Perform clustering (Both hierarchical and K means clustering) for the airlines data

library(readxl)  
Airlines\_data<-read\_xlsx("E:\\assignments data\\clustering\\EastWestAirlines.xlsx", sheet="data")  
View(Airlines\_data)  
attach(Airlines\_data)  
Airlines\_data$Award<-Airlines\_data$`Award?`  
Airlines\_data1<-Airlines\_data[, c(2:11,13)]  
colnames(Airlines\_data1)

## [1] "Balance" "Qual\_miles" "cc1\_miles"   
## [4] "cc2\_miles" "cc3\_miles" "Bonus\_miles"   
## [7] "Bonus\_trans" "Flight\_miles\_12mo" "Flight\_trans\_12"   
## [10] "Days\_since\_enroll" "Award"

summary(Airlines\_data1)

## Balance Qual\_miles cc1\_miles cc2\_miles   
## Min. : 0 Min. : 0.0 Min. :1.00 Min. :1.000   
## 1st Qu.: 18528 1st Qu.: 0.0 1st Qu.:1.00 1st Qu.:1.000   
## Median : 43097 Median : 0.0 Median :1.00 Median :1.000   
## Mean : 73601 Mean : 144.1 Mean :2.06 Mean :1.015   
## 3rd Qu.: 92404 3rd Qu.: 0.0 3rd Qu.:3.00 3rd Qu.:1.000   
## Max. :1704838 Max. :11148.0 Max. :5.00 Max. :3.000   
## cc3\_miles Bonus\_miles Bonus\_trans Flight\_miles\_12mo  
## Min. :1.000 Min. : 0 Min. : 0.0 Min. : 0.0   
## 1st Qu.:1.000 1st Qu.: 1250 1st Qu.: 3.0 1st Qu.: 0.0   
## Median :1.000 Median : 7171 Median :12.0 Median : 0.0   
## Mean :1.012 Mean : 17145 Mean :11.6 Mean : 460.1   
## 3rd Qu.:1.000 3rd Qu.: 23801 3rd Qu.:17.0 3rd Qu.: 311.0   
## Max. :5.000 Max. :263685 Max. :86.0 Max. :30817.0   
## Flight\_trans\_12 Days\_since\_enroll Award   
## Min. : 0.000 Min. : 2 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:2330 1st Qu.:0.0000   
## Median : 0.000 Median :4096 Median :0.0000   
## Mean : 1.374 Mean :4119 Mean :0.3703   
## 3rd Qu.: 1.000 3rd Qu.:5790 3rd Qu.:1.0000   
## Max. :53.000 Max. :8296 Max. :1.0000

str(Airlines\_data1)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 3999 obs. of 11 variables:  
## $ Balance : num 28143 19244 41354 14776 97752 ...  
## $ Qual\_miles : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ cc1\_miles : num 1 1 1 1 4 1 3 1 3 3 ...  
## $ cc2\_miles : num 1 1 1 1 1 1 1 1 2 1 ...  
## $ cc3\_miles : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ Bonus\_miles : num 174 215 4123 500 43300 ...  
## $ Bonus\_trans : num 1 2 4 1 26 0 25 4 43 28 ...  
## $ Flight\_miles\_12mo: num 0 0 0 0 2077 ...  
## $ Flight\_trans\_12 : num 0 0 0 0 4 0 0 1 12 3 ...  
## $ Days\_since\_enroll: num 7000 6968 7034 6952 6935 ...  
## $ Award : num 0 0 0 0 1 0 0 1 1 1 ...

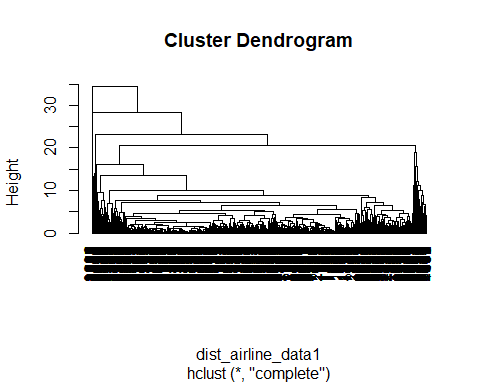
#Normalizing data  
norm\_airline\_data1<-scale(Airlines\_data1)  
  
#distance matrix  
dist\_airline\_data1<-dist(norm\_airline\_data1, method="euclidean")  
str(dist\_airline\_data1)

