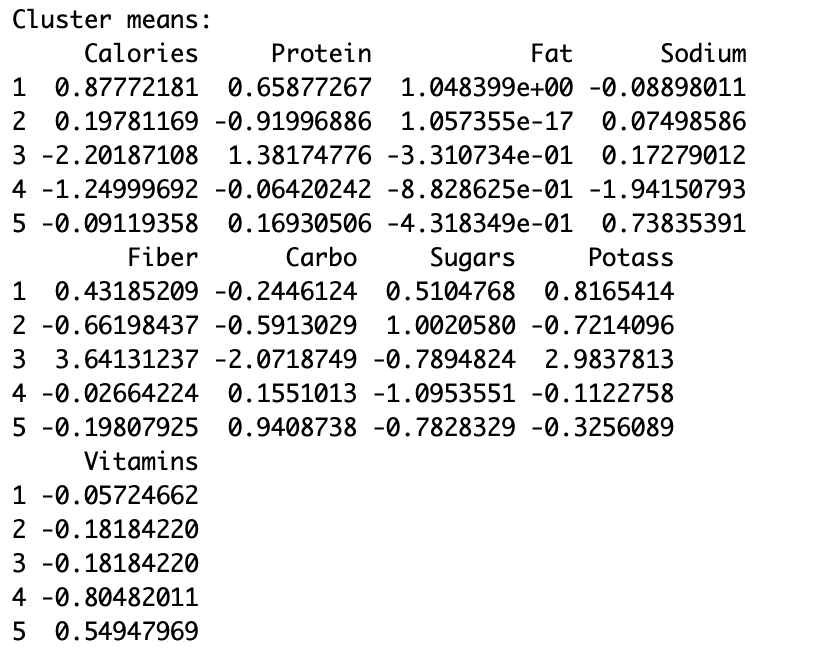
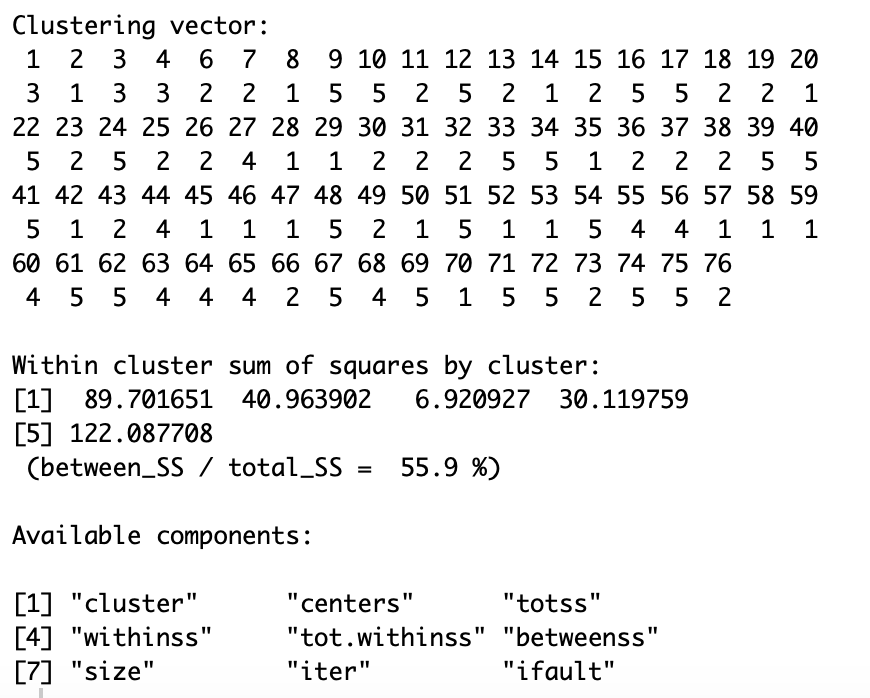
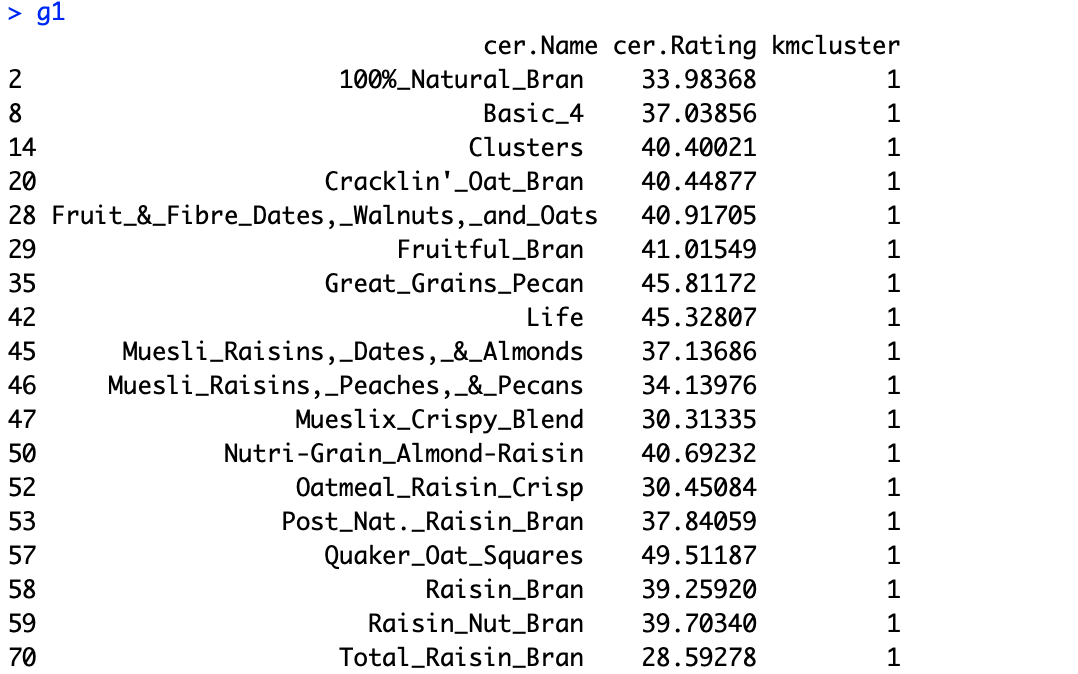
