Clustering assignment

Problem statement 1

Perform Clustering for the crime data and identify the number of clusters formed and draw inferences.

Data Description:

Murder -- Muder rates in different places of United States

Assualt- Assualt rate in different places of United States

UrbanPop - urban population in different places of United States

Rape - Rape rate in different places of United States

Answer:

Rcode:

```
crime_data <- read.csv(file.choose())
View(crime_data)
attach(crime_data)
ncol(crime_data)
names(crime_data)
crime_data1 <- crime_data[,2:5]
norm_crime_data1 <- scale(crime_data1)
distance <- dist(norm_crime_data1,method = "euclidean")
str(distance)
crime_clust <- hclust(distance,method = "complete")
plot(crime_clust,hang=-1)
rect.hclust(crime_clust,plot(crime_clust,hang=-1),k=4,border="blue")
group <- cutree(crime_clust,k=4)
crime_data_final <- cbind(crime_data,group)
View(crime_data_final)
aggregate(crime_data_final[,2:6],by=list(group),FUN = max)
```

Console:

```
> crime_data <- read.csv(file.choose())
> View(crime_data)
```

n_Rules.R × clustering1.R × crime_data_final × crime_data ×

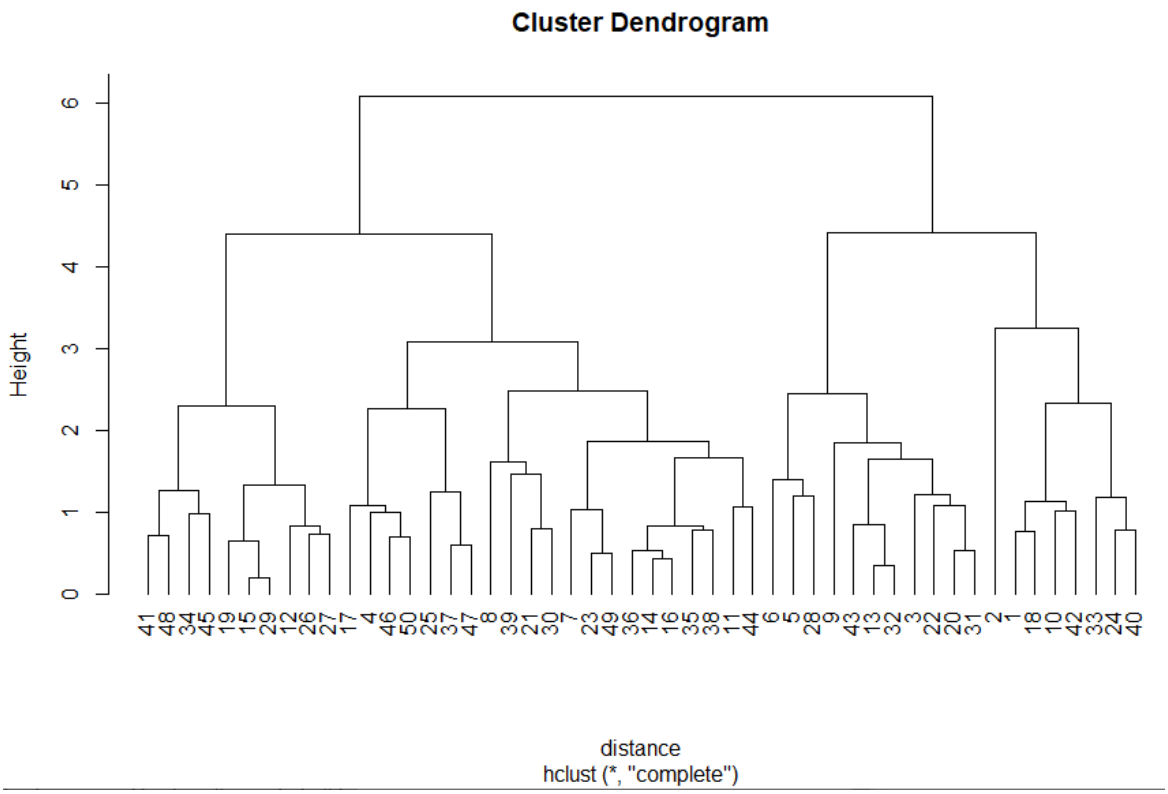
Filter

	X	Murder	Assault	UrbanPop	Rape
