Sharon Kasturiarachi

MIS 64060 Fundamentals of Machine Learning

Dr. Rouzbeh Razavi

Kasturiarachi-Assignment 5

Column variables

```
> str(cereals_data)
spec_tbl_df [77 x 16] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ name    : chr [1:77] "100%_Bran" "100%_Natural_Bran" "All-Bran" "All-Bran_with_Extra_Fiber"
 ...
 $ mfr     : chr [1:77] "N" "Q" "K" "K" ...
 $ type    : chr [1:77] "C" "C" "C" "C" ...
 $ calories: num [1:77] 70 120 70 50 110 110 110 130 90 90 ...
 $ protein : num [1:77] 4 3 4 4 2 2 2 3 2 3 ...
 $ fat     : num [1:77] 1 5 1 0 2 2 0 2 1 0 ...
 $ sodium  : num [1:77] 130 15 260 140 200 180 125 210 200 210 ...
 $ fiber   : num [1:77] 10 2 9 14 1 1.5 1 2 4 5 ...
 $ carbo   : num [1:77] 5 8 7 8 14 10.5 11 18 15 13 ...
 $ sugars  : num [1:77] 6 8 5 0 8 10 14 8 6 5 ...
 $ potass  : num [1:77] 280 135 320 330 NA 70 30 100 125 190 ...
 $ vitamins: num [1:77] 25 0 25 25 25 25 25 25 25 25 ...
 $ shelf   : num [1:77] 3 3 3 3 3 1 2 3 1 3 ...
 $ weight  : num [1:77] 1 1 1 1 1 1 1 1.33 1 1 ...
 $ cups    : num [1:77] 0.33 1 0.33 0.5 0.75 0.75 1 0.75 0.67 0.67 ...
 $ rating  : num [1:77] 68.4 34 59.4 93.7 34.4 ...
 - attr(*, "spec")=
  .. cols(
```

```
> summary(cereals_data)
     name               mfr                type              calories        protein
 Length:77          Length:77          Length:77          Min.   : 50.0   Min.   :1.000
 Class :character   Class :character   Class :character   1st Qu.:100.0   1st Qu.:2.000
 Mode  :character   Mode  :character   Mode  :character   Median :110.0   Median :3.000
                                                          Mean   :106.9   Mean   :2.545
                                                          3rd Qu.:110.0   3rd Qu.:3.000
                                                          Max.   :160.0   Max.   :6.000

      fat            sodium          fiber            carbo          sugars
 Min.   :0.000   Min.   :  0.0   Min.   : 0.000   Min.   : 5.0   Min.   : 0.000
 1st Qu.:0.000   1st Qu.:130.0   1st Qu.: 1.000   1st Qu.:12.0   1st Qu.: 3.000
 Median :1.000   Median :180.0   Median : 2.000   Median :14.5   Median : 7.000
 Mean   :1.013   Mean   :159.7   Mean   : 2.152   Mean   :14.8   Mean   : 7.026
 3rd Qu.:2.000   3rd Qu.:210.0   3rd Qu.: 3.000   3rd Qu.:17.0   3rd Qu.:11.000
 Max.   :5.000   Max.   :320.0   Max.   :14.000   Max.   :23.0   Max.   :15.000
                                                  NA's   :1      NA's   :1
     potass          vitamins          shelf           weight          cups
 Min.   : 15.00   Min.   :  0.00   Min.   :1.000   Min.   :0.50   Min.   :0.250
 1st Qu.: 42.50   1st Qu.: 25.00   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.670
 Median : 90.00   Median : 25.00   Median :2.000   Median :1.00   Median :0.750
 Mean   : 98.67   Mean   : 28.25   Mean   :2.208   Mean   :1.03   Mean   :0.821
 3rd Qu.:120.00   3rd Qu.: 25.00   3rd Qu.:3.000   3rd Qu.:1.00   3rd Qu.:1.000
 Max.   :330.00   Max.   :100.00   Max.   :3.000   Max.   :1.50   Max.   :1.500
 NA's   :2
     rating
 Min.   :18.04
 1st Qu.:33.17
 Median :40.40
 Mean   :42.67
 3rd Qu.:50.83
 Max.   :93.70
```

Finding the number of missing values

```
> # Missing values
> sum(is.na(cereals_data))
[1] 4
```

