Unsupervised Learning Assignment

Yanshen Wang

March 2023

1 Exploratory data analysis

The wholesale dataset includes the annual spending in monetary units (m.u.) on diverse product categories for 440 clients of a wholesale distributor. We will perform dimension reduction and cluster analysis on this data, but before that, we need to conduct an exploratory data analysis.

Firstly, we need to check whether there are any missing values in the dataset. Secondly, we need to check for the presence of outliers. The most common visualization method for detecting outliers is boxplot. Therefore, we will use boxplot to identify the outliers.

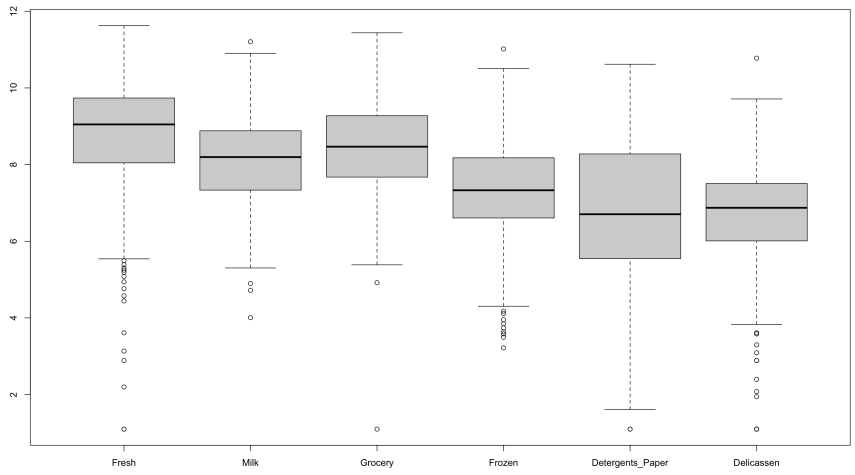


Figure 1: Boxplot of Wholesale data

After checking, there are no missing values in the dataset. However, boxplots show that all categories have outliers, with Fresh and Delicassen having the most outliers and the largest range. Frozen also has many outliers, but with a smaller range. In contrast, Milk, Grocery, and Detergents_Paper have fewer outliers, with Grocery having an extremely small minimum value.

Since the dataset is not large, processing outliers may affect the accuracy of the results. Therefore, we do not process the outliers. We prioritize the use of analysis methods that are not sensitive to outliers. If, during the analysis process, we find that outliers are affecting the analysis, we will adjust the analysis parameters according to the actual situation.

From the pairs plot, we can see that there are high correlations (correlation=0.76, 0.68) between Milk and Grocery, Detergents_Paper, respectively. Also, there is a high correlation (correlation=0.80) between Grocery and Detergents_Paper. This indicates the existence of redundant information among Milk, Grocery, and Detergents_Paper. PCA can remove this redundancy by reducing the dimensionality of the data.

```
## sd
## Fresh       Milk         Grocery     Frozen    Detergents_Paper    Delicassen
## 1.480071    1.081365     1.116172    1.284540      1.721020       1.310832
```

By assessing the standard deviation of the variables, it is important to note that for this dataset, data normalization should be performed before data processing to scale the data to a consistent level.

