Clustering Analysis of Wholesale Distribution

Madison Meinke, Natalia Mendoza-Orr, Cody Ortloff, Claire Tsao, Qianer Wu

7/3/2023

Load the required packages and data set

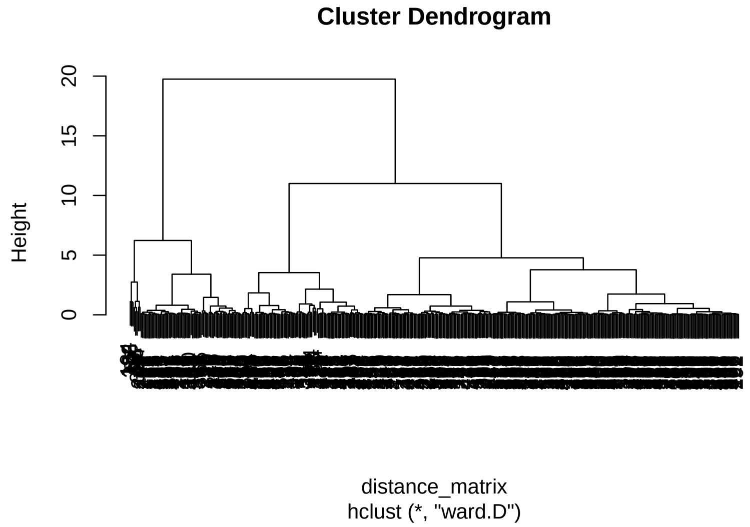
```
library(dplyr)
library(stats)
library(cluster)
library(ggplot2)
library(factoextra)
library(knitr)

wholesale <- read.csv("Wholesale_customers_data.csv")
```

Normalize the data set

```
normalize <- function(x){
  return ((x - min(x)) / (max(x) - min(x)))
}

wholesale_normalized <- wholesale %>% mutate_at(c(3:8), normalize)
```

Build the distance matrix and look at the dendrogram using hierachical clustering

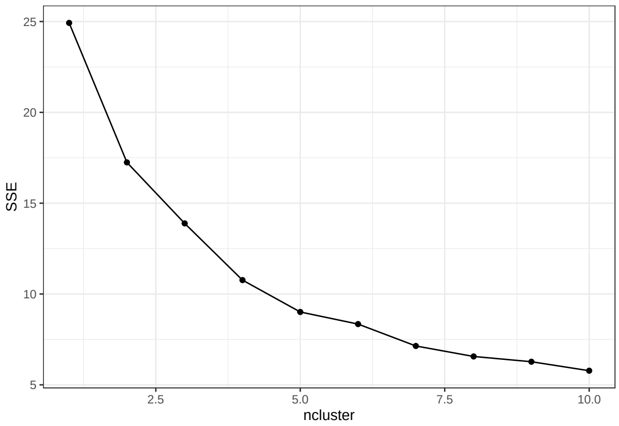
```
distance_matrix <- dist(wholesale_normalized[,3:8], method = "euclidean")

hierarchical <- hclust(distance_matrix, method = "ward.D")

plot(hierarchical)
```

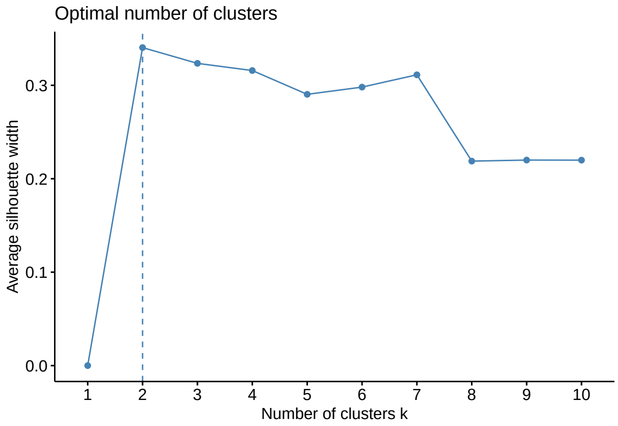


Deciding on Cluster Solution Amount (SSE Curve and Silhoutte Scores)

We can see that the "elbow" of the SSE curve is somewhere between 2-8 clusters. The silhouette score appears to have the greatest average value at 2 clusters. However, it is possible for a data set to have a high average silhouette score for a certain number of clusters, but it might not be the most meaningful or informative solution. (1)

```
SSE_curve = c()
for (n in 1:10){
  kcluster = kmeans(wholesale_normalized[,3:8], centers = n)
  SSE_curve[n] = kcluster$tot.withinss
}

plot_data = data.frame(ncluster = 1:10, SSE = SSE_curve)
ggplot(plot_data, aes(x = ncluster, y = SSE)) +
  geom_line() + geom_point() + theme_bw()

fviz_nbclust(wholesale_normalized[,3:8], hcut, method='silhouette')
```