## 'dist' num [1:7994001] 0.137 0.377 0.135 4.774 0.159 ...  
## - attr(\*, "Size")= int 3999  
## - attr(\*, "Diag")= logi FALSE  
## - attr(\*, "Upper")= logi FALSE  
## - attr(\*, "method")= chr "euclidean"  
## - attr(\*, "call")= language dist(x = norm\_airline\_data1, method = "euclidean")

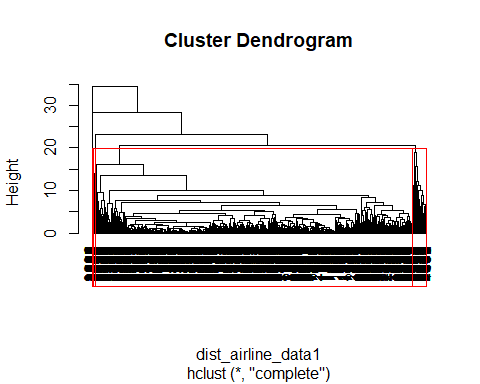
#heirarchical clustering  
fit<-hclust(dist\_airline\_data1, method="complete")  
fit

##   
## Call:  
## hclust(d = dist\_airline\_data1, method = "complete")  
##   
## Cluster method : complete   
## Distance : euclidean   
## Number of objects: 3999

#Dendrogram  
plot(fit, hang=-1)



rect.hclust(fit,plot(fit, hang=-1), k=5, border="red")



#cut tree  
group\_airline<-cutree(fit, k=5)  
membership<-as.matrix(group\_airline)  
final<-data.frame(Airlines\_data, membership)  
final1<-final[,c(ncol(final),1:(ncol(final)-1))]  
aggregate(final1[,-2],by=list(final$membership),mean)

## Group.1 membership Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles  
## 1 1 1 65902.07 137.3707 2.033580 1.000000 1.000793  
## 2 2 2 117123.66 255.7529 2.252941 1.341176 1.000000  
## 3 3 3 806433.29 383.2143 3.571429 1.000000 1.000000  
## 4 4 4 138061.40 78.8000 3.466667 1.000000 4.066667  
## 5 5 5 131999.50 347.0000 2.500000 1.000000 1.000000  
## Bonus\_miles Bonus\_trans Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll  
## 1 15571.37 10.72448 270.5854 0.8183501 4072.295  
## 2 37437.17 26.72941 4066.6235 11.8823529 4701.688  
## 3 58412.32 21.21429 1344.3929 5.6071429 6835.893  
## 4 93927.87 28.06667 506.6667 1.6000000 4613.867  
## 5 65634.25 69.25000 19960.0000 49.2500000 2200.250  
## Award. Award  
## 1 0.3503437 0.3503437  
## 2 0.7058824 0.7058824  
## 3 0.8571429 0.8571429  
## 4 0.5333333 0.5333333  
## 5 1.0000000 1.0000000

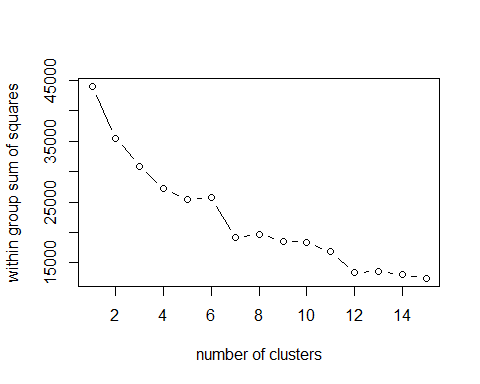
#k means clustering  
library(stats)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
library(ggfortify)  
data<-Airlines\_data1  
  
#normlizing data  
data<-scale(data)  
  
#choosing optimum number of clusters  
wss<-(nrow(data)-1)\*sum(apply(data,2,var))  
for(i in 2:15){  
 wss[i]<-sum(kmeans(data, centers=i)$withinss)  
}  
plot(1:15, wss, type="b", xlab="number of clusters", ylab="within group sum of squares")



#kmeans clustering  
km<-kmeans(data, 5)  
str(km)

