Clustering Analysis on Wholesale Customer Data

MSBA 6130 Homework 1

Rebecca Meyer, Narae Kang, Shubham Garg, Pranvi Setia, Yun-Chien Yen

7/6/2022

Introduction to Wholesale Customer Data

The wholesale customer data was provided by Professor Mochen Yang in his MSBA 6130 course, Introduction to Data Analytics in R, Carlson management of School, University of Minnesota. The data contains information on the clients of Company XYZ. For each client, information is provided on their channels, regions, and annual spendings across six product categories.

For the Channel column, 1 means Horeca (Hotel/Restaurant/Cafe) and 2 means Retail. For the Region column, 1 means Lisbon, 2 means Oporto, and 3 means other regions. The six product categories are fresh products, milk products, grocery products, frozen products, detergent and paper products, and delicatessen products.

Company XYZ hired us to analyze this data to gain a better understanding of their client spending patterns and use this information to more effiecently meet clients' demand.To help Company XYZ accomplish this goal, we decidied to employ the exploratory analysis technique of clustering analysis.

Dataset observation

Loading Packages and Data

Our analysis was preformed in Jupyter Lab - RStudio. The packages needed to preform the following analysis are shown below.

```
library(dplyr)
library(cluster)
library(cluster.datasets)
library(stats)
library(ggplot2)
library(GGally)
library(gridExtra)
library(corrplot)
library(RSNNS)
library(psych)
library(data.table)
library(factoextra)
```

We then uploaded the data using the read.csv() function.

```
XYZ_clients <- read.csv("Wholesale customers data.csv")
```

Exploratory Data Analysis

We explored the dataset first to get a preliminary understanding of trends. We used the glimpse function as a starting point.

```
glimpse(XYZ_clients)
```

```
## Rows: 440
## Columns: 8
## $ Channel          <int> 2, 2, 2, 1, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1,~
## $ Region           <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,~
## $ Fresh            <int> 12669, 7057, 6353, 13265, 22615, 9413, 12126, 7579, 5~
## $ Milk             <int> 9656, 9810, 8808, 1196, 5410, 8259, 3199, 4956, 3648,~
## $ Grocery          <int> 7561, 9568, 7684, 4221, 7198, 5126, 6975, 9426, 6192,~
## $ Frozen           <int> 214, 1762, 2405, 6404, 3915, 666, 480, 1669, 425, 115~
## $ Detergents_Paper <int> 2674, 3293, 3516, 507, 1777, 1795, 3140, 3321, 1716, ~
## $ Delicatessen     <int> 1338, 1776, 7844, 1788, 5185, 1451, 545, 2566, 750, 2~
```

We found the mean annual spending for each product category in each channel and region.

```
#by channel
XYZ_clients %>%
  group_by(Channel) %>%
  summarise(avg_Fresh = mean(Fresh), avg_milk = mean(Milk),
         avg_grocery = mean(Grocery), avg_frozen = mean(Frozen),
         avg_paper = mean(Detergents_Paper), avg_del = mean(Delicatessen))
```

```
## # A tibble: 2 x 7
##   Channel avg_Fresh avg_milk avg_grocery avg_frozen avg_paper avg_del
##     <int>     <dbl>    <dbl>       <dbl>      <dbl>     <dbl>   <dbl>
## 1       1    13476.    3452.       3962.      3748.      791.   1416.
## 2       2     8904.   10716.      16323.      1653.     7270.   1753.
```

We can see that clients in the Horeca (Channel=1) buy 1.5 times more fresh than retail (Channel=2). While clients in retail channel buy more milk (3 times), grocery(4 times), and especially paper products (9 times).

