Group Homework 1

Group Members: Aastha
Ameer Akhtar
Anantha Narayanan Balaji
Ko Jen Kang

Importing required libraries for the analysis

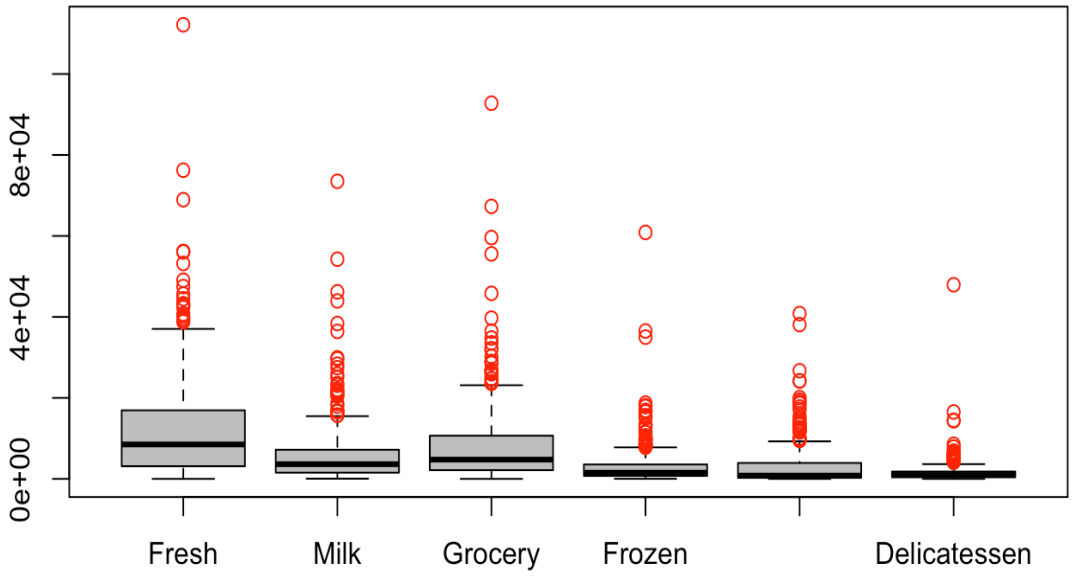
```
library(dplyr)
library(stats)
library(cluster)
library(ggplot2)
library(factoextra)
```

Reading the Wholesale customers datafile

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

Boxplot for numerical columns

```
boxplot(customers[3:8], col = "grey", outcol = "red")
```



Removing outliers (1.5 * IQR)

```
identify_outliers <- function(x) {
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  return(x < lower_bound | x > upper_bound)
}
```

```r
# Identify outliers for each column
outliers <- apply(customers[, 3:8], 2, identify_outliers)

# Identify rows with one or more outliers
outlier_rows <- apply(outliers, 1, any)

customers_outliers <- customers[outlier_rows, ]
customers_clean <- customers[!outlier_rows, ]
```

We observed numerous outliers in our dataset, which could potentially skew the clustering results. To address this, we removed data points that fell outside the range defined by 1.5 times the Interquartile Range (IQR) below the first quartile (Q1) and above the third quartile (Q3). This step was crucial for obtaining a clearer understanding of the underlying data patterns.

