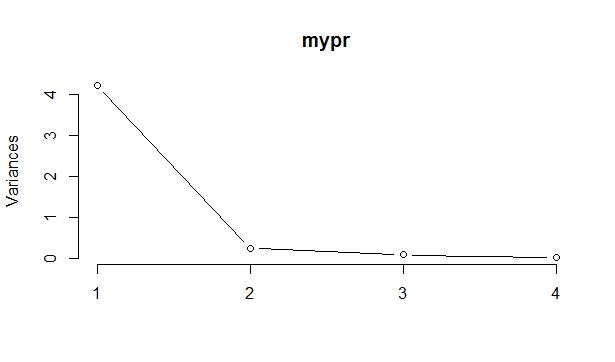
****

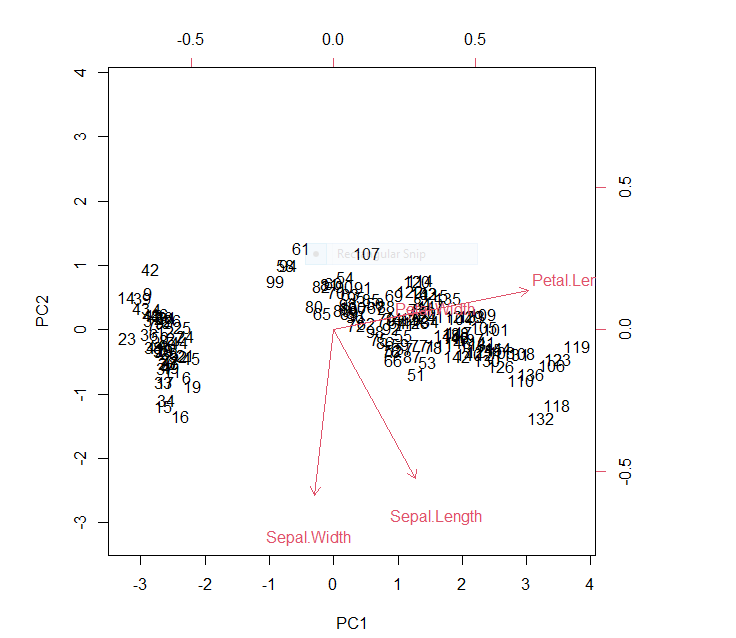
****

**Graphs :**

****

****

****

****

# 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.