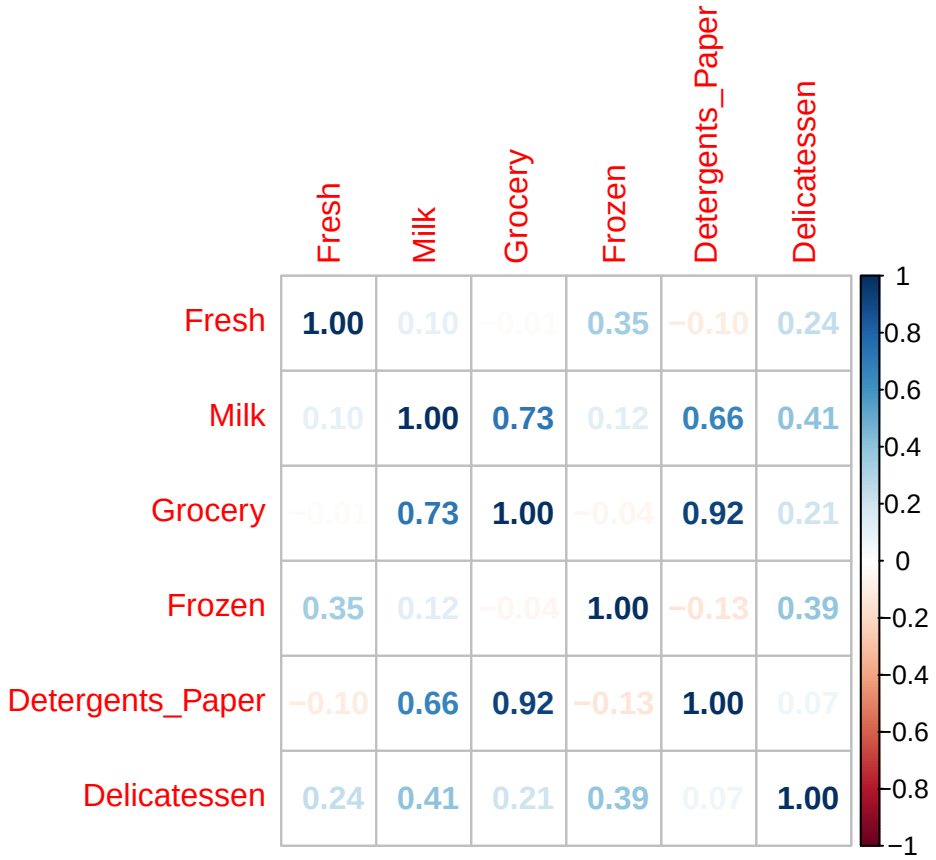
```
#by region
XYZ_clients %>%
  group_by(Region) %>%
  summarise(avg_Fresh = mean(Fresh), avg_milk = mean(Milk),
            avg_grocery = mean(Grocery), avg_frozen = mean(Frozen),
            avg_paper = mean(Detergents_Paper), avg_del = mean(Delicatessen))
```

```
## # A tibble: 3 x 7
##   Region avg_Fresh avg_milk avg_grocery avg_frozen avg_paper avg_del
##    <int>     <dbl>    <dbl>       <dbl>      <dbl>     <dbl>   <dbl>
## 1      1    11102.    5486.       7403.      3000.     2651.   1355.
## 2      2     9888.    5088.       9219.      4045.     3687.   1160.
## 3      3    12533.    5977.       7896.      2945.     2818.   1621.
```

Here we can see the clients in Lisbon (Region=1) tend to spend an average amount within all categories compared to other regions(Region=2,3). Clients in Oporto(Region=2) tend to spend more on grocery, frozen, and paper products. We can also see that clients in other regions(Region=3) buy the most fresh products than any other categories.

Also, we looked at the correlation matrix for spending patterns.

```
# correlation matrix
corrmatrix <- cor(XYZ_clients[,3:8])
corrplot(corrmatrix, method = 'number')
```

