**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 Description:**

**The file EastWestAirlinescontains information on passengers who belong to an airline’s frequent flier program. For each passenger the data include information on their mileage history and on different ways they accrued or spent miles in the last year. The goal is to try to identify clusters of passengers that have similar characteristics for the purpose of targeting different segments for different types of mileage offers**

**ID --Unique ID**

**Balance--Number of miles eligible for award travel**

**Qual\_mile--Number of miles counted as qualifying for Topflight status**

**cc1\_miles -- Number of miles earned with freq. flyer credit card in the past 12 months:**

**cc2\_miles -- Number of miles earned with Rewards credit card in the past 12 months:**

**cc3\_miles -- Number of miles earned with Small Business credit card in the past 12 months:**

**1 = under 5,000**

**2 = 5,000 - 10,000**

**3 = 10,001 - 25,000**

**4 = 25,001 - 50,000**

**5 = over 50,000**

**Bonus\_miles--Number of miles earned from non-flight bonus transactions in the past 12 months**

**Bonus\_trans--Number of non-flight bonus transactions in the past 12 months**

**Flight\_miles\_12mo--Number of flight miles in the past 12 months**

**Flight\_trans\_12--Number of flight transactions in the past 12 months**

**Days\_since\_enrolled--Number of days since enrolled in flier program**

**Award--whether that person had award flight (free flight) or not**

**K means clustering:**

install.packages("readxl")

library(readxl)

data <- read\_xlsx(file.choose(),2)# Read xlsx file

View(data)

head(data)

#display the structure of the data

str(data)

**# here we can see that variable cc1\_miles,cc2\_miles,cc3\_miles and Award? are categorical form**

**#1 = under 5,000**

**#2 = 5,000 - 10,000**

**#3 = 10,001 - 25,000**

**#4 = 25,001 - 50,000**

**#5 = over 50,000**

**# we have to take average of cc1\_miles,cc2\_miles,cc3\_miles above respective range**

data$cc1\_miles = ifelse(data$cc1\_miles==1,3500,

ifelse(data$cc1\_miles==2,8000,

ifelse(data$cc1\_miles==3,20000,

ifelse(data$cc1\_miles==4,38000,

ifelse(data$cc1\_miles==5,65000,0)))))

data$cc2\_miles = ifelse(data$cc2\_miles==1,3500,

ifelse(data$cc2\_miles==2,8000,

ifelse(data$cc2\_miles==3,20000,

ifelse(data$cc2\_miles==4,38000,

ifelse(data$cc2\_miles==5,65000,0)))))

data$cc3\_miles = ifelse(data$cc3\_miles==1,3500,

ifelse(data$cc3\_miles==2,8000,

ifelse(data$cc3\_miles==3,20000,

ifelse(data$cc3\_miles==4,38000,

ifelse(data$cc3\_miles==5,65000,0)))))

View(data)

head(data)

**#Normalize the data**

**#here i am taking 2 to 11 columns**

normalized\_Airline\_data <- scale(data[2:11]) #excluding the ID and Award column before normalizing

View(normalized\_Airline\_data)

summary(normalized\_Airline\_data)

Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles

Min. :-0.7303 Min. :-0.1863 Min. :-0.6475 Min. :-0.08692 Min. :-0.05704

1st Qu.:-0.5465 1st Qu.:-0.1863 1st Qu.:-0.6475 1st Qu.:-0.08692 1st Qu.:-0.05704

Median :-0.3027 Median :-0.1863 Median :-0.6475 Median :-0.08692 Median :-0.05704

Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.00000 Mean : 0.00000

3rd Qu.: 0.1866 3rd Qu.:-0.1863 3rd Qu.: 0.2574 3rd Qu.:-0.08692 3rd Qu.:-0.05704

Max. :16.1868 Max. :14.2231 Max. : 2.7252 Max. :15.26910 Max. :23.55849

Bonus\_miles Bonus\_trans Flight\_miles\_12mo Flight\_trans\_12

Min. :-0.7099 Min. :-1.20805 Min. :-0.3286 Min. :-0.36212

1st Qu.:-0.6581 1st Qu.:-0.89568 1st Qu.:-0.3286 1st Qu.:-0.36212

Median :-0.4130 Median : 0.04145 Median :-0.3286 Median :-0.36212

Mean : 0.0000 Mean : 0.00000 Mean : 0.0000 Mean : 0.00000

3rd Qu.: 0.2756 3rd Qu.: 0.56208 3rd Qu.:-0.1065 3rd Qu.:-0.09849

Max. :10.2083 Max. : 7.74673 Max. :21.6803 Max. :13.61035

Days\_since\_enroll

Min. :-1.99336

1st Qu.:-0.86607

Median :-0.01092

Mean : 0.00000

3rd Qu.: 0.80960

Max. : 2.02284

**#lets create clusters using kmeans function**

**#initially we are going to create three clusters, k is the number of clusters**

fit1 <- kmeans(normalized\_Airline\_data, 3) 3 cluster solution

str(fit1)

List of 9

$ cluster : int [1:3999] 2 2 2 2 3 2 2 2 3 3 ...