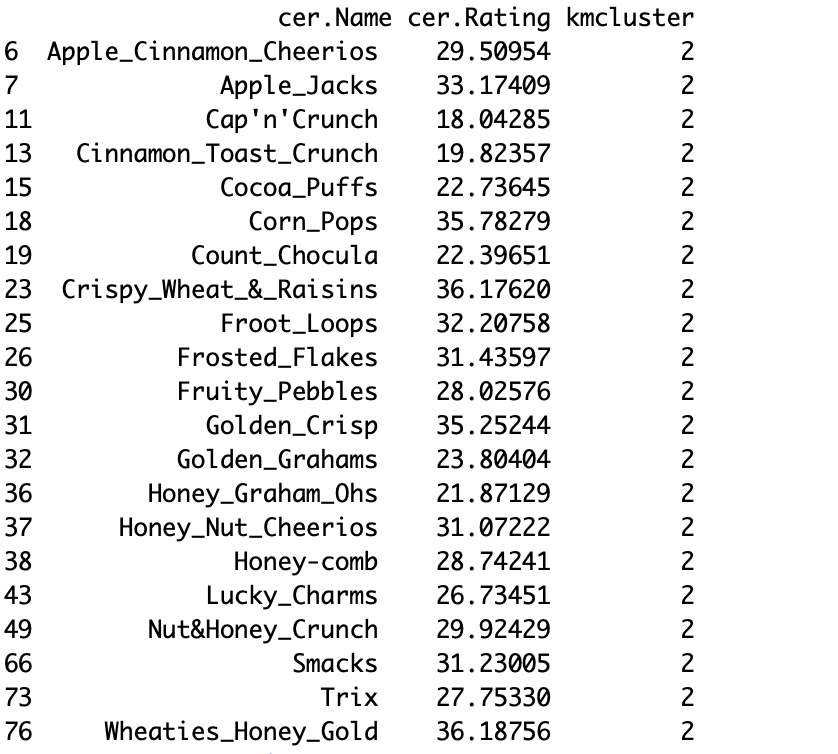
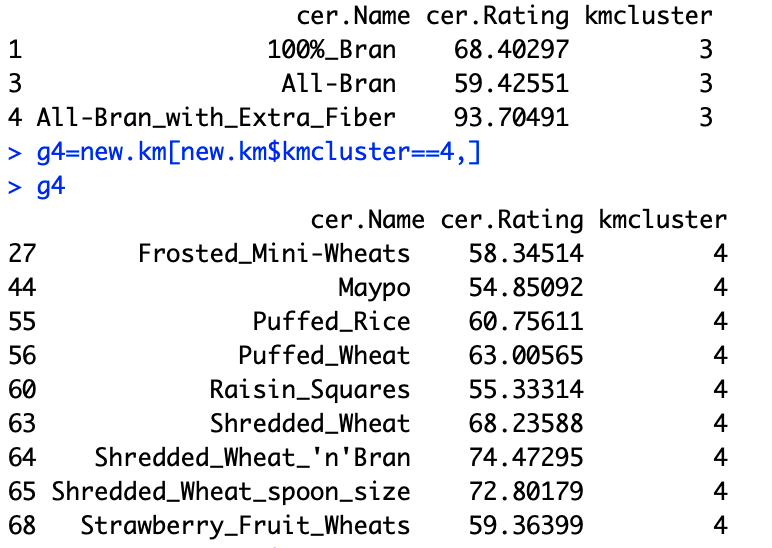
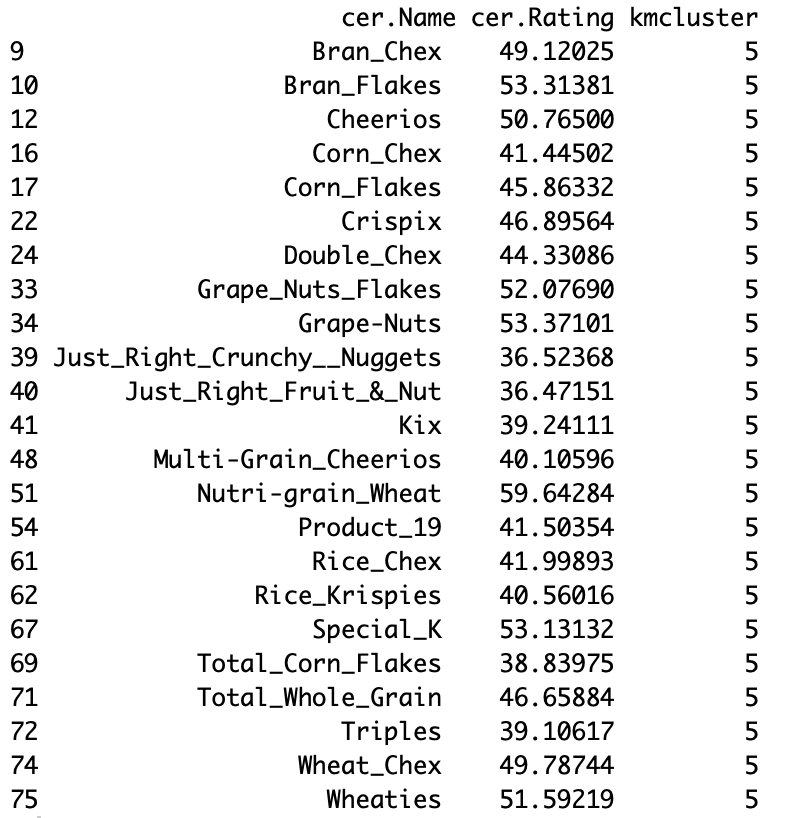
BSAN 450 Assignment 20

