UCI_ML_Repository_WholeSale_Customers

Thomas Young

2025-05-05

Website: https://archive.ics.uci.edu/dataset/292/wholesale+customers

Only want continous variables for this analysis

```
WHOLE <- read_csv("wholesale.csv")
```

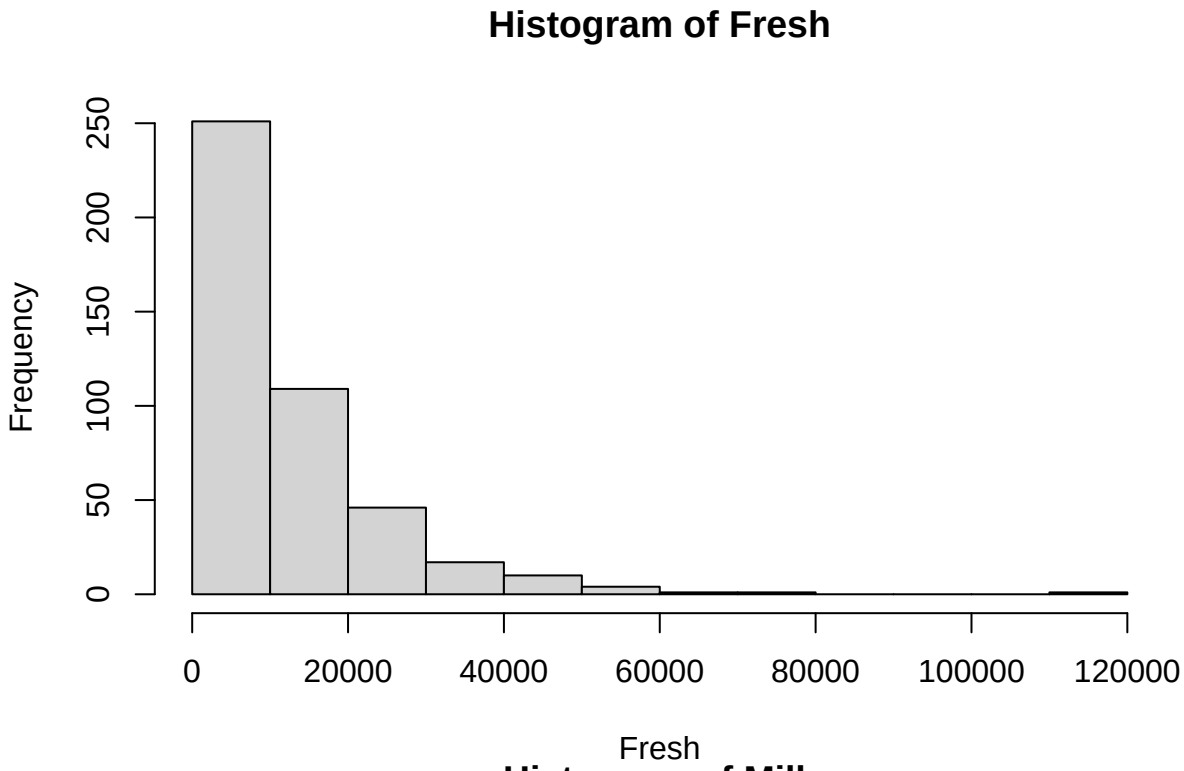
```
## Rows: 440 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (8): Channel, Region, Fresh, Milk, Grocery, Frozen, Detergents_Paper, De...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

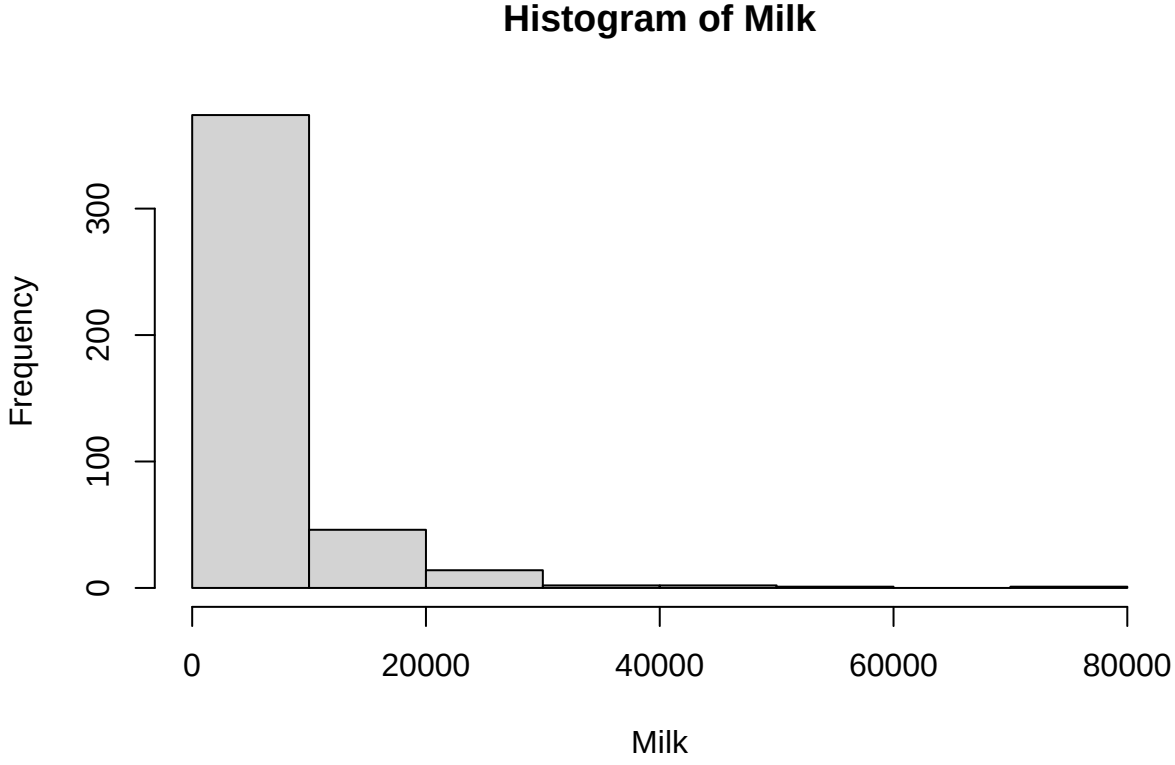
```
WHOLE <- WHOLE |>
  select(-Region,-Channel)
```

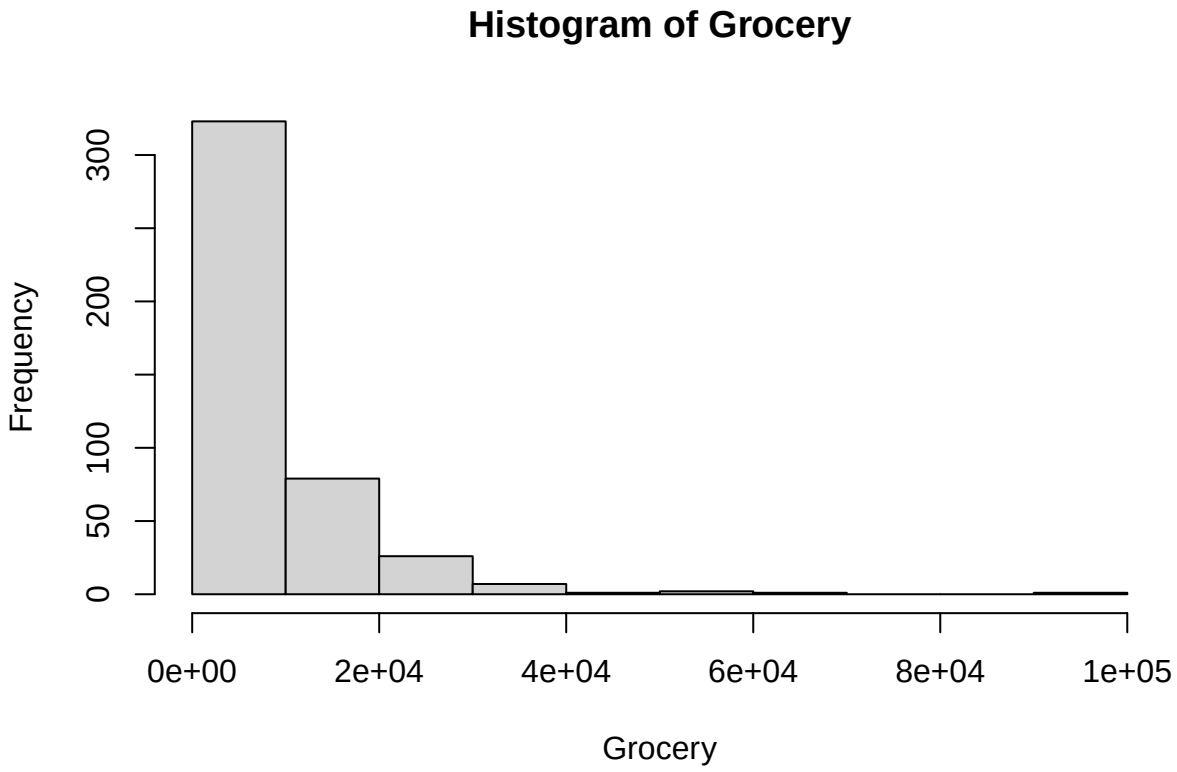
```
# Check for Missing Values
missing_values <- WHOLE |>
  summarise(across(everything(), ~ sum(is.na(.))))