$ centers : num [1:3, 1:10] -0.3525 -0.0932 0.8204 -0.0671 -0.013 ...

..- attr(\*, "dimnames")=List of 2

.. ..$ : chr [1:3] "1" "2" "3"

.. ..$ : chr [1:10] "Balance" "Qual\_miles" "cc1\_miles" "cc2\_miles" ...

$ totss : num 39980

$ withinss : num [1:3] 4855 7210 17732

$ tot.withinss: num 29796

$ betweenss : num 10184

$ size : int [1:3] 1684 1429 886

$ iter : int 4

$ ifault : int 0

- attr(\*, "class")= chr "kmeans"

**#here we have total withiness between the observation and cluster centre should be very small**

**#and betweenss between the observation and cluster centre should be very large**

**#our objective of k means clustering into keep these values large and small**

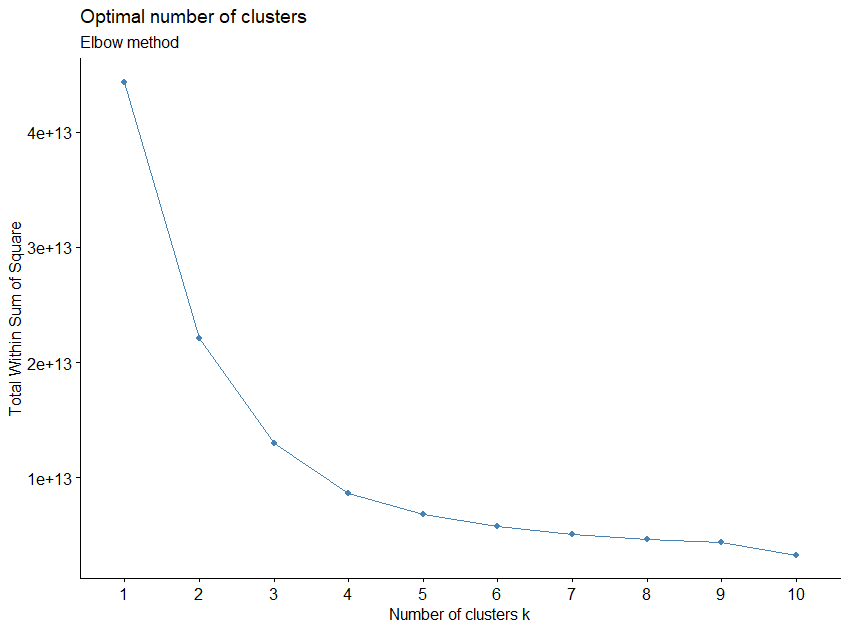
**#so we have to try different k values to get higher betweeness and fewer withiness between the clusters**

**#elbow curve & k ~ sqrt(n/2) to decide the k value**

install.packages("factoextra")

library(factoextra)

fviz\_nbclust(data[2:11],kmeans,method="wss")+labs(subtitle = "Elbow method")



**#the elbow curve we can see that 10 data points first data pont contain a big slope**

**#similarly data ponts2,3,4,5,6 also contain some slope.**

**#but the data points 7,8,9,10 contains no slope. this is our cutoff**

**# we can try k=7,8,9,10, which k value will give the result of smaller withiness and higher betweeness**

**#considerd as final fit**

fit2 <- kmeans(normalized\_Airline\_data, 7) **# 7 cluster solution**

str(fit2)

List of 9

$ cluster : int [1:3999] 3 3 3 3 6 3 6 3 1 6 ...

$ centers : num [1:7, 1:10] 1.027 0.275 -0.254 -0.146 1.029 ...

..- attr(\*, "dimnames")=List of 2

.. ..$ : chr [1:7] "1" "2" "3" "4" ...

.. ..$ : chr [1:10] "Balance" "Qual\_miles" "cc1\_miles" "cc2\_miles" ...

$ totss : num 39980

$ withinss : num [1:7] 3319 4106 1704 1701 6219 ...

