Big Data Analytics: Assignment 3
Christopher Kwan
Student ID: 250830704

Recommendation

First, it was determined that the appropriate approach to clustering cereals was through complete linkage. Cereals through the complete linkage approach were then segmented based on a cut off Euclidean distance of approximately six, yielding nine distinct clusters. Further examination of these clusters defined them as Bran and High Fibre, Bran, High Sugar Content/Marketed to Children, Nut and Raisin Based, Wheat Based, Healthier, Dieting, Corn and Rice Based, and Puffed.

Methodology

The first step was preparing the data, this consisted of creating dummy variables as well as normalizing all numerical variables. The second step created the clusters through single and complete linkage, the results were then plotted as dendrograms (provided in Exhibits 2 & 3). The third step consisted of determining the cut off Euclidean distance and the resultant number of clusters generated. The fourth step calculated the mean of all variables measured in each cluster and combined them into a new data frame (provided in Exhibit 4). This allowed easier comparison between mean variable values in different clusters. The final step was to determine the meaning of each cluster through evaluating qualitative and quantitative measures.

Analysis

After creating the dendrograms using complete and single linkage (provided in Exhibits 2 & 3), it is obvious that the complete linkage dendrogram has clearer and more segmented clusters compared to the single linkage dendrogram, which has many long thin clusters. This is due to complete linkage clusters being more spherical since it compares maximum distances while single linkage clusters are more jagged since it forms long chains. As such, complete linkage is less sensitive to outliers, this is important in this dataset as cereal products can vary tremendously depending on the value proposition to consumers.

After selecting the complete linkage dendrogram, the Euclidean distance cut off of six was selected. This value was selected since qualitatively it provided an acceptable amount of specificity in clusters while clustering many similar cereals together. The clusters and associated variable means are provided Exhibit 4.

The first cluster generally contained cereals that specify they have bran and are high in fibre, this is confirmed after reviewing the average fibre content of this cluster which is more than four times greater than the next highest fibre content cluster. The second cluster contained a cereal that had bran, but was lower in fibre. The third cluster of cereals mostly contained brands that were marketed to children and therefore were high in calories and contained the most sugar. The fourth cluster contained cereals which had nuts and raisins, this resulted in the highest calorie mean since nuts are extremely high in energy. The fifth cluster contained wheat based cereals which is attributed with higher calories, carbohydrates, potassium, and being relatively light. The sixth cluster contained brands which are perceived as being healthier, these cereals are nutritionally well rounded with a lower sugar content. The seventh cluster are cereals which are branded as being great for diets and have a very low caloric density (calories/cups). The eighth cluster contains corn and rice based cereals which accordingly has the highest mean carbohydrate content. The ninth cluster were puffed cereals, as the name suggests, these cereals are the lightest, the least calorically dense, and provide the least nutritional value.

After classifying the clusters, they were named accordingly as Bran and High Fibre, Bran, High Sugar Content/Marketed to Children, Nut and Raisin Based, Wheat Based, Healthier, Dieting, Corn and Rice Based, and Puffed respectively.