print(missing_values) # No missing values so no imputation
```

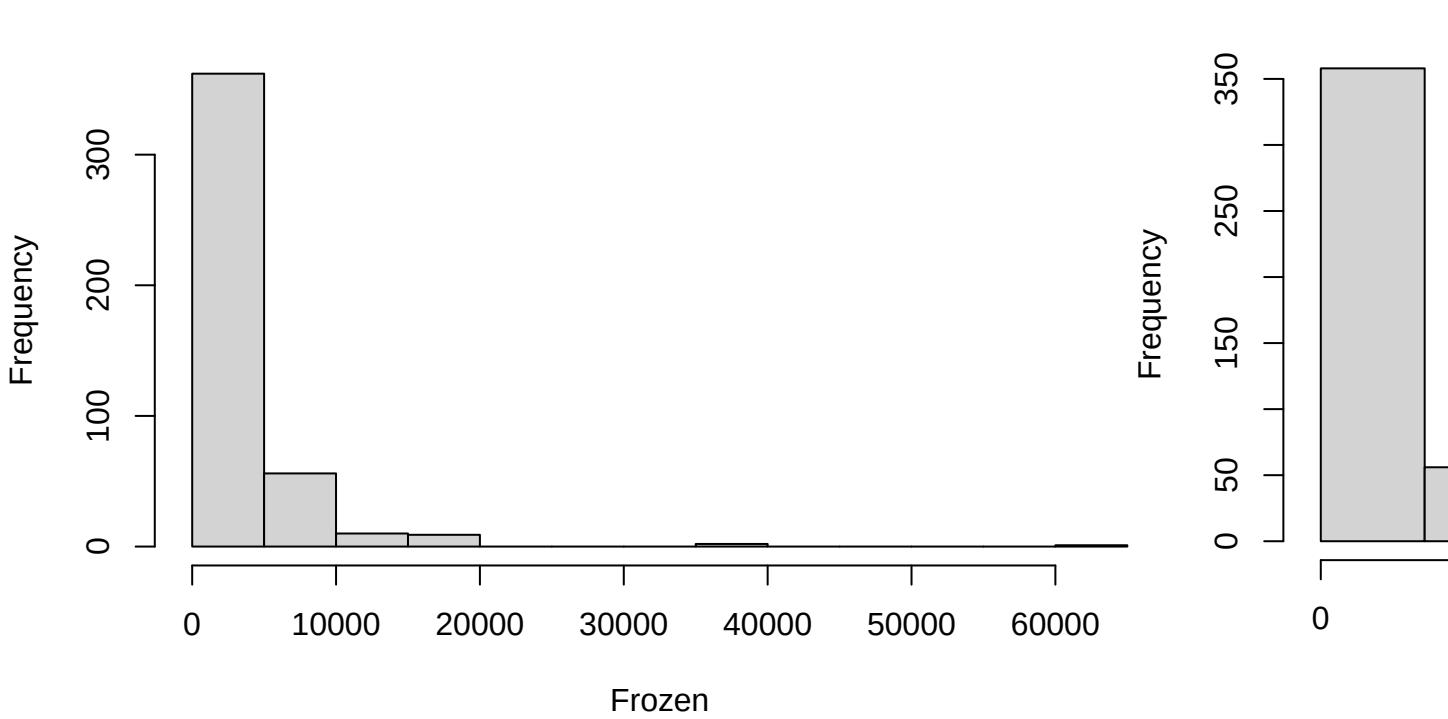
```
## # A tibble: 1 x 6
##   Fresh  Milk Grocery Frozen Detergents_Paper Delicassen
##   <int> <int>   <int>  <int>            <int>      <int>
## 1     0     0       0      0                0          0
```

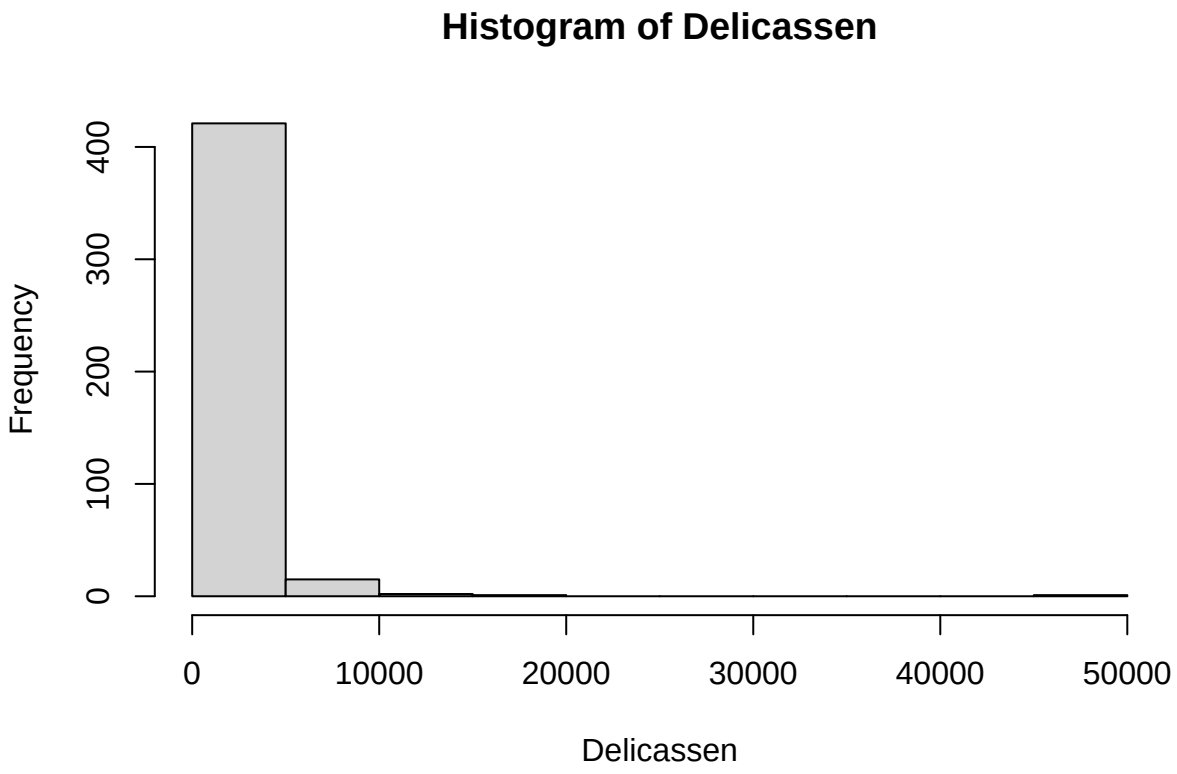
```
numeric_vars <- names(WHOLE)[sapply(WHOLE, is.numeric)]

