**Assignment No. 22.1**

2. Perform the below given activities:

a. apply K-means clustering to identify similar recipies

b. apply K-means clustering to identify similar attributes

> sum(is.na(epi\_r))

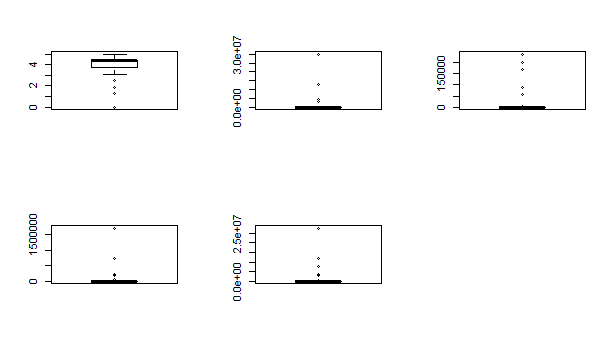
[1] 16581

> library(mice)

> tempData <- mice(epi\_r,m=5,maxit=50,meth='pmm',seed=500)

**> par(mfrow=c(2,3))**

**(boxplot(compepi\_r$rating)$out);(boxplot(compepi\_r$calories)$out);(boxplot(compepi\_r$protein)$out);(boxplot(compepi\_r$fat)$out);(boxplot(compepi\_r$sodium)$out)**



> qntcalories <- quantile(compepi\_r$calories, probs=c(.25, .75))

> qntprotein <- quantile(compepi\_r$protein, probs=c(.25, .75))

> qntfat <- quantile(compepi\_r$fat, probs=c(.25, .75))

> qntsodium <- quantile(compepi\_r$sodium, probs=c(.25, .75))

> Hcalories <- 1.5 \* IQR(compepi\_r$calories)

> Hprotein <- 1.5 \* IQR(compepi\_r$protein)

> Hfat <- 1.5 \* IQR(compepi\_r$fat)

> Hsodium <- 1.5 \* IQR(compepi\_r$sodium)

> outcompepi\_r <- compepi\_r

> qntcalories[1]

25%

201

> qntcalories[2]

75%

591

> qntprotein[1]

25%

3

> qntprotein[2]

75%

28

> qntfat[1]

25%

8

> qntfat[2]

75%

34

> qntsodium[1]

25%

86

> qntsodium[2]

75%

731

> summary(compepi\_r[,c(3,4,5,6)])

calories protein fat sodium

Min. : 0 Min. : 0.00 Min. : 0.0 Min. : 0

1st Qu.: 201 1st Qu.: 3.00 1st Qu.: 8.0 1st Qu.: 86

Median : 337 Median : 8.00 Median : 18.0 Median : 302

Mean : 5137 Mean : 84.62 Mean : 281.9 Mean : 5128

3rd Qu.: 591 3rd Qu.: 28.00 3rd Qu.: 34.0 3rd Qu.: 731

Max. :30111218 Max. :236489.00 Max. :1722763.0 Max. :27675110

> outcompepi\_r$calories[outcompepi\_r$calories < (qntcalories[1] - Hcalories)] <- qntcalories[1]

> outcompepi\_r$protein[outcompepi\_r$protein < (qntprotein[1] - Hprotein)] <- qntprotein[1]

> outcompepi\_r$fat[outcompepi\_r$fat < (qntfat[1] - Hfat)] <- qntfat[1]

> outcompepi\_r$sodium[outcompepi\_r$sodium < (qntsodium[1] - Hsodium)] <- qntsodium[1]

> outcompepi\_r$sodium[outcompepi\_r$sodium > (qntsodium[2] + Hsodium)] <- qntsodium[2]

> outcompepi\_r$fat[outcompepi\_r$fat > (qntfat[2] + Hfat)] <- qntfat[2]

> outcompepi\_r$protein[outcompepi\_r$protein > (qntprotein[2] + Hprotein)] <- qntprotein[2]

> outcompepi\_r$calories[outcompepi\_r$calories > (qntcalories[2] + Hcalories)] <- qntcalories[2]

> summary(outcompepi\_r[,c(3,4,5,6)])

calories protein fat sodium

Min. : 0.0 Min. : 0.00 Min. : 0.00 Min. : 0.0

1st Qu.: 201.0 1st Qu.: 3.00 1st Qu.: 8.00 1st Qu.: 86.0

Median : 337.0 Median : 8.00 Median :18.00 Median : 302.0

Mean : 395.5 Mean :15.25 Mean :21.25 Mean : 420.9

3rd Qu.: 591.0 3rd Qu.:28.00 3rd Qu.:34.00 3rd Qu.: 731.0

Max. :1176.0 Max. :65.00 Max. :73.00 Max. :1698.0

> km1 = kmeans(compdf\_train1,1)