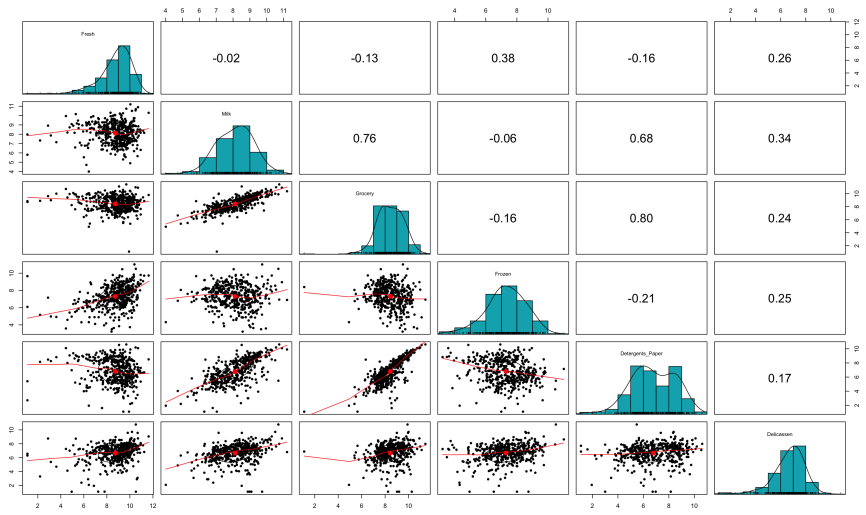


Figure 2: Panel pairs of Wholesale data

2 Dimension reduction

As we discovered in exploratory data analysis, there are high correlations among Milk Grocery, and Detergents_Paper, which suggests that there might be redundant information among the three variables, which could interfere with data analysis and modeling. By reducing dimensions, we can map high-dimensional data to a lower-dimensional space while preserving the key information of the original data, thus reducing noise and redundant information and making data analysis and modeling more convenient and effective, and dimension reduction can help us better understand and interpret data, discover latent structures and patterns, and extract valuable information from them.

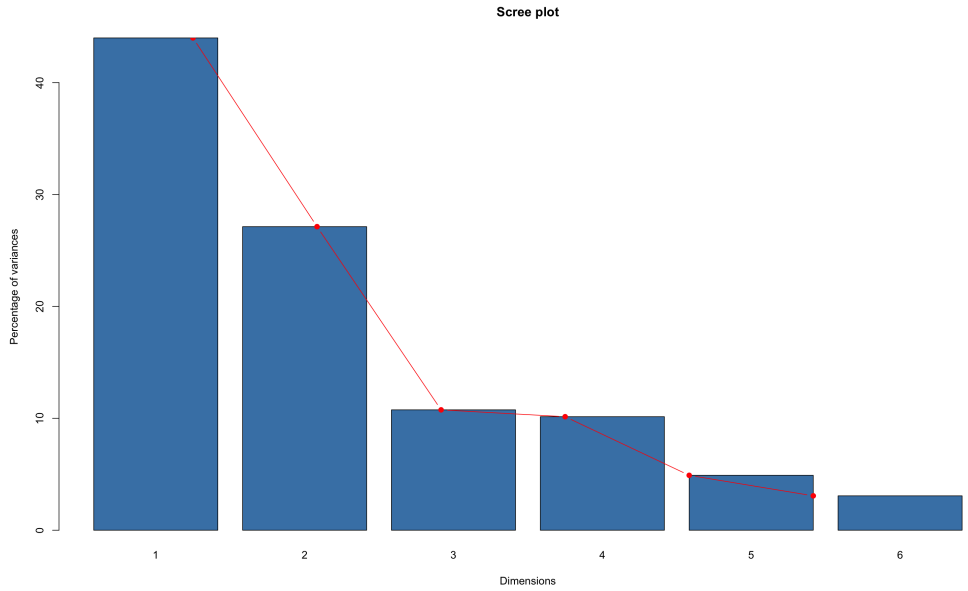
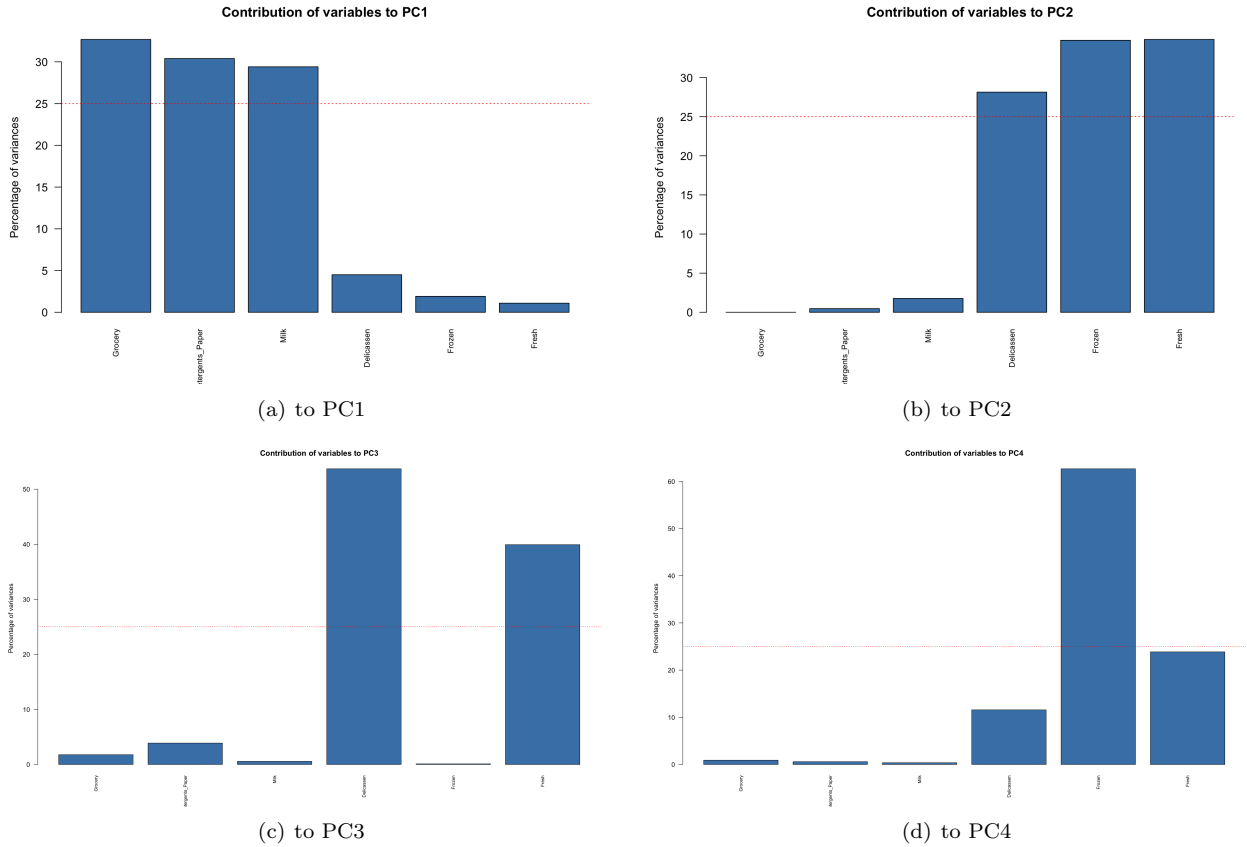