# Visual INspection
for (var in numeric_vars) {
  hist(WHOLE[[var]], main = paste("Histogram of", var), xlab = var)
}
```


Histogram of Fresh
Frequency
0
50
100
150
200
250
0
20000
40000
60000
80000
100000
120000
Fresh


Histogram of Milk
Frequency
0
100
200
300
0
20000
40000
60000
80000
Milk

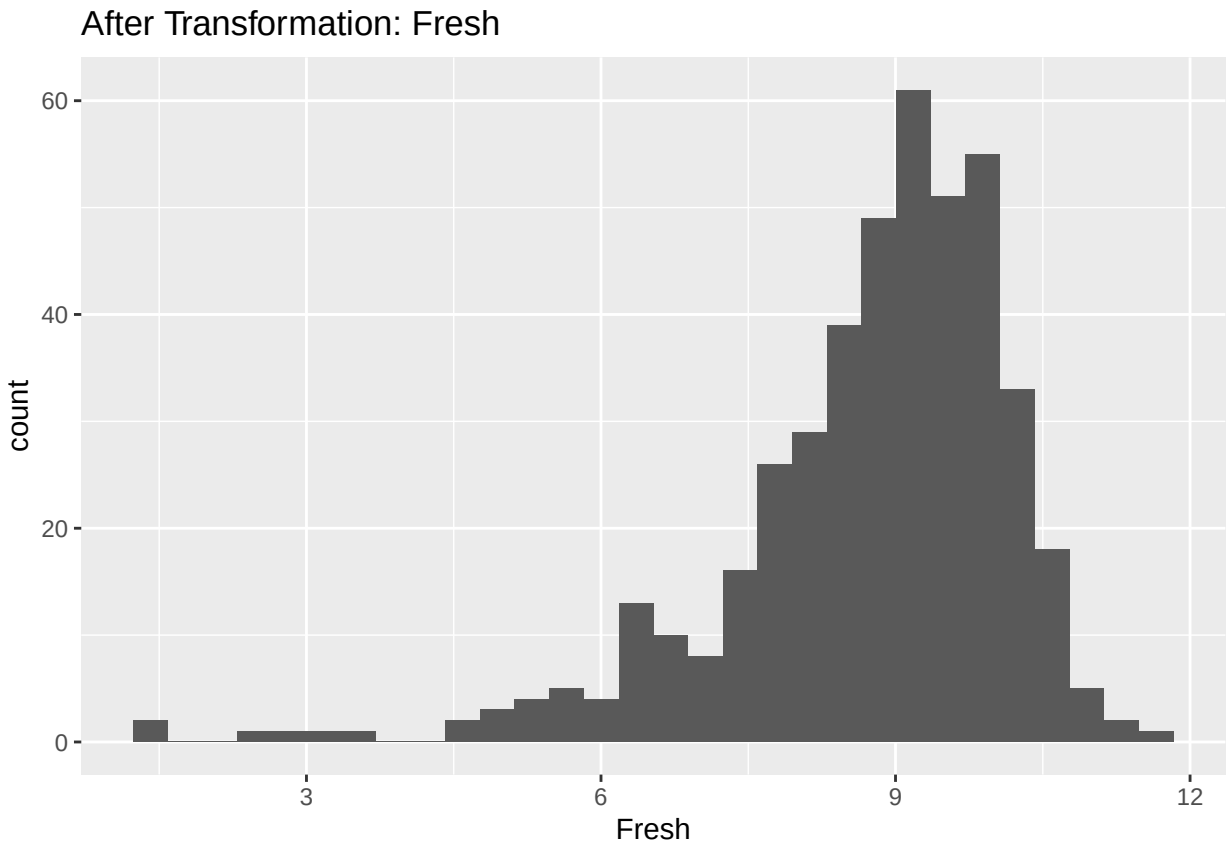

Histogram of Grocery
Frequency
0
50
100
200
300
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Grocery


Histogram of Frozen
Frequency
0
100
200
300
0
10000
20000
30000
40000
50000
60000
Frozen
Frequency
0
50
150
250
350
0

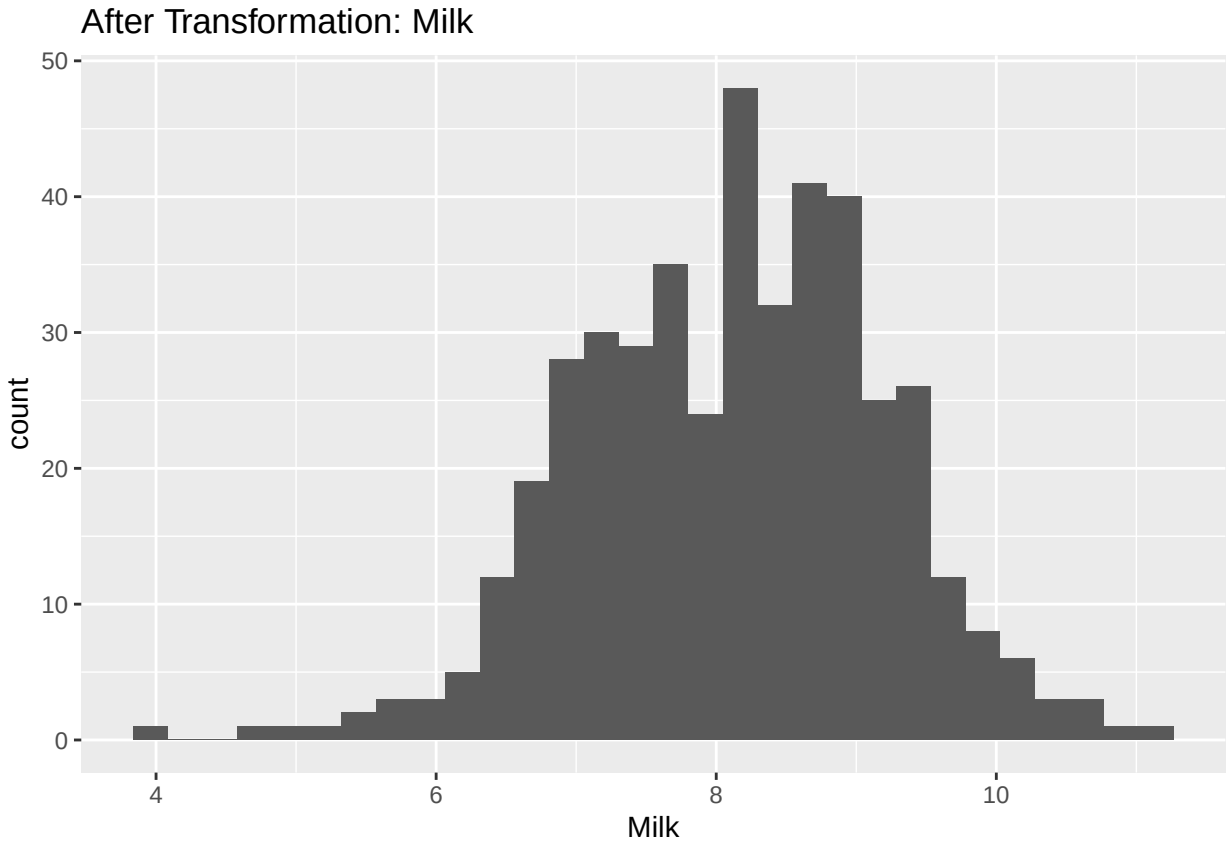



