# SDM HW -1

Sashank M

50376160

## Radial Velocity of Galaxy NGC7531

## Description

The galaxy data frame records the radial velocity of a spiral galaxy measured at 323 points in the area of sky which it covers. All the measurements lie within seven slots crossing at the origin. The positions of the measurements given by four variables (columns).

## DATA DESCRIPTION:

East-West: The east-west coordinate. The origin, (0,0), is near the center of the galaxy, east is negative, west is positive.

North-South: The north-south coordinate. The origin, (0,0), is near the center of the galaxy, south is negative, north is positive.

Angle: Degrees of counter-clockwise rotation from the horizontal of the slot within which the observation lies.

Radial-position: signed distance from origin; negative if east-west coordinate is negative.

Velocity: Radial velocity measured in km/sec

Names: Values starting from 3 (dropped column)

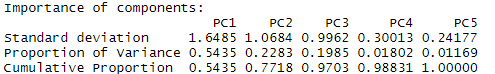
Goal :

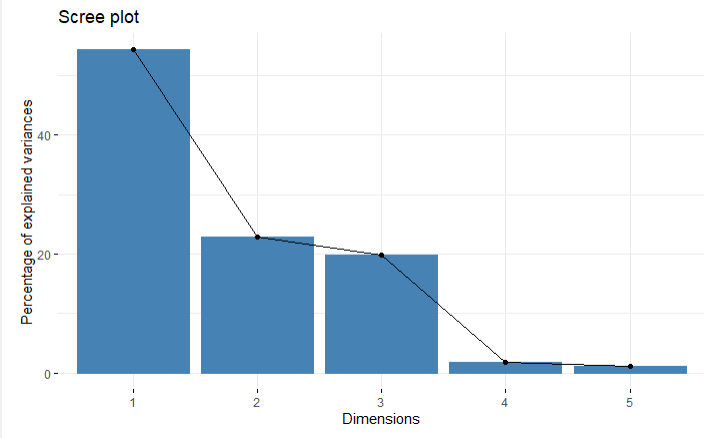
Perform Dimensional Reduction, distance based clustering and spectral clustering methods

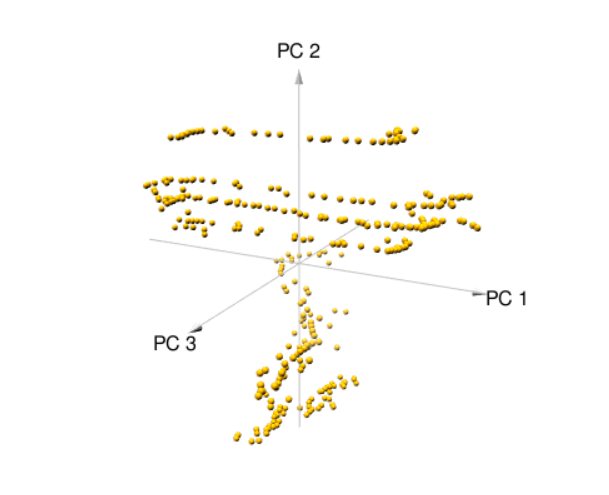
EDA

We observe the Names Columns is unnecessary and drop it from the data set, as its just starts from 3 and

PCA Results after scaling and centering the data







We can see from this, that the combination of PC1, PC2, PC3 capture good amount of variance and hence we select the three components,

After performing the below clustering methods, we observed in scatter plots that the principle components are very similar to the other clustering techniques

## Distance Based Clustering

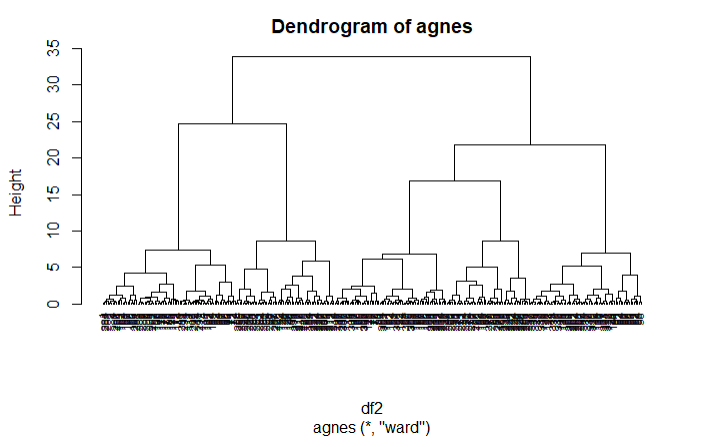
## Algometric Hierarchical clustering

# The agglomerative coefficient, which measures the amount of clustering structure found (values closer to 1 suggest strong clustering structure).

