MCDA5580 Assignment 1

Team Member

Hemalatha Srinivasan	A00452621
Ajay Jain	A00455849
Kin Wa, Chan	A00467755

Table of Content

Executive Summary

We took 2000 records from the sample data received from Sobey's transactions and analyzed using K-Means Algorithm for clustering the customer and products based on the result we received.

Customer Segments

Based on the K-Means Algorithm we formed 5 clusters for the total record and divided customers into 5 different categories based on customers behaviours. The results are given below.

Cluster – 1 Elite Customers

Cluster – 2 Loyal Customers

Cluster – 3 Undecided Customers

Cluster – 4 Weekend Bees

Cluster – 5 Need-Based Customers

Product Segments

Products were clustered into 5 categories. One cluster required attention which is having low price but not many sales. There are premium categories which have high prices and sales and are performing well for stores. Staples are also performing well in terms of sales. The results are given below.

Cluster – 1 Premium Staples

Cluster – 2 Premium Products

Cluster – 3 Problem Area

Cluster – 4 Semi Premium

Cluster – 5 Staples

The detailed behavior of the clusters we have mentioned below in Cluster Analysis segment.

Objectives

Transaction data of a Sobey store is collected to improve sales performance. The data is analyzed by K-mean clustering method (product clustering and customer clustering) in order to discover customer behavior and buying preference of product and produce recommendations and strategy to improve sales.

Main objectives

- Find out product and customer groups lead to highest revenue and try to maintain the strength.
- Find out potential product and customer groups that have space to improve the revenue from them.

Design/ Methodology/ Approach

K mean clustering method is used to discover the potential clusters for product and customer because the feature of clusters is unknown. It is used to identify the group share common characteristics in the complex data set.

Before performing the K mean clustering, data cleaning will be carried out. Data will be visualized by ggpair function in R and with help of box plot, outliners will be found and removed. After that, elbow plot will be used to define the optimal clusters. Finally, the resulted data will be used for K mean analysis.

Customer Cluster Analysis

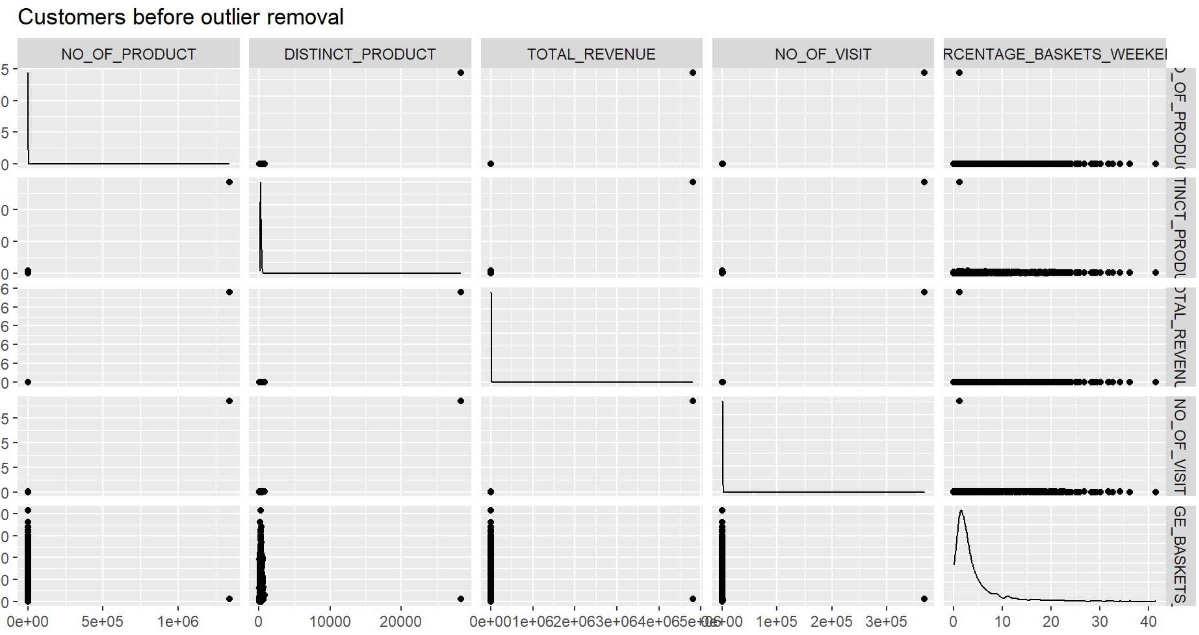
Selected Columns

Feature Name	**Measurement**	**Description**
CUSTOMER_SK	N/A	Unique field for identifying customer
NO_OF_PRODUCT	COUNT	Total number of products brought by a customer
DISTINCT_PRODUCT	COUNT DISTINCT	Total number of distinct product brought by a customer
TOTAL_REVENUE	SUM	Total revenue of distinct generated from a customer
NO_OF_VISIT	COUNT	Total number visit of a customer, assume each basket created is a visit for the customer
PERCENTAGE_BASKETS_WEEKEND (Additional attribute)	COUNT PERCENT	Percentage of basket created on weekend (Saturday & Sunday). Use to evaluate customer purchase behavior.