1a) Use the continuous variables from the cereals data set except for the variables Name and Rating. Run the k-means algorithm with k = 5 to identify clusters within the data. Note that there are some missing data so these need to be removed before the k-means command can be executed. Also the variables should be standardized so the scaling for the different variables is roughly the same.

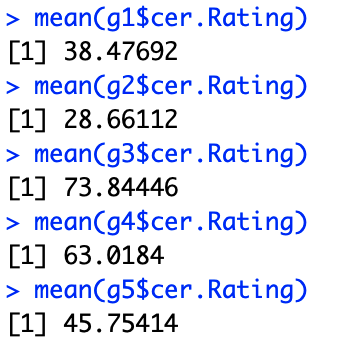
 

b) Develop clustering profiles that clearly describe the characteristics of the cereals within each cluster. R commands to print out the names of the cereals within each cluster follow.

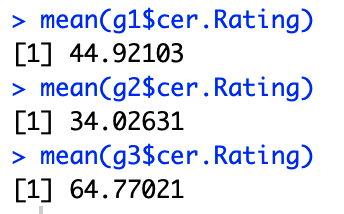
 

c) Suppose you wanted to predict the variable Rating based upon the results of the k-means clustering. One way to do this is to compute the average rating within each cluster. The R commands to compute these means are below. What Rating would you predict for each cluster?



d) Repeat parts a) through c) with k = 3 clusters. Which clustering solution do you prefer? Why?



