**Business Problem:**

Perform K means clustering for the airlines data to obtain optimum number of clusters. Draw the inferences from the clusters obtained.

**Data:**

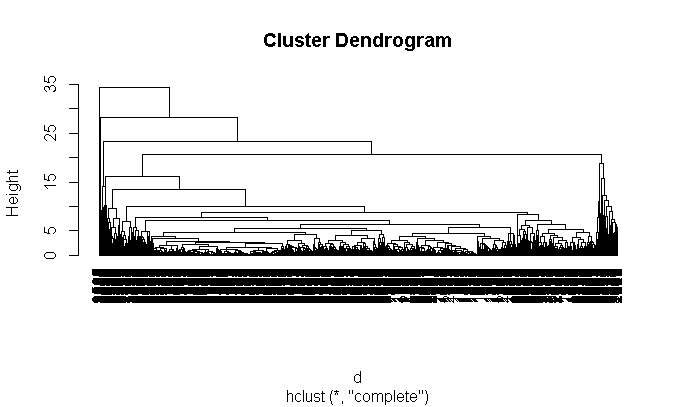
Data in the form of mixed data. It contains the numerical data and dummied data (CC1\_Miles, CC2\_Miles, CC3\_Miles and Award).

**Pre-processing Data:**

All the features are not in same scale. So first of all convert them all into single scale. There was no outlier and NA in the data. Delete the unused feature like ID# from the processing.

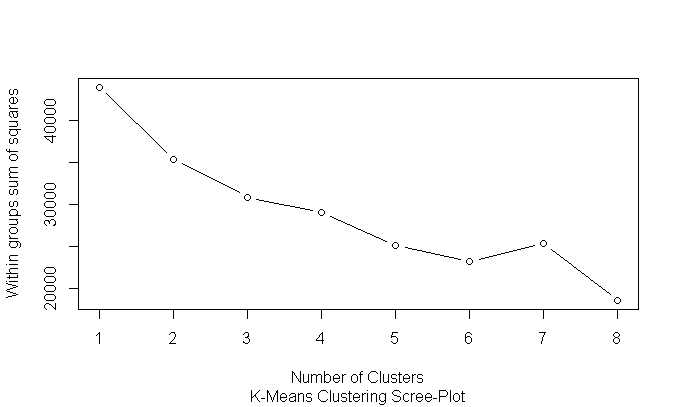
**Building the Model:**

Build the model using Euclidean distance and complete linkage functions. Please find the Dendrogram.



Now treat them as Non Hierarchical cluster. Then we will find out the K by using different techniques.

**Elbow Diagram:**



Elbow Curve value is subjective**. As per my analysis 3** will be good but it subjective.

**K-Selection:**

k <- kselection(normalized\_data, parallel = TRUE,k\_threshold = 0.9, max\_centers=20)

k It is giving the value as 9.

**KMeans:**

Calculate the tot.withinss and betweenss. For a good model tot.withinss should be more and betweenss should be less.

Trails :

km <- kmeans(normalized\_data,2)

str(km)

### $ tot.withinss: num 35401

### $ betweenss : num 8577

### $ size : int [1:2] 1299 2700

km <- kmeans(normalized\_data,3)

str(km)

### $ tot.withinss: num 30891

### $ betweenss : num 13087

### $ size : int [1:3] 1259 166 2574

km <- kmeans(normalized\_data,4)

str(km)

### $ tot.withinss: num 28902

### $ betweenss : num 15076

### $ size : int [1:4] 1105 1246 1492 156

km <- kmeans(normalized\_data,5)

str(km)

### $ tot.withinss: num 26969

### $ betweenss : num 17009

### $ size : int [1:5] 843 990 1185 837 144

km <- kmeans(normalized\_data,6)

str(km)

### $ tot.withinss: num 23276

### $ betweenss : num 20702

### $ size : int [1:5] 843 990 1185 837 144

In the cluster 6 , there will no big different in tot.withinss and betweenss. As per the standards betweenss should be less and tot.withinss should be more**. As per my analysis 3 will** be good value.