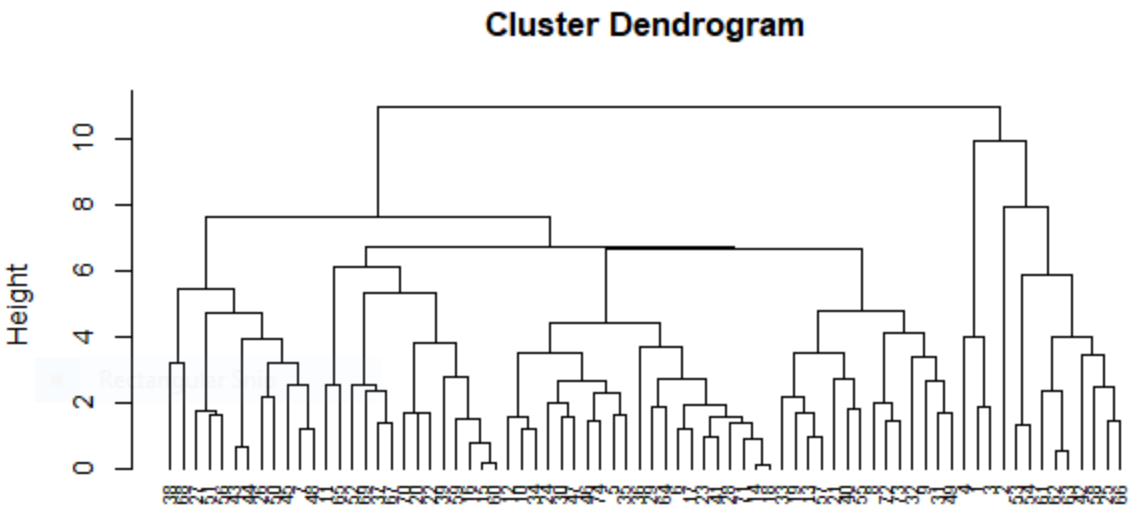
after converting the variables

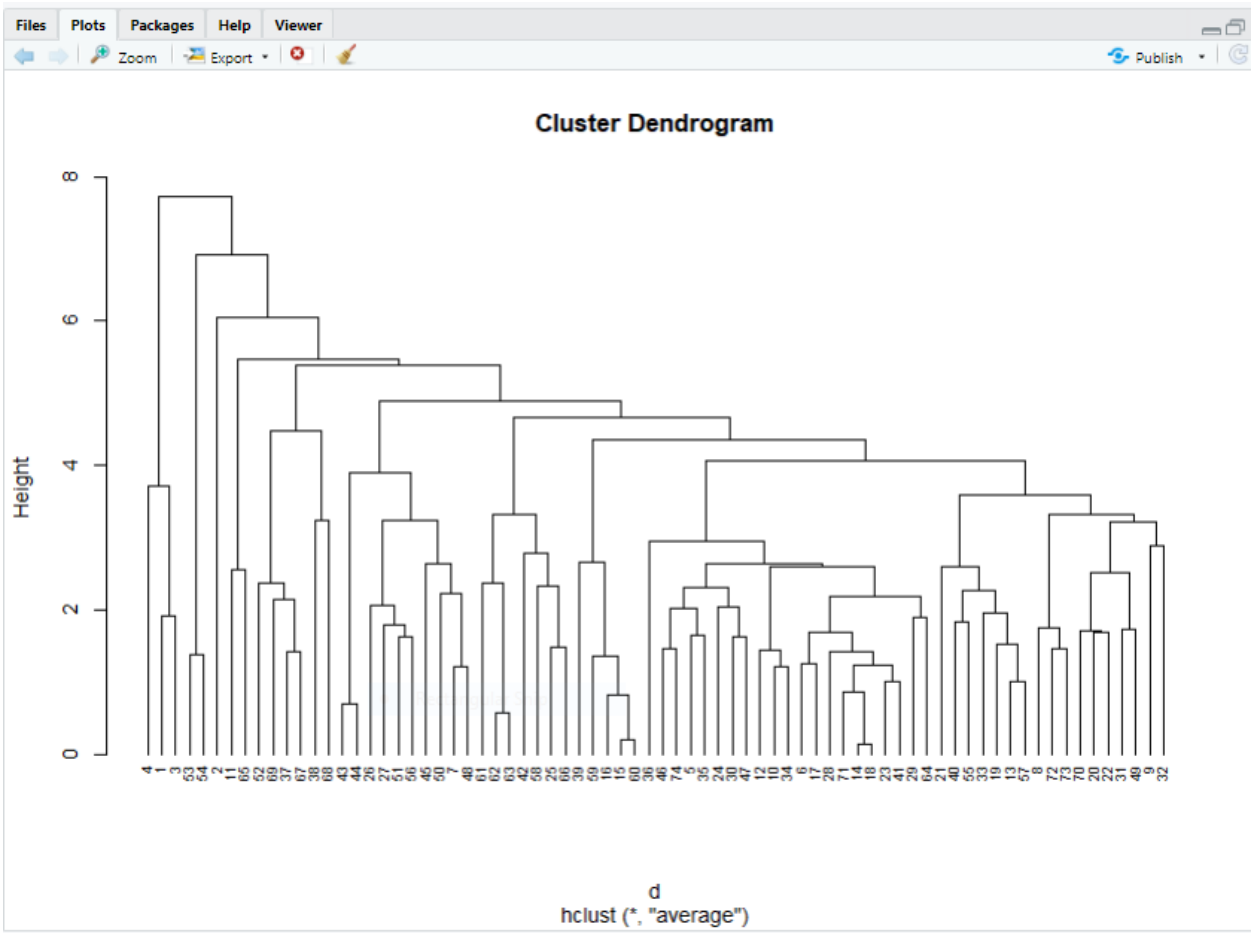
```
'data.frame':   77 obs. of  16 variables:
 $ name    : Factor w/ 77 levels "100%_Bran","100%_Natural_Bran",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ mfr     : Factor w/ 7 levels "A","G","K","N",..: 4 6 3 3 7 2 3 2 7 5 ...
 $ type    : Factor w/ 2 levels "C","H": 1 1 1 1 1 1 1 1 1 1 ...
 $ calories: int  70 120 70 50 110 110 110 130 90 90 ...
 $ protein : int  4 3 4 4 2 2 2 3 2 3 ...
 $ fat     : int  1 5 1 0 2 2 0 2 1 0 ...
 $ sodium  : int  130 15 260 140 200 180 125 210 200 210 ...
 $ fiber   : num  10 2 9 14 1 1.5 1 2 4 5 ...
 $ carbo   : num  5 8 7 8 14 10.5 11 18 15 13 ...
 $ sugars  : int  6 8 5 0 8 10 14 8 6 5 ...
 $ potass  : int  280 135 320 330 NA 70 30 100 125 190 ...
 $ vitamins: int  25 0 25 25 25 25 25 25 25 25 ...
 $ shelf   : int  3 3 3 3 3 1 2 3 1 3 ...
 $ weight  : num  1 1 1 1 1 1 1 1.33 1 1 ...
 $ cups    : num  0.33 1 0.33 0.5 0.75 0.75 1 0.75 0.67 0.67 ...
 $ rating  : num  68.4 34 59.4 93.7 34.4 ...
> rmr
```