1	Alabama	13.2	236	58	21.2
2	Alaska	10.0	263	48	44.5
3	Arizona	8.1	294	80	31.0
4	Arkansas	8.8	190	50	19.5
5	California	9.0	276	91	40.6
6	Colorado	7.9	204	78	38.7
7	Connecticut	3.3	110	77	11.1
8	Delaware	5.9	238	72	15.8
9	Florida	15.4	335	80	31.9
10	Georgia	17.4	211	60	25.8

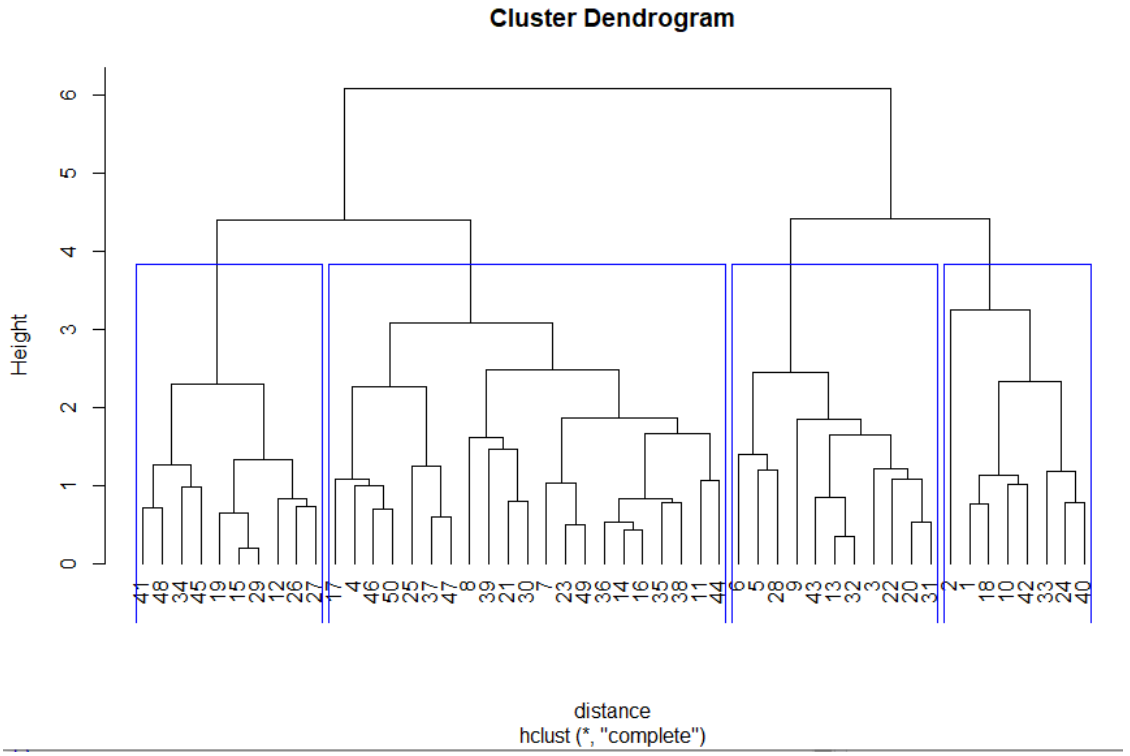
Showing 1 to 11 of 50 entries, 5 total columns

```
> attach(crime_data)

> ncol(crime_data)
[1] 5
> names(crime_data)
[1] "X"        "Murder"   "Assault"  "UrbanPop" "Rape"
> crime_data1 <- crime_data[,2:5]
> norm_crime_data1 <- scale(crime_data1)
> distance <- dist(norm_crime_data1,method = "euclidean")
> str(distance)
 'dist' num [1:1225] 2.7 2.29 1.29 3.26 2.65 ...
 - attr(*, "Size")= int 50
 - attr(*, "Diag")= logi FALSE
 - attr(*, "Upper")= logi FALSE
 - attr(*, "method")= chr "euclidean"
 - attr(*, "call")= language dist(x = norm_crime_data1, method = "eucl
idean")
> crime_clust <- hclust(distance,method = "complete")
> plot(crime_clust,hang=-1)
```




```
> rect.hclust(crime_clust,plot(crime_clust,hang=-1),k=4,border="blue")
```



```
> group <- cutree(crime_clust,k=4)
> crime_data_final <- cbind(crime_data,group)
> View(crime_data_final)
```