$ tot.withinss: num 20612

$ betweenss : num 19368

$ size : int [1:7] 129 66 1046 729 299 608 1122

$ iter : int 5

$ ifault : int 0

- attr(\*, "class")= chr "kmeans"

**#tot.withinss: num 18002**

**#betweenss : num 21978**

fit3 <- kmeans(normalized\_Airline\_data, 8) **# 8 cluster solution**

str(fit3)

List of 9

$ cluster : int [1:3999] 2 2 2 2 7 2 7 2 1 7 ...

$ centers : num [1:8, 1:10] 0.594 -0.199 0.712 -0.182 5.995 ...

..- attr(\*, "dimnames")=List of 2

.. ..$ : chr [1:8] "1" "2" "3" "4" ...

.. ..$ : chr [1:10] "Balance" "Qual\_miles" "cc1\_miles" "cc2\_miles" ...

$ totss : num 39980

$ withinss : num [1:8] 2890 1784 5302 4843 1018 ...

$ tot.withinss: num 19440

$ betweenss : num 20540

$ size : int [1:8] 120 1024 284 660 46 1093 718 54

$ iter : int 4

$ ifault : int 0

- attr(\*, "class")= chr "kmeans"

**#tot.withinss: num 17177**

**# betweenss : num 22803**

fit4 <- kmeans(normalized\_Airline\_data, 9) **# 9 cluster solution**

str(fit4)

List of 9

$ cluster : int [1:3999] 6 6 6 6 2 6 2 6 9 2 ...

$ centers : num [1:9, 1:10] -0.177 0.29 0.431 -0.212 5.995 ...

..- attr(\*, "dimnames")=List of 2

.. ..$ : chr [1:9] "1" "2" "3" "4" ...

.. ..$ : chr [1:10] "Balance" "Qual\_miles" "cc1\_miles" "cc2\_miles" ...

$ totss : num 39980

$ withinss : num [1:9] 1760.4 1701.2 709.4 89.5 1017.6 ...

$ tot.withinss: num 16053

$ betweenss : num 23927

$ size : int [1:9] 732 684 54 15 46 1011 1058 282 117

$ iter : int 5

$ ifault : int 0

- attr(\*, "class")= chr "kmeans"

**#tot.withinss: num 13775**

**# betweenss : num 26205**

fit5 <- kmeans(normalized\_Airline\_data, 10) # **10 cluster solution**

str(fit5)

List of 9

$ cluster : int [1:3999] 3 3 3 3 6 3 5 3 1 5 ...

$ centers : num [1:10, 1:10] 0.358 0.419 -0.352 0.739 0.289 ...

..- attr(\*, "dimnames")=List of 2

.. ..$ : chr [1:10] "1" "2" "3" "4" ...

.. ..$ : chr [1:10] "Balance" "Qual\_miles" "cc1\_miles" "cc2\_miles" ...

$ totss : num 39980

$ withinss : num [1:10] 1391 693 899 1536 1142 ...

$ tot.withinss: num 15137

$ betweenss : num 24843

$ size : int [1:10] 183 53 868 49 524 647 11 609 66 989

$ iter : int 6

$ ifault : int 0

- attr(\*, "class")= chr "kmeans"

**#tot.withinss: num 12826**

**# betweenss : num 27154**

**#here k=10, we got small withiness and higher betweeness**

**#to view centers of each clusters using following methods**

fit$centers

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

1 0.3581881 0.11176005 -0.22902062 -0.01826702 -0.02241486 0.02508722 0.83462559

2 0.4191562 7.21784750 -0.04530281 -0.08692286 -0.05703711 0.07647738 0.06895705

3 -0.3516838 -0.12094349 -0.62837941 -0.08692286 -0.05703711 -0.58750656 -0.76144048

4 0.7389089 0.53027273 0.31779702 0.22646536 -0.05703711 0.96277000 2.60633558

5 0.2891281 -0.11515795 -0.03008675 -0.06294573 -0.05373946 -0.08391169 0.40092312

6 0.3741784 -0.09460016 1.77100734 -0.08692286 -0.02231756 1.43419390 0.82343837

7 0.7203925 -0.18627539 0.97528321 -0.08692286 17.90332710 3.47034725 1.53707074

8 -0.2104630 -0.13565154 -0.09222051 0.41738003 -0.05703711 -0.08506326 0.57079879

9 4.8236056 0.12162595 1.82657114 -0.08692286 -0.05703711 2.74272913 1.32408710

10 -0.4149803 -0.12721927 -0.63804117 -0.08692286 -0.05703711 -0.60385080 -0.82703098