Big Data Analytics: Assignment 3

Christopher Kwan

Student ID: 250830704

Exhibit 1: R Code

```
library(dummies)
cereal.df <-read.csv("/Users/chriskwan/Documents/R/RLabs/Cereals-1.csv")
#Remove rows with NA
cereal.df <-na.omit(cereal.df)
#Name rows
row.names(cereal.df) <- cereal.df[,1]
#Remove names row
cereal.df <- subset(cereal.df, select = -c(name))
#Create dummy variables
cereal.df.dummy <- cbind(cereal.df,dummy(cereal.df$mfr, sep = "_"),dummy(cereal.df$type, sep = "_"))
#normalize values
cereal.df.norm <- sapply(cereal.df.dummy[3:15],scale)
#insert mfr and type back into cereal.df.norm
cereal.df.norm<-cbind(cereal.df.dummy[16:24],cereal.df.norm)
head(cereal.df.norm)
#compute norm distances
d.norm<- dist(cereal.df.norm, method="euclidean")
#clustering
hc1<-hclust(d.norm,method="single")
hc2<-hclust(d.norm,method="complete")
plot(hc1, hang=-1,ann=FALSE)
plot(hc2, hang=-1,ann=FALSE)
clusters2<-data.frame(cutree(hc2,k=9))
#Seperating clusters
colnames(clusters2) <- "cluster"
clusters2<-cbind(clusters2,cereal.df)
clusters21<-subset(clusters2,clusters2$cluster==1)
clusters22<-subset(clusters2,clusters2$cluster==2)
clusters23<-subset(clusters2,clusters2$cluster==3)
clusters24<-subset(clusters2,clusters2$cluster==4)
clusters25<-subset(clusters2,clusters2$cluster==5)
clusters26<-subset(clusters2,clusters2$cluster==6)
clusters27<-subset(clusters2,clusters2$cluster==7)
clusters28<-subset(clusters2,clusters2$cluster==8)
clusters29<-subset(clusters2,clusters2$cluster==9)
clusters21
clusters22
clusters23
clusters24
clusters25
clusters26
clusters27
clusters28
clusters29
#Averaging cluster variable values
x1<-colMeans(clusters21[4:16])
x2<-colMeans(clusters22[4:16])
x3<-colMeans(clusters23[4:16])
x4<-colMeans(clusters24[4:16])
x5<-colMeans(clusters25[4:16])
x6<-colMeans(clusters26[4:16])
x7<-colMeans(clusters27[4:16])
x8<-colMeans(clusters28[4:16])
x9<-colMeans(clusters29[4:16])
#Making a new data frame to hold mean variable cluster values
cluster2mean.df<-data.frame(x1,x2,x3,x4,x5,x6,x7,x8,x9)
colnames(cluster2mean.df)<-c("Bran and Fibre","Bran","Higher Sugar/For Kids","Nut based and Raisin","Wheat
Based","Healthier","Diet","Corn and Rice Based","Puffed")
cluster2mean.df
```

