Assignment 5 Clustering

Franklin Ngochi

2023-04-09

#Libraries

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stats)
library(cluster)
library(fastDummies)
library(factoextra)
```

```
## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

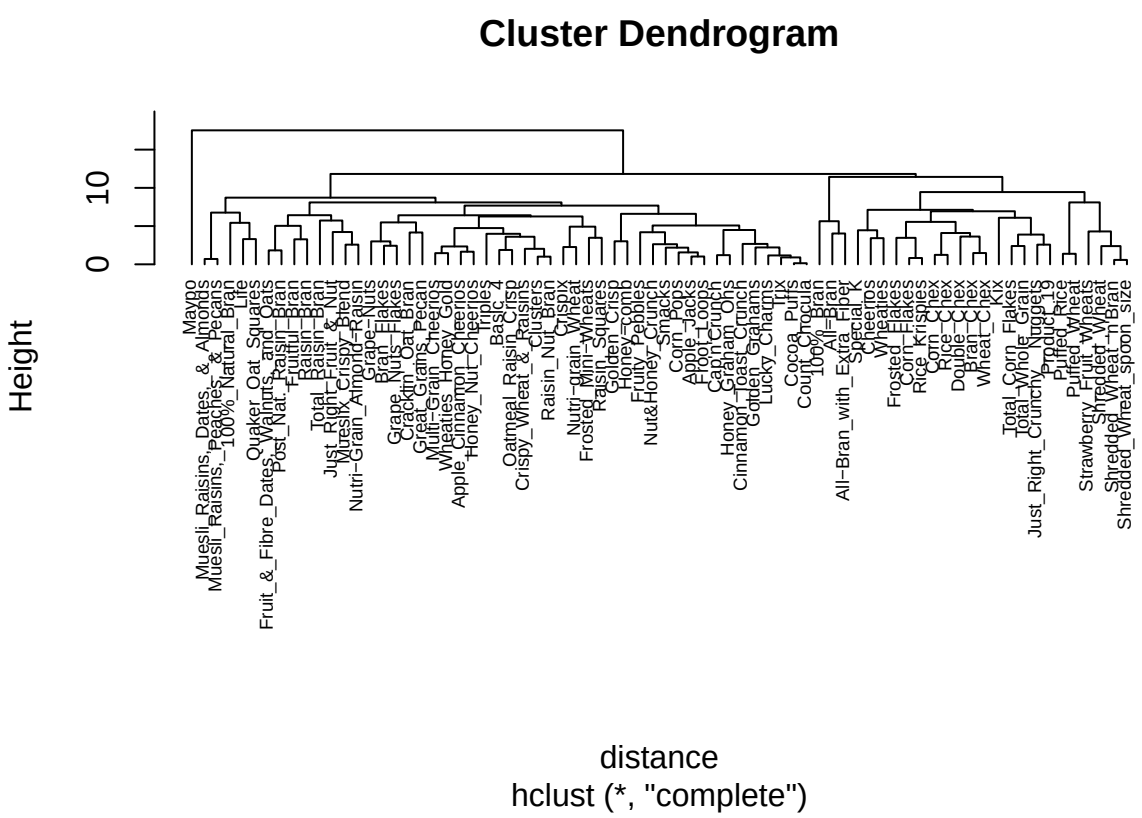
#Data Preprocessing

```
Cereals<-read.csv("C:\\Users\\ngoch\\OneDrive\\Documents\\KSU\\Fundamentals of Machine Learning\\Cereals
#removing missing values
Clean_Cereal<-na.omit(Cereals)
#Transforming categorical variables to dummies
dummy_cereals<-fastDummies::dummy_cols(Clean_Cereal, select_columns = c("mfr", "type", "shelf"), remove_
#Scaling dataset to normal
Scaled_cereals<- dummy_cereals%>%mutate(across(where(is.numeric), scale))
#Assigning row names for greater lisibility of dendrogram
rownames(Scaled_cereals)<-Scaled_cereals$name
Cereals_Data<-Scaled_cereals[, -c(colnames(Scaled_cereals)%in%("name"))]
head(Cereals_Data)
```

```
##                             calories    protein        fat     sodium
## 100%_Bran                 -1.8659155  1.3817478  0.0000000 -0.3910227
## 100%_Natural_Bran          0.6537514  0.4522084  3.9728810 -1.7804186
## All-Bran                  -1.8659155  1.3817478  0.0000000  1.1795987
## All-Bran_with_Extra_Fiber -2.8737823  1.3817478 -0.9932203 -0.2702057
## Apple_Cinnamon_Cheerios    0.1498180 -0.4773310  0.9932203  0.2130625
## Apple_Jacks                0.1498180 -0.4773310 -0.9932203 -0.4514312
##                                 fiber      carbo     sugars     potass
## 100%_Bran                  3.22866747 -2.5001396 -0.2542051  2.5605229
## 100%_Natural_Bran         -0.07249167 -1.7292632  0.2046041  0.5147738
## All-Bran                   2.81602258 -1.9862220 -0.4836096  3.1248675
## All-Bran_with_Extra_Fiber  4.87924705 -1.7292632 -1.6306324  3.2659536
## Apple_Cinnamon_Cheerios   -0.27881412 -1.0868662  0.6634132 -0.4022862
## Apple_Jacks               -0.48513656 -0.9583868  1.5810314 -0.9666308
##                             