## List of 9  
## $ cluster : int [1:3999] 4 4 4 4 2 4 2 3 5 2 ...  
## $ centers : num [1:5, 1:11] -0.387 0.645 -0.152 -0.139 1.218 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : chr [1:5] "1" "2" "3" "4" ...  
## .. ..$ : chr [1:11] "Balance" "Qual\_miles" "cc1\_miles" "cc2\_miles" ...  
## $ totss : num 43978  
## $ withinss : num [1:5] 2937 10029 6686 2654 4675  
## $ tot.withinss: num 26981  
## $ betweenss : num 16997  
## $ size : int [1:5] 1173 845 846 991 144  
## $ iter : int 4  
## $ ifault : int 0  
## - attr(\*, "class")= chr "kmeans"

km$centers

## Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles  
## 1 -0.3873897 -0.06471939 -0.5880722 -0.04626468 -0.06275873 -0.5262241  
## 2 0.6451686 0.00074222 1.5476707 -0.09021450 0.22818835 1.4044642  
## 3 -0.1520187 0.06167135 -0.2732911 0.19797865 -0.06275873 -0.2628248  
## 4 -0.1387748 -0.09494251 -0.4103834 -0.06405825 -0.05759034 -0.4504253  
## 5 1.2178769 0.81390783 0.1383415 0.18396880 -0.06275873 0.6889574  
## Bonus\_trans Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll Award  
## 1 -0.51521442 -0.22708847 -0.24974766 -0.98464184 -0.7668234  
## 2 0.88394330 -0.05385681 -0.05387644 0.39678909 0.5318827  
## 3 -0.08507358 -0.03267595 -0.03211569 0.07021378 1.2768327  
## 4 -0.32703167 -0.20324069 -0.22458724 0.72976906 -0.7668234  
## 5 1.76024342 3.75652193 4.08482939 0.25761184 0.9011426

km$size

## [1] 1173 845 846 991 144

km$withinss

## [1] 2936.613 10029.235 6686.335 2653.546 4675.144

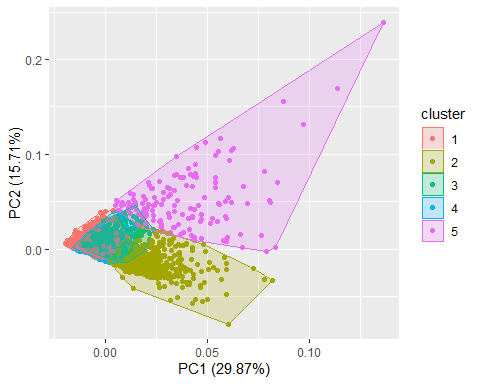
km$tot.withinss

## [1] 26980.87

km$totss

## [1] 43978

autoplot(km, data, frame=TRUE)



final1<-data.frame(norm\_airline\_data1,km$cluster )  
colnames(final1)

## [1] "Balance" "Qual\_miles" "cc1\_miles"   
## [4] "cc2\_miles" "cc3\_miles" "Bonus\_miles"   
## [7] "Bonus\_trans" "Flight\_miles\_12mo" "Flight\_trans\_12"   
## [10] "Days\_since\_enroll" "Award" "km.cluster"

aggregate(final1[, 1:12], by=list(final1$km.cluster), FUN = mean)

## Group.1 Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles  
## 1 1 -0.3873897 -0.06471939 -0.5880722 -0.04626468 -0.06275873 -0.5262241  
## 2 2 0.6451686 0.00074222 1.5476707 -0.09021450 0.22818835 1.4044642  
## 3 3 -0.1520187 0.06167135 -0.2732911 0.19797865 -0.06275873 -0.2628248  
## 4 4 -0.1387748 -0.09494251 -0.4103834 -0.06405825 -0.05759034 -0.4504253  
## 5 5 1.2178769 0.81390783 0.1383415 0.18396880 -0.06275873 0.6889574  
## Bonus\_trans Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll Award  
## 1 -0.51521442 -0.22708847 -0.24974766 -0.98464184 -0.7668234  
## 2 0.88394330 -0.05385681 -0.05387644 0.39678909 0.5318827  
## 3 -0.08507358 -0.03267595 -0.03211569 0.07021378 1.2768327  
## 4 -0.32703167 -0.20324069 -0.22458724 0.72976906 -0.7668234  
## 5 1.76024342 3.75652193 4.08482939 0.25761184 0.9011426  
## km.cluster  
## 1 1  
## 2 2  
## 3 3  
## 4 4  
## 5 5