_Rules.R × | clustering1.R × | crime_data_final × | crime_data × | clustering

Filter

	X	Murder	Assault	UrbanPop	Rape	group
1	Alabama	13.2	236	58	21.2	1
2	Alaska	10.0	263	48	44.5	1
3	Arizona	8.1	294	80	31.0	2
4	Arkansas	8.8	190	50	19.5	3
5	California	9.0	276	91	40.6	2
6	Colorado	7.9	204	78	38.7	2
7	Connecticut	3.3	110	77	11.1	3
8	Delaware	5.9	238	72	15.8	3
9	Florida	15.4	335	80	31.9	2
10	Georgia	17.4	211	60	25.8	1

Showing 1 to 11 of 50 entries, 6 total columns

```
> aggregate(crime_data_final[,2:6],by=list(group),FUN = mean)
  Group.1    Murder  Assault UrbanPop     Rape group
1       1 14.087500 252.7500 53.50000 24.53750     1
2       2 11.054545 264.0909 79.09091 32.61818     2
3       3  5.871429 134.4762 70.76190 18.58095     3
4       4  3.180000  78.7000 49.30000 11.63000     4
```

Conclusion – as per summary we can say group 2 have higher rate of crime

Problem statement 2

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

Answer:

Rcode:

```
library(readxl)
EastWestAirlines <- read_xlsx("E:/data
science/assignments/clustering/EastWestAirlines.xlsx",sheet = "data")
View(EastWestAirlines)
names(EastWestAirlines)
ncol(EastWestAirlines)
attach(EastWestAirlines)

EastWestAirlines_1 <- EastWestAirlines[,2:12]
norm_EastWestAirlines_1 <- scale(EastWestAirlines_1)

#hirerachical clutering

dist_airline <- dist(norm_EastWestAirlines_1,method = "euclidean")
str(dist_airline)
Airline_clust <- hclust(dist_airline,method = "complete")
plot(Airline_clust,hang=-1)
group_Airline <- cutree(Airline_clust,k=5)

EastWestAirlines_2 <- cbind(EastWestAirlines,group_Airline)
View(EastWestAirlines_2)

attach(EastWestAirlines_2)
aggregate(EastWestAirlines_2[,2:12],by=list(group_Airline),FUN = mean)
```

K-MEANS CLustering

```
kmeans_airline <- kmeans(norm_EastWestAirlines_1,5)
str(kmeans_airline)
EastWestAirlines_2 <- cbind(EastWestAirlines_2,kmeans_airline$cluster)
names(EastWestAirlines_2)
View(EastWestAirlines_2)
aggregate(EastWestAirlines_2[,2:12],by=list(kmeans_airline$cluster),FUN = mean)
kmeans_airline$centers

library(cluster)
clusplot(clara(norm_EastWestAirlines_1,5))
clusplot(pam(norm_EastWestAirlines_1,5))
rm(clust1,clust2)
```

Console:

```
> library(readxl)
> EastWestAirlines <- read_xlsx("E:/data science/assignments/clustering/EastW
estAirlines.xlsx",sheet = "data")
> View(EastWestAirlines)
```