vitamins     weight       cups     rating
## 100%_Bran                 -0.1818422 -0.2008324 -2.0856582  1.8549038
## 100%_Natural_Bran         -1.3032024 -0.2008324  0.7567534 -0.5977113
## All-Bran                  -0.1818422 -0.2008324 -2.0856582  1.2151965
## All-Bran_with_Extra_Fiber -0.1818422 -0.2008324 -1.3644493  3.6578436
## Apple_Cinnamon_Cheerios   -0.1818422 -0.2008324 -0.3038480 -0.9165248
## Apple_Jacks               -0.1818422 -0.2008324  0.7567534 -0.6553998
##                                mfr_A      mfr_G      mfr_K      mfr_N
## 100%_Bran                 -0.1162476 -0.6460338 -0.6669978  3.6896495
## 100%_Natural_Bran         -0.1162476 -0.6460338 -0.6669978 -0.2673659
## All-Bran                  -0.1162476 -0.6460338  1.4789951 -0.2673659
## All-Bran_with_Extra_Fiber -0.1162476 -0.6460338  1.4789951 -0.2673659
## Apple_Cinnamon_Cheerios   -0.1162476  1.5269890 -0.6669978 -0.2673659
## Apple_Jacks               -0.1162476 -0.6460338  1.4789951 -0.2673659
##                                mfr_P      mfr_Q      mfr_R    type_C     type_H
## 100%_Bran                 -0.3695814 -0.3210386 -0.3210386 0.1162476 -0.1162476
## 100%_Natural_Bran         -0.3695814  3.0727976 -0.3210386 0.1162476 -0.1162476
## All-Bran                  -0.3695814 -0.3210386 -0.3210386 0.1162476 -0.1162476
## All-Bran_with_Extra_Fiber -0.3695814 -0.3210386 -0.3210386 0.1162476 -0.1162476
## Apple_Cinnamon_Cheerios   -0.3695814 -0.3210386 -0.3210386 0.1162476 -0.1162476
## Apple_Jacks               -0.3695814 -0.3210386 -0.3210386 0.1162476 -0.1162476
##                              shelf_1    shelf_2    shelf_3
## 100%_Bran                 -0.583769 -0.6044546  1.0484407
## 100%_Natural_Bran         -0.583769 -0.6044546  1.0484407
## All-Bran                  -0.583769 -0.6044546  1.0484407
## All-Bran_with_Extra_Fiber -0.583769 -0.6044546  1.0484407
## Apple_Cinnamon_Cheerios    1.689858 -0.6044546 -0.9409083
## Apple_Jacks               -0.583769  1.6320274 -0.9409083
```