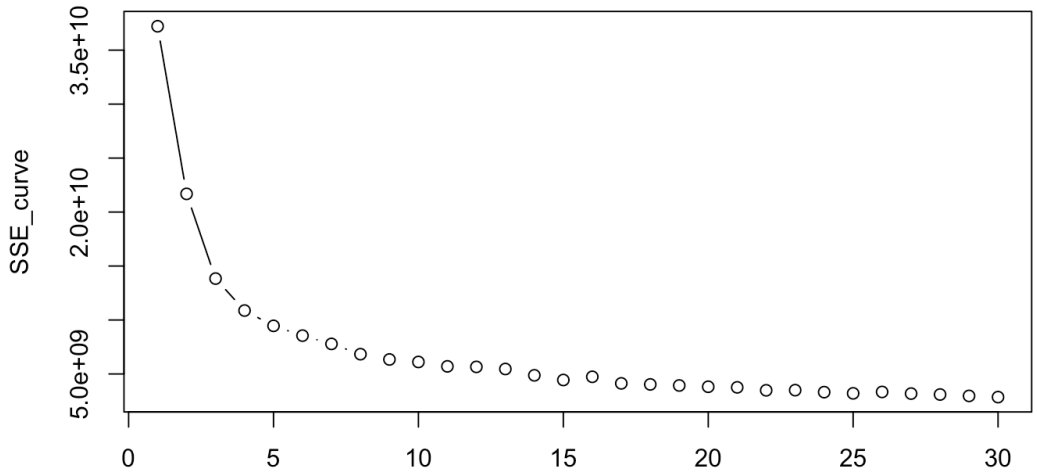
Calculating distance matrix

```r
distance_matrix = dist(customers_clean[,3:8], method = "euclidean")
```

Plotting SSE to observe elbow point

```r
SSE_curve <- c()
for (n in 1:30) {
  kcluster = kmeans(customers_clean[,3:8], n)
  sse = kcluster$tot.withinss
  SSE_curve[n] = sse}
# plot SSE against number of clusters
plot(1:30, SSE_curve, type = "b")
```

After cleaning the data, we evaluated different clustering solutions using K-means. We experimented with both 2 and 3 clusters after observing elbow point from the SSE curve. Although both configurations provided similar insights, the 2-cluster solution exhibited poor silhouette width. Therefore, we opted for a 3-cluster solution, which offered better-defined groupings and improved interpretability. This explanation clearly outlines why outliers were removed and justifies the choice of a 3-cluster solution based on silhouette analysis.



Clustering using K-Means method

```
kcluster = kmeans(customers_clean[,3:8], centers = 3)
kcluster$centers
```

```
##      Fresh     Milk   Grocery   Frozen Detergents_Paper Delicatessen
## 1  4569.759 8035.723 12536.566 1258.916       5156.6627    1276.0241
## 2 21232.964 3531.048  5004.265 2160.446       1176.1325    1265.9759
## 3  6193.434 2426.976  2980.813 2016.416        751.5542     762.2048
```

```
customers_clean$K_Cluster <- kcluster$cluster

write.csv(customers_clean,"customer_clust.csv", row.names = TRUE)
```

K-means clustering was applied to the dataset, resulting in three clusters. The cluster centers indicate distinct purchasing patterns among customers. Cluster 1 has high Grocery and Milk values, Cluster 2 shows high Fresh values, and Cluster 3 has moderate values across all categories. The results were saved for further analysis.

Evaluating the clustering solution using Silhouette Coefficient

```
sc = silhouette(customers_clean$K_Cluster, dist = distance_matrix)
summary(sc)
```

```
## Silhouette of 332 units in 3 clusters from silhouette.default(x = customers_clean$K_Cluster, dist = distance_m
atrix) :
##  Cluster sizes and average silhouette widths:
##       83       83      166
## 0.3235352 0.3532104 0.4574921
## Individual silhouette widths:
##    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.07511  0.30048  0.44504  0.39793  0.52243  0.64819
```

We evaluated our clustering solution using the Silhouette Coefficient, which measures how similar an object is to its own cluster compared to other clusters. A higher average silhouette width indicates a better-defined clustering structure.

The results indicate that while Cluster 2 has a relatively good silhouette width, Clusters 0 and 1 show moderate cohesion and separation. The presence of negative silhouette values suggests some data points may be misclassified or lie between clusters.

Plotting datapoints after dimensionality reduction using PCA

```
customers_clean$K_Cluster = as.factor(customers_clean$K_Cluster)
pca_result <- prcomp(customers_clean[, 3:8], scale = TRUE)
fviz_pca_ind(pca_result,
         geom.ind = "point",
         col.ind = customers_clean$K_Cluster,
         palette = "jco",
         addEllipses = TRUE,
         ellipse.level = 0.90,
         legend.title = "Cluster") +
```

```
labs(title = "PCA of Wholesale Customers Clustering") +
theme_minimal()
```