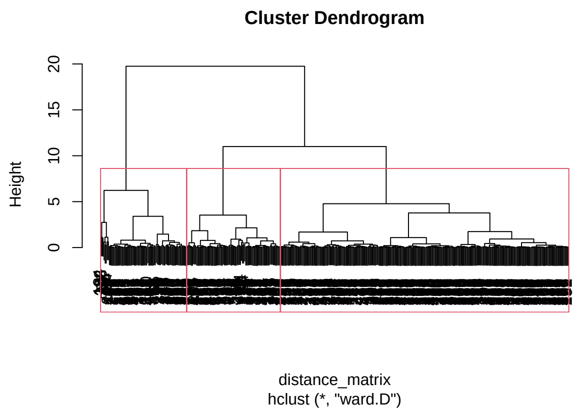


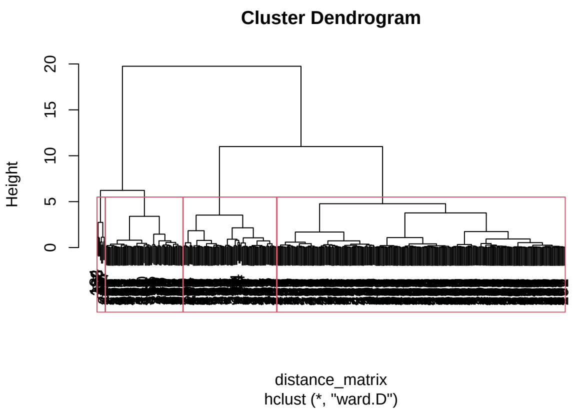


Deciding on our Cluster Solution Amount (Comparisons)

We decided that we would start by looking at 3 clusters and work our way up to see if we gain any additional information from adding more clusters.

```
plot(hierarchical)
rect.hclust(hierarchical, k = 3)

plot(hierarchical)
rect.hclust(hierarchical, k = 4)
```





As shown in the above dendrograms, the 4th cluster added is a very small one. We are interested to see if adding the 4th cluster is meaningful or not. We will look at the means for each attribute's spending amount in the cluster solutions for the 3 cluster solution vs the 4 cluster solution:

```
# 3 cluster solution
wholesale_normalized$hcluster = cutree(hierarchical, k = 3)
wholesale_normalized %>% group_by(hcluster) %>% summarize_at(3:8, mean)
```

```
## # A tibble: 3 x 7
##   hcluster  Fresh   Milk Grocery Frozen Detergents_Paper Delicatessen
##      <int>  <dbl>  <dbl>   <dbl>  <dbl>            <dbl>        <dbl>
## 1        1 0.0727 0.0439  0.0471 0.0318           0.0309       0.0218
## 2        2 0.243  0.0740  0.0627 0.110            0.0252       0.0480
## 3        3 0.0737 0.197   0.240  0.0464           0.252        0.0473
```

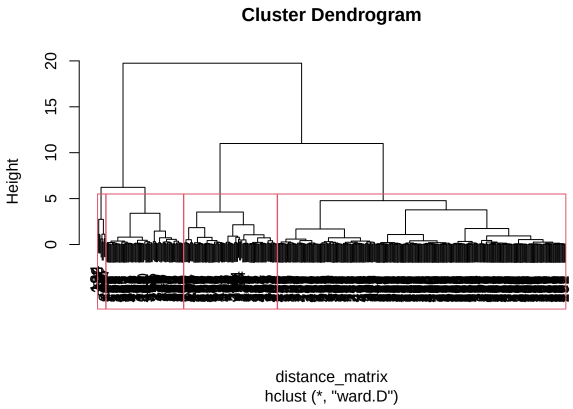
```
# 4 cluster solution
wholesale_normalized$hcluster = cutree(hierarchical, k = 4)
wholesale_normalized %>% group_by(hcluster) %>% summarize_at(3:8, mean)
```

