HW1 Group Part

Group 9 Kasturi Pal, Deepanshu Kataria, Alan Chen, Ssu Hsien Lee

HW1 Group Part

Procedure

To help Company XYZ deeply understand their clients' spending patterns, we decided to leverage R to do the cluster analysis.

For the cluster analysis, we first normalize the annual spending of the six different products to make sure all attributes of the data take value from the same range.

Then we choose the Hierarchical Clustering to form our clusters. While applying the Hierarchical Clustering, we use the Euclidean method calculating the distance matrix to find clusters that are nearest to each other.

From the dendrogram showing the Hierarchical Clustering output, we decided to cut the trees into four parts, showing four clusters in the end.

Import packages

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

```
Wholesale_customers = read.csv("Wholesale customers data.csv")
```

Normalize the data

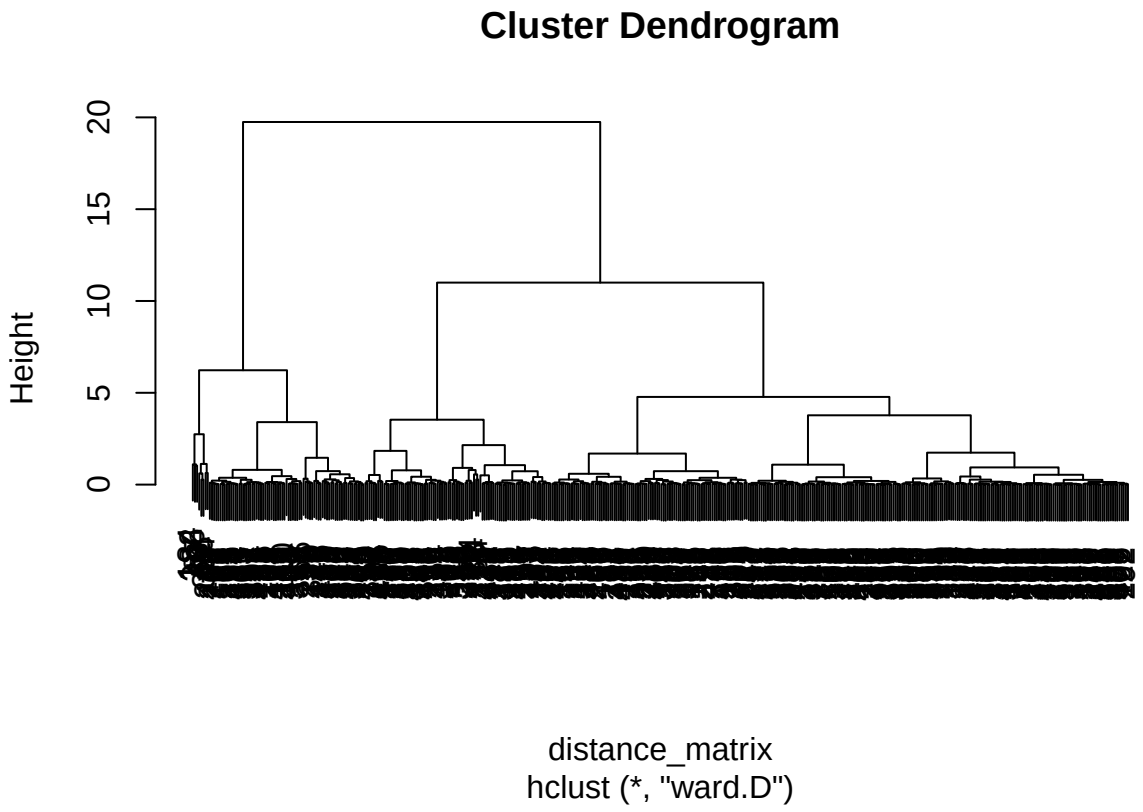
```
normalize = function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}

## just adding new columns
Wholesale_customers_normalized = Wholesale_customers %>%
  mutate_at(c(3:8), normalize)
```

Euclidean Distance method

```
distance_matrix = dist(Wholesale_customers_normalized[,3:8], method = "euclidean")
## View(as.matrix(distance_matrix))
```