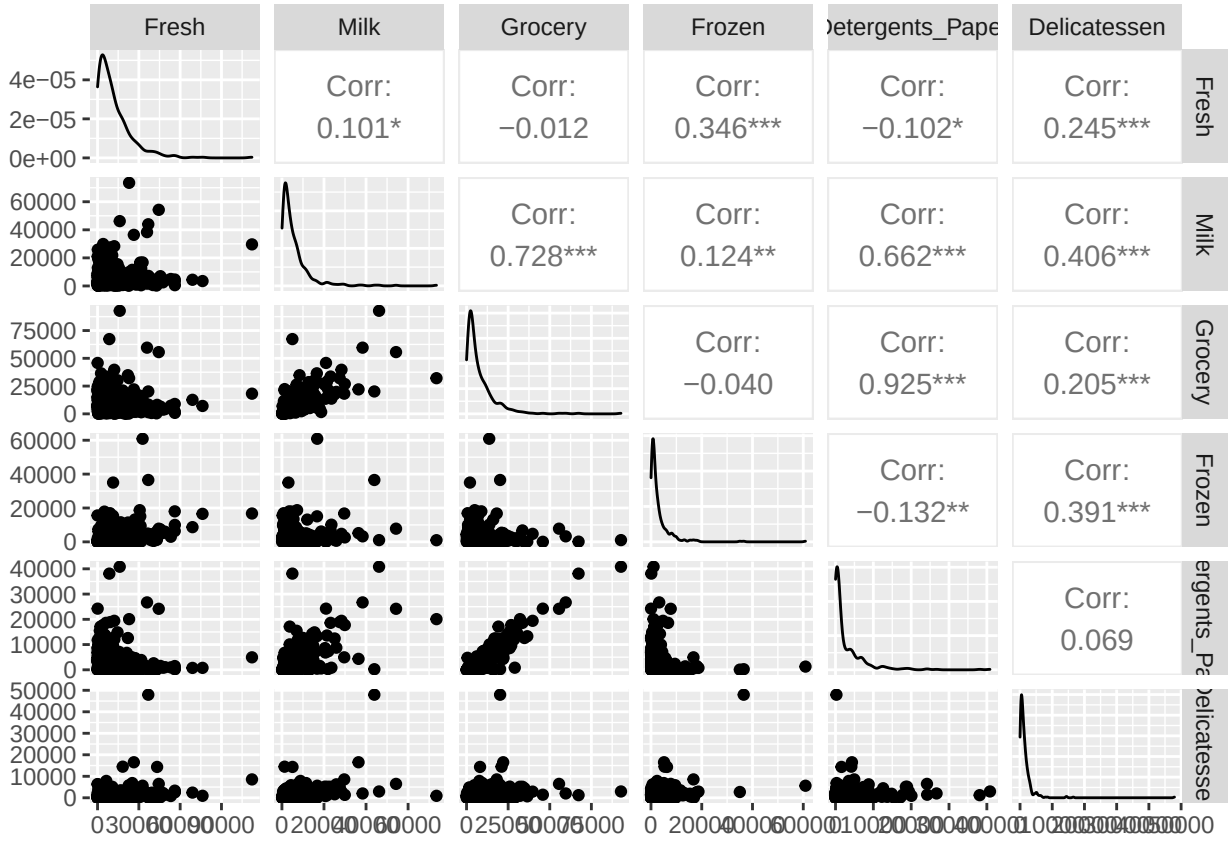


Then, we used ggpairs and ggplot to visualize product categories in scatter plots and box plots.