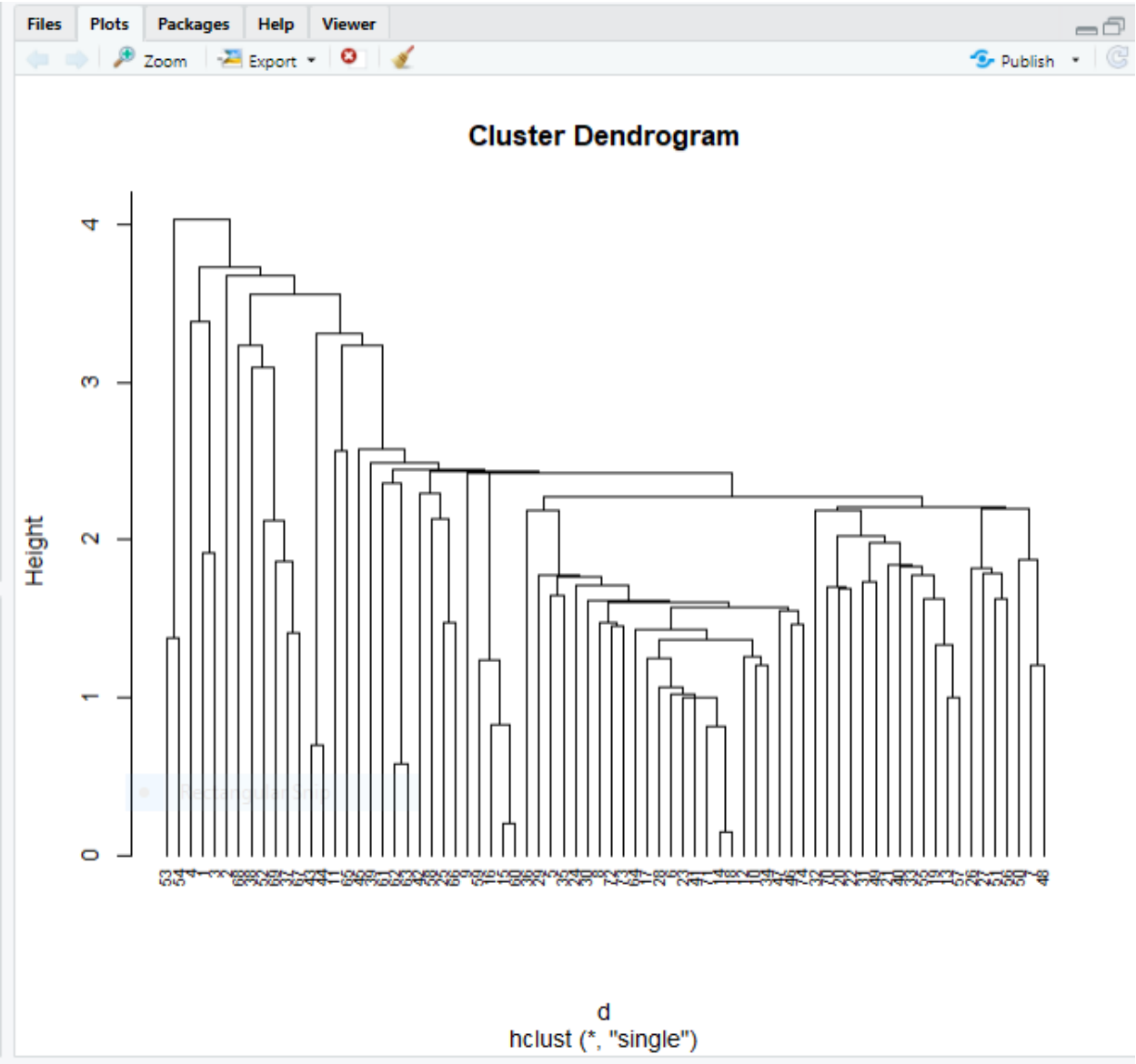
```
> sum(is.na(cereals_data))
[1] 4
> colSums(is.na(cereals_data))
    name      mfr     type calories  protein      fat   sodium    fiber    carbo   sugars
       0        0        0        0        0        0        0        0        1        1
  potass vitamins    shelf   weight     cups   rating
       2        0        0        0        0        0
> cereals_data.clean<-na.omit(cereals_data)
> nrow(cereals_data.clean)
[1] 74
> nrow(cereals_data.clean)
[1] 74
> sum(is.na(cereals_data.clean))
[1] 0
```

