Assignment 5 - Clustering

- Creating the dendrogram and clusters
- The elementary public schools would like to choose a set of cereals to include in their daily cafeterias. Every day a different cereal is offered, but all cereals should support a healthy diet. For this goal, you are requested to find a cluster of “healthy cereals.” Should the data be normalized? If not, how should they be used in the cluster analysis?
- How do you compare hierarchical clustering and k-means? What are they main advantages of hierarchical clustering compared to k-means?

First, we’ll clean and scale our data

```
CerealData <- read.csv("/Users/tpagliar/Downloads/Cereals.csv")
CerealData <- na.omit(CerealData) #remove missing data
CerealDataScaled <- scale(CerealData[,4:16]) #Scale the numeric data
head(CerealDataScaled)
```

```
##     calories    protein        fat     sodium       fiber      carbo     sugars
## 1 -1.8659155  1.3817478  0.0000000 -0.3910227  3.22866747 -2.5001396 -0.2542051
## 2  0.6537514  0.4522084  3.9728810 -1.7804186 -0.07249167 -1.7292632  0.2046041
## 3 -1.8659155  1.3817478  0.0000000  1.1795987  2.81602258 -1.9862220 -0.4836096
## 4 -2.8737823  1.3817478 -0.9932203 -0.2702057  4.87924705 -1.7292632 -1.6306324
## 6  0.1498180 -0.4773310  0.9932203  0.2130625 -0.27881412 -1.0868662  0.6634132
## 7  0.1498180 -0.4773310 -0.9932203 -0.4514312 -0.48513656 -0.9583868  1.5810314
##       potass   vitamins      shelf     weight       cups     rating
## 1  2.5605229 -0.1818422  0.9419715 -0.2008324 -2.0856582  1.8549038
## 2  0.5147738 -1.3032024  0.9419715 -0.2008324  0.7567534 -0.5977113
## 3  3.1248675 -0.1818422  0.9419715 -0.2008324 -2.0856582  1.2151965
## 4  3.2659536 -0.1818422  0.9419715 -0.2008324 -1.3644493  3.6578436
## 6 -0.4022862 -0.1818422 -1.4616799 -0.2008324 -0.3038480 -0.9165248
## 7 -0.9666308 -0.1818422 -0.2598542 -0.2008324  0.7567534 -0.6553998
```

Next we perform some observations using agnes:

```
hc_single <- agnes(CerealDataScaled, method = "single")
hc_complete <- agnes(CerealDataScaled, method = "complete")
hc_average <- agnes(CerealDataScaled, method = "average")
hc_ward <- agnes(CerealDataScaled, method = "ward")

print(hc_single$ac)
```

```
## [1] 0.6067859
```

```
print(hc_complete$ac)
```

```
## [1] 0.8353712
```

```
print(hc_average$ac)
```