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

1 2.0069944 2.1099636 -0.02241315

2 0.3380094 0.3939529 -0.08368633

3 -0.2041889 -0.2169430 0.68917774

4 6.0576411 6.2017655 0.10475762

5 -0.1963409 -0.2177290 0.98418618

6 -0.1097673 -0.1233465 0.38893101

7 -0.0493831 -0.0505580 0.16150595

8 -0.2371156 -0.2521679 -0.71455538

9 0.7311172 1.0159515 1.09519668

10 -0.2367927 -0.2443014 -1.01218368

**#create final clusters using fit5$cluster**

final<- data.frame(data, fit5$cluster) # append cluster membership

View(final)

head(final)

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

1 1 28143 0 3500 3500 3500 174 1

2 2 19244 0 3500 3500 3500 215 2

3 3 41354 0 3500 3500 3500 4123 4

4 4 14776 0 3500 3500 3500 500 1

5 5 97752 0 38000 3500 3500 43300 26

6 6 16420 0 3500 3500 3500 0 0

Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll Award. fit5.cluster

1 0 0 7000 0 3

2 0 0 6968 0 3

3 0 0 7034 0 3

4 0 0 6952 0 3

5 2077 4 6935 1 6

6 0 0 6942 0 3

**#change the position fit$cluster last to first**

final\_Airline\_data <- final[,c(ncol(final),1:(ncol(final)-1))]

View(final\_Airline\_data)

head(final\_Airline\_data)

fit5.cluster ID. Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles Bonus\_trans

1 3 1 28143 0 3500 3500 3500 174 1

2 3 2 19244 0 3500 3500 3500 215 2

3 3 3 41354 0 3500 3500 3500 4123 4

4 3 4 14776 0 3500 3500 3500 500 1

5 6 5 97752 0 38000 3500 3500 43300 26

6 3 6 16420 0 3500 3500 3500 0 0

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

1 0 0 7000 0

2 0 0 6968 0

3 0 0 7034 0

4 0 0 6952 0

5 2077 4 6935 1

6 0 0 6942 0

**#find the average values of each clusters**

aggregate(data[,2:11], by=list(fit5$cluster), FUN=mean)