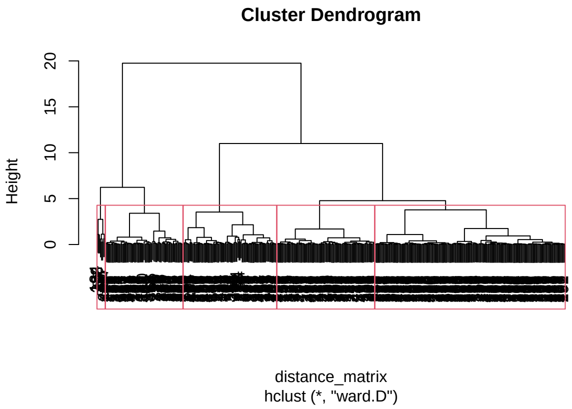
```
## # A tibble: 4 x 7
##   hcluster  Fresh   Milk Grocery Frozen Detergents_Paper Delicatessen
##      <int>  <dbl>  <dbl>   <dbl>  <dbl>            <dbl>        <dbl>
## 1        1 0.0727 0.0439  0.0471 0.0318           0.0309       0.0218
## 2        2 0.243  0.0740  0.0627 0.110            0.0252       0.0480
## 3        3 0.0440 0.162   0.213  0.0229           0.227        0.0308
## 4        4 0.345  0.523   0.484  0.261            0.479        0.197
```

It appears meaningful to add the 4th cluster because despite being very small (8 data points), this cluster has the highest average spending for every attribute.

Now, we will check to see if adding a 5th cluster would be meaningful. First Look at the dendrograms of 4 clusters vs 5 clusters:

```
plot(hierarchical)
rect.hclust(hierarchical, k = 4)

plot(hierarchical)
rect.hclust(hierarchical, k = 5)
```





As shown, adding a 5th cluster just splits our large cluster into 2 seperate clusters. We will explore to see if splitting that cluster tells us anything meaningful:

```
# 4 cluster solution
wholesale_normalized$hcluster = cutree(hierarchical, k = 4)
wholesale_normalized %>% group_by(hcluster) %>% summarize_at(3:8, mean)
```

```
## # A tibble: 4 x 7
##   hcluster  Fresh   Milk Grocery Frozen Detergents_Paper Delicatessen
##      <int>  <dbl>  <dbl>   <dbl>  <dbl>            <dbl>        <dbl>
## 1        1 0.0727 0.0439  0.0471 0.0318           0.0309       0.0218
## 2        2 0.243  0.0740  0.0627 0.110            0.0252       0.0480
## 3        3 0.0440 0.162   0.213  0.0229           0.227        0.0308
## 4        4 0.345  0.523   0.484  0.261            0.479        0.197
```

```
# 5 cluster solution
wholesale_normalized$hcluster = cutree(hierarchical, k = 5)
wholesale_normalized %>% group_by(hcluster) %>% summarize_at(3:8, mean)
```

```
## # A tibble: 5 x 7
##   hcluster  Fresh   Milk Grocery Frozen Detergents_Paper Delicatessen
##      <int>  <dbl>  <dbl>   <dbl>  <dbl>            <dbl>        <dbl>
## 1        1 0.0706 0.0749  0.0858 0.0206           0.0711       0.0304
## 2        2 0.243  0.0740  0.0627 0.110            0.0252       0.0480
## 3        3 0.0440 0.162   0.213  0.0229           0.227        0.0308
## 4        4 0.0737 0.0279  0.0272 0.0376           0.0102       0.0174
## 5        5 0.345  0.523   0.484  0.261            0.479        0.197
```

By adding the 5th cluster, nothing jumps out at us as meaningful new information versus the 4 cluster solution. As a result, we will revert to using 4 clusters.

Getting information out of our clusters

Now that we've decided on 4 clusters, lets investigate the details of our clusters:

Cluster Totals:

```
##       Cluster1 Cluster2 Cluster3 Cluster4
## Total      271       88       73        8
```

Regions by Cluster (Percentage):

```
##        Cluster1 Cluster2 Cluster3 Cluster4
## Lisbon    18.45    15.91    17.81        0
## Oporto    10.33     7.95    13.70       25
## Other     71.22    76.14    68.49       75
```

Channels by Cluster (Percentage):

```
##        Cluster1 Cluster2 Cluster3 Cluster4
## Horeca    80.07    84.09     5.48     37.5
## Retail    19.93    15.91    94.52     62.5
```

Cluster 1 is dominated by the Hotel, Restaurant, and Cafe (Horeca) industry and based on it's average spending across all attributes is the lowest spending cluster with its key product being Fresh items.

Cluster 2 is also primarily based in the Horeca industry and has key products of Fresh and Frozen. Compared to the other clusters, its spending is on the low-medium side.