# methods to assess between average single complete and ward  cluster agglomeration methods



From the above result we can see ward linkage method is closer to one and we prefer this to perform hierarchical clustering and below we see the resultant dendrogram

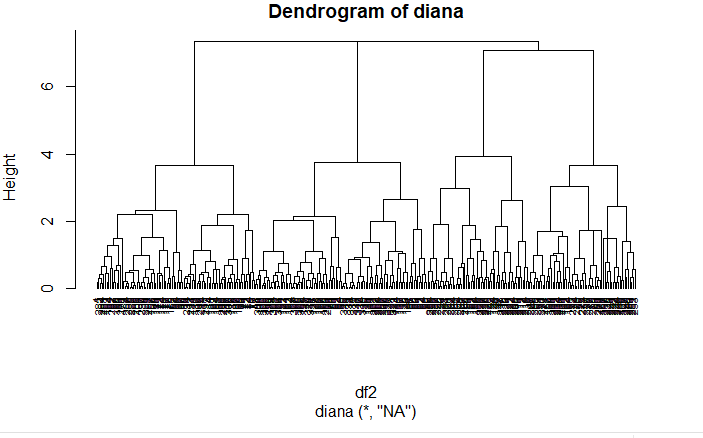


Divisive Hierarchical Clustering

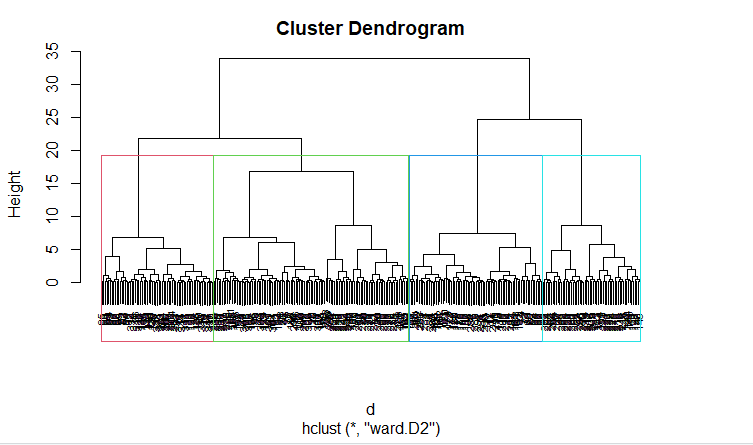
A top down approach for hierarchal clustering

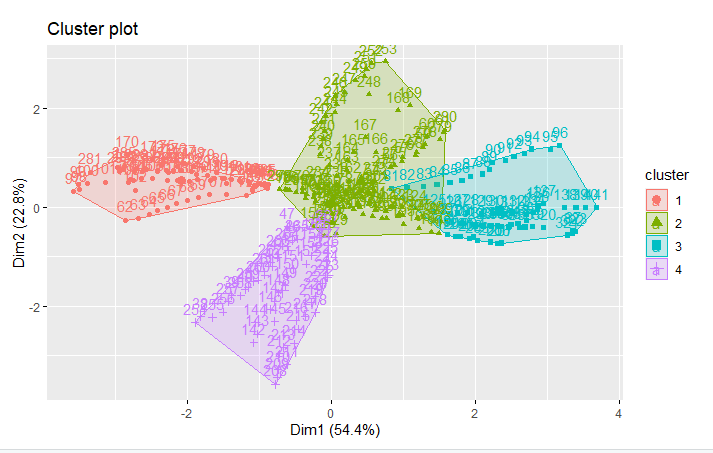
# Divise coefficient; amount of clustering structure

0.9716346



We can distinctly see from the dendrogram that there are 4 sub-groups this also matches with K means clustering



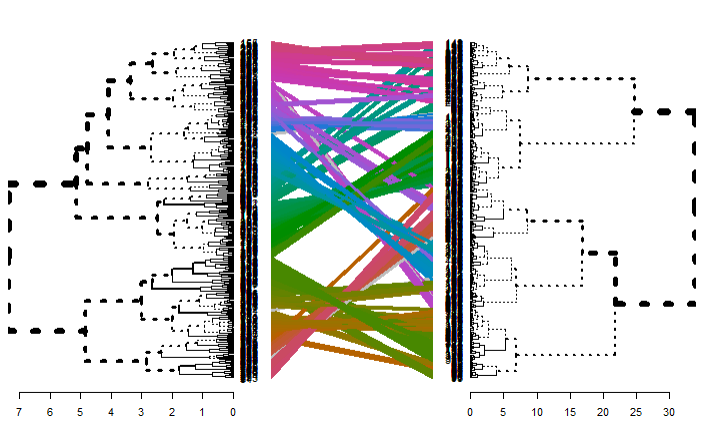


Cluster plot here gives us a clearer understanding of the 4 unique clusters in the dataset

## we can also compare two dendrograms. Here we compare hierarchical clustering with complete linkage versus Ward’s method.