#Applying Hierarchical Clustering using Euclidean distance

```
distance<-dist(Cereals_Data, method="euclidean")#dissimilary matrix
hc1<-hclust(distance, method="complete")
plot(hc1, cex=0.6, hang=-1)
```



#Using AGNES to compare clustering for different methods

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

#Compare Agglomerative coefficients to select best method

```
hc_single$ac
```

```
## [1] 0.8403456
```

```
hc_complete$ac
```

```
## [1] 0.8488671
```

```
hc_average$ac
```

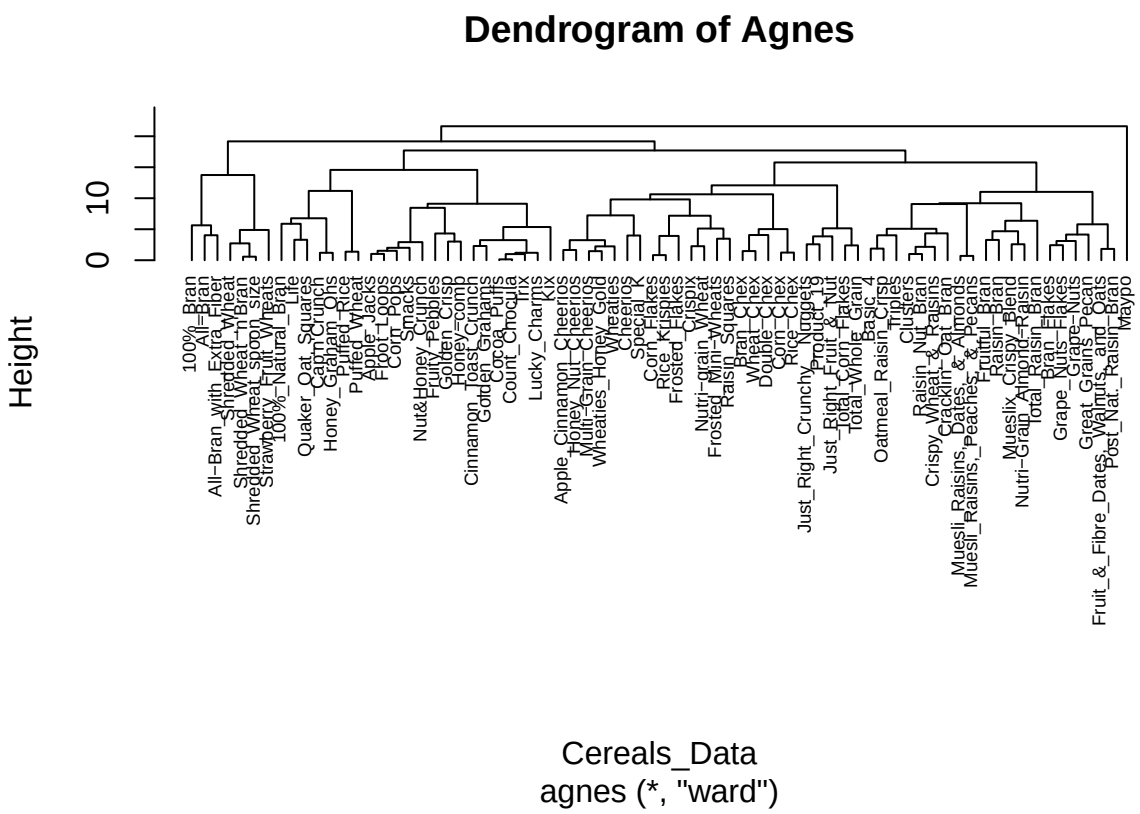
```
## [1] 0.8403212
```

```
hc_ward$ac
```

```
## [1] 0.8741867
```