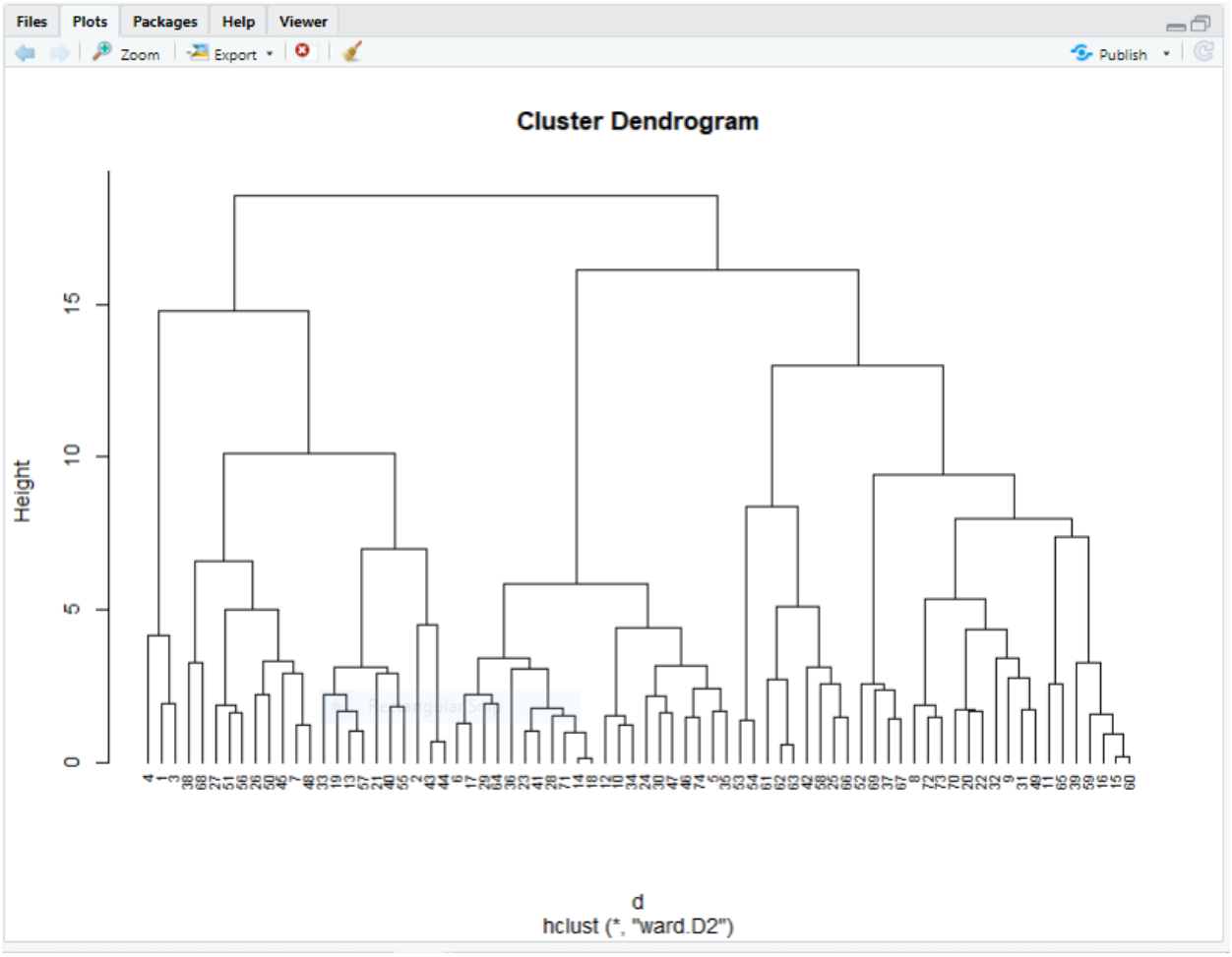
Normalizing the data for each variable to be treated equally by the distance measure

```
> cereals_data.clean.norm <- sapply(cereals_data.clean, scale)
> summary(cereals_data.clean.norm)
    calories          protein              fat              sodium            fiber
 Min.   :-2.8738   Min.   :-1.40687   Min.   :-0.9932   Min.   :-1.9616   Min.   :-0.89778
 1st Qu.:-0.3541   1st Qu.:-0.47733   1st Qu.:-0.9932   1st Qu.:-0.3306   1st Qu.:-0.79462
 Median : 0.1498   Median :-0.01256   Median : 0.0000   Median : 0.2131   Median :-0.07249
 Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
 3rd Qu.: 0.1498   3rd Qu.: 0.45221   3rd Qu.: 0.0000   3rd Qu.: 0.6661   3rd Qu.: 0.34015
 Max.   : 2.6695   Max.   : 3.24083   Max.   : 3.9729   Max.   : 1.9045   Max.   : 4.87925
     carbo             sugars             potass           vitamins           shelf
 Min.   :-2.50014   Min.   :-1.6306   Min.   :-1.1783   Min.   :-1.3032   Min.   :-1.4617
 1st Qu.:-0.70143   1st Qu.:-0.9424   1st Qu.:-0.8079   1st Qu.:-0.1818   1st Qu.:-1.1612
 Median :-0.05903   Median :-0.0248   Median :-0.1201   Median :-0.1818   Median :-0.2599
 Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.58337   3rd Qu.: 0.8928   3rd Qu.: 0.3031   3rd Qu.:-0.1818   3rd Qu.: 0.9420
 Max.   : 2.12512   Max.   : 1.8104   Max.   : 3.2660   Max.   : 3.1822   Max.   : 0.9420
     weight             cups              rating
 Min.   :-3.4600   Min.   :-2.4251   Min.   :-1.7336
 1st Qu.:-0.2008   1st Qu.:-0.6432   1st Qu.:-0.7071
 Median :-0.2008   Median :-0.3038   Median :-0.1510
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.:-0.2008   3rd Qu.: 0.7568   3rd Qu.: 0.5807
 Max.   : 3.0583   Max.   : 2.8780   Max.   : 3.6578
>
```