R × | crime_data_final × | crime_data × | clustering2.R × | EastWestAirlines_2 × | EastWestAirlines × | Kmeans_clustering+_+finding+k.R × | Hclustering_University.R × | Kme

Filter

	ID#	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award?
1	1	28143	0	1	1	1	174	1	0	0	7000	0
2	2	19244	0	1	1	1	215	2	0	0	6968	0
3	3	41354	0	1	1	1	4123	4	0	0	7034	0
4	4	14776	0	1	1	1	500	1	0	0	6952	0
5	5	97752	0	4	1	1	43300	26	2077	4	6935	1
6	6	16420	0	1	1	1	0	0	0	0	6942	0
7	7	84914	0	3	1	1	27482	25	0	0	6994	0
8	8	20856	0	1	1	1	5250	4	250	1	6938	1
9	9	443003	0	3	2	1	1753	43	3850	12	6948	1
10	10	104860	0	3	1	1	28426	28	1150	3	6931	1
11	11	40091	0	2	1	1	7278	10	0	0	6959	0
12	12	96522	0	5	1	1	61105	19	0	0	6924	1
13	13	43382	0	2	1	1	11150	20	0	0	6924	0
14	14	43097	0	1	1	1	3258	6	0	0	6918	0
15	15	17648	0	1	1	1	0	0	0	0	6912	0
16	16	28495	0	4	1	1	49442	15	0	0	6912	1
17	17	51800	0	4	1	1	48063	16	0	0	6910	1

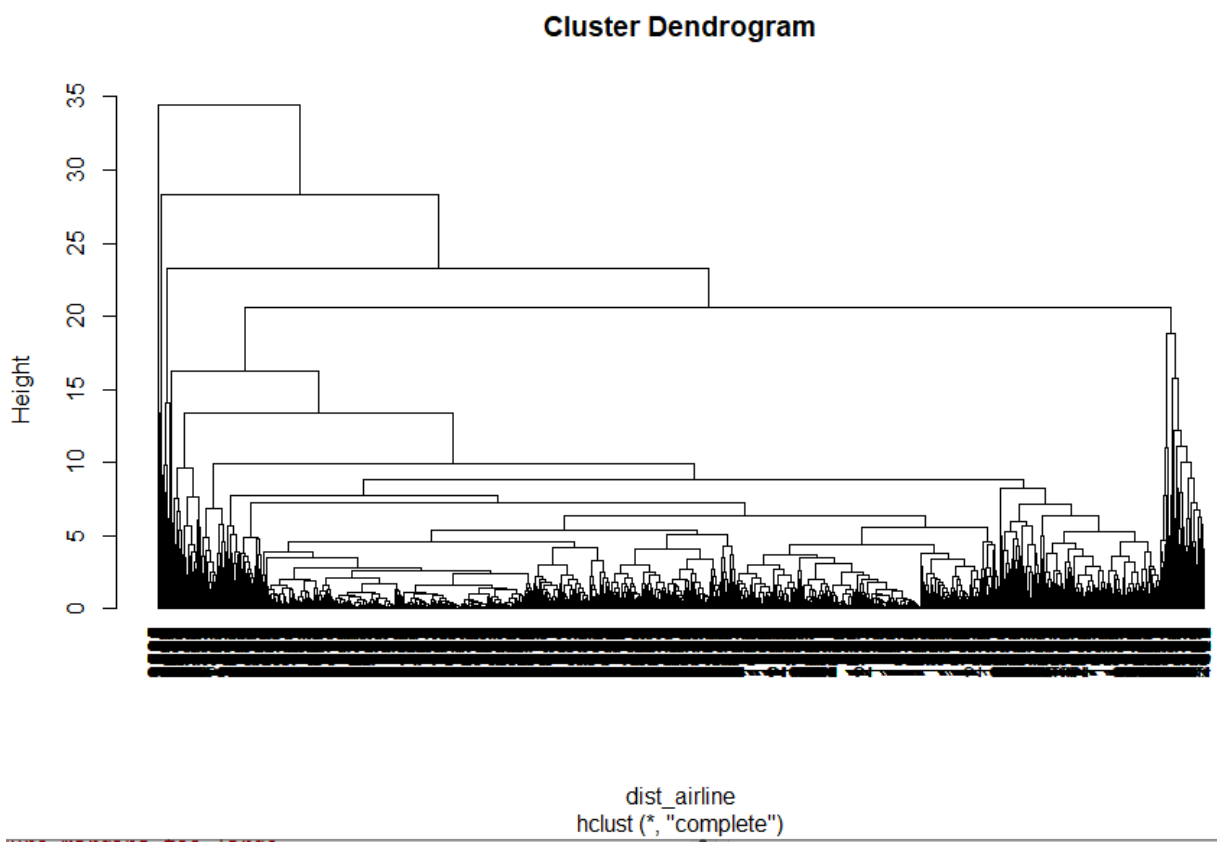
Showing 1 to 17 of 3,999 entries, 12 total columns

```
> names(EastWestAirlines)
 [1] "ID#"                "Balance"            "Qual_miles"         "cc1_miles"
 [5] "cc2_miles"          "cc3_miles"          "Bonus_miles"        "Bonus_trans"
 [9] "Flight_miles_12mo" "Flight_trans_12"   "Days_since_enroll" "Award?"
> ncol(EastWestAirlines)
```

```
[1] 12
> attach(EastWestAirlines)

> EastWestAirlines_1 <- EastWestAirlines[,2:12]
> norm_EastWestAirlines_1 <- scale(EastWestAirlines_1)

> #hirerachical clutering
> dist_airline <- dist(norm_EastWestAirlines_1,method = "euclidean")
> str(dist_airline)
 '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_EastWestAirlines_1, method = "eucl
idean")
> Airline_clust <- hclust(dist_airline,method = "complete")
> plot(Airline_clust,hang=-1)
```




```
> group_Airline <- cutree(Airline_clust,k=5)
> EastWestAirlines_2 <- cbind(EastWestAirlines,group_Airline)
> View(EastWestAirlines_2)
```