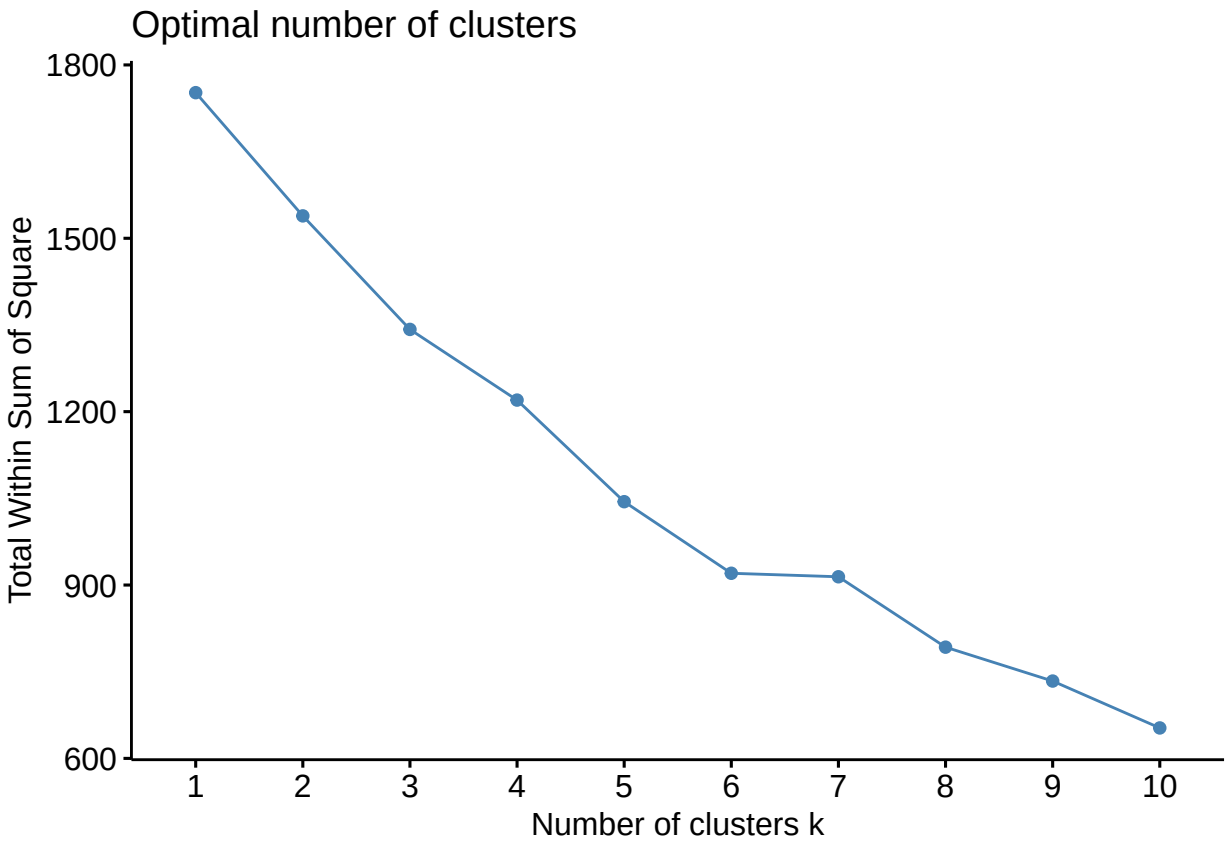
#Ward's method has the highest agglomerative coefficient. This makes it the best method

```
pltree(hc_ward, cex=0.6, hang=-1, main="Dendrogram of Agnes")
```



#The optimal number of clusters

```
fviz_nbclust(Cereals_Data, kmeans, method='wss')
```




```
#The optimal number of clusters is 6 because that is where the graph shows an elbow
```

#Structure and stability of clusters: Clustering partition A

```
Cereals_A<-Cereals_Data[1:55,]
Cereals_B<-Cereals_Data[56:74,]
hc_A<-agnes(Cereals_A, method="ward")
groups_A<-cutree(hc_A, k=6)
```

#Assign each record in partition B to the closest centroid in partition A

```
centroids_A<-aggregate(Cereals_A, by=list(groups_A), mean)[,-1]
centroids_A
```