Cluster Dendrogram
Height
0
2
4
6
8
10
d
hclust (*, "complete")


Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Cluster Dendrogram
Height
0
2
4
6
8
d
hclust (*, "average")


Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Cluster Dendrogram
Height
0
1
2
3
4
d
hclust (*, "single")



Complete – considers the maximum distance between clusters

Single – considers the smallest distance between clusters

Average- considers the average distance between clusters

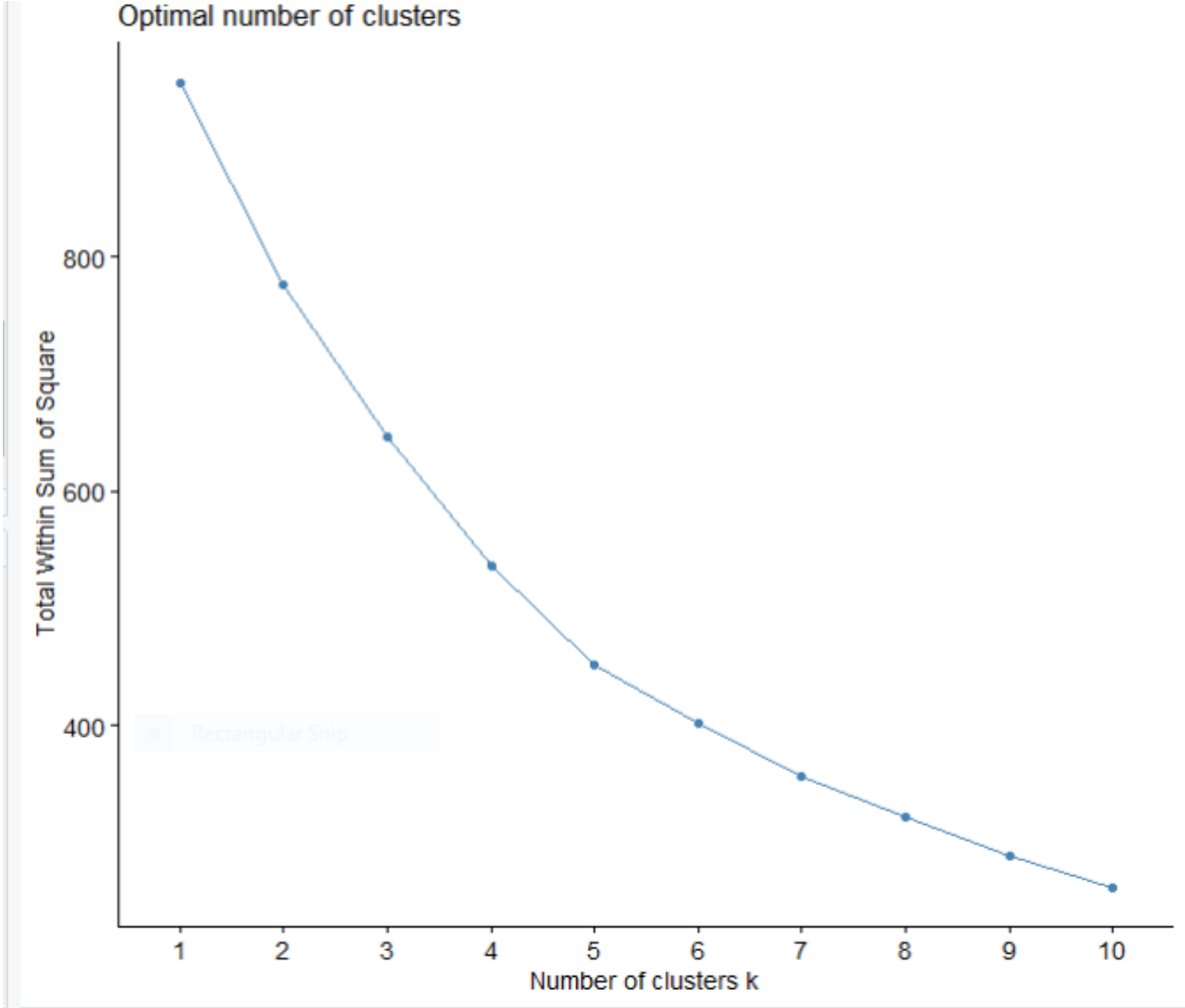
Ward - the pair of clusters with the lowest distance is merged