```
# Transform to be centered
DATA.PROCESSED <- WHOLE |>
  mutate(across(where(is.numeric), ~ log(. + 1)))

# Loop for log-transformed histograms
for (var in numeric_vars) {
  print(ggplot(DATA.PROCESSED, aes(x = .data[[var]])) +
          geom_histogram() +
          labs(title = paste("After Transformation:", var)))
}
```

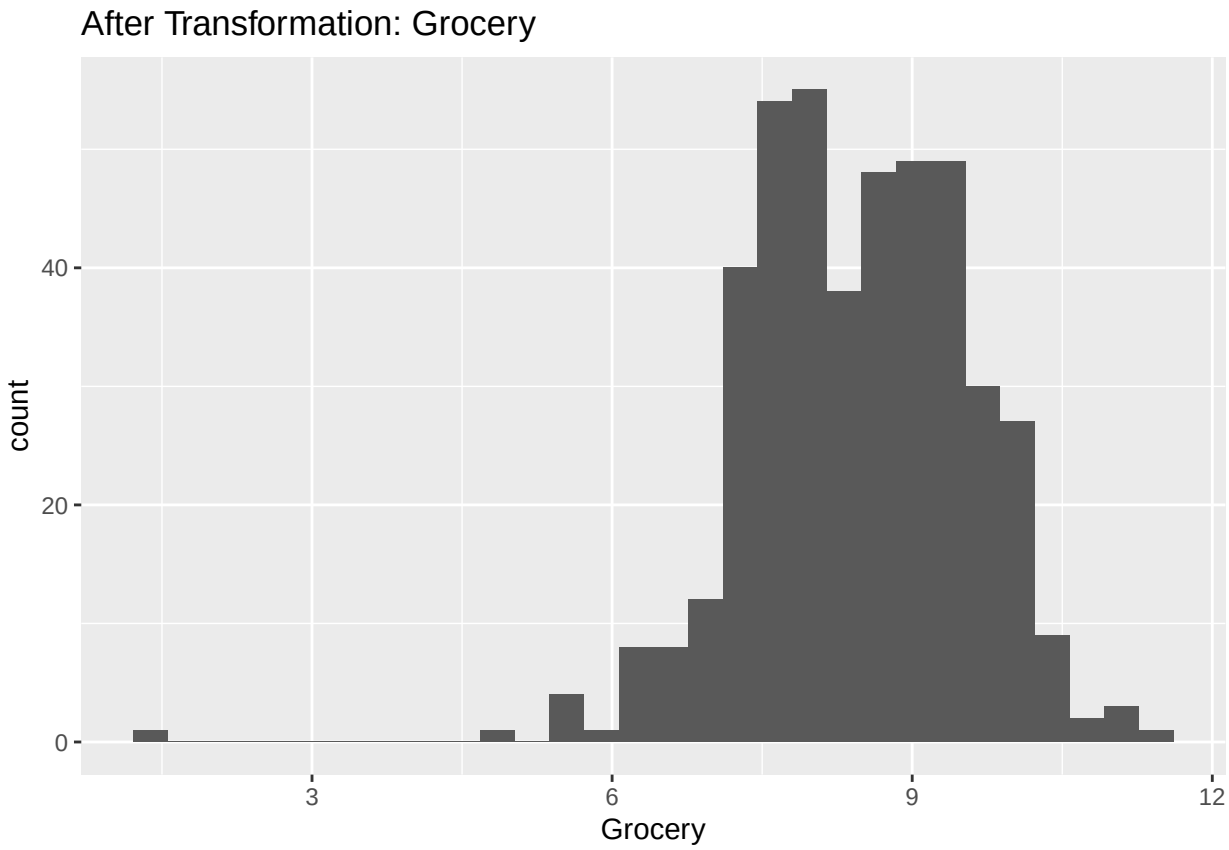
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



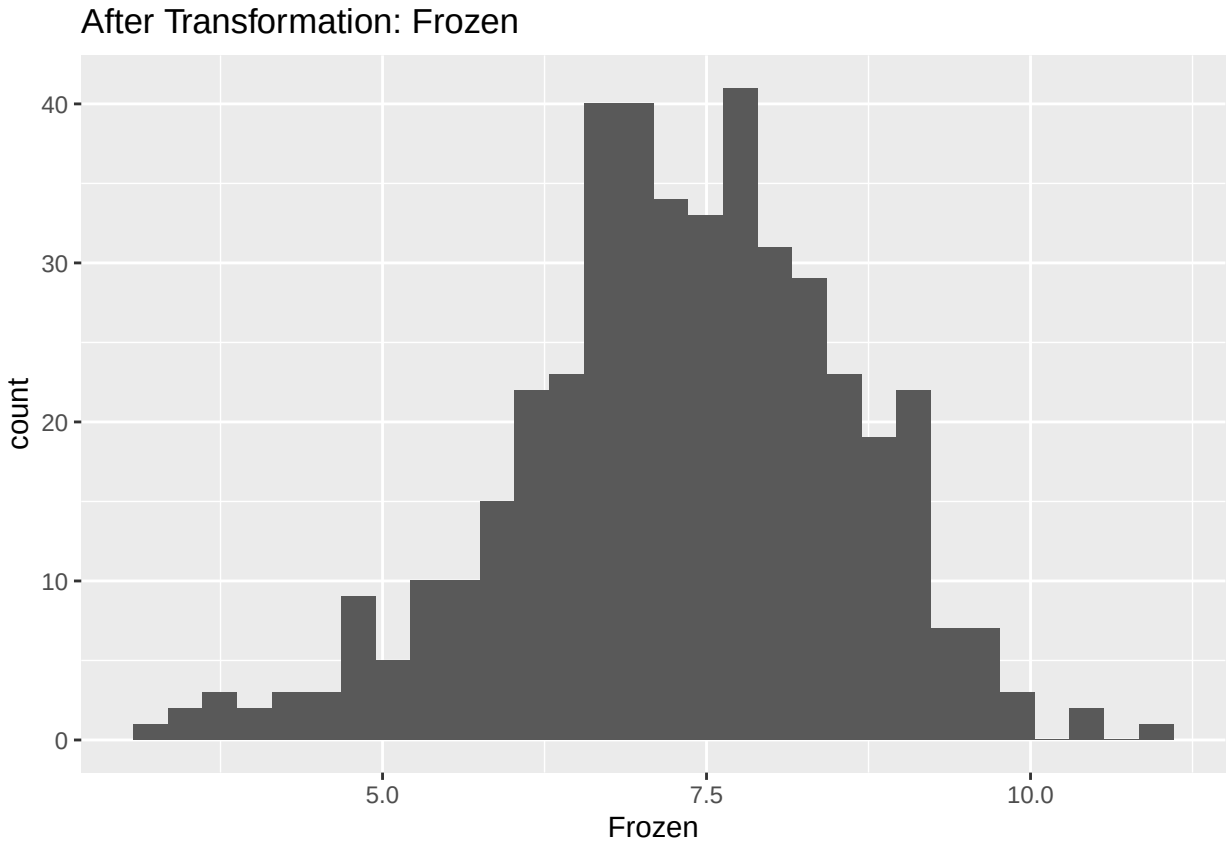

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



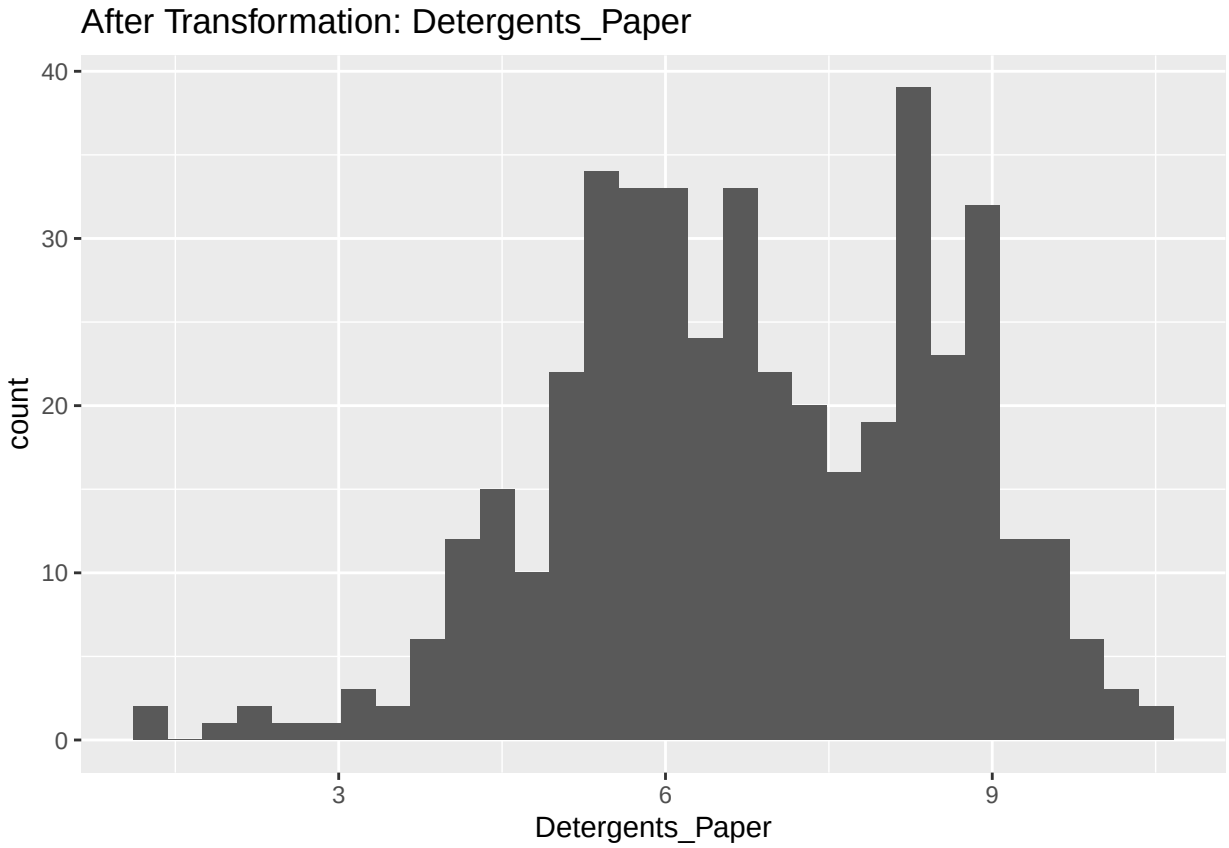

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