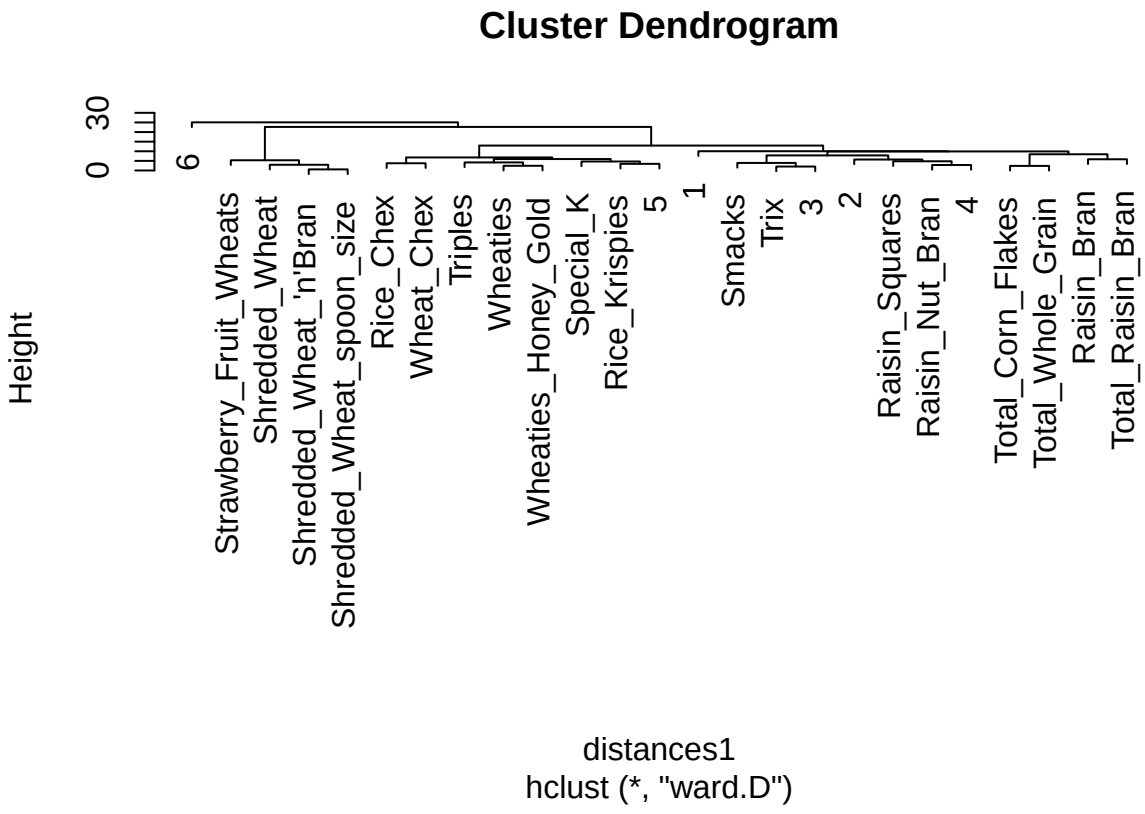
```
##     calories     protein         fat      sodium      fiber      carbo
## 1 -2.2018711  1.38174776 -0.3310734  0.17279012  3.6413124 -2.0718749
## 2 -0.6420773 -0.21174836  0.7094430 -0.68443548 -0.4261873 -0.8115532
## 3  0.1218217 -0.52897212 -0.0551789  0.13251778 -0.5997601 -0.3445408
## 4  0.5817609  0.38581269  0.4966101  0.07930075  0.5022637 -0.1416250
## 5  0.3177958 -0.01256134 -0.2483051  0.51007100 -0.1756529  0.8831512
## 6 -0.3541153  1.38174776  0.0000000 -1.96164410 -0.8977815  0.3264071
##       sugars      potass   vitamins      weight       cups      rating
## 1 -0.7894824  2.98378133 -0.1818422 -0.20083243 -1.8452553  2.24264794
## 2 -0.2214330 -0.41236375 -0.6624252 -1.13201038  0.1022108 -0.04176658
```

```
## 3  0.6506685 -0.62175351 -0.1818422 -0.20083243  0.5634883 -0.69296910
## 4  0.2537622  0.63570476 -0.1818422  0.87933398 -0.9159664 -0.06788856
## 5 -0.2542051 -0.27295721  0.6591780 -0.03787629  0.3749369 -0.02346833
## 6 -0.9424187 -0.04957081 -0.1818422 -0.20083243  0.7567534  0.88922515
##        mfr_A       mfr_G       mfr_K      mfr_N      mfr_P      mfr_Q
## 1 -0.1162476 -0.64603380  0.76366415  1.0516392 -0.3695814 -0.3210386
## 2 -0.1162476 -0.64603380 -0.66699780 -0.2673659 -0.3695814  3.0727976
## 3 -0.1162476  0.56120108 -0.07088865 -0.2673659  0.1368820 -0.3210386
## 4 -0.1162476 -0.02517015 -0.05385697 -0.2673659  0.9327531 -0.3210386
## 5 -0.1162476 -0.64603380  0.58483140 -0.2673659 -0.3695814 -0.3210386
## 6  8.4860776 -0.64603380 -0.66699780 -0.2673659 -0.3695814 -0.3210386
##        mfr_R     type_C     type_H    shelf_1    shelf_2    shelf_3
## 1 -0.3210386  0.1162476 -0.1162476 -0.5837690 -0.6044546  1.0484407
## 2 -0.3210386  0.1162476 -0.1162476 -0.5837690  0.3540377  0.1958625
## 3 -0.3210386  0.1162476 -0.1162476  0.1741065  0.8865334 -0.9409083
## 4 -0.3210386  0.1162476 -0.1162476 -0.5837690 -0.6044546  1.0484407
## 5  1.0930598  0.1162476 -0.1162476  0.1741065 -0.6044546  0.3853243
## 6 -0.3210386 -8.4860776  8.4860776 -0.5837690  1.6320274 -0.9409083
```

```
distances1<-dist(rbind(Cereals_B, centroids_A))
hcb<-hclust(distances1, method="ward")
```

```
## The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
plot(hcb)
```



#A visual comparison of the dendrograms show that the cluster assignments are consistent compared to the assignment based on all the data.