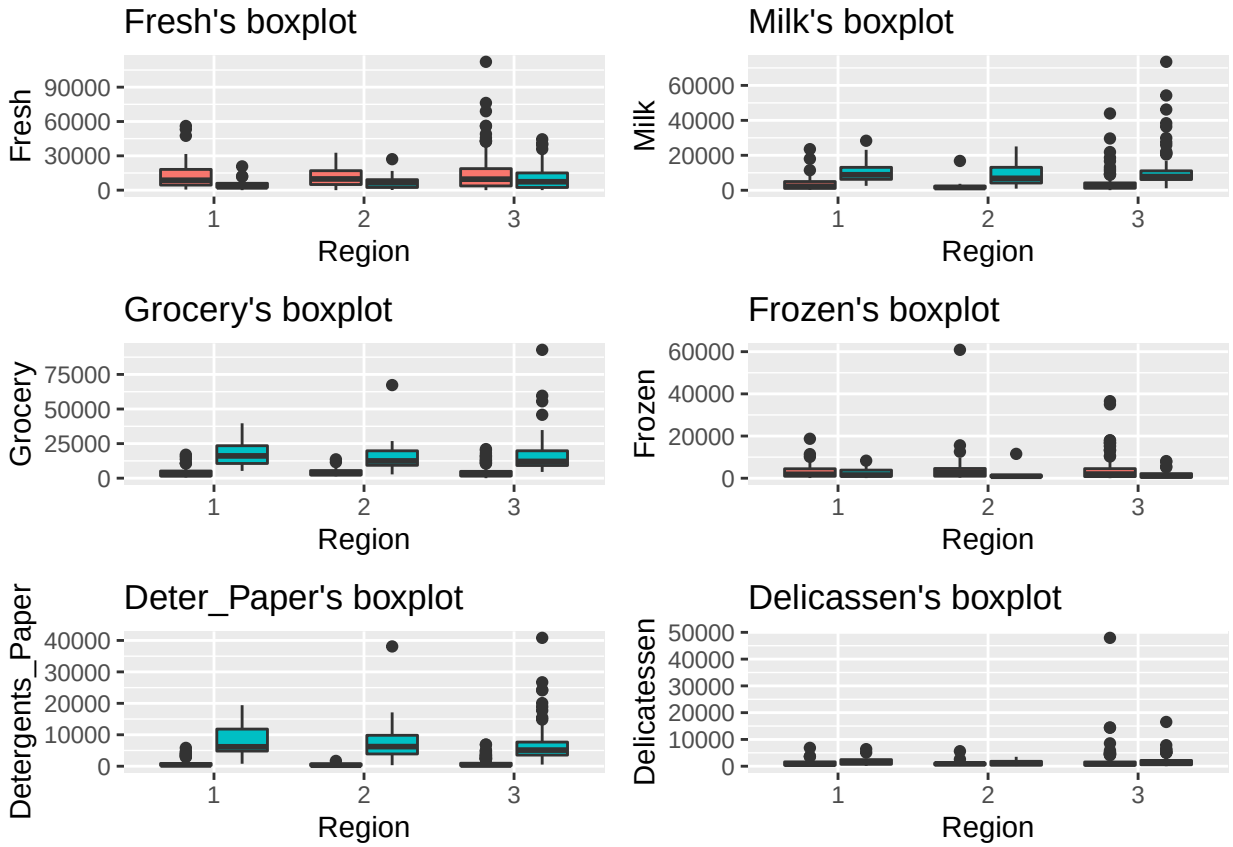
```
XYZ_clients$Region <- as.factor(XYZ_clients$Region)
XYZ_clients$Channel <- as.factor(XYZ_clients$Channel)
summary(XYZ_clients)
```

```
##  Channel Region      Fresh             Milk          Grocery
##  1:298   1: 77   Min.   :     3   Min.   :   55   Min.   :    3
##  2:142   2: 47   1st Qu.:  3128   1st Qu.: 1533   1st Qu.: 2153
##          3:316   Median :  8504   Median : 3627   Median : 4756
##                  Mean   : 12000   Mean   : 5796   Mean   : 7951
##                  3rd Qu.: 16934   3rd Qu.: 7190   3rd Qu.:10656
##                  Max.   :112151   Max.   :73498   Max.   :92780
##      Frozen        Detergents_Paper   Delicatessen
##  Min.   :   25.0   Min.   :    3.0   Min.   :    3.0
##  1st Qu.:  742.2   1st Qu.:  256.8   1st Qu.:  408.2
##  Median : 1526.0   Median :  816.5   Median :  965.5
##  Mean   : 3071.9   Mean   : 2881.5   Mean   : 1524.9
##  3rd Qu.: 3554.2   3rd Qu.: 3922.0   3rd Qu.: 1820.2
##  Max.   :60869.0   Max.   :40827.0   Max.   :47943.0
```

```
ggpairs(XYZ_clients[,3:8])
```




```
#boxplot
b1 <- ggplot(XYZ_clients, aes(x = Region, y = Fresh, fill = Channel)) +
geom_boxplot() + theme_grey() + ggtitle("Fresh's boxplot") + theme(legend.position = "none")
b2 <- ggplot(XYZ_clients, aes(x = Region, y = Milk, fill = Channel)) +
geom_boxplot() + theme_grey() + ggtitle("Milk's boxplot") + theme(legend.position = "none")
b3 <- ggplot(XYZ_clients, aes(x = Region, y = Grocery, fill = Channel)) +
geom_boxplot() + theme_grey() + ggtitle("Grocery's boxplot") + theme(legend.position = "none")
b4 <- ggplot(XYZ_clients, aes(x = Region, y = Frozen, fill = Channel)) +
geom_boxplot() + theme_grey() + ggtitle("Frozen's boxplot") + theme(legend.position = "none")
b5 <- ggplot(XYZ_clients, aes(x = Region, y = Detergents_Paper, fill = Channel)) +
geom_boxplot() + theme_grey() + ggtitle("Deter_Paper's boxplot") + theme(legend.position = "none")
b6 <- ggplot(XYZ_clients, aes(x = Region, y = Delicatessen, fill = Channel)) +
geom_boxplot() + theme_grey() + ggtitle("Delicassen's boxplot") + theme(legend.position = "none")
grid.arrange(b1, b2, b3, b4, b5, b6, nrow=3)
```



From the above visualization, we can see that in other regions (Region = 3) we have more clients and more outliers.