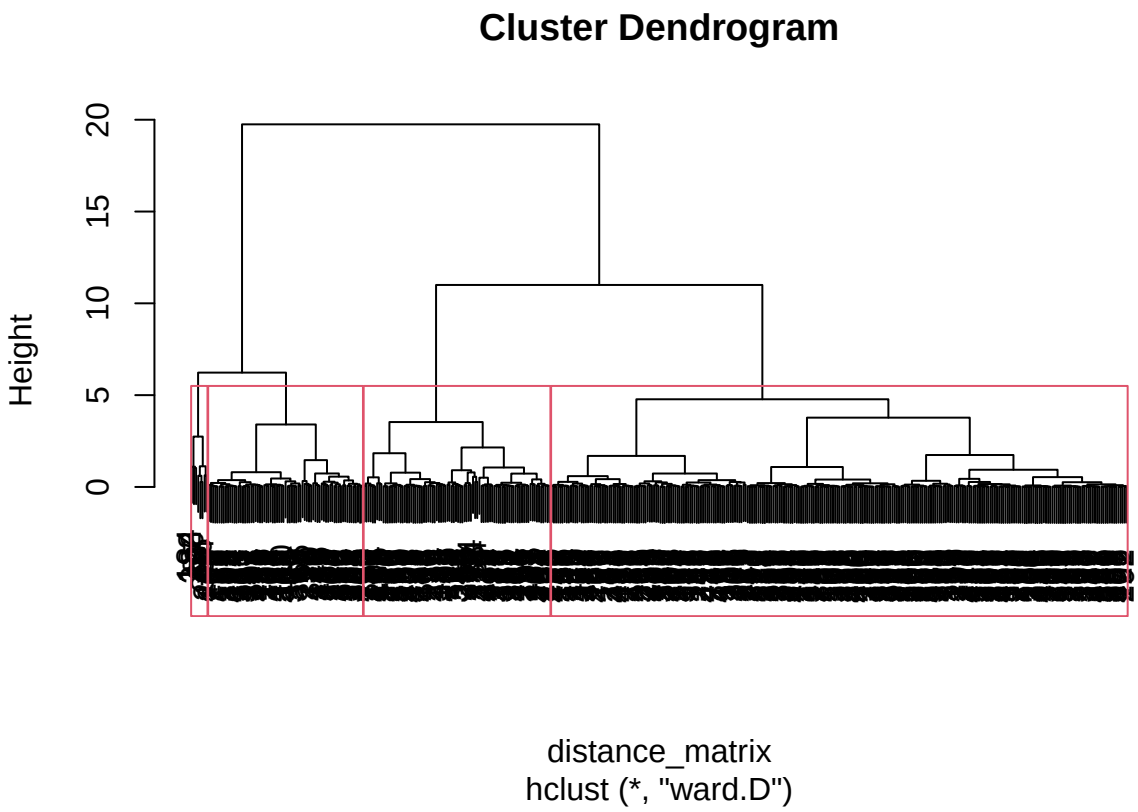
Ward's method

```
hierarchical_ward = hclust(distance_matrix, method = "ward.D")
plot(hierarchical_ward, labels = Wholesale_customers_normalized$Name)
```



k = 4 (other choice?)

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



showing the 4-cluster solution!

```
Wholesale_customers_normalized$cluster = cutree(hierarchical_ward, k = 4)
# let's check out cluster centroids
Wholesale_customers_normalized %>% group_by(cluster) %>%
  summarise_at(c(3:8), mean)
```

```
## # A tibble: 4 x 7
##   cluster  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
```

```
Wholesale_customers_normalized %>% group_by(cluster) %>%
  summarise_at(c(3:8), sum)
```

```
## # A tibble: 4 x 7
##   cluster Fresh  Milk Grocery Frozen Detergents_Paper Delicatessen
##     <int> <dbl> <dbl>   <dbl>  <dbl>            <dbl>        <dbl>
## 1       1 19.7  11.9    12.8    8.62             8.37         5.92
## 2       2 21.4   6.51    5.52   9.65             2.22         4.22
```

```
## 3       3  3.21 11.8    15.5    1.67            16.6          2.25
## 4       4  2.76  4.18    3.87   2.09             3.83          1.58
```

```
write.csv(Wholesale_customers_normalized, "Wholesale_customers_normalized.csv", row.names=FALSE)
```