#Choice of healthy cereal

```
df_opt<-cutree(hc_ward, k=6)
centroids_Cereals<-aggregate(Cereals_Data, by=list(df_opt), mean)
centroids_Cereals#Cluster 1 contains healthy cereals with low calories, low fat and high fibers and pro
```

```
##   Group.1    calories     protein        fat      sodium      fiber       carbo
## 1       1 -1.50596308  0.58499970 -0.7094430 -1.02099721  1.8138850 -0.48117765
## 2       2 -0.64207728 -0.21174836  0.7094430 -0.68443548 -0.4261873 -0.81155324
## 3       3 -0.08115143  0.06490027 -0.4138418  0.52517313 -0.2444270  0.67972544
## 4       4  0.18341359 -0.97308540 -0.1324294 -0.03662606 -0.7877428 -0.44446926
## 5       5  0.72934142  0.45220836  0.5959322  0.09224544  0.4433144 -0.04618313
## 6       6 -0.35411535  1.38174776  0.0000000 -1.96164410 -0.8977815  0.32640711
##       sugars      potass    vitamins     weight       cups      rating
## 1 -1.1062791  1.38144582 -0.66242516 -0.3591327 -0.7583914  2.03393380
## 2 -0.2214330 -0.41236375 -0.66242516 -1.1320104  0.1022108 -0.04176658
## 3 -0.5027267 -0.37583252  0.51900795 -0.1193544  0.3183715  0.11934643
## 4  0.9998732 -0.83495035 -0.18184220 -0.2008324  0.5814006 -0.94414489
## 5  0.3651873  0.74051161 -0.01363817  0.8257913 -0.6262708 -0.17682683
## 6 -0.9424187 -0.04957081 -0.18184220 -0.2008324  0.7567534  0.88922515
##        mfr_A       mfr_G       mfr_K      mfr_N      mfr_P      mfr_Q
## 1 -0.1162476 -0.64603380 -0.05385697  2.5590737 -0.3695814 -0.3210386
## 2 -0.1162476 -0.64603380 -0.66699780 -0.2673659 -0.3695814  3.0727976
## 3 -0.1162476  0.07830713  0.31658229 -0.2673659 -0.3695814 -0.3210386
## 4 -0.1162476  0.36804350  0.04833317 -0.2673659  0.2381747 -0.3210386
## 5 -0.1162476  0.11452417 -0.13049957 -0.2673659  0.5420528 -0.3210386
## 6  8.4860776 -0.64603380 -0.66699780 -0.2673659 -0.3695814 -0.3210386
##         mfr_R     type_C     type_H    shelf_1    shelf_2     shelf_3
## 1 -0.32103855  0.1162476 -0.1162476  0.3906424 -0.2849572 -0.08833016
## 2 -0.32103855  0.1162476 -0.1162476 -0.5837690  0.3540377  0.19586254
## 3  0.38601064  0.1162476 -0.1162476  0.7425132 -0.5112679 -0.19490243
## 4 -0.32103855  0.1162476 -0.1162476 -0.2806188  1.3338298 -0.94090828
## 5  0.01834506  0.1162476 -0.1162476 -0.5837690 -0.4926305  0.94897320
## 6 -0.32103855 -8.4860776  8.4860776 -0.5837690  1.6320274 -0.94090828
```

```
Clustered_Cereal<-as.data.frame(cbind(Cereals_Data, df_opt))#Assigning each Cereal to cluster
Healthy_Cereals<-Clustered_Cereal%>%filter(Clustered_Cereal$df_opt==1)
Healthy_Cereals
```

```
##                              calories    protein        fat     sodium     fiber
## 100%_Bran                  -1.8659155  1.3817478  0.0000000 -0.3910227 3.2286675
## All-Bran                   -1.8659155  1.3817478  0.0000000  1.1795987 2.8160226
## All-Bran_with_Extra_Fiber -2.8737823  1.3817478 -0.9932203 -0.2702057 4.8792470
## Shredded_Wheat             -1.3619821 -0.4773310 -0.9932203 -1.9616441 0.3401532
## Shredded_Wheat_'n'Bran     -0.8580487  0.4522084 -0.9932203 -1.9616441 0.7527981
## Shredded_Wheat_spoon_size -0.8580487  0.4522084 -0.9932203 -1.9616441 0.3401532
## Strawberry_Fruit_Wheats    -0.8580487 -0.4773310 -0.9932203 -1.7804186 0.3401532
##                                  carbo     sugars      potass   vitamins
## 100%_Bran                  -2.50013957 -0.2542051  2.56052289 -0.1818422
## All-Bran                   -1.98622199 -0.4836096  3.12486748 -0.1818422
```