.R × | crime_data_final × | crime_data × | clustering2.R × | EastWestAirlines_2 × | EastWest

Filter

Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award?	group_Airline
1	0	0	7000	0	1
2	0	0	6968	0	1
4	0	0	7034	0	1
1	0	0	6952	0	1
26	2077	4	6935	1	1
0	0	0	6942	0	1
25	0	0	6994	0	1
4	250	1	6938	1	1
43	3850	12	6948	1	2
28	1150	3	6931	1	1
10	0	0	6959	0	1
19	0	0	6924	1	1
20	0	0	6924	0	1
6	0	0	6918	0	1
0	0	0	6912	0	1
15	0	0	6912	1	1

```
> attach(EastWestAirlines_2)
> aggregate(EastWestAirlines_2[,2:12],by=list(group_Airline),FUN = mean)
  Group.1   Balance Qual_miles cc1_miles cc2_miles cc3_miles Bonus_miles Bonu
s_trans
1       1  65902.07   137.3707  2.033580  1.000000  1.000793    15571.37     1
0.72448
2       2 117123.66   255.7529  2.252941  1.341176  1.000000    37437.17     2
6.72941
3       3 806433.29   383.2143  3.571429  1.000000  1.000000    58412.32     2
1.21429
4       4 138061.40    78.8000  3.466667  1.000000  4.066667    93927.87     2
8.06667
5       5 131999.50   347.0000  2.500000  1.000000  1.000000    65634.25     6
9.25000
  Flight_miles_12mo Flight_trans_12 Days_since_enroll    Award?
1          270.5854       0.8183501          4072.295 0.3503437
2         4066.6235      11.8823529          4701.688 0.7058824
3         1344.3929       5.6071429          6835.893 0.8571429
4          506.6667       1.6000000          4613.867 0.5333333
5        19960.0000      49.2500000          2200.250 1.0000000
>
> # K-MEANS CLustering
>
> kmeans_airline <- kmeans(norm_EastWestAirlines_1,5)
> str(kmeans_airline)
List of 9
 $ cluster     : int [1:3999] 2 2 2 2 1 2 1 4 5 1 ...
 $ centers     : num [1:5, 1:11] 0.655 -0.138 -0.387 -0.154 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] 10049 3121 4215 4910 4675
 $ tot.withinss: num 26970
 $ betweenss   : num 17008
 $ size        : int [1:5] 837 995 1180 843 144
 $ iter        : int 5
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
> EastWestAirlines_2 <- cbind(EastWestAirlines_2,kmeans_airline$cluster)
> names(EastWestAirlines_2)
 [1] "ID#"                    "Balance"                "Qual_miles"
 [4] "cc1_miles"              "cc2_miles"              "cc3_miles"
 [7] "Bonus_miles"            "Bonus_trans"            "Flight_miles_12mo"
[10] "Flight_trans_12"        "Days_since_enroll"      "Award?"
[13] "group_Airline"          "kmeans_airline$cluster"
> View(EastWestAirlines_2)
```