**KMeans Centers and Animation Daigram:**

km <- kmeans.ani(normalized\_data, 3)

km$centers

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

##[1,] 0.5211265 0.23527228 0.9826914 -0.01563495 -0.03673291 0.9572350 0.9346656

##[2,] -0.3394993 -0.06877284 -0.6950264 0.03416583 -0.06275873 -0.5585523 -0.5906727

##[3,] 0.3769130 -0.10265477 0.8877736 -0.08873064 0.25331807 0.4822836 0.6171354

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

##[1,] 0.5781578 0.6423030 0.3935208 1.2301065

##[2,] -0.1801135 -0.1998437 -0.2332890 -0.2937025

##[3,] -0.2163900 -0.2412144 0.2101077 -0.7493992

km <- kmeans.ani(normalized\_data, 4)

km$centers

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

##[1,] -0.1381418 -0.041382806 -0.5094780 0.14696290 -0.05863815 -0.4496403 -0.3345047

##[2,] 0.4623084 0.005207654 1.3168737 -0.07989180 -0.03502291 1.0525418 0.7940116

##[3,] 1.1793333 0.778254961 0.1776275 0.13941116 1.19524712 0.8751779 1.6806652

##[4,] -0.3670911 -0.059182702 -0.5796813 -0.07988766 -0.06275873 -0.5125043 -0.5087130

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

##[1,] -0.14077618 -0.14875705 0.7695733 -0.1454833

##[2,] -0.07514787 -0.08707015 0.3156184 0.6609040

##[3,] 3.30901047 3.61085143 0.2610349 0.8920611

##[4,] -0.20825468 -0.22754019 -0.9146390 -0.4766340

km <- kmeans.ani(normalized\_data, 5)

km$centers

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

##[1,] 0.6564207 0.001209454 1.5526353 -0.08202682 0.23132056 1.4155076 0.88666103

##[2,] 1.1928611 0.786875849 0.1138058 0.17634182 -0.06275873 0.6629405 1.74967064

##[3,] -0.1552108 0.066485684 -0.2467497 -0.03334087 -0.06275873 -0.2588067 -0.09484595

##[4,] -0.3883714 -0.067156604 -0.5886759 -0.08125526 -0.06275873 -0.5287557 -0.52814902

##[5,] -0.1330921 -0.094199206 -0.4111550 0.17047610 -0.05754827 -0.4399343 -0.29380380

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

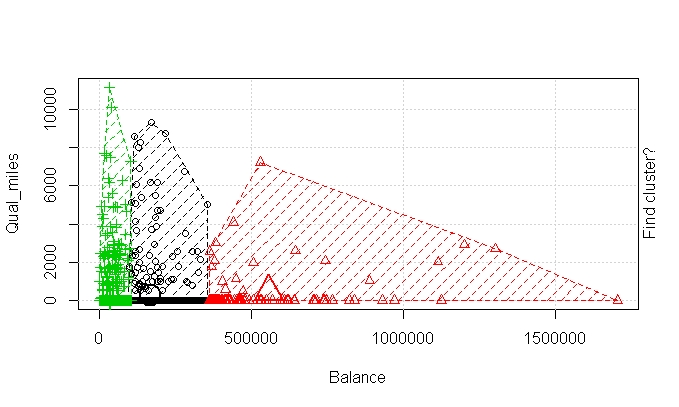
##[1,] -0.05324787 -0.05371149 0.39905434 0.5235730

##[2,] 3.70591227 4.03054939 0.27263524 0.9120240

##[3,] -0.04125950 -0.04071308 0.06344509 1.3037551

##[4,] -0.22770840 -0.25023884 -0.96160730 -0.7668234

##[5,] -0.20034706 -0.22185879 0.73663056 -0.7562915



There will be less variation in all centres when we increase the number of clusters. There will be no over lapping for the centres of the cluster when cluster3.

**Hence as per my analysis K will be 3 as good value.**