Big Data Analytics: Assignment 3
Christopher Kwan
Student ID: 250830704

Exhibit 2: Compete Linkage

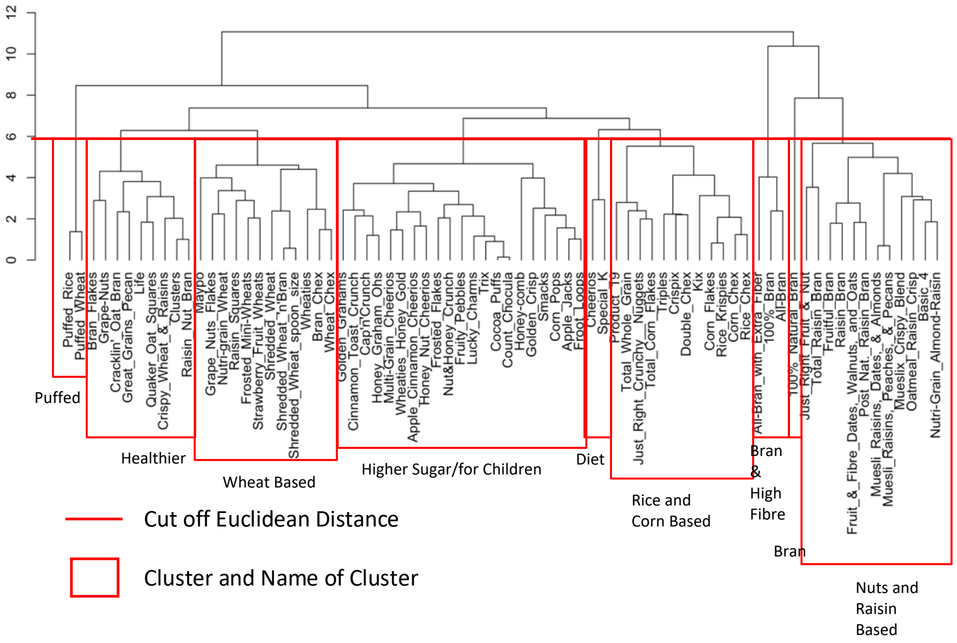

Exhibit 3: Single Linkage

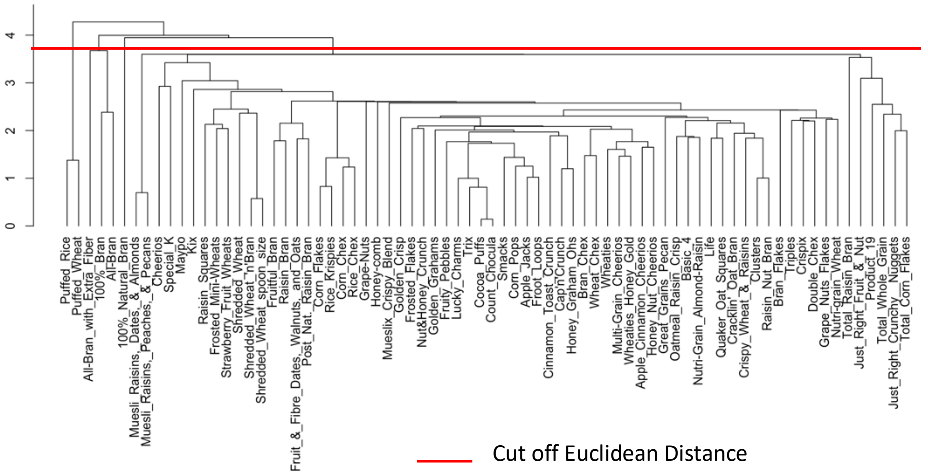

Exhibit 4: Mean Variable Value Comparison for Complete Clustering

	Bran and Fibre	Bran	Higher Sugar/For Children	Nut based and Raisin	Wheat Based	Healthier	Diet	Corn and Rice Based	Puffed
calories	63.3333333	120.00000	110.9523810	135.0000000	93.3333333	104.4444444	110.00000	106.6666667	50.00000
protein	4.0000000	3.00000	1.5238095	3.1666667	2.7500000	3.1111111	6.00000	2.0833333	1.50000
fat	0.6666667	5.00000	1.0000000	1.6666667	0.4166667	1.5555556	1.00000	0.4166667	0.00000
sodium	176.6666667	15.00000	172.3809524	180.4166667	79.5833333	144.4444444	260.00000	242.5000000	0.00000
fiber	11.0000000	2.00000	0.5714286	3.5416667	2.8333333	2.8333333	1.50000	0.6666667	0.50000
carbo	6.6666667	8.00000	12.6190476	15.6250000	16.4166667	12.6111111	16.50000	20.2500000	11.50000
sugars	3.6666667	8.00000	11.2857143	10.9166667	3.3333333	6.2222222	2.00000	3.2500000	0.00000
potass	310.0000000	135.00000	45.9523810	172.0833333	106.2500000	123.3333333	80.00000	48.7500000	32.50000
vitamins	25.0000000	0.00000	25.0000000	37.5000000	18.7500000	25.0000000	25.00000	50.0000000	0.00000
shelf	3.0000000	3.00000	1.6666667	2.9166667	1.7500000	2.8888889	1.00000	2.2500000	3.00000
weight	1.0000000	1.00000	1.0000000	1.2875000	0.9858333	1.0000000	1.00000	1.0000000	0.50000
cups	0.3866667	1.00000	0.8871429	0.7583333	0.8216667	0.5188889	1.12500	1.0108333	1.00000
rating	73.8444633	33.98368	28.8482485	36.1556942	58.8019532	44.8961178	51.94816	41.9139197	61.88088