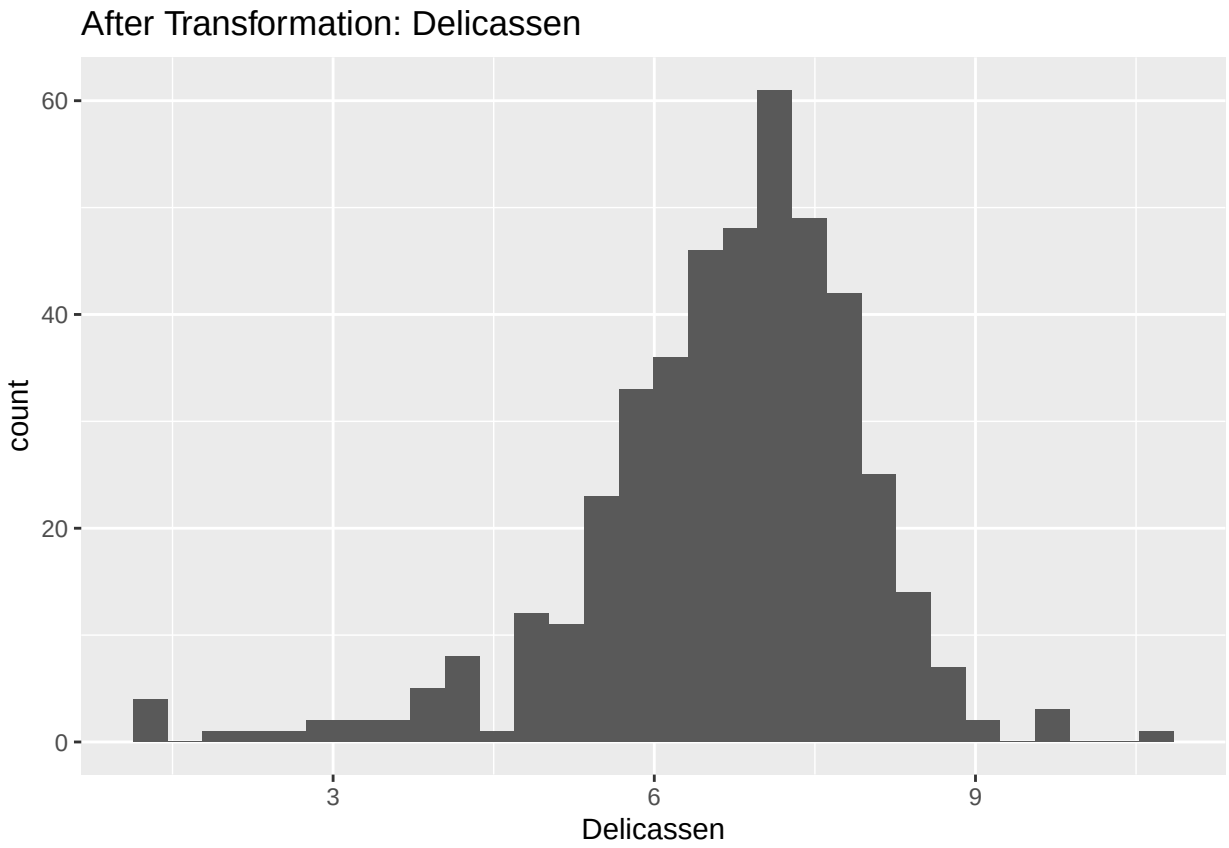

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



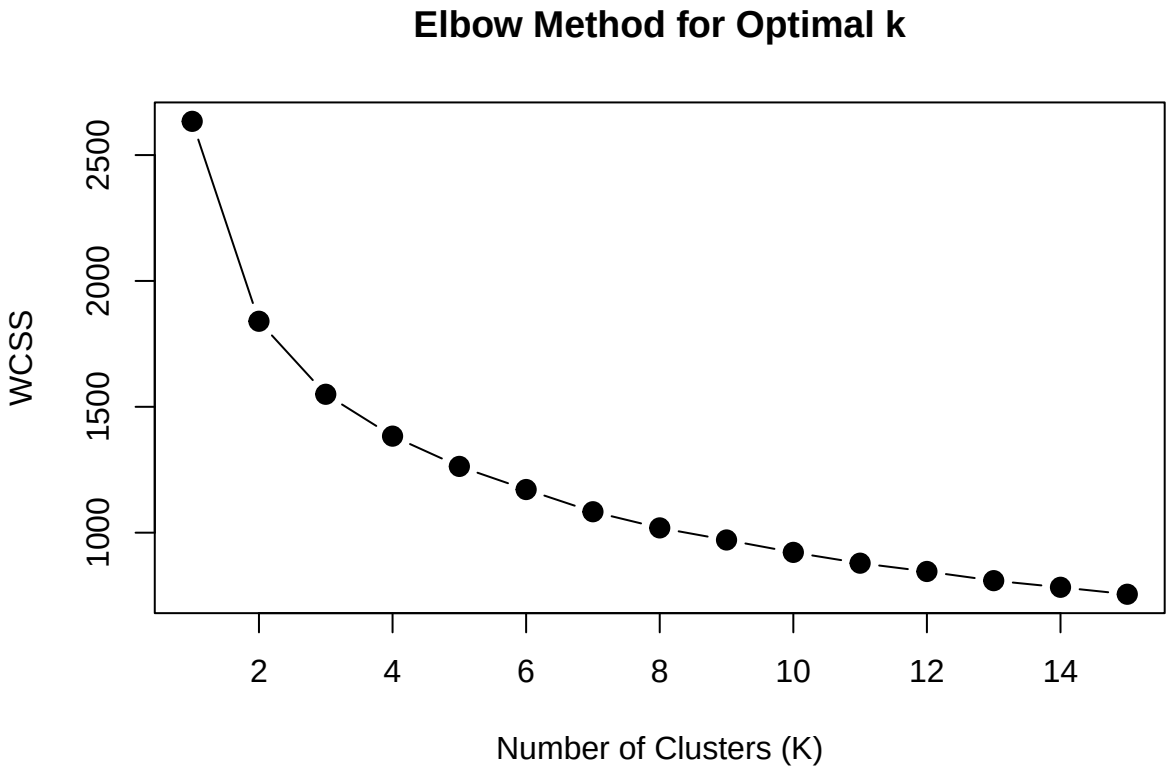

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Scale
DATA.PROCESSED <- DATA.PROCESSED |>
  mutate(across(where(is.numeric), ~ scale(.)[,1]))
```