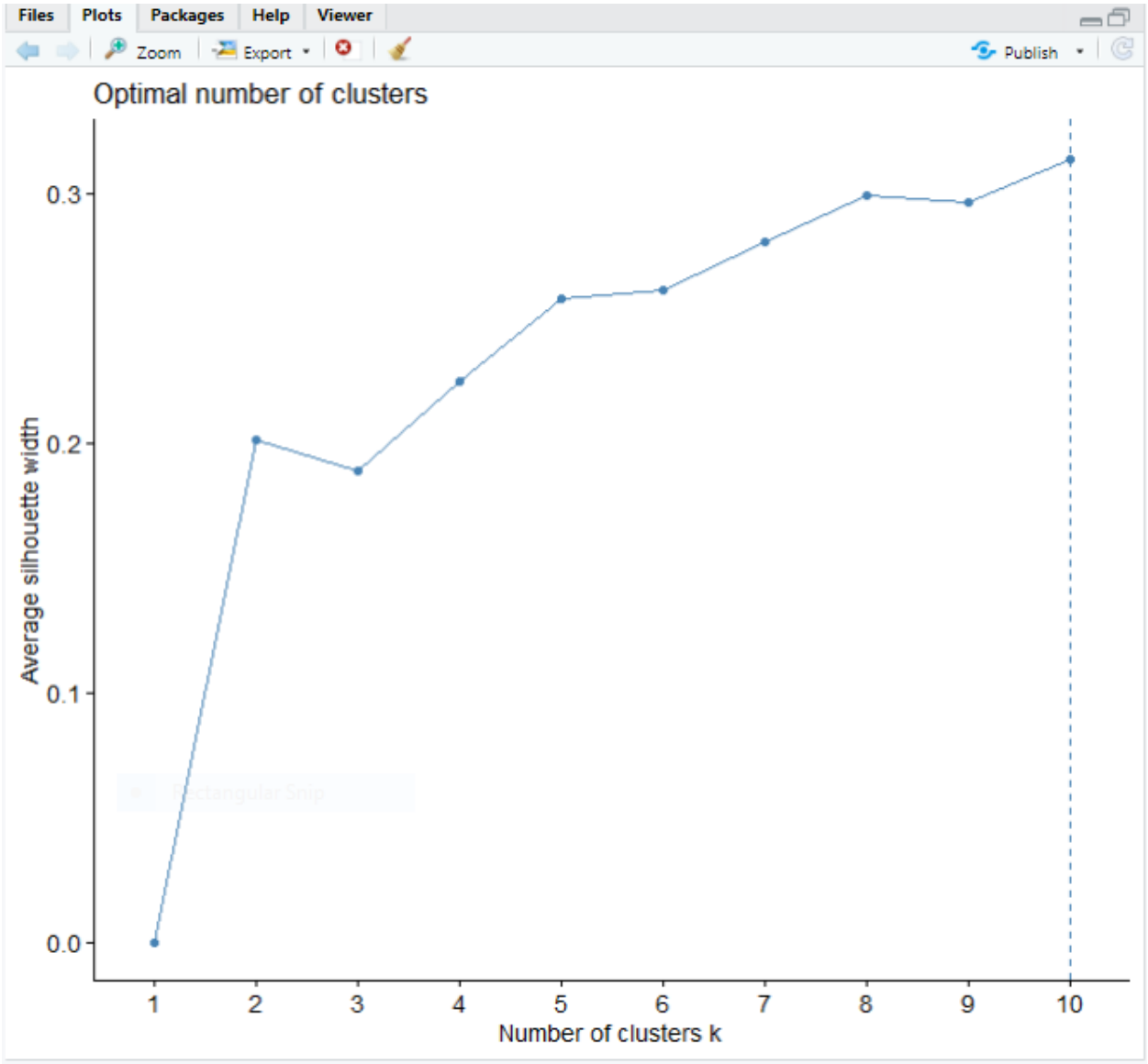
The agglomerative coefficient is highest in the ward method.

The ward method will be used in the hierarchical clustering analysis.