The output displays unique nodes, the dashed lines represent a combination of labels/items which is not present in the other tree

## Comparing two dendrograms. Here we compare hierarchical clustering with complete linkage versus Ward’s method.



The output displays unique nodes, with a combination of labels/items not present in the other tree, highlighted with dashed lines. The quality of the alignment of the two trees can be measured using the Entanglement which is between 1 (full entanglement) and 0 (no entanglement). A lower entanglement coefficient corresponds to a good alignment

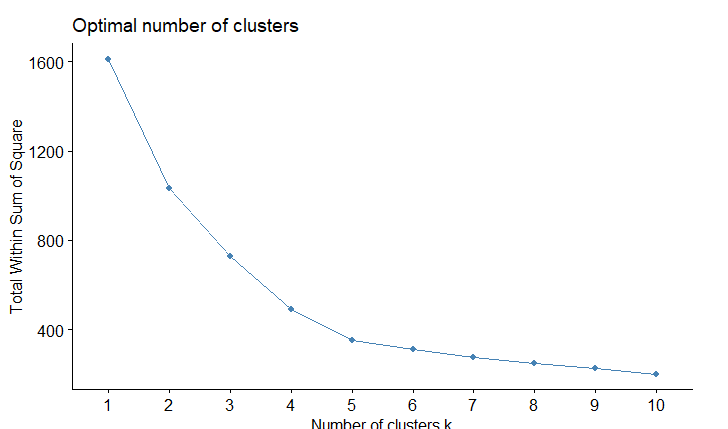
For the above graph between Euclidean and ward linkage method we get an entanglement score of 0.34

which indicates a good alignment

## Determining Optimal Clusters

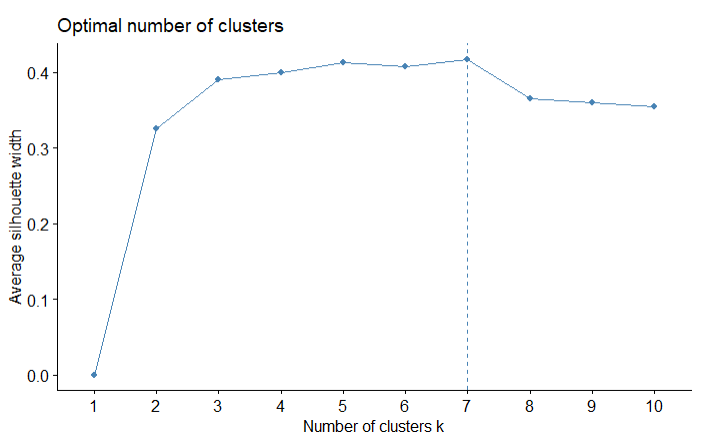
Using elbow method

Algometric Hierarchal Clustering

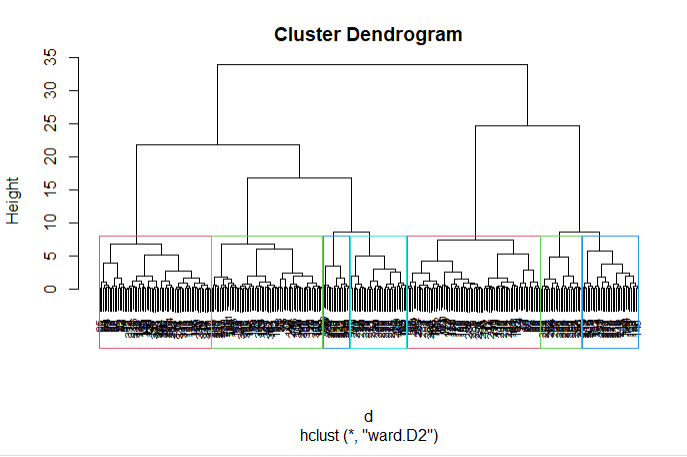


### Average Silhouette Method

It measures the quality of a clustering by determining how well each point lies within its cluster. A high average silhouette width indicates a good clustering