K Means for cluster analysis

```
WCSS <- c(); possible.k <- 1:15; set.seed(577)
for (i in possible.k) {
  result <- kmeans(DATA.PROCESSED, centers = i, iter.max = 50, nstart = 25)
  WCSS[i] <- result$tot.withinss
}
plot(WCSS ~ possible.k, type = "b", pch = 20, cex = 2,
     xlab = "Number of Clusters (K)", ylab = "WCSS", main = "Elbow Method for Optimal k")
```



Determine with 3 cluster

```
set.seed(577)
KMEANS <- kmeans(DATA.PROCESSED, centers = 3, iter.max = 75, nstart = 50)
table(KMEANS$cluster)
```

```
##
##   1   2   3
## 212 147  81
```

```
round(KMEANS$centers, 2)
```

```
##   Fresh  Milk Grocery Frozen Detergents_Paper Delicassen
## 1  0.17 -0.73   -0.77   0.23            -0.75      -0.21
## 2  0.41  0.80    0.74   0.30             0.66       0.68
## 3 -1.19  0.47    0.67  -1.13             0.76      -0.69
```

There are a few cluster defining characteristics we can look at for the three clusters created.

- Cluster 1 has positive values for Fresh and Frozen products meaning these clients buy more of these items on average while these same clients buy less of Milk, Grocery, Detergents_Paper, and Delicassen compared to the average customer.
- Cluster 2 has positive values for every category indicating these clients spend more than the average consumer at the grocery store, especially for milk and groceries.
- Cluster 3 has large positive values for Milk, Grocery, and Detergents_Paper indicating they spend more than the average consumer, but also has very large negative values for Fresh, Frozen, and Delicassen items indicating significantly lower spending when compared to the average consumer.

Try Again with 4 Clusters

```
set.seed(577)
KMEANS4 <- kmeans(DATA.PROCESSED, centers = 4, iter.max = 75, nstart = 50)
table(KMEANS4$cluster)
```

```
##
##   1   2   3   4
## 130  61 119 130
```

```
round(KMEANS4$centers, 2)
```

```
##   Fresh  Milk Grocery Frozen Detergents_Paper Delicassen
## 1  0.54 -0.08   -0.24   0.75            -0.36       0.45
## 2 -1.31  0.52    0.70  -1.24             0.79      -0.90
## 3  0.12  0.91    0.95  -0.08             0.98       0.56
## 4 -0.04 -1.00   -0.95  -0.09            -0.91      -0.54
```

There are a few cluster defining characteristics we can look at for the four cluster set up. I believe this is a better set up than the 3 cluster method because cluster 4 provides additional information, all negative values, which we do not see represented in the 3 cluster set up.

- Cluster 1 has positive values for Fresh and Frozen products, like before, but also Delicassen meaning these clients buy more of these items on average while these same clients buy less of Milk, Grocery, and Detergents_Paper compared to the average customer.
- Cluster 2 has positive values for Milk, Grocery, and Detergents_Paper indicating these clients spend more than the average consumer for these items, but also has very large negative values for Fresh, Frozen, and Delicassen items indicating significantly lower spending when compared to the average consumer with the overall report looking similar to cluster 3 during the 3 cluster set up.
- Cluster 3 has very positive values for Milk, Grocery, and Detergents_Paper and relatively positive values for Fresh and Delicassen indicating these clients spend more than the average consumer at the grocery store while being slightly below average for Frozen and this cluster has very similar patterns to Cluster 2 from the 3 cluster set up.
- Cluster 4 has very large negative values for Milk, Grocery, Detergents_Paper, and Delicassen and moderatly negative values for Fresh and Frozen items indicating these clients spend less than the average consumer and represents a new cluster that is different from the 3 cluster set up (or could provide more inssights to the analysis).

Maybe check for a 5 cluster to see if it can provide any meaningful insights

```
set.seed(577)
KMEANS5 <- kmeans(DATA.PROCESSED, centers = 5, iter.max = 75, nstart = 50)
table(KMEANS5$cluster)
```

```
##
##   1   2   3   4   5
##  54  88 100 100  98
```