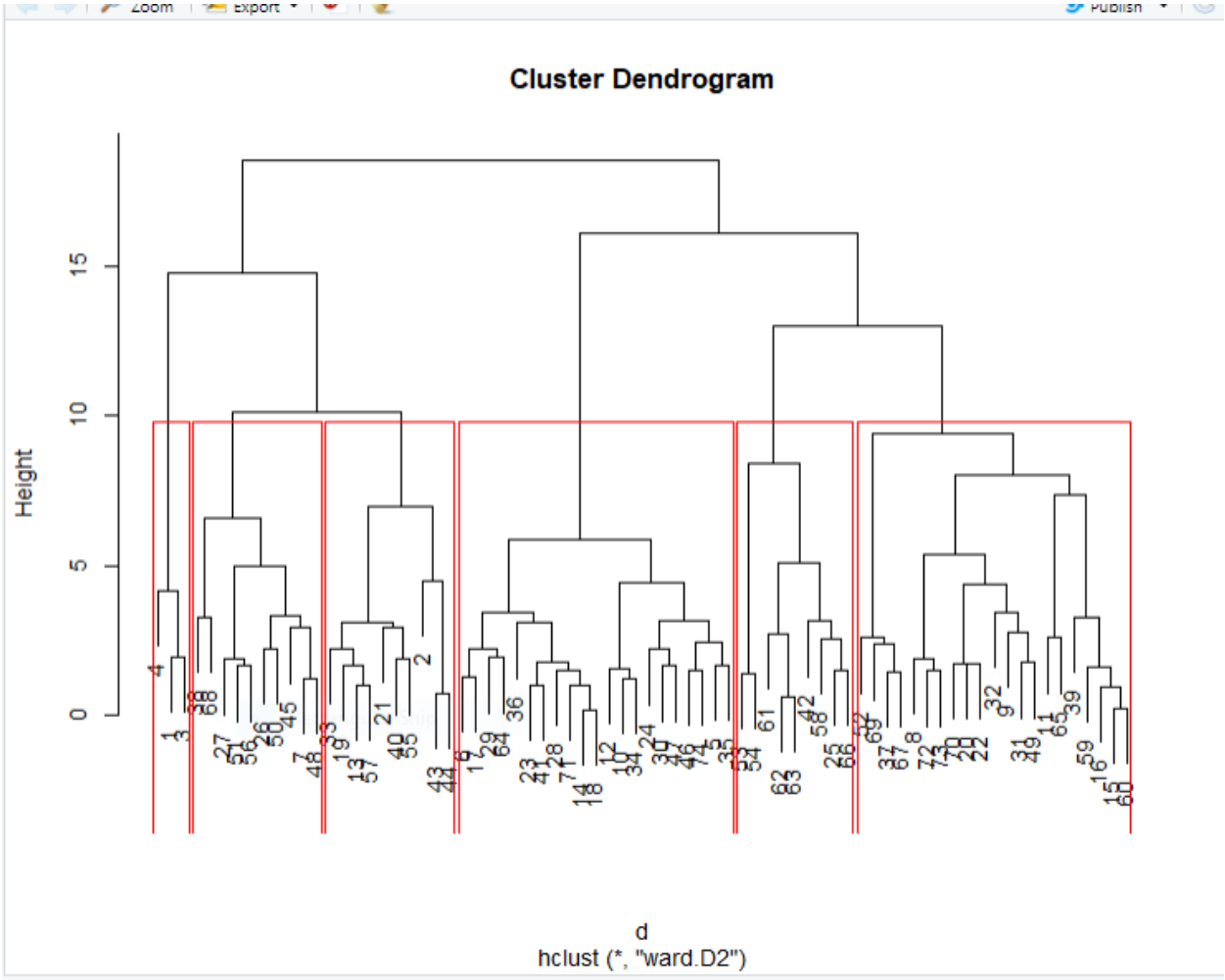
```
> library(cluster)
> #Using Agnes function
> df <- cereals_data.clean.norm
> hc_single <- agnes(df, method = "single")
> hc_complete <- agnes(df, method = "complete")
> hc_average <- agnes(df, method = "average")
> print(hc_single$ac)
[1] 0.6067859
> print(hc_complete$ac)
[1] 0.8353712
> print(hc_average$ac)
[1] 0.7766075
> hc_ward <- agnes(df, method = "ward")
> print(hc_ward$ac)
[1] 0.9046042
```

Determining the optimal number of clusters using the elbow method





I chose 6 clusters




```
> clusters <- cutree(fit.ward, k=6)
> table(clusters)
clusters
 1  2  3  4  5  6
 3 10 21 10 21  9
> 
```