Figure 3: Scree plot

```
> eig2
        eig  variance cumvariance
1 2.6394546 43.990911    43.99091
2 1.6276597 27.127662    71.11857
3 0.6454053 10.756755    81.87533
4 0.6086835 10.144726    92.02005
5 0.2944039  4.906732    96.92679
6 0.1843929  3.073214   100.00000
```

Based on both the proportion of variance explained and the scree plot, we have four dimensions that cumulatively explain about 92% of the variability in the data. This value is well beyond the rule of thumb which suggested retaining components that explain 80-90% of the variability in the original data set.

```
> contrib
                         PC1          PC2         PC3         PC4
Fresh              1.094673 34.865937036 39.9289508 23.8656918
Milk              29.406125  1.772769023  0.5787568  0.3768231
Grocery           32.683405  0.003946809  1.7808851  0.9153050
Frozen             1.914087 34.755143806  0.1130986 62.6644648
Detergents_Paper  30.397337  0.470928598  3.8910883  0.5982610
Delicassen         4.504372 28.131274728 53.7072203 11.5794544
```



(a) to PC1

(b) to PC2

(c) to PC3

(d) to PC4

Figure 4: Contribution of variables

According to contrib and figure4, we can see that Grocery, Detergent_Paper, and Milk make the most significant contributions to PC1, while Fresh, Frozen, and Delicassen make the most significant contributions

to PC2. Delicassen and Fresh have the most significant contributions to PC3, while Frozen has the most significant contribution to PC4.