**I prefer the 5 class solution because it appears that there might be some significant differences in the averages, which indicates noteworthy differences. In the 3 class solution the averages are all kind of close which makes me thing that they could be broken up more.**

2) Using the numeric variables in the cereal data (the same variables as in problem 1), do a hierarchical clustering with complete linkage. Plot the dendrogram. How many groups would you suggest based upon this dendrogram?

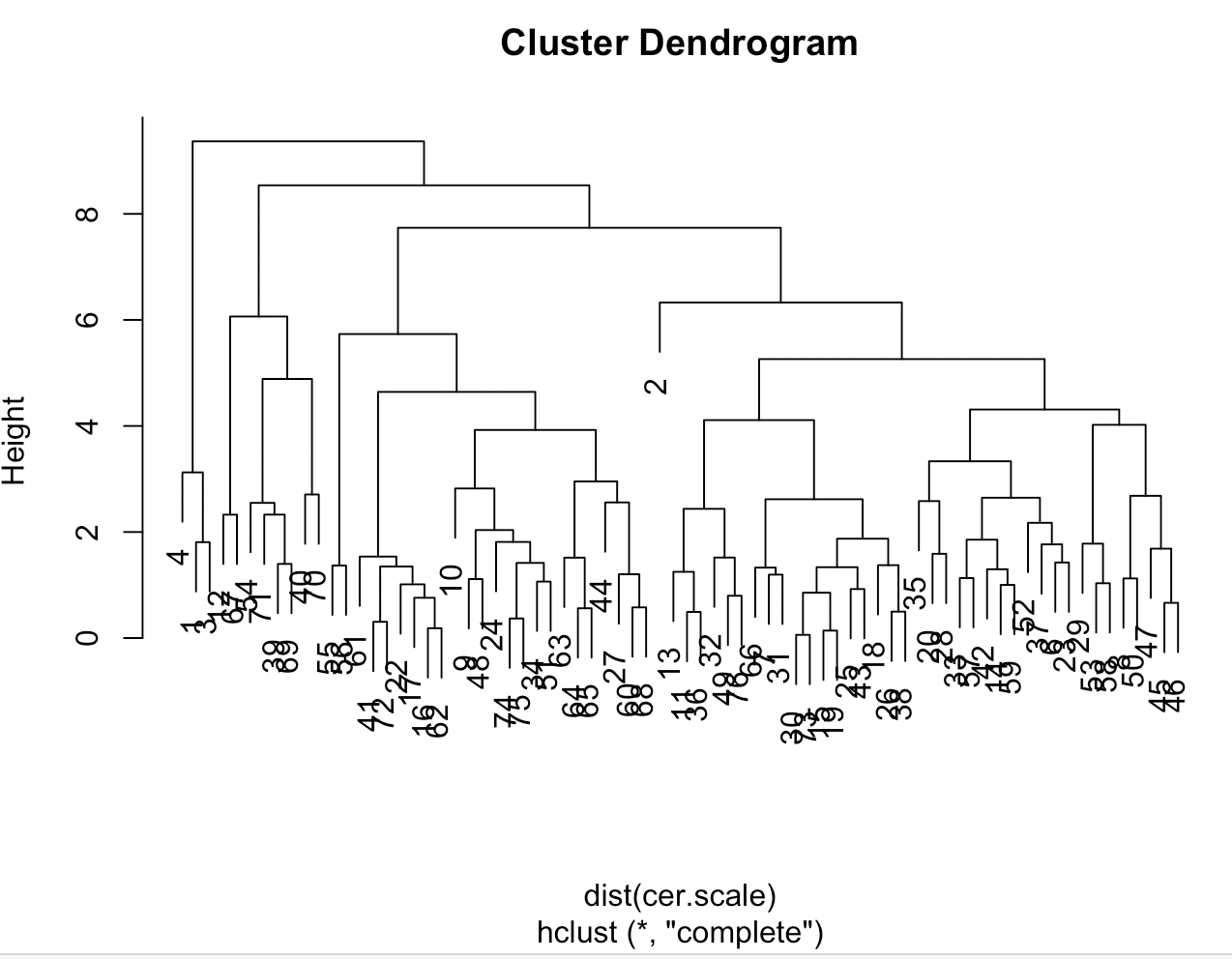
cer=read.csv("cereals.csv")

cer=na.omit(cer)

cer.dat=cer[,-c(1,2,3,13)]

cer.scale=scale(cer.dat)

cer.com=hclust(dist(cer.scale),method="complete")



**I would say 4 classes looks good for this data.**