.R × | crime_data_final × | crime_data × | clustering2.R × | EastWestAirlines_2 × | EastWest »

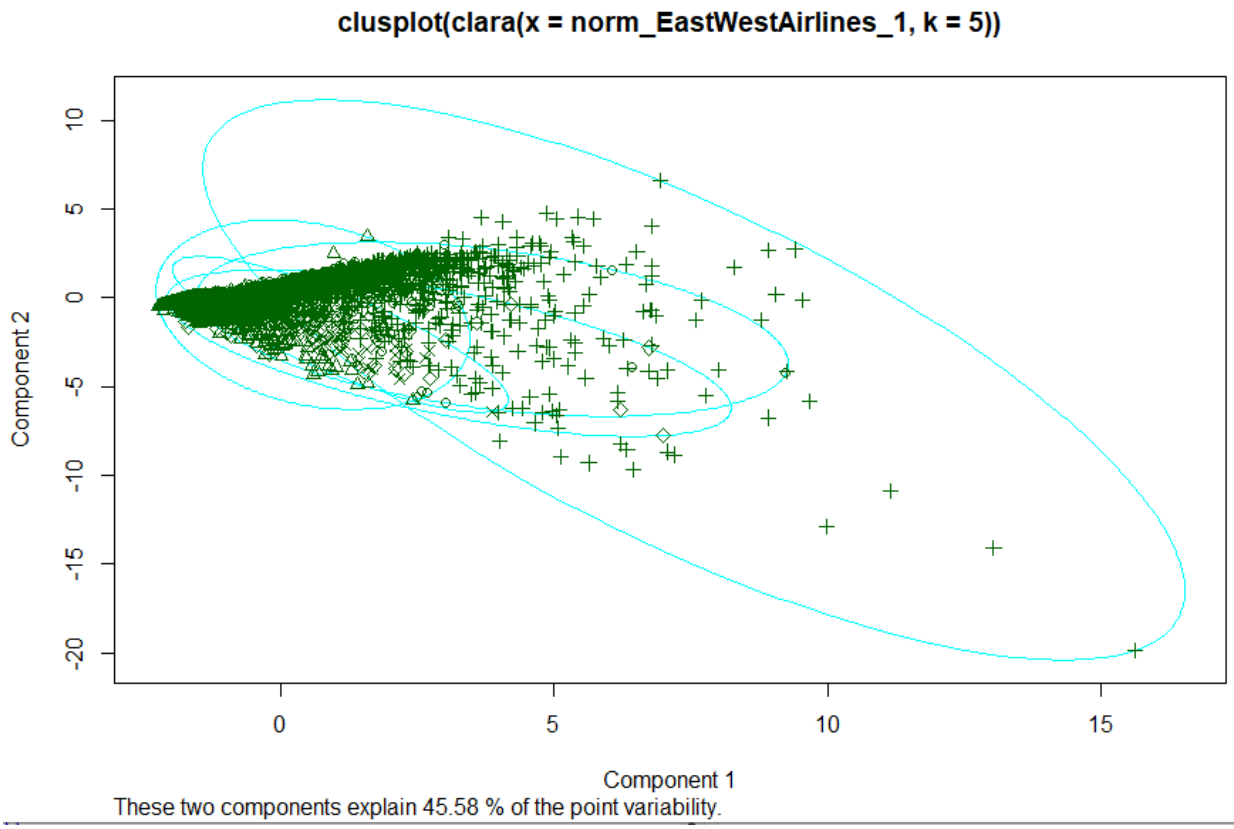
Filter

niles_12mo	Flight_trans_12	Days_since_enroll	Award?	group_Airline	kmeans_airline$cluster
	0	7000	0	1	5
	0	6968	0	1	5
	0	7034	0	1	5
	0	6952	0	1	5
	4	6935	1	1	3
	0	6942	0	1	5
	0	6994	0	1	4
	1	6938	1	1	2
	12	6948	1	2	1
	3	6931	1	1	3
	0	6959	0	1	4
	0	6924	1	1	3
	0	6924	0	1	4
	0	6918	0	1	5
	0	6912	0	1	5
	0	6912	1	1	3

Showing 1 to 16 of 3,999 entries, 14 total columns

```
> aggregate(EastWestAirlines_2[,2:12],by=list(kmeans_airline$cluster),FUN = m
ean)
  Group.1   Balance Qual_miles cc1_miles cc2_miles cc3_miles Bonus_miles Bonu
s_trans
```

```
1       1 139612.51  146.07168  4.197133  1.002389  1.057348   51295.927   20
.113501
2       2  59713.05   70.37688  1.492462  1.011055  1.001005    6299.114    8
.497487
3       3  34590.82   93.48559  1.248305  1.019492  1.000000    4505.752    6
.701695
4       4  58092.93  192.51008  1.709371  1.018980  1.000000   10887.630   10
.786477
5       5 196333.68  773.80556  2.250000  1.041667  1.000000   33783.833   28
.506944
  Flight_miles_12mo Flight_trans_12 Days_since_enroll    Award?
1          385.1565       1.1696535          4943.025 0.6236559