```
## All-Bran_with_Extra_Fiber -1.72926320 -1.6306324  3.26595362 -0.1818422
## Shredded_Wheat             0.32640711 -1.6306324 -0.04957081 -1.3032024
## Shredded_Wheat_'n'Bran     1.09728348 -1.6306324  0.58531685 -1.3032024
## Shredded_Wheat_spoon_size  1.35424227 -1.6306324  0.30314456 -1.3032024
## Strawberry_Fruit_Wheats    0.06944832 -0.4836096 -0.12011388 -0.1818422
##                               weight       cups   rating      mfr_A      mfr_G
## 100%_Bran                 -0.2008324 -2.0856582 1.854904 -0.1162476 -0.6460338
## All-Bran                  -0.2008324 -2.0856582 1.215196 -0.1162476 -0.6460338
## All-Bran_with_Extra_Fiber -0.2008324 -1.3644493 3.657844 -0.1162476 -0.6460338
## Shredded_Wheat            -1.3089342  0.7567534 1.842998 -0.1162476 -0.6460338
## Shredded_Wheat_'n'Bran    -0.2008324 -0.6432404 2.287432 -0.1162476 -0.6460338
## Shredded_Wheat_spoon_size -0.2008324 -0.6432404 2.168350 -0.1162476 -0.6460338
## Strawberry_Fruit_Wheats   -0.2008324  0.7567534 1.210813 -0.1162476 -0.6460338
##                                mfr_K      mfr_N      mfr_P      mfr_Q
## 100%_Bran                 -0.6669978  3.6896495 -0.3695814 -0.3210386
## All-Bran                   1.4789951 -0.2673659 -0.3695814 -0.3210386
## All-Bran_with_Extra_Fiber  1.4789951 -0.2673659 -0.3695814 -0.3210386
## Shredded_Wheat            -0.6669978  3.6896495 -0.3695814 -0.3210386
## Shredded_Wheat_'n'Bran    -0.6669978  3.6896495 -0.3695814 -0.3210386
## Shredded_Wheat_spoon_size -0.6669978  3.6896495 -0.3695814 -0.3210386
## Strawberry_Fruit_Wheats   -0.6669978  3.6896495 -0.3695814 -0.3210386
##                                mfr_R    type_C     type_H   shelf_1    shelf_2
## 100%_Bran                 -0.3210386 0.1162476 -0.1162476 -0.583769 -0.6044546
## All-Bran                  -0.3210386 0.1162476 -0.1162476 -0.583769 -0.6044546
## All-Bran_with_Extra_Fiber -0.3210386 0.1162476 -0.1162476 -0.583769 -0.6044546
## Shredded_Wheat            -0.3210386 0.1162476 -0.1162476  1.689858 -0.6044546
## Shredded_Wheat_'n'Bran    -0.3210386 0.1162476 -0.1162476  1.689858 -0.6044546
## Shredded_Wheat_spoon_size -0.3210386 0.1162476 -0.1162476  1.689858 -0.6044546
## Strawberry_Fruit_Wheats   -0.3210386 0.1162476 -0.1162476 -0.583769  1.6320274
##                              shelf_3 df_opt
## 100%_Bran                  1.0484407      1
## All-Bran                   1.0484407      1
## All-Bran_with_Extra_Fiber  1.0484407      1
## Shredded_Wheat            -0.9409083      1
## Shredded_Wheat_'n'Bran    -0.9409083      1
## Shredded_Wheat_spoon_size -0.9409083      1
## Strawberry_Fruit_Wheats   -0.9409083      1
```