3 Cluster analysis

Clustering can discover potential structures and patterns in the dataset, providing a better understanding of the dataset while reducing the representation. By discovering and removing the redundancies among different products in the wholesale dataset, we can perform better modeling and analysis, cluster similar customers together, discover their needs and behaviors, and provide targeted marketing strategies and services.

As we discovered in the exploratory data analysis, there are many outliers in this dataset. Therefore, when performing clustering analysis, we prefer to choose the clustering method that is less sensitive to outliers and has higher robustness, which is K-medoids. Additionally, hierarchical clustering has the advantage of not requiring the pre-specification of the number of clusters K. Therefore, in this case, we choose to use both K-medoids and Hierarchical clustering methods for analysis.

3.1 K-medoids

Firstly, we use the silhouette measure to determine optimal number of clusters for `wholesale` dataset using the `K-medoids` algorithm.

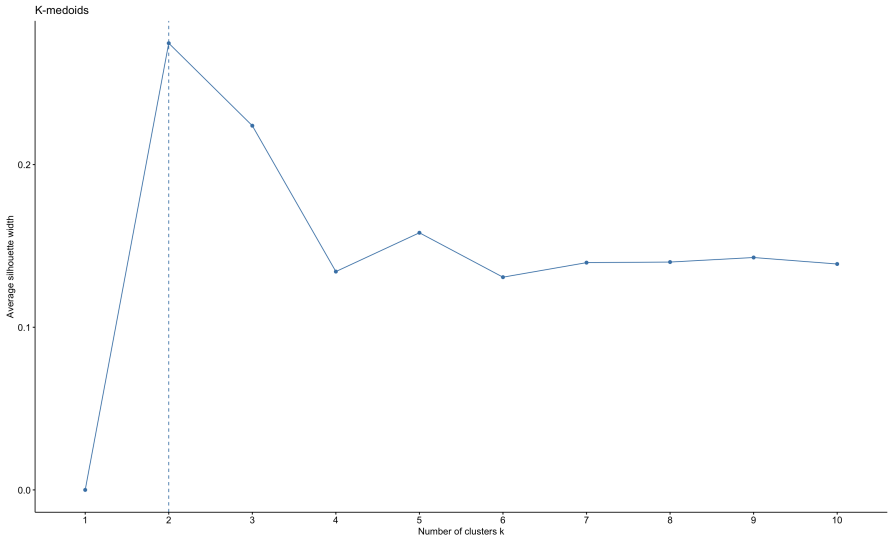


Figure 5: Optimal number of cluster

Based on the silhouette measure results, we perform K-medoids clustering analysis with $k = 2$. The results are shown in Figure 6. However, we can observe that the lower half of the right classification seems to have a tendency to cluster together. Therefore, we attempt to perform the clustering analysis again with $k = 3$, and the results are shown in Figure 7. As we predicted, $k = 3$ appears to be more reasonable.

For the selection of the K value, the results of PCA can also be used as a reference. Therefore, ultimately, we choose $k = 3$.

According to the results, the percentage for Dimension1 is 44%, and the percentage for Dimension2 is 27.1%. Dimension1 and Dimension2 together can explain 71.1% of the total variance, which does not seem to be very high.

Additionally, the cluster plot shows a considerable overlap among the three clusters, which could be due to the data itself being close or overlapping, unclear cluster boundaries, or other reasons. These factors require further investigation and analysis.