2          177.3337       0.5276382          5625.304 0.0000000
3          141.5805       0.4254237          2089.730 0.0000000
4          415.4199       1.2550415          4270.543 1.0000000
5         5719.9722      16.8680556          4650.562 0.8055556
> kmeans_airline$centers
     Balance   Qual_miles  cc1_miles   cc2_miles   cc3_miles Bonus_miles Bonu
s_trans
1  0.6550310  0.002529724  1.5524640 -0.08204618  0.23096921   1.4140668  0.8
8627324
2 -0.1378138 -0.095309673 -0.4118270 -0.02335485 -0.05761111  -0.4490807 -0.3
2324806
3 -0.3871025 -0.065440486 -0.5891484  0.03378186 -0.06275873  -0.5233370 -0.5
1023556
4 -0.1538903  0.062553727 -0.2542949  0.03031629 -0.06275873  -0.2590876 -0.0
8490626
5  1.2178769  0.813907828  0.1383415  0.18396880 -0.06275873   0.6889574  1.7
6024342
  Flight_miles_12mo Flight_trans_12 Days_since_enroll     Award?
1       -0.05349147     -0.05376499        0.39923099  0.5245051
2       -0.20191419     -0.22302050        0.72961064 -0.7668234
3       -0.22744834     -0.24996746       -0.98242012 -0.7668234
4       -0.03187798     -0.03125402        0.07359516  1.3037551
5        3.75652193      4.08482939        0.25761184  0.9011426
>
> library(cluster)
> clusplot(clara(norm_EastWestAirlines_1,5))
```




```
> clusplot(pam(norm_EastWestAirlines_1,5))
```