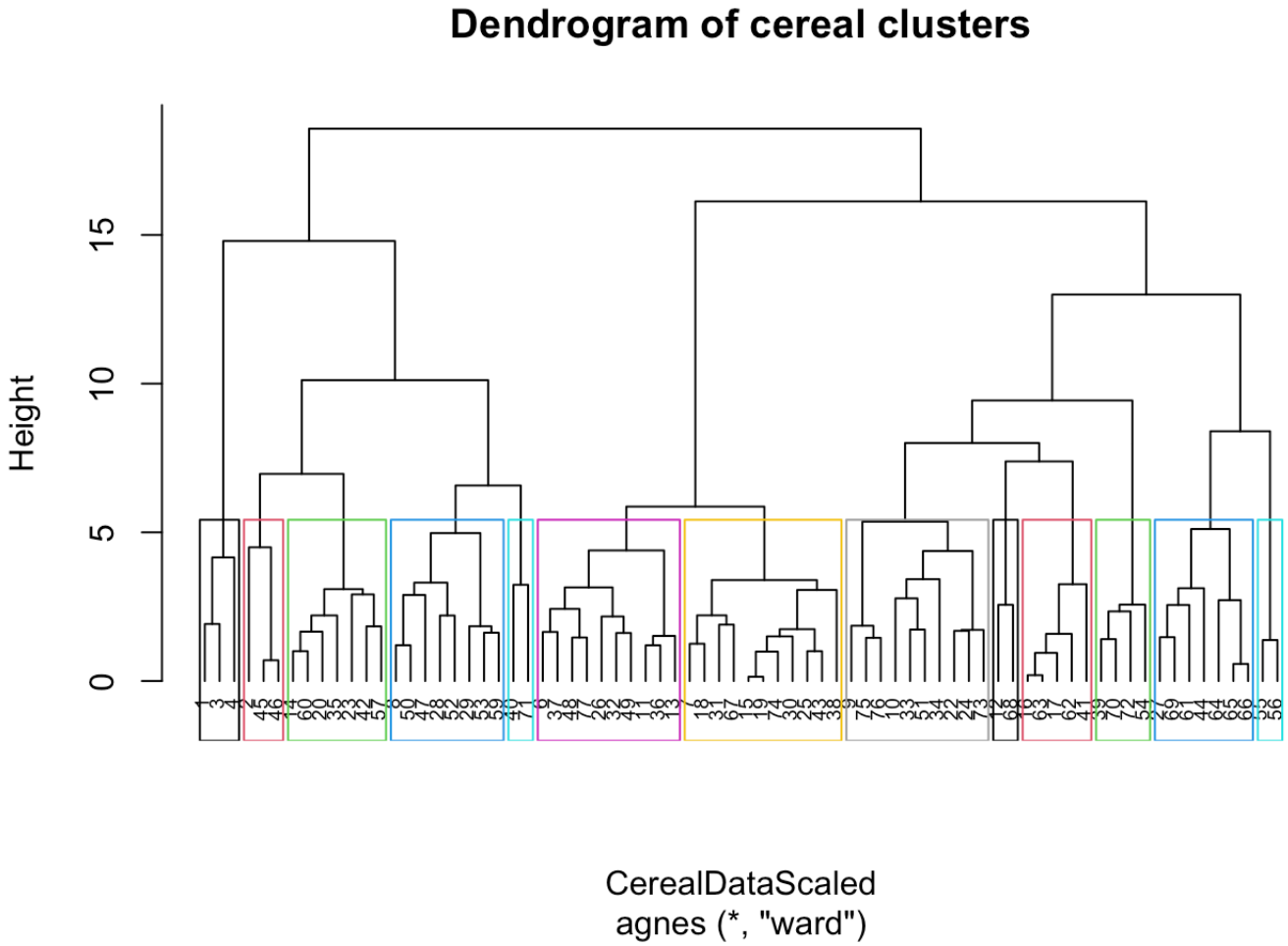
```
## [1] 0.7766075
```

```
print(hc_ward$ac)
```

```
## [1] 0.9046042
```

We see that **Ward's method has the best structure at above .90**, we'll use his method moving forward.