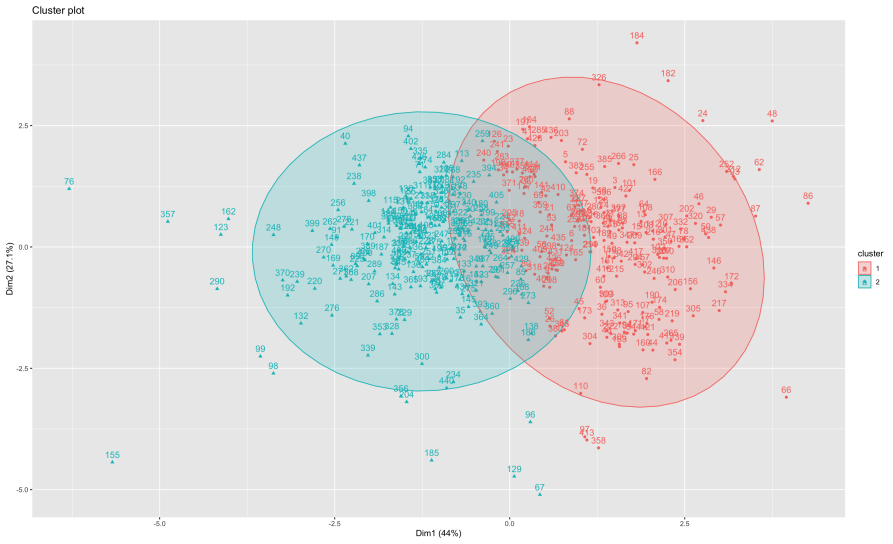


Figure 6: Cluster plot $k = 2$

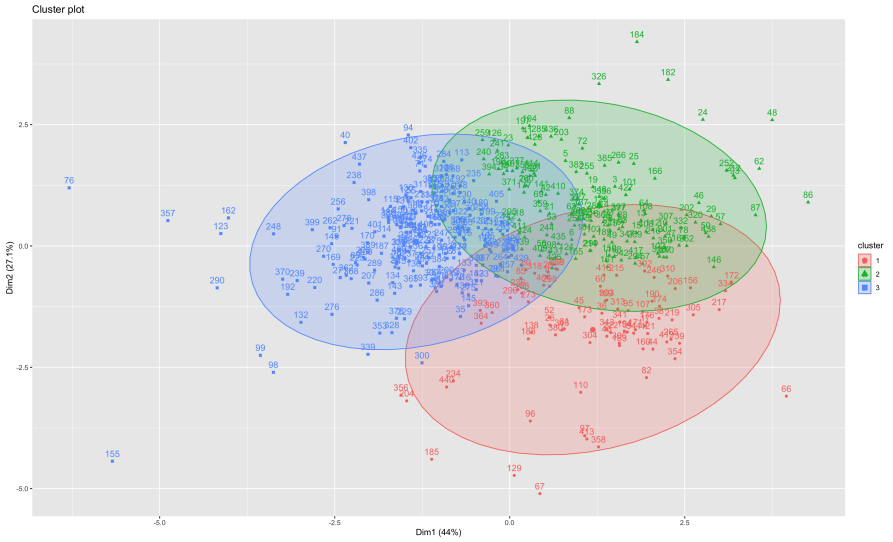


Figure 7: Cluster plot $k = 3$

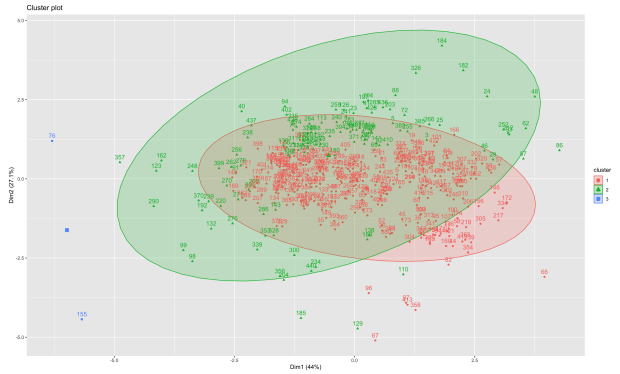
3.2 Hierarchical clustering

As mentioned earlier, Hierarchical clustering does not require pre-specifying the number of clusters, making it an alternative clustering method to K-means.

