AIRLINES DATASET

BUSSINESS PROBLEM:

Perform clustering (Both hierarchical and K means clustering) for the airlines data to obtain optimum number of clusters.

Draw the inferences from the clusters obtained.

DATA:

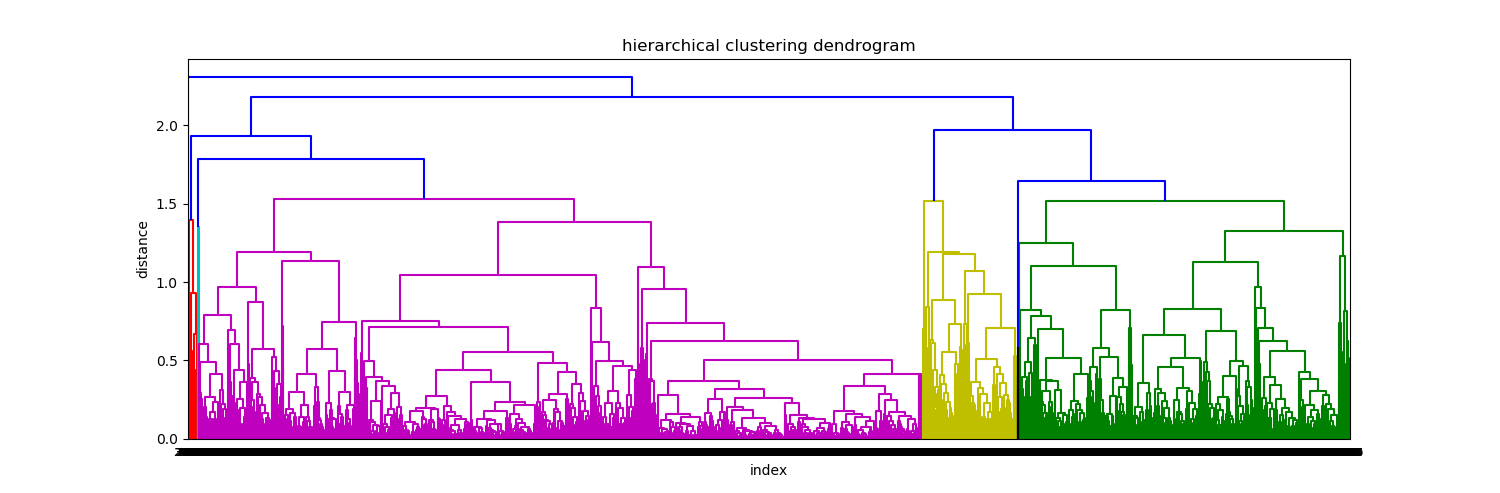
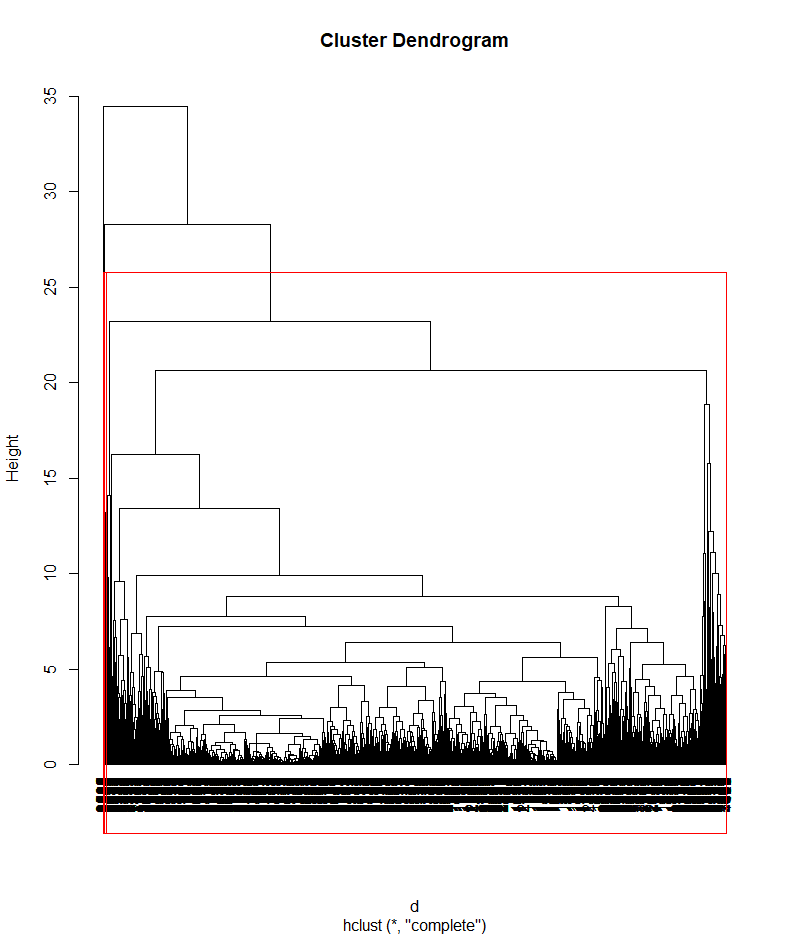
Given data is in the form of numerical and dummies.

PREPROCESSING:

Access the data which is in numerical form because we cannot access the categorical data.

Build model:

build the required model using complete linkage function and calculate the distance using the Euclidean distance.

DENDROGRAM:

now applying Agglomerative clusters and choosing clusters as 3 based on the dendrogram

I have used k=3 and processed further.

Airlines data is divided into 3 clusters.

