Assignment 11.2

1. Use the below given data set

DataSet

2. Perform the below given activities:

a. apply K-means clustering to identify similar recipies

> km$withinss

[1] 1833

> km$tot.withinss

[1] 1833

> km = kmeans(df\_train1,2)

> km$withinss

[1] 600.9150 685.8406

> km$tot.withinss

[1] 1286.756

> km = kmeans(df\_train1,3)

> km$withinss

[1] 346.0231 336.7207 298.1925

> km$tot.withinss

[1] 980.9363

> km = kmeans(df\_train1,4)

> km$withinss

[1] 298.1925 216.7761 169.1202 231.8171

> km$tot.withinss

[1] 915.9058

> km = kmeans(df\_train1,4)

> km$withinss

[1] 90.7904 298.1925 211.4158 301.2184

> km$tot.withinss

[1] 901.6171

>

> km = kmeans(df\_train1,5)

> km$withinss

[1] 94.49965 165.73861 168.75390 309.23224 112.04907

> km$tot.withinss

[1] 850.2735

>

> km = kmeans(df\_train1,6)

> km$withinss

[1] 84.21473 86.00161 153.99935 78.06204 298.19248 94.11996

> km$tot.withinss

[1] 794.5902

>

> km = kmeans(df\_train1,7)

> km$withinss

[1] 69.91732 151.56160 94.11996 153.99935 86.00161 84.21473 94.49965

> km$tot.withinss

[1] 734.3142

>

> km = kmeans(df\_train1,8)

> km$withinss

[1] 120.03869 76.66393 94.11996 118.50099 55.00107 69.91732 78.78610 86.00161

> km$tot.withinss

b. apply K-means clustering to identify similar attributes

> km$withinss

[1] 33.42481 148.18838 43.72160 72.07943 43.90180 57.36509 84.21473 54.41516

[9] 34.73034 71.15537

> km$tot.withinss

[1] 643.1967

> dev.off()

null device

1

> sumsq=NULL

> for (i in 1:15)

+ sumsq[i] = sum(kmeans(df\_train,centers=i,

+ iter.max = 1000,

+ nstart=i,

+ algorithm='Forgy')$withinss)

> plot(1:15,sumsq,type='b', main='Screeplot showing within group sum of squares')

> km = kmeans(df\_train1,3)

> km$withinss

[1] 336.7207 346.0231 298.1925

> km$tot.withinss

[1] 980.9363

> class(km$cluster)

[1] "integer"

> summary(km)

Length Class Mode

cluster 142 -none- numeric

centers 39 -none- numeric

totss 1 -none- numeric

withinss 3 -none- numeric

tot.withinss 1 -none- numeric

betweenss 1 -none- numeric

size 3 -none- numeric

iter 1 -none- numeric

ifault 1 -none- numeric

> km$centers

V2 V3 V4 V5 V6 V7 V8

1 0.7873041 -0.3366416 0.30126292 -0.6685043 0.63578373 0.81778071 0.9257427

2 -1.0688290 -0.3745716 -0.45159755 0.2834020 -0.69107412 0.03476227 0.1176195

3 0.1178079 0.7627367 0.08674128 0.5102454 -0.07309503 -0.99715372 -1.2053248

V9 V10 V11 V12 V13 V14

1 -0.57705540 0.59737684 0.08451446 0.5280108 0.7409713 1.0661743

2 -0.02704718 0.05220622 -0.94949223 0.4882226 0.4014300 -0.8160752

3 0.70608916 -0.75462324 0.82885315 -1.0992526 -1.2652089 -0.4577762

> as.numeric(km$cluster)

[1] 1 2 2 2 3 3 1 1 2 2 2 2 1 3 1 3 3 1 1 1 1 3 1 1 1 2 2 3 3 3 2 1 1 1 1 2 1 3 3 2

[41] 2 2 3 3 3 2 2 2 1 2 1 3 2 3 1 3 2 2 3 2 2 1 3 3 1 3 1 1 1 2 1 1 2 2 2 1 2 1 1 3

[81] 2 1 3 1 2 3 1 3 1 3 3 1 1 2 1 3 3 1 2 1 2 1 1 1 2 1 2 3 1 2 1 3 2 3 2 1 3 1 2 3

[121] 3 1 1 2 3 2 2 1 3 3 3 3 3 1 1 2 3 1 3 3 3 2

c. how many unique recipies that people order often

Answer

> km$withinss

[1] 298.1925 336.7207 346.0231

> km$tot.withinss

[1] 980.9363

d. what are their typical profiles

> d1 <- dist(df\_train,method='euclidean')

> summary(d1)

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.611 128.333 280.418 347.690 525.128 1402.192