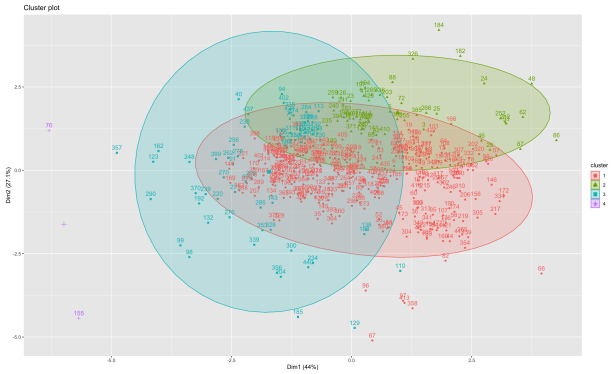
Due to the presence of some outliers in the data, we choose the complete linkage method here, as it has a lower sensitivity to noise and outliers. Referring to the clustering results from PCA and K-medoids, we set the value of k to 3 to cut the data into three groups.

From the k=3 result in Figure 8, the data is actually divided into two clusters, with the other one cluster containing only two data points. This is very likely due to the influence of outliers. Since the dataset is relatively small and outliers have not been removed, we will adjust the parameters by setting k to 4. This way, we can cluster the data without the outliers into 3 clusters.

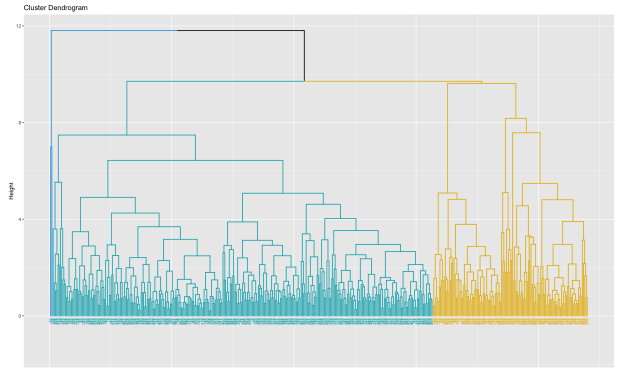
From the results in Figure 8, it can be seen that removing the two outliers and dividing the data into three clusters is quite reasonable, and it has some similarity with the K-medoids results. Both the Dendrogram and Phylogenetic-like Tree clearly show the trend of the data being divided into three clusters (excluding the cluster with outliers).