Cluster Analysis

Normalization of Wholesale Customer Data

In order to make sure all attributes of data are from the same range, we normalized it based on the below table.

```
#total
describe(XYZ_clients)
```

```
##                   vars   n     mean       sd median trimmed     mad min    max
## Channel*             1 440     1.32     0.47    1.0    1.28    0.00   1      2
## Region*              2 440     2.54     0.77    3.0    2.68    0.00   1      3
## Fresh                3 440 12000.30 12647.33 8504.0 9864.61 8776.25   3 112151
## Milk                 4 440  5796.27  7380.38 3627.0 4375.52 3647.20  55  73498
## Grocery              5 440  7951.28  9503.16 4755.5 6158.43 4586.42   3  92780
## Frozen               6 440  3071.93  4854.67 1526.0 2144.07 1607.88  25  60869
## Detergents_Paper     7 440  2881.49  4767.85  816.5 1849.73 1060.80   3  40827
## Delicatessen         8 440  1524.87  2820.11  965.5 1113.24  945.16   3  47943
##                    range  skew kurtosis     se
## Channel*               1  0.76    -1.43   0.02
## Region*                2 -1.27    -0.13   0.04
## Fresh             112148  2.54    11.33 602.94
```

```
## Milk              73443  4.03    24.25 351.85
## Grocery           92777  3.56    20.56 453.05
## Frozen            60844  5.87    53.80 231.44
## Detergents_Paper  40824  3.61    18.68 227.30
## Delicatessen      47940 11.08   167.97 134.44
```

We created a function called normalize and mutated each attribute value using that function.

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

XYZ_clients_norm = XYZ_clients %>%
  mutate_at(c(3:8), normalize)
```

Clustering Method

Choosing the Number of Clusters

To find the best number of clusters for the wholesale customer data, we created the SSE Curve to find the elbow point which shows the optimal number of clusters.

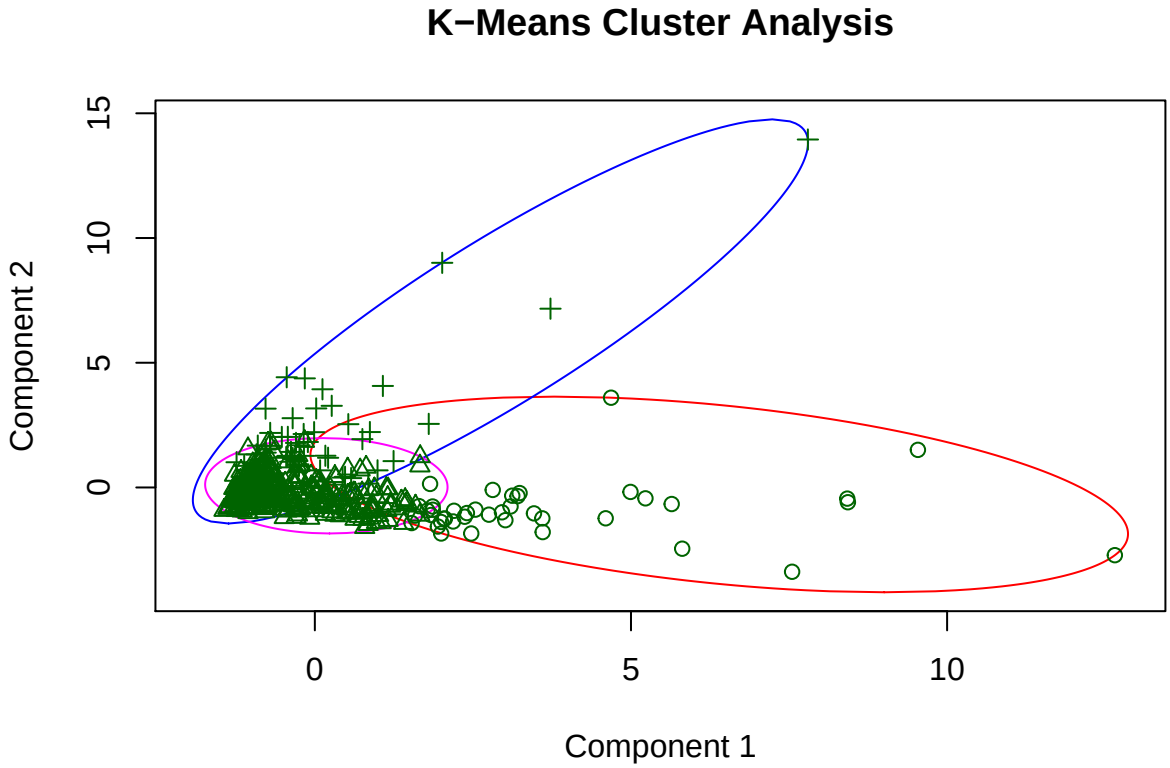
```
SSE_curve <- c()
for (n in 1:10) {
  kcluster = kmeans(XYZ_clients_norm[,3:8], n)
  sse = kcluster$tot.withinss
  SSE_curve[n] = sse}
plot(1:10, SSE_curve, type = "b",
     main = "SSE Curve",
     xlab = "Number of Clusters", ylab = "SSE",
     col = 'blue')
```

According to SSE Curve, we chose 3 clusters as the SSE curve decreases smoothly after k = 3.

K-Means Cluster Analysis

We used the kmeans() function to find the cluster groupings for clients based on their spending pattern in different categories. We then added those groupings to the XYZ clients data and found the cluster centers.