Creating the dendrogram and clusters

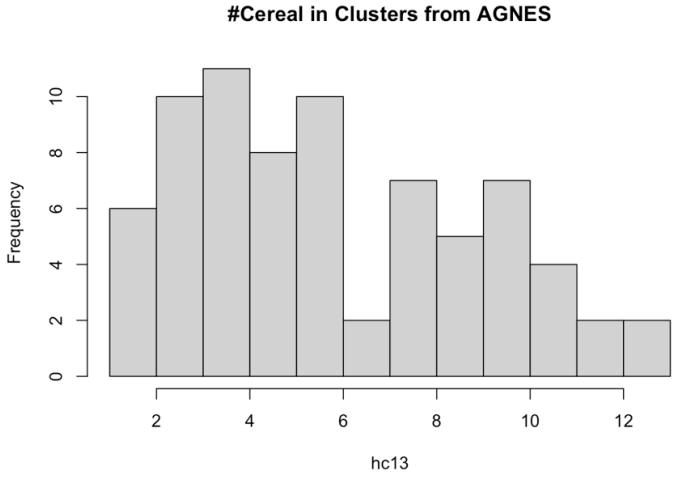


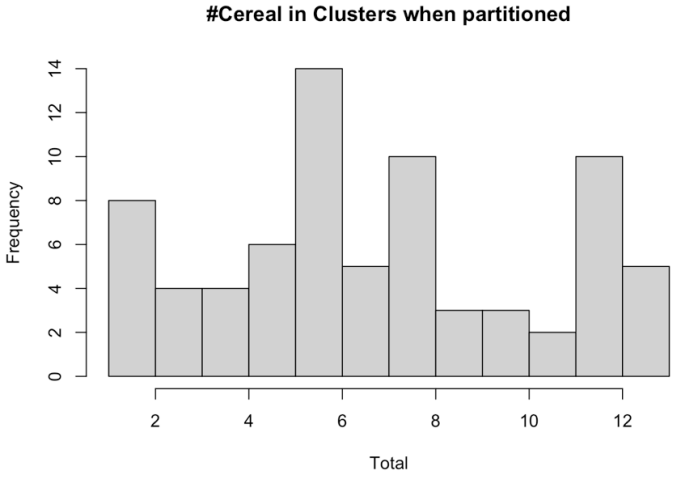
The algorithm has sorted our cereals, based on the level of plateaus and the output of the NbClust analysis, I would choose 13 clusters. Being able to choose the ideal K by observing the dendrogram is a benefit of hierarchical clustering over kmeans.

Now we'll try to check stability by partitionning our data, clustering half and then assigning clusters to the other half. Then, we'll compare these cluster counts to the results when the data was clustered in whole.

```
A_Index <- sample(row.names(CerealDataScaled), .5*dim(CerealDataScaled)[1])
B_Index <- setdiff(row.names(CerealDataScaled), A_Index)
Adata <- CerealDataScaled[A_Index,]
Bdata <- CerealDataScaled[B_Index,]
A.agnes<-agnes(Adata,method="ward")
A13<-cutree(A.agnes, k=13) ##put A into 13 clusters

##get centroids
clust.centroid = function(i, dat, A13) {
    ind = (A13 == i)
    colMeans(dat[ind,])
    }
centroids <- sapply(unique(A13), clust.centroid, CerealDataScaled, A13)
B13 <- get.knnx(centroids,Bdata,1)$nn.index[,1]
Total <-append(A13,B13)
hist(hc13,breaks=13,main="#Cereal in Clusters from AGNES")
hist(Total,breaks=13,main="#Cereal in Clusters when partitioned")
```





We see that the clustering looks quite different when we use only half of the data at a time.

The elementary public schools would like to choose a set of cereals to include in their daily cafeterias. Every day a different cereal is offered, but all cereals should support a healthy diet. For this goal, you are requested to find a cluster of “healthy cereals.” Should the data be normalized? If not, how should they be used in the cluster analysis?

If we are to select a cluster of ‘healthy cereals’, we should take the school’s definition of ‘healthy’ (high protein, high fiber, high vitamin, low fat, etc) and weight the scale of these variables to allow a more specific differentiation by those areas.

How do you compare hierarchical clustering and k-means? What are they main advantages of hierarchical clustering compared to k-means?

Both forms of clustering help to categorize data, but kmeans requires more knowledge of the dataset to select an appropriate number of clusters before performing the analysis while hierarchical clustering can be on a sliding scale of the number of clusters based on divisions visible on the dendrogram. Hierarchical clustering is also more computationally resource heavy, but is more flexible in its use. Some advantages of hierarchical clustering are that it does not rely on random starting centroids, and can be initiated as as divisive or agglomerative (as opposed to a random cluster start by kmeans).