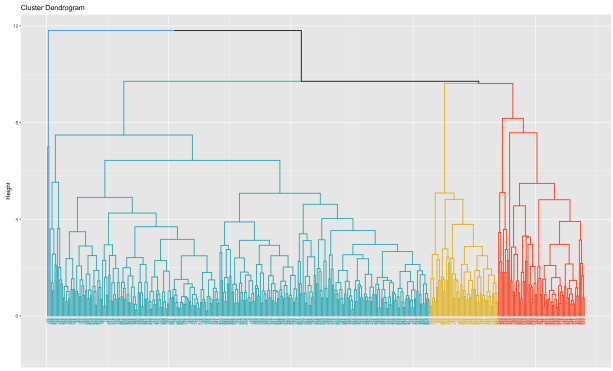
(a) Cluster Plot $k = 3$



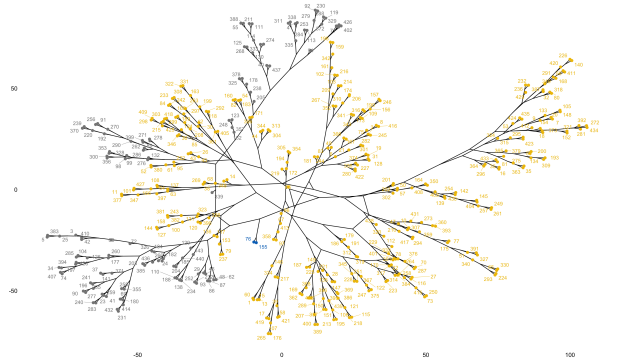
(b) Cluster Plot $k = 4$



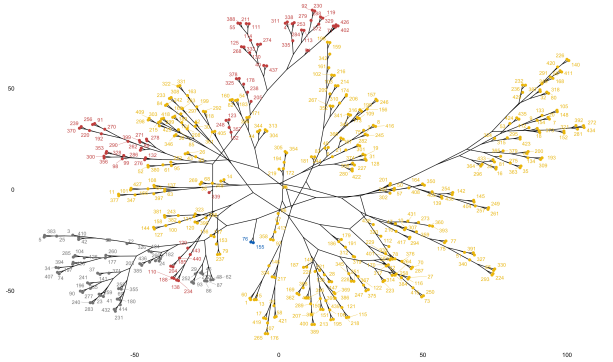
(c) Cluster Dendrogram $k = 3$



(d) Cluster Dendrogram $k = 4$



(e) Phylogenic-like Tree $k = 3$



(f) Phylogenic-like Tree $k = 4$

Figure 8: Hierarchical clustering plots

4 Discussion and conclusion

K-medoids is more robust to outliers and noise because it uses the center of the data points (medoid) as the cluster center. The results are easy to interpret, as the cluster centers are actual data points. However, it requires pre-specification of the number of clusters, has a relatively high computational complexity, may be inefficient for large datasets, and the results may be influenced by initial values, possibly requiring multiple runs to obtain the best results. On the other hand, Hierarchical clustering does not require pre-specifying the number of clusters, allows for visualization of the clustering hierarchy through a dendrogram, facilitating understanding of the data structure, and the results are not affected by initial values, providing higher certainty. However, it is sensitive to outliers and noise, and even when using the outlier-insensitive complete linkage, as we can see from the results, the influence of outliers is still considerable.

Since the dataset is unlabeled, we cannot assess cluster accuracy using known labels. Comparing cluster

plots, we can see that Dimension1 and Dimension2 explain 71.1% of the total variance. The clusters generated by K-medoids appear to have smaller boundaries, with each cluster encompassing fewer data points than the clusters formed by Hierarchical clustering. This observation suggests that K-medoids might provide a more distinct separation between clusters compared to Hierarchical clustering in this particular dataset. The overlap between the three clusters in the Hierarchical clustering is also greater than that in the K-medoids clustering. This observation further seems to support the notion that K-medoids may provide a clearer separation between clusters compared to Hierarchical clustering for this specific dataset.

For Hierarchical clustering, we can easily observe the hierarchy and similarity between data points from the Dendrogram and Phylogenetic-like Tree.

Due to the presence of outliers in this dataset, Kmedoids has higher robustness to outliers and seems to be more suitable for clustering this data.