Group.1 Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles Bonus\_trans

1 1 109697.97 230.57923 11131.148 3573.770 3590.164 17750.727 19.617486

2 2 115842.08 5728.30189 14481.132 3500.000 3500.000 18991.849 12.264151

3 3 38160.16 50.54493 3849.078 3500.000 3500.000 2955.994 4.289171

4 4 148065.37 554.36735 21102.041 3836.735 3500.000 40396.673 36.632653

5 5 102738.41 55.02099 14758.588 3525.763 3508.588 15118.298 15.452290

6 6 111309.40 70.92581 47600.464 3500.000 3590.417 51782.017 19.510046

7 7 146199.36 0.00000 33090.909 3500.000 50272.727 100957.091 26.363636

8 8 52391.78 39.16585 13625.616 4041.872 3500.000 15090.486 17.083744

9 9 559703.39 238.21212 48613.636 3500.000 3500.000 83384.409 24.318182

10 10 31781.41 45.68959 3672.902 3500.000 3500.000 2561.265 3.659252

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

1 3270.2678 9.3770492 4072.273

2 933.3396 2.8679245 3945.736

3 174.1486 0.5506912 5541.804

4 8942.0204 24.8979592 4334.898

5 185.1374 0.5477099 6151.036

6 306.3586 0.9057187 4921.754

7 390.9091 1.1818182 4452.091

8 128.0443 0.4170772 2642.906

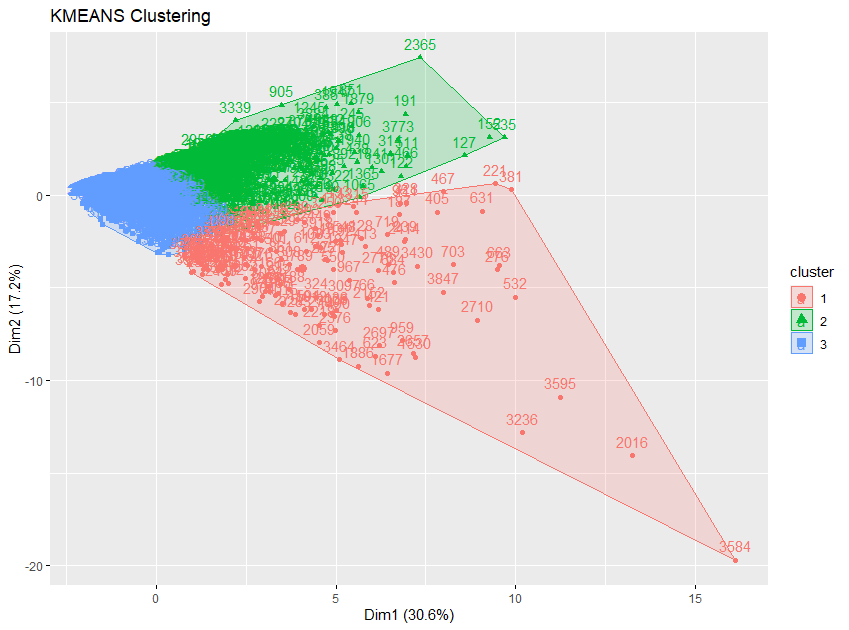
9 1483.7727 5.2272727 6380.288

10 128.4965 0.4469161 2028.264

**###VISUALIZATION**

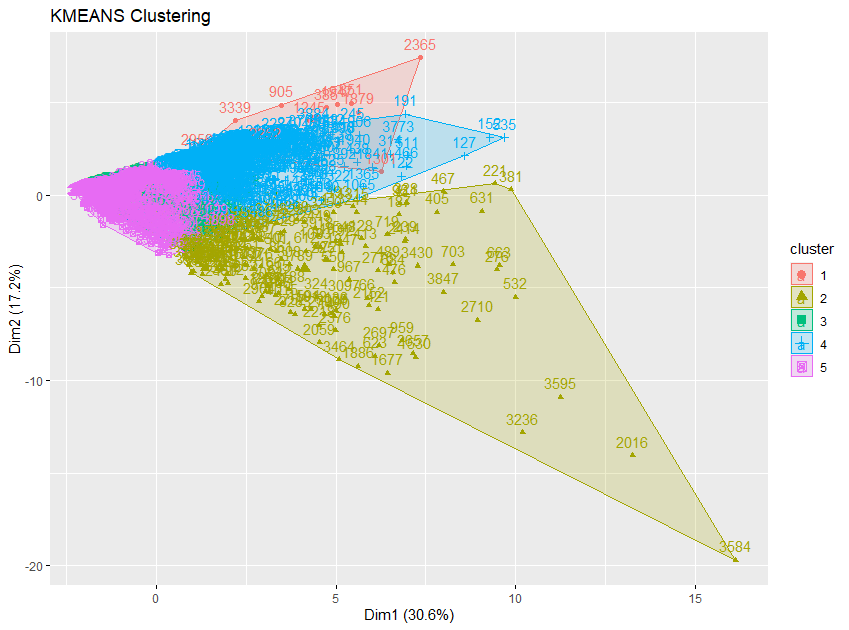
# k=3

eclust(normalized\_Airline\_data, "kmeans", k = 3, nstart = 25, graph = TRUE)



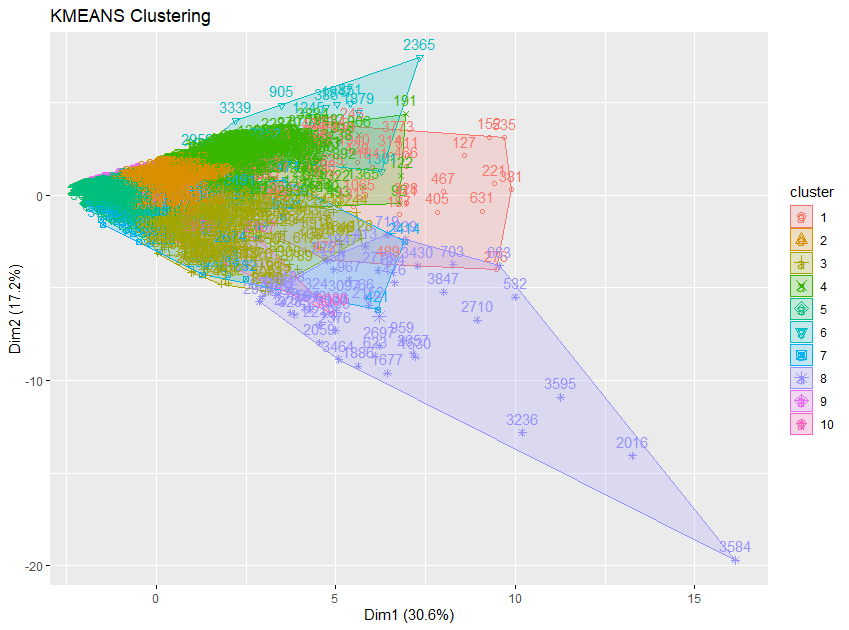
#k=5

eclust(normalized\_Airline\_data, "kmeans", k = 5, nstart = 25, graph = TRUE)



#k=10

eclust(normalized\_Airline\_data, "kmeans", k = 10, nstart = 25, graph = TRUE)

****