> km2 = kmeans(compdf\_train1,2)

> km3 = kmeans(compdf\_train1,3)

> km4 = kmeans(compdf\_train1,4)

> km5 = kmeans(compdf\_train1,5)

> km6 = kmeans(compdf\_train1,6)

> km7 = kmeans(compdf\_train1,7)

> km8 = kmeans(compdf\_train1,8)

> km9 = kmeans(compdf\_train1,9)

> km10 = kmeans(compdf\_train1,10)

> km11 = kmeans(compdf\_train1,11)

> km12 = kmeans(compdf\_train1,12)

> km13 = kmeans(compdf\_train1,13)

> km14 = kmeans(compdf\_train1,14)

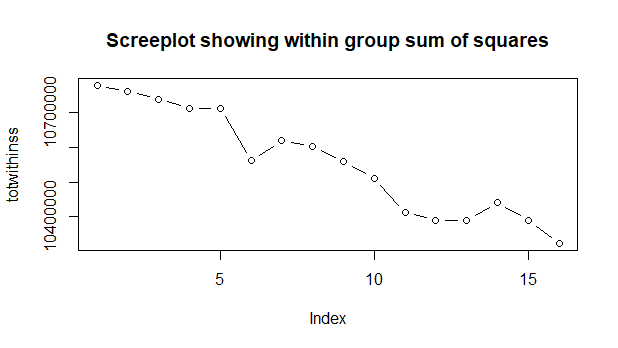
> km15 = kmeans(compdf\_train1,15)

> totwithinss <- c(km1$tot.withinss,km2$tot.withinss,km3$tot.withinss,km4$tot.withinss,km4$tot.withinss,km5$tot.withinss,km6$tot.withinss,km7$tot.withinss,km8$tot.withinss,km9$tot.withinss,km10$tot.withinss,km11$tot.withinss,km12$tot.withinss,km13$tot.withinss,km14$tot.withinss,km15$tot.withinss)

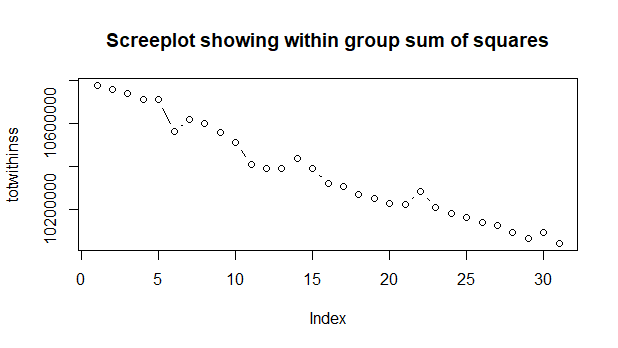
> class(totwithinss)

[1] "numeric"

> plot(totwithinss, type = 'b', main = "Screeplot showing within group sum of squares")



|  |
| --- |
| > totwithinss <- c(km1$tot.withinss,km2$tot.withinss,km3$tot.withinss,km4$tot.withinss,km4$tot.withinss,km5$tot.withinss,km6$tot.withinss,km7$tot.withinss,km8$tot.withinss,km9$tot.withinss,km10$tot.withinss,km11$tot.withinss,km12$tot.withinss,km13$tot.withinss,km14$tot.withinss,km15$tot.withinss,km16$tot.withinss,km17$tot.withinss,km18$tot.withinss,km19$tot.withinss,km20$tot.withinss,km21$tot.withinss,km22$tot.withinss,km23$tot.withinss,km24$tot.withinss,km25$tot.withinss,km26$tot.withinss,km27$tot.withinss,km28$tot.withinss,km29$tot.withinss,km30$tot.withinss)  > plot(totwithinss, type = 'b', main = "Screeplot showing within group sum of squares") |
| |  | | --- | |  | |



c. how many unique recipes that people order often

> kmc2$tot.withinss

[1] 66223.33

> kmc3$tot.withinss

[1] 41371

> kmc4$tot.withinss

[1] 36262.84

> kmc6$tot.withinss

[1] 30056.82

3 group clustering creates 3 clusters with the following sizes, and hence the people should be ordering 3 unique recipies

> kmc3$size

[1] 1602 5601 8838

Further analysis shows that at 6 group clustering, the totwithinss is lower than that of 3 group clustering. So people should be ordering 6 unique recipies

> kmc6$size

[1] 1078 2545 5874 578 3021 2945

d. what are their typical profiles