```
kcluster = kmeans(XYZ_clients_norm[,3:8], centers = 3)
XYZ_clients_norm$cluster <- kcluster$cluster
clusplot(XYZ_clients_norm[,3:8], kcluster$cluster, color = T, shade = F,
         labels=0, lines=0, main = 'K-Means Cluster Analysis')
```



We looked at the centers of the clusters formed above.

```
kcluster$centers
```

```
##        Fresh       Milk    Grocery     Frozen Detergents_Paper Delicatessen
## 1 0.07612741 0.26781048 0.31604316 0.03156052       0.34516819   0.04830937
## 2 0.07297318 0.05468315 0.06047771 0.03993169       0.04513443   0.02316561
## 3 0.31612924 0.08086923 0.07042464 0.11874307       0.02650803   0.06815213
```

We then found the demographics (region and channel) of the clients in each group and plotted this information on box plots for a visual representation.

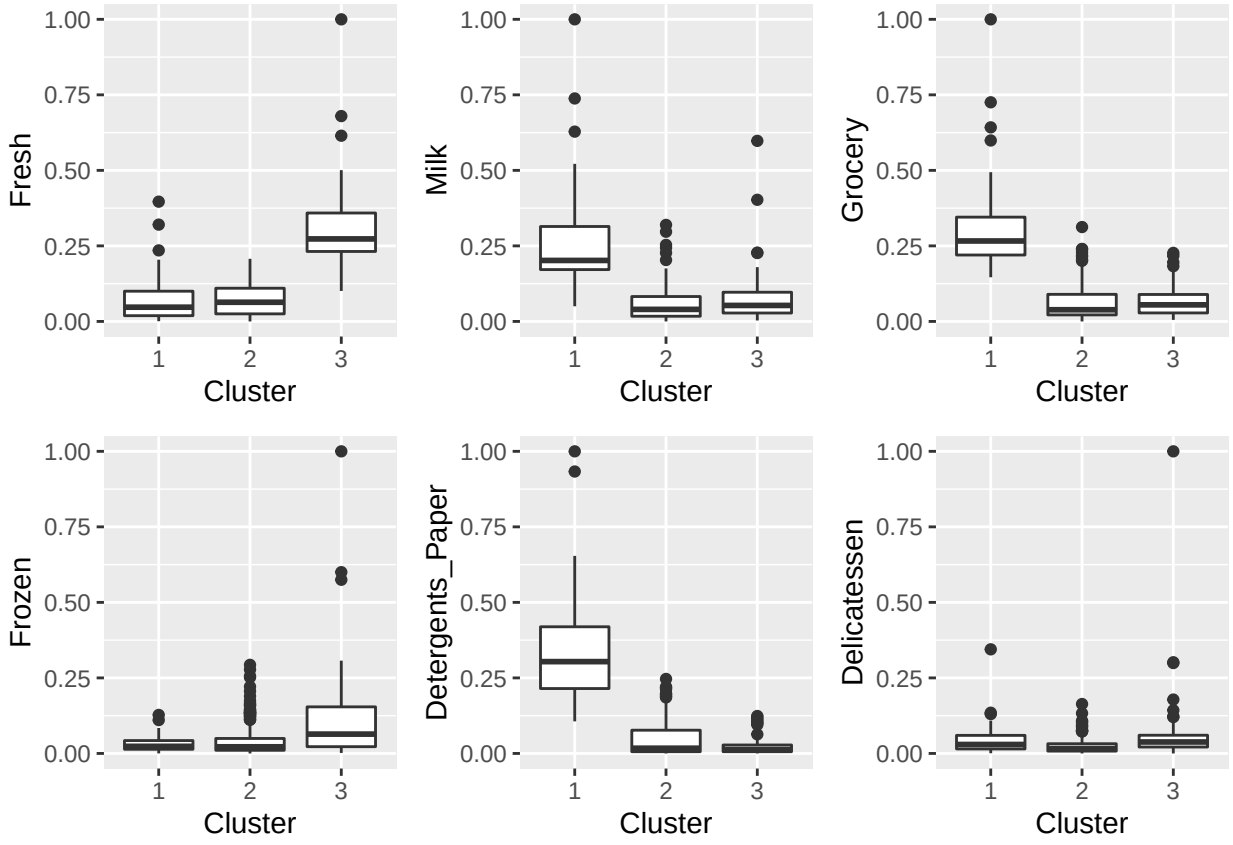
```
## Distribution and sample counts in each cluster
XYZ_clients_k <- XYZ_clients_norm %>% group_by(cluster) %>%
                 count(Channel, Region)
as.data.frame(XYZ_clients_k)
```

```
##    cluster Channel Region   n
## 1        1       2      1   7
## 2        1       2      2   8
## 3        1       2      3  26
## 4        2       1      1  49
## 5        2       1      2  25
## 6        2       1      3 171
## 7        2       2      1  11
## 8        2       2      2  10
## 9        2       2      3  72
## 10       3       1      1  10
## 11       3       1      2   3
## 12       3       1      3  40
```

```
## 13       3       2       2   1
## 14       3       2       3   7
```

```
# box plots of 6 product categories after clustering
XYZ <- XYZ_clients_norm  %>% select(-c("Channel", "Region"))
XYZ  %>% mutate(Cluster = as.factor(kcluster$cluster))  %>%
      ggplot(aes(y = Fresh, x = Cluster)) + geom_boxplot() + scale_fill_brewer(palette="Dark2")-> b1
XYZ  %>% mutate(Cluster = as.factor(kcluster$cluster))  %>%
      ggplot(aes(y = Milk, x = Cluster)) + geom_boxplot() -> b2
XYZ  %>% mutate(Cluster = as.factor(kcluster$cluster))  %>%
      ggplot(aes(y = Grocery, x = Cluster)) + geom_boxplot() -> b3
XYZ  %>% mutate(Cluster = as.factor(kcluster$cluster))  %>%
      ggplot(aes(y = Frozen, x = Cluster)) + geom_boxplot() -> b4
XYZ  %>% mutate(Cluster = as.factor(kcluster$cluster))  %>%
      ggplot(aes(y = Detergents_Paper, x = Cluster)) + geom_boxplot() -> b5
XYZ  %>% mutate(Cluster = as.factor(kcluster$cluster))  %>%
      ggplot(aes(y = Delicatessen, x = Cluster)) + geom_boxplot() -> b6

grid.arrange(b1, b2, b3, b4, b5, b6, nrow=2)
```



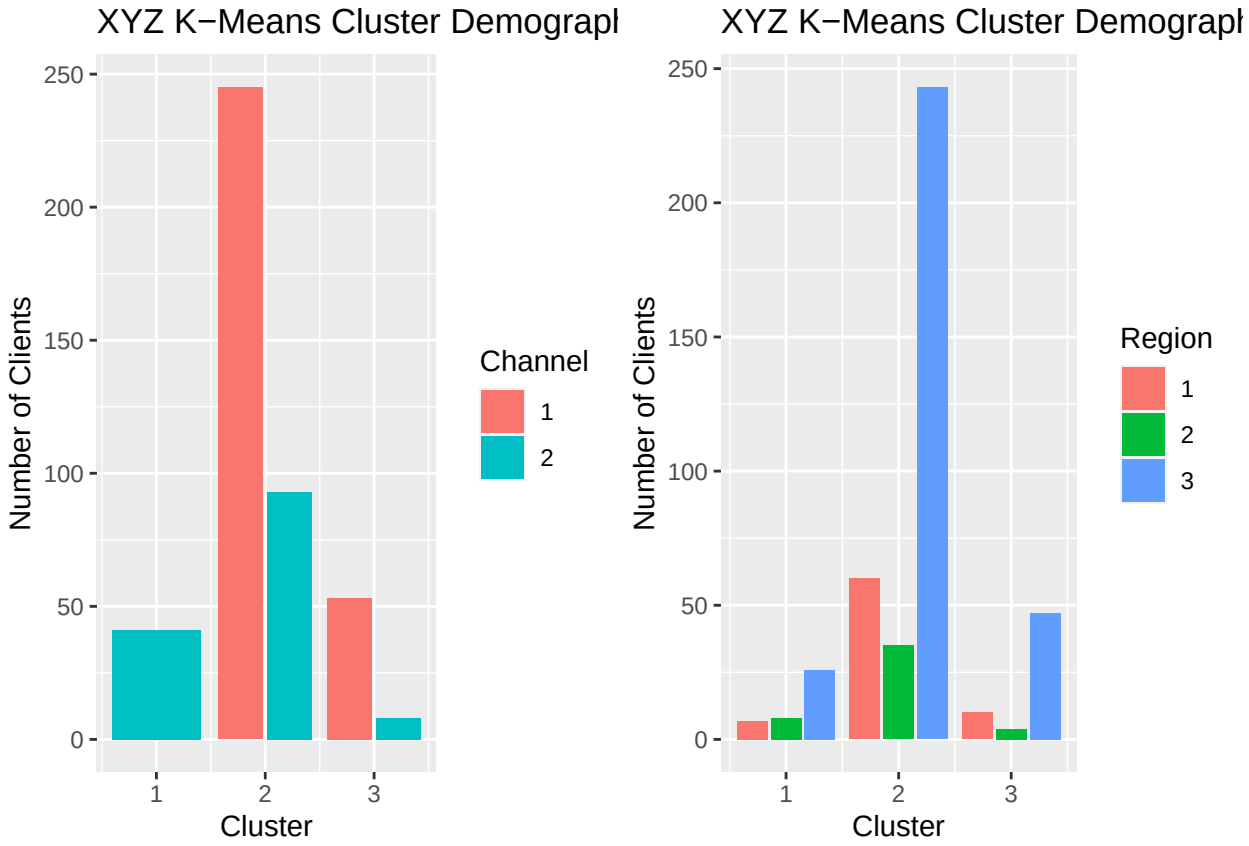
To get a better understanding of the distribution of the channels and regions across the different clusters, we used ggplot to create two bar charts. The first bar chart shows the distribution of channel type across the clusters. The second bar chart shows the distribution of region across the clusters.

```
channel <-
  ggplot(XYZ_clients_norm %>% group_by(cluster, Channel) %>% count(),
       aes(fill=as.factor(Channel), y=n, x=cluster)) +
  geom_bar(position = "dodge2", stat = "identity") +
  ggtitle('XYZ K-Means Cluster Demographics - By Channel') +
```

```
  xlab('Cluster') +
  ylab('Number of Clients')+
  scale_fill_discrete(name = "Channel")
```

```
region <-
  ggplot(XYZ_clients_norm %>% group_by(cluster, Region) %>% count(),
       aes(fill=as.factor(Region), y=n, x=cluster)) +
  geom_bar(position = "dodge2", stat = "identity") +
  ggtitle('XYZ K-Means Cluster Demographics - By Region') +
  xlab('Cluster') +
  ylab('Number of Clients') +
  scale_fill_discrete(name = "Region")

grid.arrange(channel, region, ncol=2)
```



We can see that one cluster only contains clients from channel 2. We can also see that the other two clusters contain a majority of clients from channel 1.

Data Evaluation

Lastly, we evaluated our K-means clustering results with the silhouette coefficient. We used the silhouette() function with our cluster groups and the distance matrix. The summary can be seen below.

```
XYZ_clients_norm_distance_matrix = dist(XYZ_clients_norm[,3:8], method = "euclidean")
sc_k = silhouette(XYZ_clients_norm$cluster, dist = XYZ_clients_norm_distance_matrix)
summary(sc_k)
```

```
## Silhouette of 440 units in 3 clusters from silhouette.default(x = XYZ_clients_norm$cluster, dist = X
##  Cluster sizes and average silhouette widths:
##        41       338        61
## 0.1993003 0.5320006 0.0961455
## Individual silhouette widths:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.1810  0.2999  0.5179  0.4406  0.6217  0.6735
```