Cluster 3 is mostly Retail clients its key products are Milk, Grocery and detergents paper. It is in the medium spending class across all attributes.

Cluster 4 is a small cluster, but it is the large spenders, blowing away the rest of the clusters for average spending habits. Similar to 3, its key products are Milk, Grocery, and detergents paper, and mostly in the Retail client range.

Correlation Between Attributes within Clusters

We want to look at which products have effects on the other products within each cluster:

```r
options(width = 100)
cluster1 <- wholesale_normalized %>% filter(hcluster == 1)
round(cor(cluster1[3:8]), 5)
```

Cluster 1:

```
##                     Fresh     Milk  Grocery   Frozen Detergents_Paper Delicatessen
## Fresh             1.00000 -0.18798 -0.10023  0.06353         -0.14395      0.06843
## Milk             -0.18798  1.00000  0.65627 -0.17048          0.60074      0.36864
## Grocery          -0.10023  0.65627  1.00000 -0.19149          0.76936      0.33426
## Frozen            0.06353 -0.17048 -0.19149  1.00000         -0.18224     -0.02153
## Detergents_Paper -0.14395  0.60074  0.76936 -0.18224          1.00000      0.23366
## Delicatessen      0.06843  0.36864  0.33426 -0.02153          0.23366      1.00000
```

Noticeably in Cluster 1 is the strong positive correlation between Detergents_Paper and Grocery. This indicates that as spending habits within this cluster goes up for Grocery products, it also tends to increase Detergents_Paper spending habits as well. While we cannot confirm causation, it is an interesting trend to discover.

```r
cluster2 <- wholesale_normalized %>% filter(hcluster == 2)
round(cor(cluster2[3:8]), 5)
```

Cluster 2:

```
##                     Fresh     Milk  Grocery   Frozen Detergents_Paper Delicatessen
## Fresh             1.00000 -0.18497  0.01211 -0.20374         -0.00816     -0.02141
## Milk             -0.18497  1.00000  0.56526 -0.06889          0.29288      0.43907
## Grocery           0.01211  0.56526  1.00000 -0.13774          0.63854      0.51256
## Frozen           -0.20374 -0.06889 -0.13774  1.00000         -0.23617      0.01134
## Detergents_Paper -0.00816  0.29288  0.63854 -0.23617          1.00000      0.34507
## Delicatessen     -0.02141  0.43907  0.51256  0.01134          0.34507      1.00000
```

There are no two attributes that are evidently strongly correlated in Cluster 2, however it is interesting how weak the correlation Fresh products are with the other products within this cluster.

```r
cluster3 <- wholesale_normalized %>% filter(hcluster == 3)
round(cor(cluster3[3:8]), 5)
```

Cluster 3:

```
##                    Fresh    Milk Grocery  Frozen Detergents_Paper Delicatessen
## Fresh            1.00000 0.18654 0.17568 0.30565          0.12217      0.22471
## Milk             0.18654 1.00000 0.56351 0.35835          0.50980      0.43762
## Grocery          0.17568 0.56351 1.00000 0.21576          0.72599      0.29648
## Frozen           0.30565 0.35835 0.21576 1.00000          0.26480      0.29696
## Detergents_Paper 0.12217 0.50980 0.72599 0.26480          1.00000      0.24203
## Delicatessen     0.22471 0.43762 0.29648 0.29696          0.24203      1.00000
```

Again, in Cluster 3 Grocery and Detergents_Paper appear to be strongly correlated.

```
cluster4 <- wholesale_normalized %>% filter(hcluster == 4)
round(cor(cluster4[3:8]), 5)
```

Cluster 4:

```
##                     Fresh     Milk  Grocery   Frozen Detergents_Paper Delicatessen
## Fresh             1.00000 -0.03254 -0.51669  0.16094         -0.55050      0.11863
## Milk             -0.03254  1.00000  0.02780 -0.33498          0.02212      0.07658
## Grocery          -0.51669  0.02780  1.00000 -0.71126          0.95510     -0.42754
## Frozen            0.16094 -0.33498 -0.71126  1.00000         -0.80554      0.45102
## Detergents_Paper -0.55050  0.02212  0.95510 -0.80554          1.00000     -0.57868
## Delicatessen      0.11863  0.07658 -0.42754  0.45102         -0.57868      1.00000
```

In Cluster 4, Grocery and Detergents_Paper are even more correlated then the other two clusters we pointed it out in. Also worth noting is the strong negative correlation between Frozen products and Detergents_Paper in this cluster.