#store the clusters in a data frame along with the cereals data

cereals_clusts_hc <-cbind(clusters, cereals_data.clean)

```
> colnames(cereals_clusts_hc)[1]
[1] "clusters"
> head(cereals_clusts_hc)
                        clusters calories protein fat sodium fiber carbo sugars
100%_Bran                      1       70       4   1    130  10.0   5.0       6
100%_Natural_Bran              2      120       3   5     15   2.0   8.0       8
All-Bran                       1       70       4   1    260   9.0   7.0       5
All-Bran_with_Extra_Fiber      1       50       4   0    140  14.0   8.0       0
Apple_Cinnamon_Cheerios        3      110       2   2    180   1.5  10.5      10
Apple_Jacks                    3      110       2   0    125   1.0  11.0      14
                        potass vitamins shelf weight cups   rating
100%_Bran                  280       25     3      1 0.33 68.40297
100%_Natural_Bran          135        0     3      1 1.00 33.98368
All-Bran                   320       25     3      1 0.33 59.42551
All-Bran_with_Extra_Fiber  330       25     3      1 0.50 93.70491
Apple_Cinnamon_Cheerios     70       25     1      1 0.75 29.50954
Apple_Jacks                 30       25     2      1 1.00 33.17409
> str(cereals_clusts_hc)
```

```
> tail(cereals_clusts_hc)
                  clusters calories protein fat sodium fiber carbo sugars potass
Total_Whole_Grain        5      100       3   1    200     3    16       3    110
Triples                  5      110       2   1    250     0    21       3     60
Trix                     3      110       1   1    140     0    13      12     25
Wheat_Chex               5      100       3   1    230     3    17       3    115
Wheaties                 5      100       3   1    200     3    17       3    110
Wheaties_Honey_Gold      3      110       2   1    200     1    16       8     60
                  vitamins shelf weight cups   rating
Total_Whole_Grain      100     3      1 1.00 46.65884
Triples                 25     3      1 0.75 39.10617
Trix                    25     2      1 1.00 27.75330
Wheat_Chex              25     1      1 0.67 49.78744
Wheaties                25     1      1 1.00 51.59219
Wheaties_Honey_Gold     25     1      1 0.75 36.18756
> t
```

```
> str(cereals_clusts_hc)
'data.frame':   74 obs. of  14 variables:
 $ clusters: int  1 2 1 1 3 3 4 5 5 3 ...
 $ calories: int  70 120 70 50 110 110 130 90 90 120 ...
 $ protein : int  4 3 4 4 2 2 3 2 3 1 ...
 $ fat     : int  1 5 1 0 2 0 2 1 0 2 ...
 $ sodium  : int  130 15 260 140 180 125 210 200 210 220 ...
 $ fiber   : num  10 2 9 14 1.5 1 2 4 5 0 ...
 $ carbo   : num  5 8 7 8 10.5 11 18 15 13 12 ...
 $ sugars  : int  6 8 5 0 10 14 8 6 5 12 ...
 $ potass  : int  280 135 320 330 70 30 100 125 190 35 ...
 $ vitamins: int  25 0 25 25 25 25 25 25 25 25 ...
 $ shelf   : int  3 3 3 3 1 2 3 1 3 2 ...
 $ weight  : num  1 1 1 1 1 1 1.33 1 1 1 ...
 $ cups    : num  0.33 1 0.33 0.5 0.75 1 0.75 0.67 0.67 0.75 ...
 $ rating  : num  68.4 34 59.4 93.7 29.5 ...
> summary(cereals_clusts_hc)
    clusters        calories        protein           fat        sodium
 Min.   :1.000   Min.   : 50   Min.   :1.000   Min.   :0   Min.   :  0.0
 1st Qu.:3.000   1st Qu.:100   1st Qu.:2.000   1st Qu.:0   1st Qu.:135.0
 Median :4.000   Median :110   Median :2.500   Median :1   Median :180.0
 Mean   :3.851   Mean   :107   Mean   :2.514   Mean   :1   Mean   :162.4
 3rd Qu.:5.000   3rd Qu.:110   3rd Qu.:3.000   3rd Qu.:1   3rd Qu.:217.5
 Max.   :6.000   Max.   :160   Max.   :6.000   Max.   :5   Max.   :320.0
     fiber            carbo           sugars           potass          vitamins
 Min.   : 0.000   Min.   : 5.00   Min.   : 0.000   Min.   : 15.00   Min.   :  0.00
 1st Qu.: 0.250   1st Qu.:12.00   1st Qu.: 3.000   1st Qu.: 41.25   1st Qu.: 25.00
 Median : 2.000   Median :14.50   Median : 7.000   Median : 90.00   Median : 25.00
 Mean   : 2.176   Mean   :14.73   Mean   : 7.108   Mean   : 98.51   Mean   : 29.05
 3rd Qu.: 3.000   3rd Qu.:17.00   3rd Qu.:11.000   3rd Qu.:120.00   3rd Qu.: 25.00
 Max.   :14.000   Max.   :23.00   Max.   :15.000   Max.   :330.00   Max.   :100.00
     shelf           weight           cups           rating
 Min.   :1.000   Min.   :0.500   Min.   :0.2500   Min.   :18.04
 1st Qu.:1.250   1st Qu.:1.000   1st Qu.:0.6700   1st Qu.:32.45
 Median :2.000   Median :1.000   Median :0.7500   Median :40.25
 Mean   :2.216   Mean   :1.031   Mean   :0.8216   Mean   :42.37
 3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:50.52
 Max.   :3.000   Max.   :1.500   Max.   :1.5000   Max.   :93.70
>
```

To find a cluster of healthy cereals to support a healthy diet. They should be high in fiber and low in sugar content

Lowest sugar is 0.00 and maximum fiber is 14

Let's examine the clusters

```
> clusters <- cutree(fit.ward, k=6)
> table(clusters)
clusters
 1  2  3  4  5  6
 3 10 21 10 21  9
>
```

Cluster 1 has 3 cereals - 100%-Bran, All-Bran, All-Bran-with-Extra-Fiber- fall under healthy breakfast cereals

Cluster 2 has 10 cereals - 100%-Natural-Bran, Clusters, Cracklin'-Oat-Bran, Crispy-Wheat-&-Raisins, Life, Muesli-Raisins Dates &Almonds, Muesli-Raisins Peaches & Pecans, Quaker Oat Squares, Raisin Nut Bran, Great Grains Pecan-

Cluster 3 has 21 cereals – Apple Cinnamon Cheerios, Cap'n'Crunch, Apple Jacks, Cinnamon Toast Crunch, Cocoa Puffs, Corn Pops, Count Chocula, Froot Loops, Frosted Flakes, Fruity Pebbles, Golden Crisp, Golden Grahams, Honey Graham Ohs, Honey Nut Cheerios, Honeycomb, Lucky Charms, Multi-Grain Cheerios, Nut &Honey Crunch, Smacks, Trix, Wheaties Honey Gold.

Cluster 4 has 10 kinds of cereals – Basic 4, Fruit & Fibre Dates, Walnuts, and Oats, Fruitful Bran, Just Right Fruit & Nut, Mueslix Crispy Blend, Nutri-Grain Almond Raisin, Oatmeal Raisin Crisp, Post Nat. Raisin Bran, Raisin Bran, Total Raisin Bran

Cluster 5 has 21 kinds of cereals- Bran Chex, Bran Flakes, Cheerios, Corn Chex, Cornflakes, Crispix, Double Chex, Grape Nuts Flakes, Grape-Nuts, Just Right Crunchy Nuggets, Kix, Nutri-grain Wheat, Product 19, Rice Chex, Rice Krispies, Special K, Total Corn Flakes, Total Whole Grain, Triples, Wheat Chex, Wheaties,

Cluster 6 has 9 kinds of cereal – Strawberry Fruit Wheats, Shredded Wheat spoon size, Shredded Wheat 'n' Bran, Shredded Wheat, Raisin Squares, Puffed Wheat, Puffed Rice, Maypo, Frosted Mini-Wheats