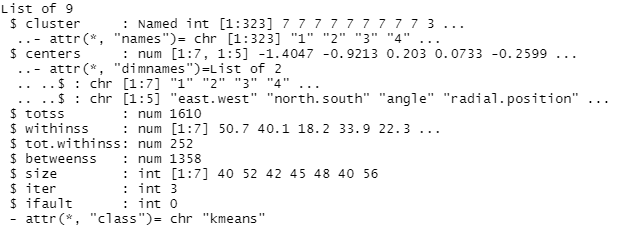


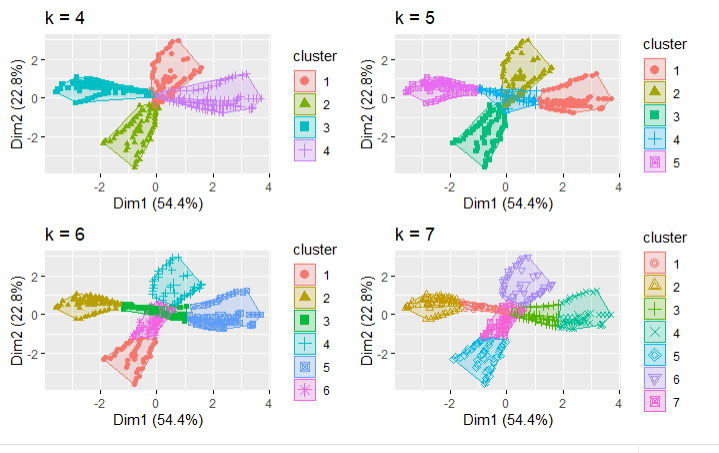
So from this inference if we take 7 clusters the dendrogram and cluster plot would look like this





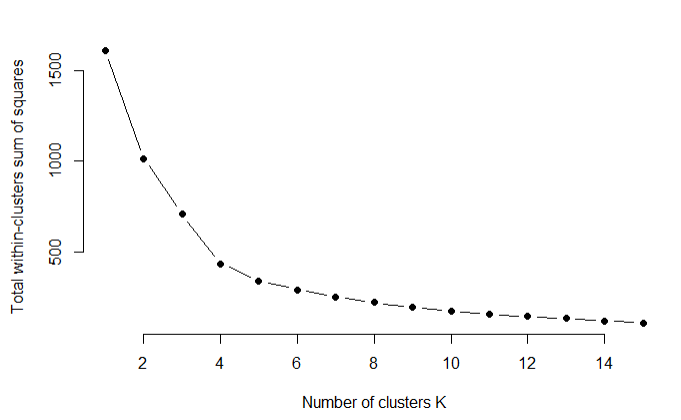
## K – Means





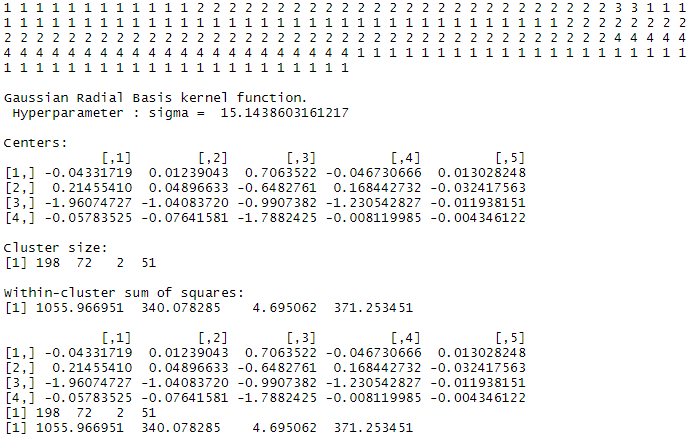
For different values of K we can see the clusters

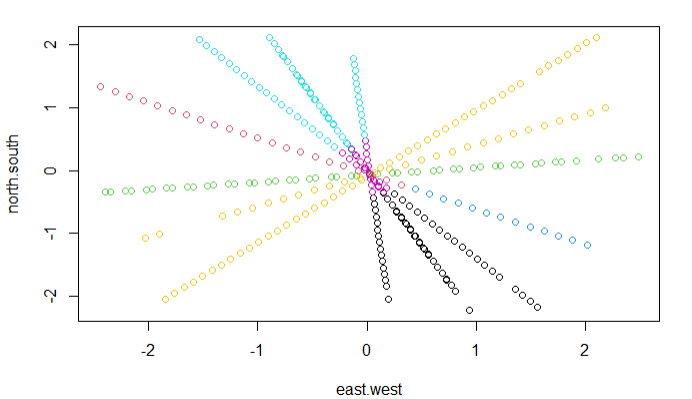
## Determining Optimal Clusters



Here when using Euclidean distance, we see optimal number of clusters is 4 but in case of AHC clustering when using wards distance, we see that 7 is the optimal number of clusters

Spectral Clustering





Data isn’t really in a geometric shapes and the best method for clustering the data set would be AHC method over K means also, as in k means you need to sense ahead of time the number of cluster you desire (usually done with domain knowledge of the dataset and knowing where the appropriate cut in the dendrogram would be) and also K means will give misleading results or ones which make no real sense if the data like we have not separated into well defined sphere like clusters. Where as in AHC method we calculate distance between each data point. And AHC joins nearby points into a cluster and then adds nearby points to the group. Which ends up as a dendrogram and thus making it easier to cut the tree into clusters

## Core inference from performing clustering on the data set

We noticed that the cluster at the center has low radial velocity and the clusters which are branching out from the central cluster have increasing radial velocity