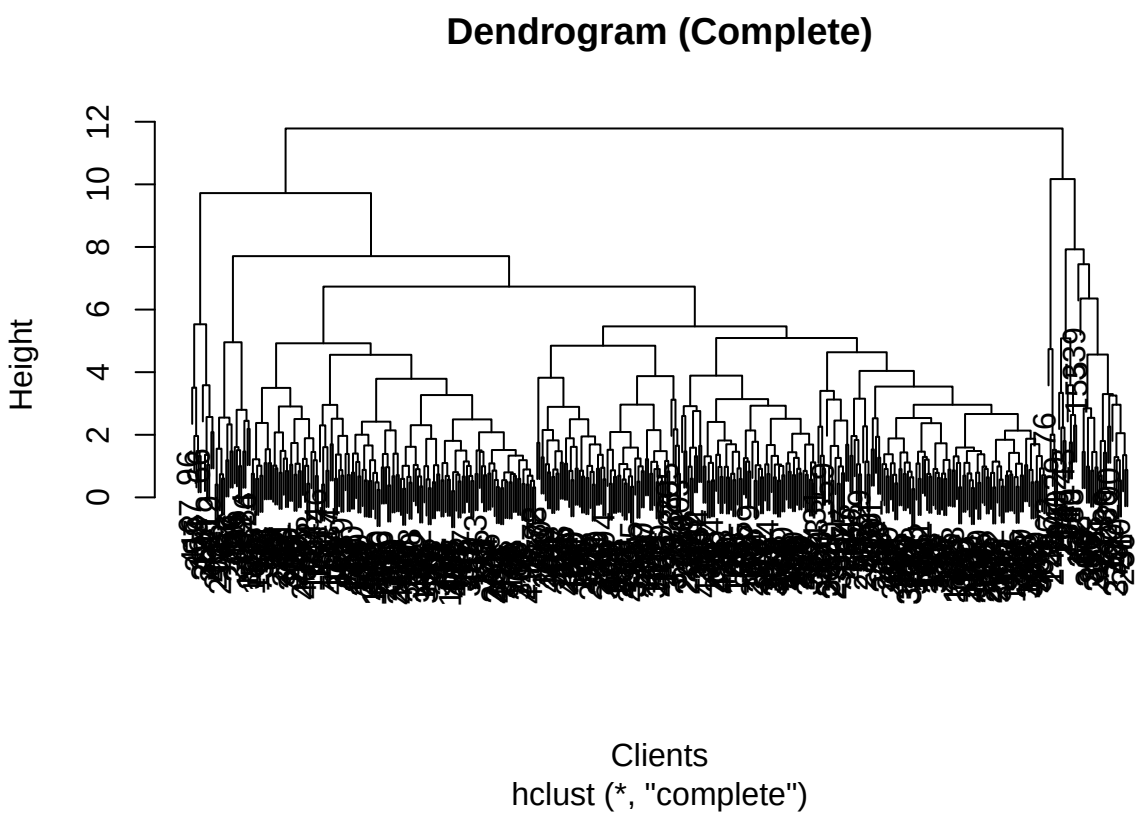
```
round(KMEANS5$centers, 2)
```

```
##   Fresh  Milk Grocery Frozen Detergents_Paper Delicassen
## 1 -1.43  0.54    0.81  -1.29             0.90      -0.94
## 2 -0.05 -1.25   -1.15   0.20            -1.05      -0.75
## 3  0.64 -0.10   -0.25   1.05            -0.37       0.55
## 4  0.11  1.06    1.04   0.00             1.09       0.62
## 5  0.06 -0.15   -0.23  -0.54            -0.28       0.00
```

The 5 cluster layout is a little better dispersion, but cluster 1 has 54 while the rest are around 100. There also exhibits some 0.00 values in cluster 4 and cluster 5 so I would probably want to stick with the 4 cluster layout over this one.

Lets try Hierical Clustering instead

```
HC <- hclust(dist(DATA.PROCESSED),method="complete")
plot(HC,main="Dendrogram (Complete)",xlab="Clients")
```




```
unusual_client_index <- which.max(HC$height)
print(DATA.PROCESSED[unusual_client_index, ])
```

```
## # A tibble: 1 x 6
##   Fresh   Milk Grocery Frozen Detergents_Paper Delicassen
##   <dbl>  <dbl>   <dbl>  <dbl>            <dbl>      <dbl>
## 1 0.344 -0.490  -0.658 -0.279           -0.972      0.766
```

```r
summary(DATA.PROCESSED)
```

```
##      Fresh              Milk             Grocery            Frozen
##  Min.   :-4.9955   Min.   :-3.79061   Min.   :-6.34796   Min.   :-3.15553
##  1st Qu.:-0.4654   1st Qu.:-0.72733   1st Qu.:-0.69016   1st Qu.:-0.53991
##  Median : 0.2146   Median : 0.06924   Median : 0.02255   Median : 0.02178
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.6829   3rd Qu.: 0.70237   3rd Qu.: 0.74829   3rd Qu.: 0.68107
##  Max.   : 1.9684   Max.   : 2.85333   Max.   : 2.69521   Max.   : 2.89680
##  Detergents_Paper     Delicassen
##  Min.   :-3.16199   Min.   :-4.0842
##  1st Qu.:-0.72523   1st Qu.:-0.5076
##  Median :-0.05004   Median : 0.1566
##  Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.86739   3rd Qu.: 0.6462
##  Max.   : 2.23767   Max.   : 3.1737
```

One of the first clients to split off was a client with Fresh as 0.34 which is close to the median in the summary statistics so not unusual. Milk (-0.49), Grocery (-0.65), and Frozen (-0.27) were both a little lower than average, but within the interquartile range. Detergents_Paper (-0.97) was unusual as its between the first quartile and minimum meaning this client spends significantly less on detergents and paper than most other clients. The last variable Delicassen (0.76) is slightly above average, but not that unusual.

Compared to K Means

```r
HC_ward <- hclust(dist(DATA.PROCESSED),method ="ward.D2")
cluster_assignments <- cutree(HC_ward, k = 4)
DATA.PROCESSED$cluster <- cluster_assignments

# Calculate mean values
aggregate(DATA.PROCESSED[, -ncol(DATA.PROCESSED)],
          by = list(cluster = DATA.PROCESSED$cluster), FUN = mean)
```

```
##   cluster      Fresh        Milk    Grocery      Frozen Detergents_Paper
## 1       1 -0.1785254  0.90309094  0.9772447 -0.36353368        1.0742503
## 2       2  0.7773502  0.33912188  0.1199850  1.36890155       -0.2300783
## 3       3  0.2109613 -0.62094528 -0.6637142  0.05867984       -0.6382889
## 4       4 -1.1988901 -0.03853453  0.1776938 -0.76762980        0.2057931
##   Delicassen
## 1  0.5607234
## 2  0.8953337
## 3 -0.1699890
## 4 -1.5675609
```

```r
# Add cluster assignments
WHOLE$cluster <- cluster_assignments

# Calculate median values
aggregate(WHOLE[, -ncol(WHOLE)], by = list(cluster = WHOLE$cluster), FUN = median)
```

```
##   cluster   Fresh   Milk Grocery Frozen Detergents_Paper Delicassen
## 1       1  5981.0 8259.0 12469.0   1148           5618.0     1682.0
## 2       2 23123.5 4674.5  4961.5   8151            666.5     2420.5
## 3       3  9581.5 1813.0  2256.5   1662            310.5      691.0
## 4       4  1689.0 4230.0  7854.0    661           2970.0      156.0
```

- Cluster 1 is most similar to the kmeans cluster 3 by having very similar values for Milk, Grocery, Detergents_Paper, and Delicassen.
- Cluster 2 is most similar to the kmeans cluster 1 by sharing high values for Fresh, Frozen, and Delicassen with negative values for Detergents_Paper. Grocery and Milk have opposite signs.
- Cluster 3 is most similar to the kmeans cluster 4 by having negative values for all with the exception of hierarchical clustering having a slight positive value for Fresh and Frozen.
- Cluster 4 is most similar to the kmeans cluster 2 by having very low values for Fresh, Frozen, and Delicassen with moderate values for Grocery and Detergents_Paper