In agglomerative clustering taking clusters as 3 and using complete linkage function find the distance between them using Euclidean method.

After finding the cluster attach the column to data to make the further processes.

K-Means clustering:

=>K means clustering is used for large datasets.

=>need measure of within cluster similarity and between cluster similarity.

K selection:

**By Using K-Selection:**

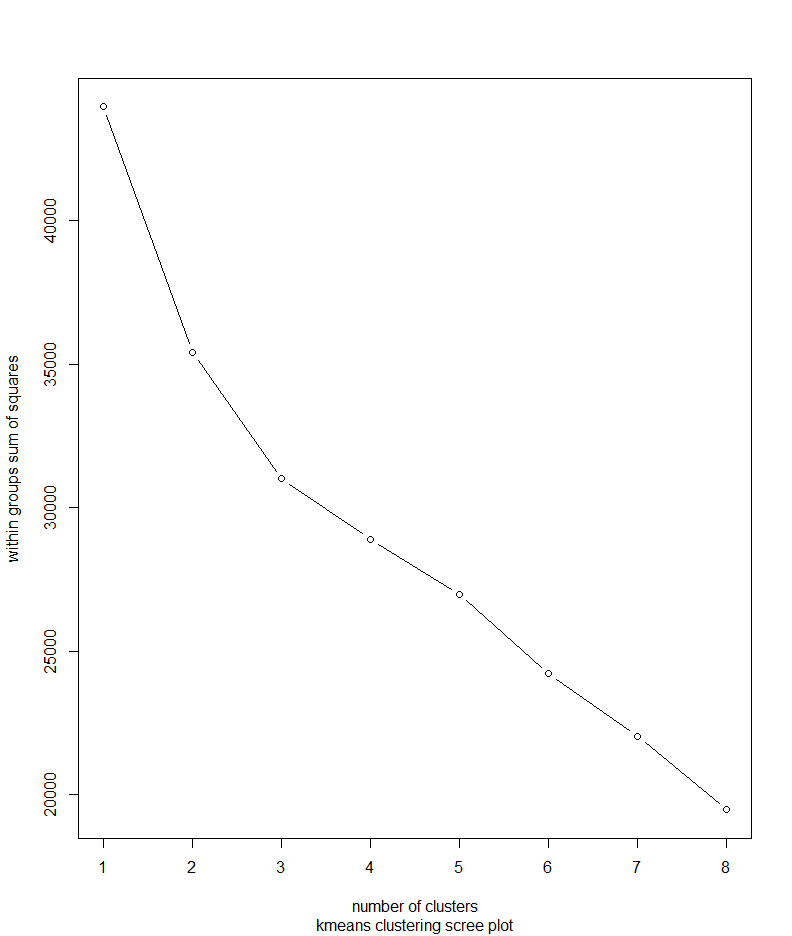
k <-k selection(mydata[,-1], parallel = TRUE, k\_threshold = 0.9, max\_centers=20)

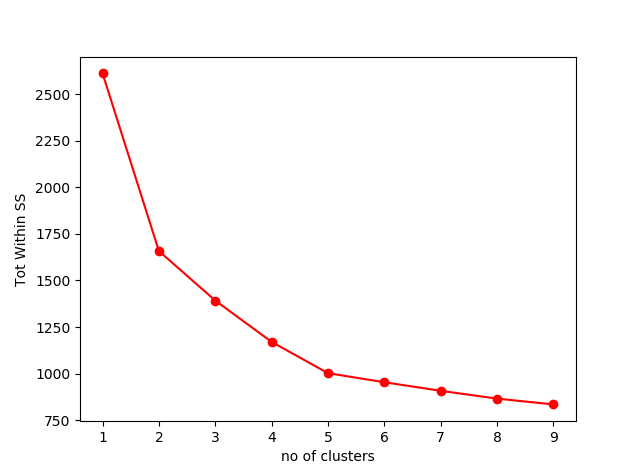
It is giving the value as 2.

ELBOW CURVE:

Using elbow curve we can find that how many clusters can be formed

Based on the scree plot by looking at it’s elbow which is formed .





#selecting 3 clusters from the above scree plot which is the optimum no of clusters

Based on the scree plot I have decided to form 3 clusters.

K value is subjective we can also take the nearest value.

km$size

[1] 1285 192 2522

Cluster 1 cluster 2 cluster 3

1285 192 2522

km$centers

Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles Bonus\_trans

1 0.4067701 0.008990352 1.1764397 -0.08241771 -0.03884333 0.8710729 0.7213056

2 1.1947579 0.718557501 0.2442543 0.11341920 1.05765273 0.9915843 1.6646144

3 -0.2982130 -0.059284553 -0.6180102 0.03335855 -0.06072785 -0.5193152 -0.4942441

Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll Award?

-0.1086584 -0.1208535 0.3658556 0.5560910

3.1487903 3.3946273 0.3160019 0.9047374

-0.1843544 -0.1968563 -0.2104666 -0.3522151

km <- kmeans(nor\_data,3)

str(km)

$ tot. withinss: num 31001

$ betweenss: num 12977

km <- kmeans(nor\_data,4)

str(km)

$ tot. withinss: num 29394

$ betweenss : num 14584

For a good model tot. withinss should be more and betweenss should be less. If there should be less variation in the difference then we can treat it as final K.

By using this clusters, we can categorize the passengers into 3 forms

Using groupby function using the mean we can calculate the clusters average and group them.

Cluster 2 -High frequent fliers:

Cluster 1 -middle class travellers

Cluster 3-non frequent travellers

Clara is used for larger datasets:

