University of Cape Town

STA4010W

Analytics

Assignment 3: Customer Segmentation

Author:
Chongo Nkalamo
Yovna Junglee

Student Number:
NKLCHO001
JNGYOV001

April 24, 2019

Customer Segmentation

Exploratory data analysis

Figure 1 shows the frequency plots of purchases made on weekends and weekdays. We can see from the top left figure that most products are purchased in similar quantities overall. We can see similar purchasing patterns over weekends and weekdays. Further, it can be observed that more purchases are made during the weekdays than over the weekend. Customers mostly bought two to four different items during one transaction as shown in Figure 2.

Customer Segmentation.

Following the exploratory data analysis, we proceed to carry out the customer segmentation of purchases of different bakery products using cluster analysis. The data provided is given in binary form and hence popular clustering methods such as the k-means algorithm fail to cluster well on this type of data. Therefore, we resort to other k-mean algorithm variants. For our purposes, we model the problem using k-medoids which involves the use of a partitioning around medoids (PAM) algorithm which has a function `PAM` in R. The `PAM` function in R requires one to specify the number of clusters or segmentations,k that will be used to partition the data. Given the high dimensionality of the data at hand, an optimal selection for this value k is often subjective. However, we choose to try values of k in the range of 1 to 8 and if need be, higher values of k may be considered in order to assess whether there will be significant differences in clusters obtained.

Finding the optimal number of cluster.

The measure adopted as a selection critera for choosing the optimal number of clusters was the avergare silhouette width. In addition, two different metrics for computation of the dissimialrity matrix was used. These are the gower distance and correlation distance. Plotted against a range of clusters, the number of clusters corresponding to the highest average silhouette width is often deemed as optimal in both methods. The figures below show that the optimal number of clusters is choosen to be 8, in both cases. However, the average silhouette widths are seen to be relatively small, which implies that the clustering done may not accurately represent the data at hand. Furthermore, if we consider a wider range say for values of k ranging from 1 to 25, the average silhouette begins to taper off after k = 18 under both distance measures. Thus it is quite evident that the optimal number of clusters using these method would be 18, but this will provide us with alot of segments and thus analysis may be difficult to carry out and in addition, most of the clusters will still be similar to the clusters produced when k = 8. Therefore, we choose to use a value of k = 8 and correlation distance as a distance metric.

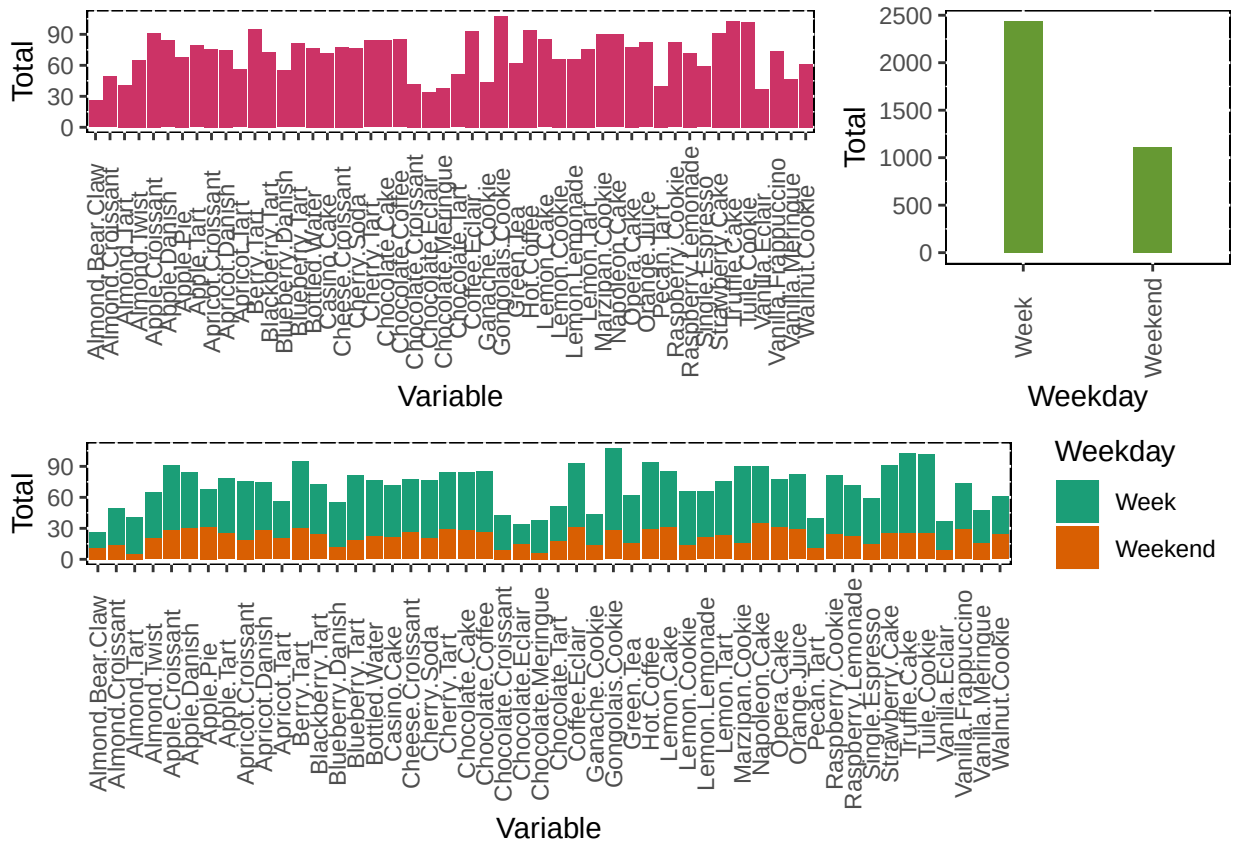


Figure 1: Analysis of purchases of products made on Weekends and Weekdays

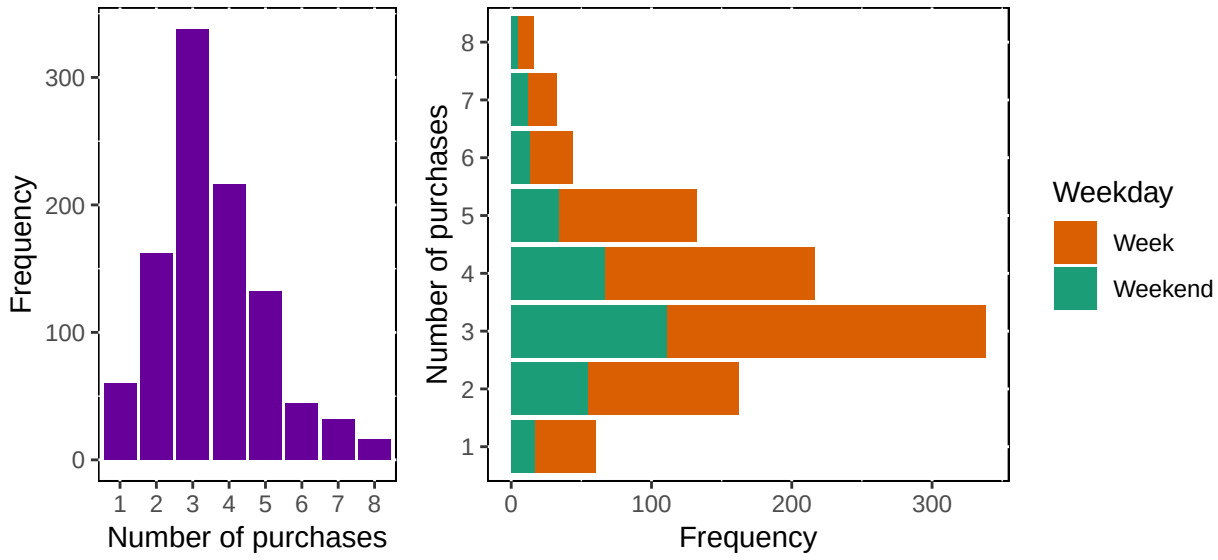
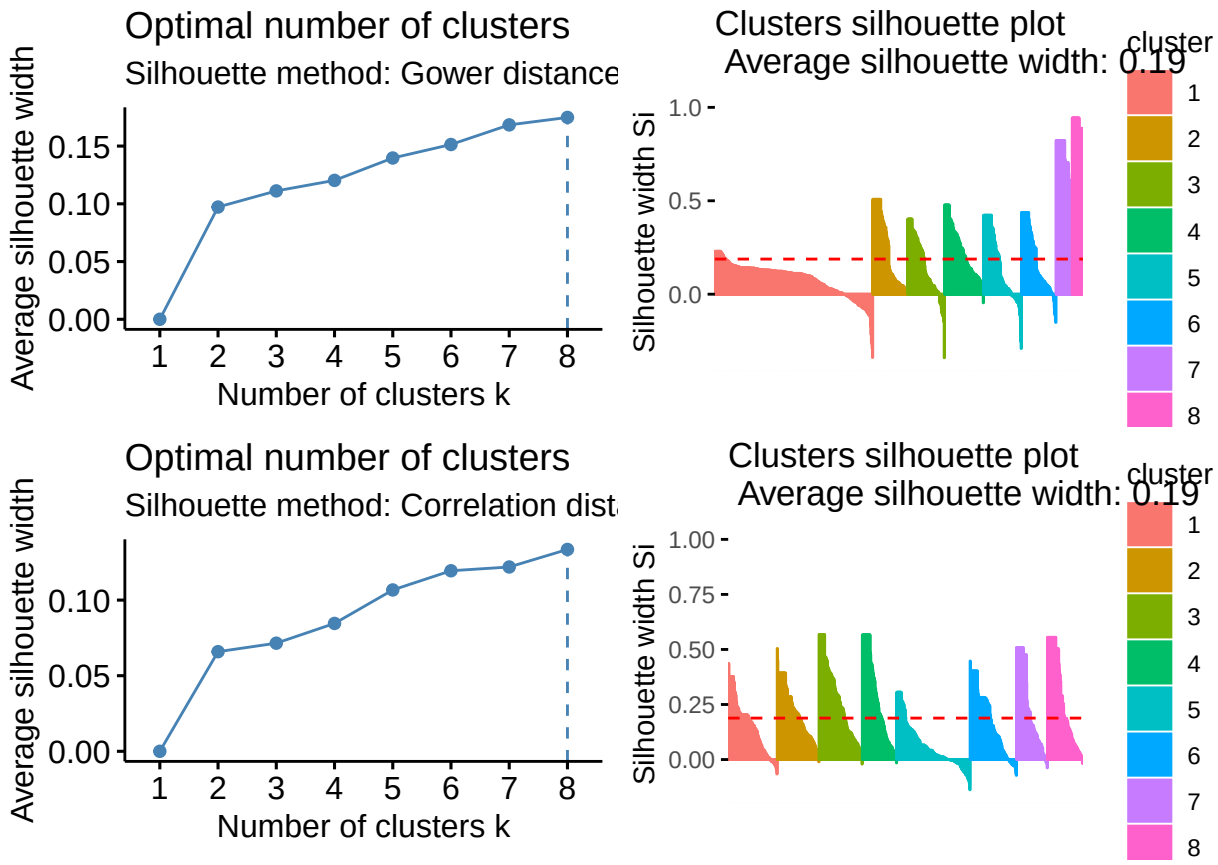


Figure 2: Analysis of customer behaviour on weekdays and weekends



Cluster analysis.

Having selected the optimal number of clusters we will use to partition or carry out customer segementation, we then proceed to cluster our data. Clustering involves finding homogenous groups in the data and as outlined in the introduction we will apply the PAM algorithm. Using a number of clusters, k = 8, we obtained 8 different customer segmentations. The hope at this point is that each segement contains customers with different spending patterns. We can view this by comparing the most dominant/common items bought in each of these segments. If there is a difference in items then the clustering algorithm is indeed partitioning the customers. Figure 3 shows the top five items purchased by individuals in each of the 8 different customer segementations.

From Figure 3, we can see that the most dominant items purchased by customers forming cluster 2 are Opera Cake, Cherry Tart and Apricot Danish. In cluster 5 we see Marzipan Cookies and Tuile Cookies being purchases most by customers in that group. Cluster 8 saw Cherry soda, Apple Danish, Apple croissants and Apple Tart being purchased most by customers. Following this analysis we can see that customers purchasing certain similar products are placed in similar groupings and hence the algorithm is performing the clustering we expect.

Characterizing the Customer segments

A closer look into the eight different segments formed is considered in this section. From the various plots produced, it is quite evident that each segment is characterized by certain items purchased. Most clusters have dominant items that were always purchased by majority of individuals that fell into that cluster/segment.

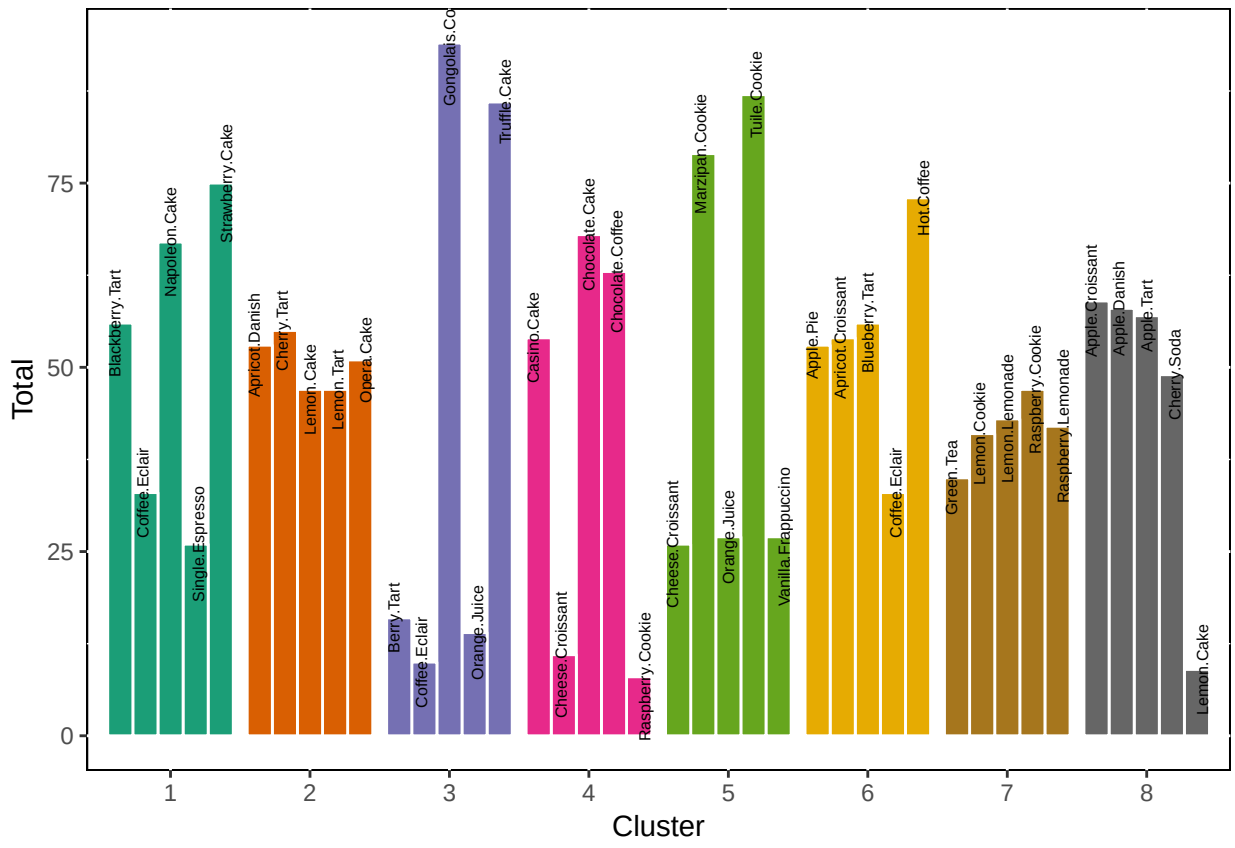


Figure 3: Top 5 products purchased under each cluster

Table 1: Summary of most commonly purchased items.

Cluster	Dominant item (Weekdays and Weekend)
1	Blackberry Tart, Napoleon Cake, Strawberry Cake
2	Opera Cake, Cherry Tart, Apricot Danish
3	Gongolais Cookie, Truffle Cake
4	Casino Cake, Chocolate Cake, Chocolate Coffee
5	Marzipan Cookie, Tuile Cookie
6	Apple Pie, Apple Croissant, Blueberry Tart, Hot Coffee
7	Green Tea, Lemon Lemonade, Raspberry Lemonade, Lemonade Cookie, Raspberry Cookie
8	Apple Croissant, Apple Danish, Apple Tart, Cherry Soda

Table 1 below shows for each cluster/segment, what the most dominant items purchased where for customers that fell into their particular group.

As can be seen in the table, in each cluster or segment the **most common** item(s) purchased in a particular segment or cluster do not appear as most commonly purchased in other clusters. Furthermore, the items in table 1 obtained for each segment are the items that seem strongly associated with each customer segment meaning that majority of customers falling into the respective cluster/segment buys some form of combination of these items. As an example, if we consider cluster 4 and from the frequency plot for this cluster, we can see that proportion of Casino Cake, Chocolate Cake and Chocolate Coffee are very high indicating that almost everyone in this group segment will always purchase these three items together in some combination and purchase all other items in very small proportion. However, these particular items do not appear as most commonly bought in other clusters and therefore they characterize this particular cluster.

Comparision between different clusters sees that there is not much of an overlap between the items, especially most common items purchased in each cluster. Therefore, each cluster is know for its unique characteristic. For instance, consider cluster 8, customers in this cluster can be known for having a strong liking of different apple pastry products along with Cherry soda while consuming the apple item. The same can not be said for cluster 7 as customers in this cluster can be considered to like Lemon and Raspberry cookies with a serving of green tea or Lemon or Raspberry Lemonade to go with. The same analysis can be considered for all characterization of all the other segments.

Weekend purchases analysis

In this section, we further analyse weekend purchases and determine which customer segments predominantly purchased items over the weekend. Figure 4 shows another frequency graph that indicates totals of weekend purchases in each cluster/segment. Cluster 5 produced the most weekend purchasers relative to the other clusters although the distribution of weekend buyers appear to be almost the same accross all clusters. As such, there is no distinct cluster that primarily makes purchases over the weekends.

However, a closer look into what items customers purchased items over the weekend was carried out (Figure 5 and 6).

Table 2 summarises the dominant items purchased over the weekends in each cluster: We note that there is no significant difference between weekends and weekdays in the purchasing behaviour in the segments identified previously. Taking a look at figure 6, but in particular the cluster with the highest amount of items purchased (cluster 5), we can see that predominantly Marzipan Cookies, and Tuile Cookies were purchased mostly. These also happen to be the dominant items purchased by individuals in this segment. The same analysis can be used to picture the kind of items that the rest of the individuals in the other customer segments purchase over the weekend.

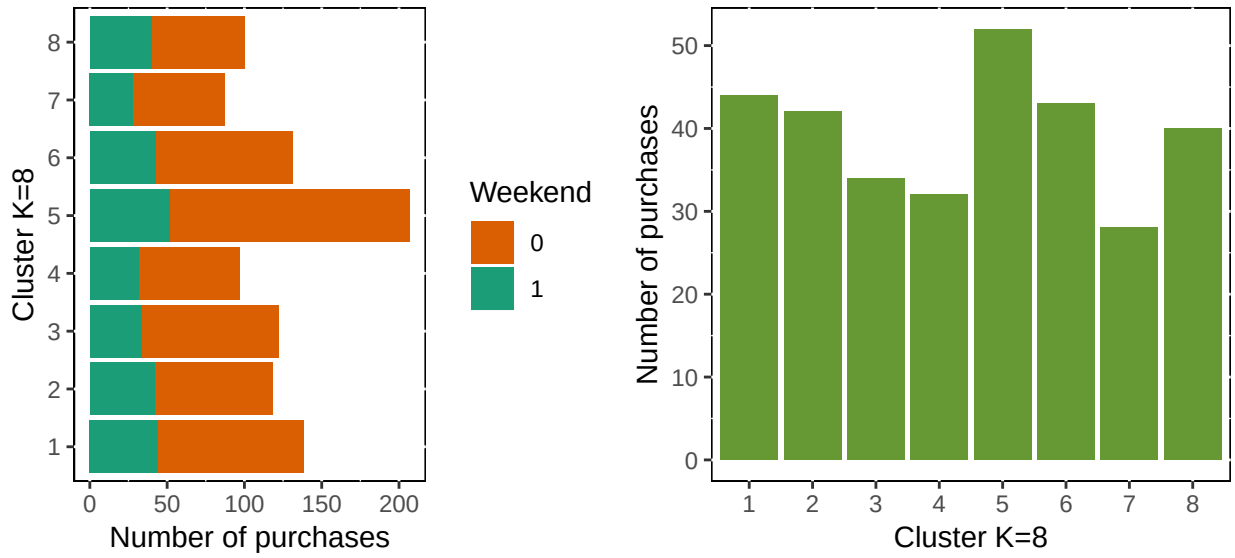


Figure 4: Analysis of total purchases made during weekends in clusters

Table 2: Summary of weekend purchases

Cluster	**Dominant item (Weekend Only)**
1	Blackberry Tart, Napoleon Cake, Strawberry Cake
2	Opera Cake, Cherry Tart, Apricot Danish, Lemon Cake, Lemon Tart
3	Gongolais Cookie, Truffle Cake
4	Casino Cake, Chocolate Cake, Chocolate Coffee
5	Tuile Cookie
6	Apple Pie, Hot Coffee
7	Green Tea, Lemon Lemonade, Raspberry Lemonade, Lemonade Cookie, Raspberry Cookie
8	Apple Croissant, Apple Danish, Apple Tart, Cherry Soda

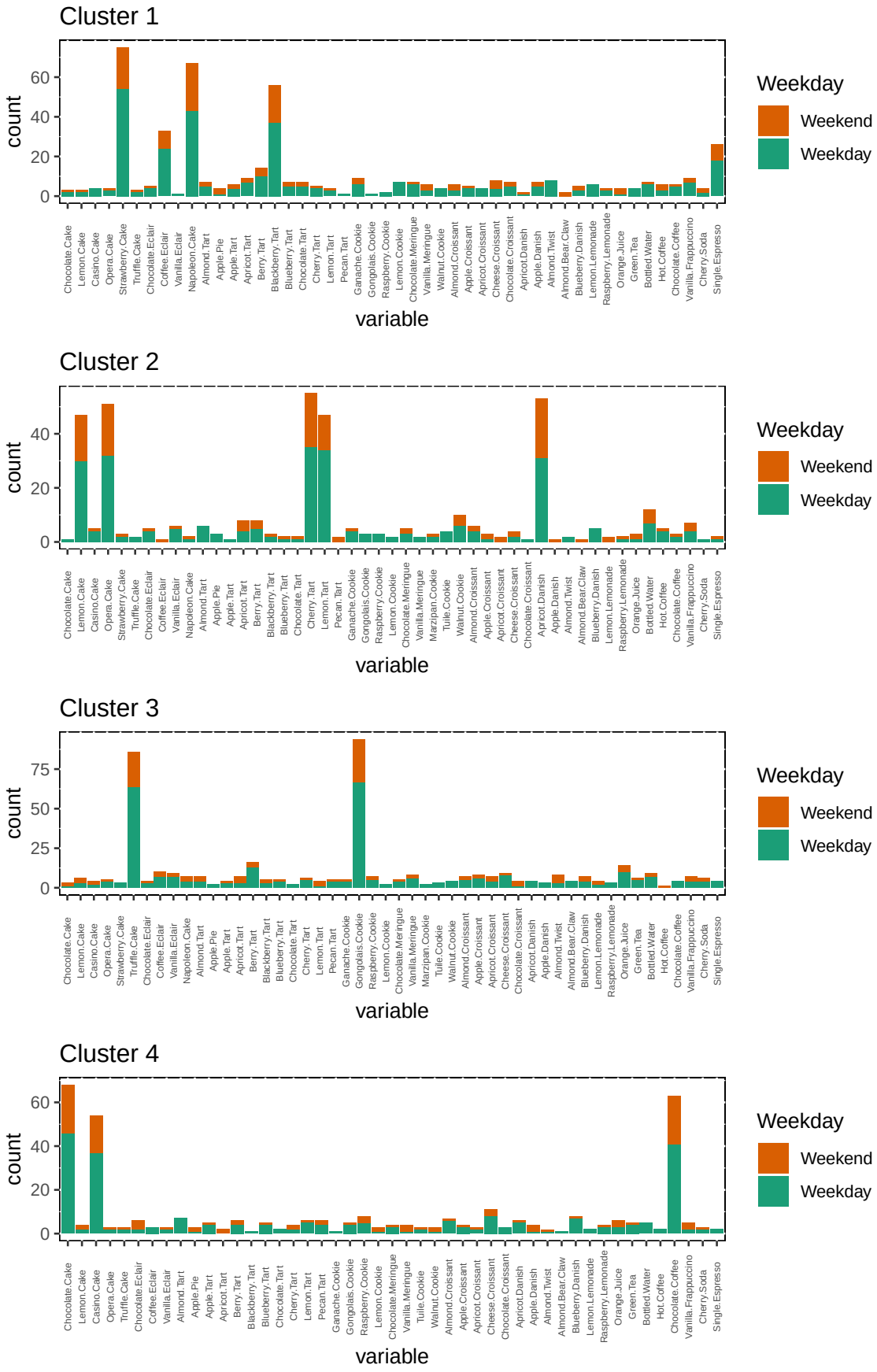


Figure 5: Cluster Analysis(Part I)

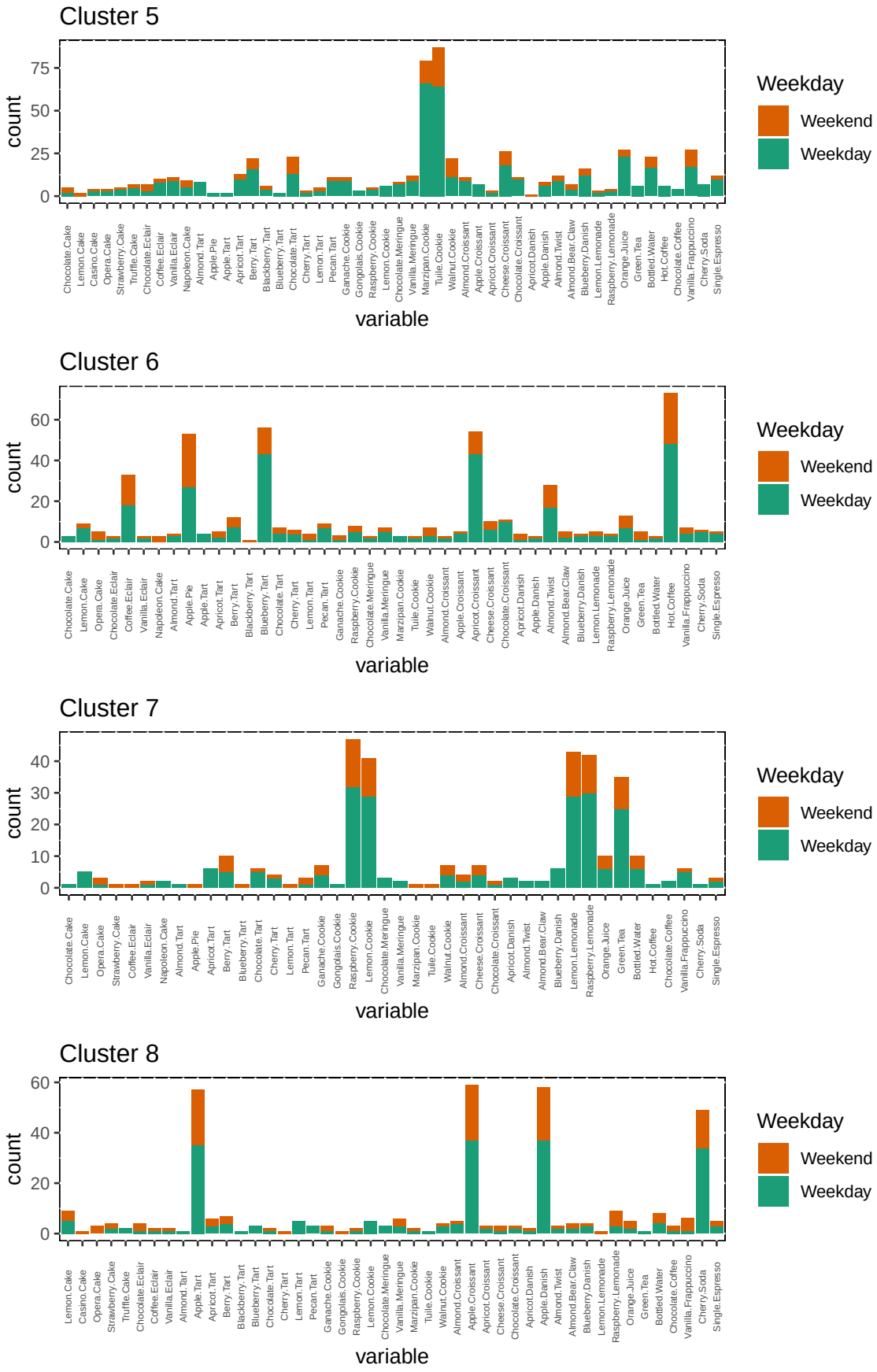


Figure 6: Cluster Analysis(Part II)

Recommendations

From our analysis above, we have identified the items most commonly bought by customers in the same segments and we can note that it is mostly common 2-4 dominant items in each cluster. Thus the bakery can implement different strategies in order to boost sales.

Below are 3 opportunities that the bakery can exploit:

- If the strategy is to boost the sales of less popular products, then the bakery should consider offering combination deals of 2 dominant products within the same segment and 1 less popular product. For example, the bakery can consider a *combo deal* of *Gongolais Cookie* and *Truffle Cake* (Cluster 3) with a *Vanilla Frappucino*.
- If the strategy is to increase current sales and customer loyalty, a *cake+drink combo deal* could be offered. Examples include *Apple Pie and Hot Coffee* (cluster 6) and *Apple Tart and Cherry Soda* (cluster 8)
- The bakery can also implement a strategy that will ensure that most cakes are sold on a particular day to reduce losses. We can note that Cluster 4 is mostly dominated by cakes made of chocolate whereas in cluster 8, apple products are the most common. Thus, the bakery can attract those segments by introducing *Chocolate deals on Day X* and *Apple deals on Day Y*, thereby inducing those customers to buy those products on that particular day.

Appendix

```
knitr::opts_chunk$set(echo = FALSE, message=FALSE, warning=FALSE)
library(reshape2)
library(ggplot2)
library(tidyverse)
library(cluster)
library(klaR)
library(gridExtra)
library(NbClust)
library(factoextra)
dat <- read.csv("bakery.csv", header = T)

# EXPLORATORY DATA ANALYSIS

overall <- as.data.frame(colSums(dat))
overall <- cbind( rownames(overall),overall)
colnames(overall) <-c("Variable", "Total")

overall <- overall[-51,]
eda1 <- ggplot(aes(x=Variable, y=Total), data=overall) +
  geom_col(fill="#CC3366") +
  theme( axis.text.x  = element_text(angle=90, vjust=0.5),
         panel.background = element_rect(fill = "white", colour='black'))


# PART 2


week <- dat[which(dat$Weekend==0),-51]
week <- as.data.frame(colSums(week))
names(week) <- "Week"
#week <- cbind(Variable=rownames(week),Total=week, Weekday="Week")


weekend <- dat[which(dat$Weekend==1),-51]
weekend <- as.data.frame(colSums(weekend))
names(weekend) <- "Weekend"
#weekend <- cbind(Variable=rownames(weekend),weekend, Weekday="Weekend")

total_week <- cbind(Variable=rownames(week),week,weekend)

total_week1 <- as.data.frame(melt(total_week))
names(total_week1) <- c("Variable","Weekday","Total")
```

```r
eda2<- ggplot(total_week1,aes(x=Variable,y=Total, fill=Weekday))+geom_col()+
  theme( axis.text.x  = element_text(angle=90, vjust=0.5),
         panel.background = element_rect(fill = "white", colour='black'))+
  scale_fill_brewer(palette = "Dark2")


eda3 <- ggplot(aes(x=Weekday, y=Total), data=total_week1) +
  geom_col(width = .3, fill="#669933") +
  theme( axis.text.x  = element_text(angle=90, vjust=0.5),
         panel.background = element_rect(fill = "white", colour='black'))

grid.arrange(
  grobs = list(eda1,eda2,eda3),
  widths = c(2,1),
  layout_matrix = rbind(c(1, 3),
                        c(2, 2))
)
## BY CUSTOMER EDA

week1 <- dat[which(dat$Weekend==0),-51]
week1 <- as.data.frame(cbind(rowSums(week1),"Week"))
names(week1) <- c("Purchases","Weekday")
#week <- cbind(Variable=rownames(week),Total=week, Weekday="Week")

weekend1 <- dat[which(dat$Weekend==1),-51]
weekend1 <- as.data.frame(cbind(rowSums(weekend1),"Weekend"))
names(weekend1) <- c("Purchases","Weekday")

#weekend <- cbind(Variable=rownames(weekend),weekend, Weekday="Weekend")

total_week_cus <- rbind(week1,weekend1)


eda5 <- ggplot(total_week_cus,aes(x=Purchases, fill=Weekday))  +
  geom_bar() + labs(y="Frequency", x="Number of purchases")+
  theme(panel.background = element_rect(fill = "white", colour='black'))+
  scale_fill_brewer(palette = "Dark2",direction=-1)+coord_flip()


eda4 <- ggplot(total_week_cus,aes(x=Purchases))  +
  geom_bar(fill="#660099") + labs(y="Frequency", x="Number of purchases")+
  theme(panel.background = element_rect(fill = "white", colour='black'))
```

```r
grid.arrange(
  grobs = list(eda4,eda5),
  widths = c(1,2))

bakery<- read.csv("bakery.csv", sep = ",", header = T)
bakery.data <- bakery[,-51]

# Using gower distance as dissimilarity matrix
dis.matrix <- daisy(bakery.data, metric = "gower") #dissimilarity matrix
k.meds <- pam(dis.matrix, k=8, metric = "euclidean")
# Using correlation as dissimilarity matrix
corr <- as.dist(1-cor(t(bakery.data))) # Get correlation as dissimilirity
clusB8 <- pam(corr,8, metric="euclidean") # Clustering with K=8

s1 <- fviz_nbclust(bakery.data, pam, diss = dis.matrix, k.max = 8,
method = "silhouette") + labs(subtitle = "Silhouette method: Gower distance.")
s11 <- fviz_silhouette(k.meds, print.summary = F)

s2<- fviz_nbclust(bakery.data, pam, diss = corr, k.max = 8,
method = "silhouette", nboot = 5) +
  labs(subtitle = "Silhouette method: Correlation distance.") # K=8
s22 <- fviz_silhouette(clusB8, print.summary = F)

grid.arrange(grobs=list(s1,s11,s2,s22),nrow=2,ncol=2)

# Method B
dat_clus_B <- data.frame(cbind(bakery,clusB8$clustering))
names(dat_clus_B) <- c(colnames(bakery),"K8")

## VISUALIZE FOR K=8

## Get TOP 5 sums of purchases/total purchases for each product in each cluster number

dat_clus_B1 <- data.frame(Total=colSums(dat_clus_B[which(dat_clus_B$K8=="1"),1:50]))
dat_clus_B1 <- cbind(Variable=rownames(dat_clus_B1),dat_clus_B1,Cluster=1)
dat_clus_B1 <- dat_clus_B1[order(dat_clus_B1$Total,decreasing=T)[1:5],]


dat_clus_B2 <- data.frame(Total=colSums(dat_clus_B[which(dat_clus_B$K8=="2"),1:50]))
dat_clus_B2 <- cbind(Variable=rownames(dat_clus_B2),dat_clus_B2,Cluster=2)
dat_clus_B2 <- dat_clus_B2[order(dat_clus_B2$Total,decreasing=T)[1:5],]
```

```r
dat_clus_B3 <- data.frame(Total=colSums(dat_clus_B[which(dat_clus_B$K8=="3"),1:50]))
dat_clus_B3 <- cbind(Variable=rownames(dat_clus_B3),dat_clus_B3,Cluster=3)
dat_clus_B3 <- dat_clus_B3[order(dat_clus_B3$Total,decreasing=T)[1:5],]


dat_clus_B4 <- data.frame(Total=colSums(dat_clus_B[which(dat_clus_B$K8=="4"),1:50]))
dat_clus_B4 <- cbind(Variable=rownames(dat_clus_B4),dat_clus_B4,Cluster=4)
dat_clus_B4 <- dat_clus_B4[order(dat_clus_B4$Total,decreasing=T)[1:5],]


dat_clus_B5 <- data.frame(Total=colSums(dat_clus_B[which(dat_clus_B$K8=="5"),1:50]))
dat_clus_B5 <- cbind(Variable=rownames(dat_clus_B5),dat_clus_B5,Cluster=5)
dat_clus_B5 <- dat_clus_B5[order(dat_clus_B5$Total,decreasing=T)[1:5],]

## GATHER ALL CLUSTERS AND SUMS
dat_clus_Bk8 <- rbind(dat_clus_B5,dat_clus_B2,dat_clus_B3,dat_clus_B4,
                      dat_clus_B1,dat_clus_B6,dat_clus_B7,dat_clus_B8)
dat_clus_Bk8$Cluster <- as.factor(dat_clus_Bk8$Cluster)

ggplot(dat_clus_Bk8, aes(x=Cluster,y=Total,group=Variable, fill=Cluster))+
  geom_bar(position="dodge", stat="identity",  colour="white")+
  theme(panel.background = element_rect(fill = "white", colour='black'),
        legend.position = "none")+
  geom_text(aes(label=Variable),angle=90,position = position_dodge(width = 1),size=2)+
  scale_fill_brewer(palette = "Dark2")

dat_clus_B$Weekend <- as.factor(dat_clus_B$Weekend )

w1 <- ggplot(dat_clus_B, aes(x=as.factor(K8), fill=Weekend, group=Weekend))+geom_bar()+
  labs(x="Cluster K=8", y="Number of purchases")+
  scale_fill_brewer(palette = "Dark2",direction=-1)+
  theme(panel.background = element_rect(fill = "white", colour='black'))+coord_flip()

dat_clusB_weekend <- dat_clus_B[which(dat_clus_B$Weekend==1),]
w2 <- ggplot(dat_clusB_weekend, aes(x=as.factor(K8)))+geom_bar(fill="#669933")+
  labs(x="Cluster K=8", y="Number of purchases")+
  theme(panel.background = element_rect(fill = "white", colour='black'))
grid.arrange(w1,w2, widths=c(1,1))


## ANALYSE WEEKEND SALE

clusterB.weekend <-melt(dat_clus_B[which(dat_clus_B$Weekend==1),-51],
id.vars = c("K8"))
```

```
clusterB.weekend <- cbind(clusterB.weekend, Weekday="Weekend")



clusterB.weekday <-melt(dat_clus_B[which(dat_clus_B$Weekend==0),-51],
id.vars = c("K8"))
clusterB.weekday<- cbind(clusterB.weekday, Weekday="Weekday")


clusterB <- rbind(clusterB.weekend,clusterB.weekday)
clusterB <- clusterB[which(clusterB$value==1),]


clus1<- ggplot(clusterB[which(clusterB$K8==1),], aes(x=variable, fill=Weekday))
+ geom_bar()+
  theme( axis.text.x  = element_text(angle=90, vjust=0.5,size=5),
         panel.background = element_rect(fill = "white", colour='black'))+
  labs(title="Cluster 1")+
  scale_fill_brewer(palette = "Dark2",direction=-1)


clus2<- ggplot(clusterB[which(clusterB$K8==2),], aes(x=variable, fill=Weekday))
+ geom_bar()+
  theme( axis.text.x  = element_text(angle=90, vjust=0.5,size=5),
         panel.background = element_rect(fill = "white", colour='black'))+
  labs(title="Cluster 2")+
  scale_fill_brewer(palette = "Dark2",direction=-1)


clus3<- ggplot(clusterB[which(clusterB$K8==3),], aes(x=variable, fill=Weekday))
+ geom_bar()+
  theme( axis.text.x  = element_text(angle=90, vjust=0.5,size=5),
         panel.background = element_rect(fill = "white", colour='black'))+
  labs(title="Cluster 3")+
  scale_fill_brewer(palette = "Dark2",direction=-1)

clus4<- ggplot(clusterB[which(clusterB$K8==4),], aes(x=variable, fill=Weekday))
+ geom_bar()+
  theme( axis.text.x  = element_text(angle=90, vjust=0.5,size=5),
         panel.background = element_rect(fill = "white", colour='black'))+
  labs(title="Cluster 4")+
  scale_fill_brewer(palette = "Dark2",direction=-1)

clus5<- ggplot(clusterB[which(clusterB$K8==5),], aes(x=variable, fill=Weekday))
```

```
+ geom_bar()+
  theme( axis.text.x  = element_text(angle=90, vjust=0.5,size=5),
         panel.background = element_rect(fill = "white", colour='black'))+
  labs(title="Cluster 5")+
  scale_fill_brewer(palette = "Dark2",direction=-1)



grid.arrange(
  grobs=list(clus1,clus2,clus3,clus4,clus5),
  nrow=4)
```