A Appendix

```
# Seciton 1 EDA
wholesale = read.csv('~/Documents/DDS/DS/DVUL/Assignment2/wholesale.csv')
View(wholesale)
summary(wholesale)
boxplot(wholesale)
sapply(wholesale, function(x) sum(is.na(x)))
library(psych)
pairs.panels(wholesale,
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             ellipses = TRUE # show correlation ellipses
)

sd <- apply(wholesale, 2, sd)
sd
```

```
# Section 2 PCA
S <- cor(wholesale)
eigdec <- eigen(S)
eigdec

eig <- eigdec$values
eig
sum(eig)

pc_loading <- eigdec$vectors
rownames(pc_loading) <- colnames(wholesale)
pc_loading

# Variances in percentage
eig <- eigdec$values
variance <- eig*100/sum(eig)
# Cumulative variances
cumvar <- cumsum(variance)
eig2 <- data.frame(eig = eig, variance = variance,
                   cumvariance = cumvar)
eig2

barplot(eig2[, 2], names.arg=1:nrow(eig2),
        main = "Scree␣plot",
        xlab = "Dimensions",
        ylab = "Percentage␣of␣variances",
        col ="steelblue")
# Add connected line segments to the plot
lines(x = 1:nrow(eig2), eig2[, 2],
      type="b", pch=19, col = "red")

pc_score <- as.matrix(scale(wholesale))%*% pc_loading
colnames(pc_score) <- paste0("PC", 1:6)
pc_score[1:6,]

library(psych)
pairs.panels(pc_score,
```

```
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             ellipses = TRUE # show correlation ellipses
)

cor(wholesale, pc_score[,1:4])
t(t(pc_loading)*sqrt(eig))

cos2 <- (cor(wholesale, pc_score[,1:4]))^2
cos2

comp.cos2 <- apply(cos2, 2, sum)
comp.cos2 # same as the corresponding eigenvalues

contrib2 <- function(cos2, comp.cos2){cos2*100/comp.cos2}
contrib <- t(apply(cos2,1, contrib2, comp.cos2))
contrib

names1 = c("Grocery", "Detergents_Paper","Milk",
"Delicassen","Frozen","Fresh")
barplot(contrib[order(contrib[, 1],decreasing = T), 1], names.arg=names1,
        main = "Contribution␣of␣variables␣to␣PC1",
        xlab = "␣",
        ylab = "Percentage␣of␣variances",
        col ="steelblue",las=2,cex.names=0.7)
abline(h=25, col="red",lty=3, lwd =1)

barplot(contrib[order(contrib[, 1],decreasing = T), 2], names.arg=names1,
        main = "Contribution␣of␣variables␣to␣PC2",
        xlab = "␣",
        ylab = "Percentage␣of␣variances",
        col ="steelblue",las=2,cex.names=0.7)
abline(h=25, col="red",lty=3, lwd =1)

barplot(contrib[order(contrib[, 1],decreasing = T), 3], names.arg=names1,
        main = "Contribution␣of␣variables␣to␣PC3",
        xlab = "␣",
        ylab = "Percentage␣of␣variances",
        col ="steelblue",las=2,cex.names=0.7)
abline(h=25, col="red",lty=3, lwd =1)

barplot(contrib[order(contrib[, 1],decreasing = T), 4], names.arg=names1,
        main = "Contribution␣of␣variables␣to␣PC4",
        xlab = "␣",
        ylab = "Percentage␣of␣variances",
        col ="steelblue",las=2,cex.names=0.7)
abline(h=25, col="red",lty=3, lwd =1)

ws.scaled <- scale(wholesale)
```

```
# Section 3 K-medoids
fviz_nbclust(ws.scaled, clara, method = "silhouette")+
  labs(title = "K-medoids")
```

```
library(cluster)
# k=2
pam.res <- clara(ws.scaled, k=2)
names(pam.res)

pam.res$medoids
pam.res$i.med
fviz_cluster(pam.res, wholesale, ellipse.type = "norm")

#k=3
pam.res <- clara(ws.scaled, k=3)
names(pam.res)

pam.res$medoids
pam.res$i.med
fviz_cluster(pam.res, wholesale, ellipse.type = "norm")
```

```
#Section 4 Hierarchical clustering
# Compute distances and hierarchical clustering
# k = 4
dd <- dist(ws.scaled, method = "euclidean")
hc <- hclust(dd, method = "complete")
hc

dendros <- as.dendrogram(hc)
plot(dendros, main = "Wholesale␣data␣-␣Complete␣linkage",
     ylab = "Height")
abline(h=0.2, lty = 2, col="red")
abline(h=0.428, lty = 2, col="blue")#point 116 and 142 are merged
Hs <- hc$height[(length(hc$height)-4):length(hc$height)]
abline(h=Hs, col=3, lty=2)

fviz_cluster(list(data=ws.scaled, cluster=cutree(hc, 4)),
 ellipse.type = "norm")

fviz_dend(hc, k = 4, # Cut in three groups
          cex = 0.5, # label size
          k_colors = c("#2E9FDF", "#00AFBB", "#E7B800","#FF4500"),
          color_labels_by_k = TRUE, # color labels by groups
          ggtheme = theme_gray() # Change theme
)

hcut <- cutree(hc, k = 4)
table(hcut)

# install.packages("igraph")
library(igraph)
fviz_dend(hc, k = 4, k_colors = "jco", type = "phylogenic", repel = TRUE)
```