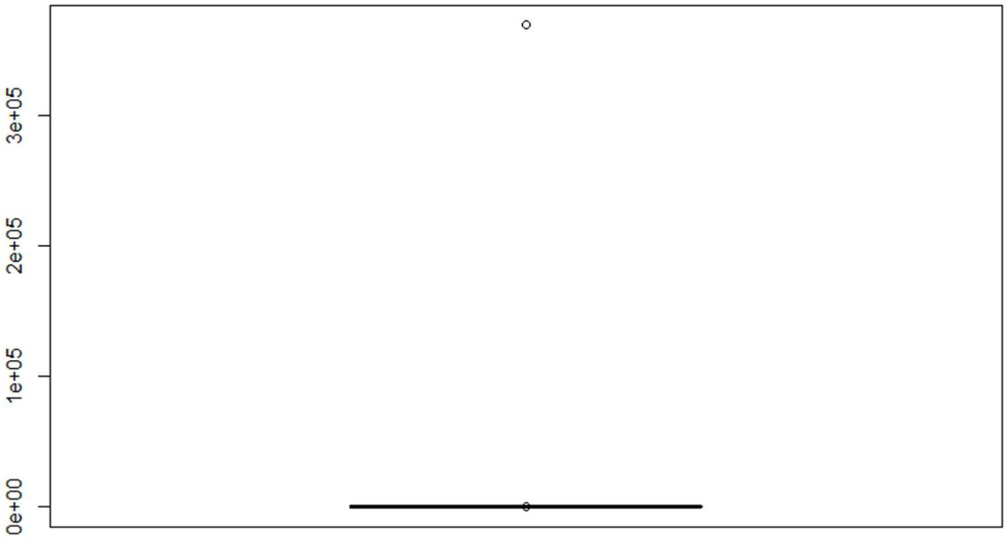
Data Cleaning

Data distribution (Before outliner removal)

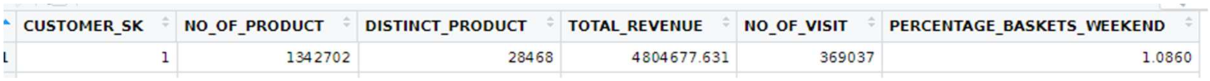
There is obvious outliner in “NO_OF_VISIT” dimension, use box plot for further analysis.



Box Plot for field “NO_OF_VISIT”



Outliner: CUSTOMER_SK = 1

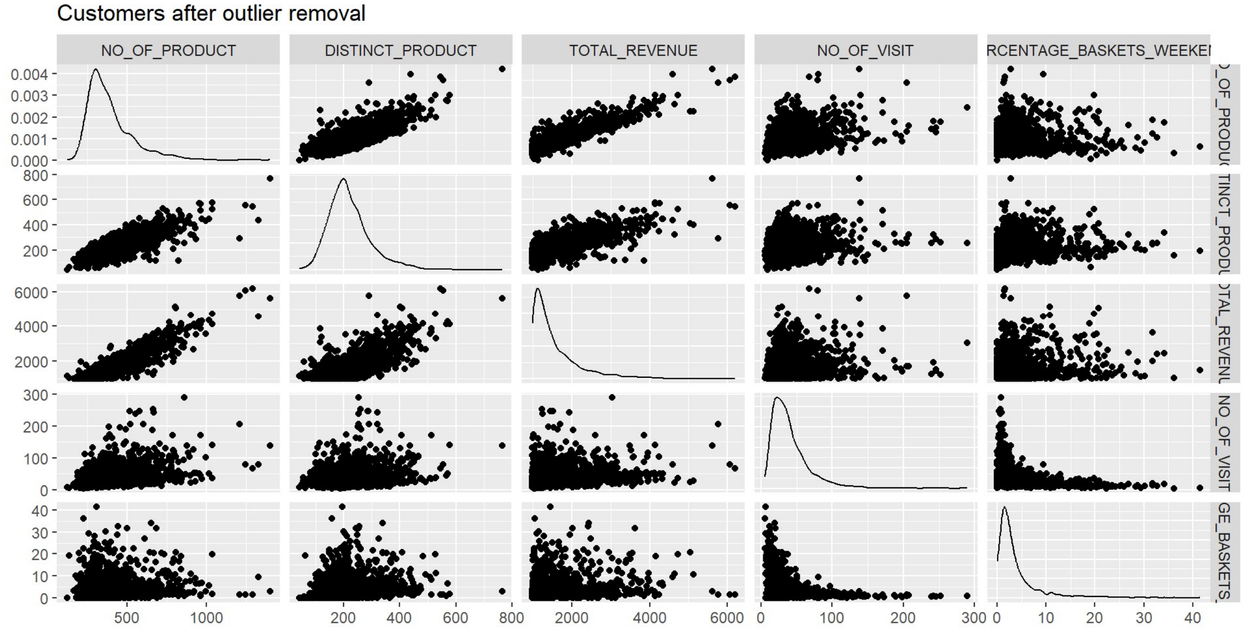


CUSTOMER_SK	NO_OF_PRODUCT	DISTINCT_PRODUCT	TOTAL_REVENUE	NO_OF_VISIT	PERCENTAGE_BASKETS_WEEKEND
1	1342702	28468	4804677.631	369037	1.0860

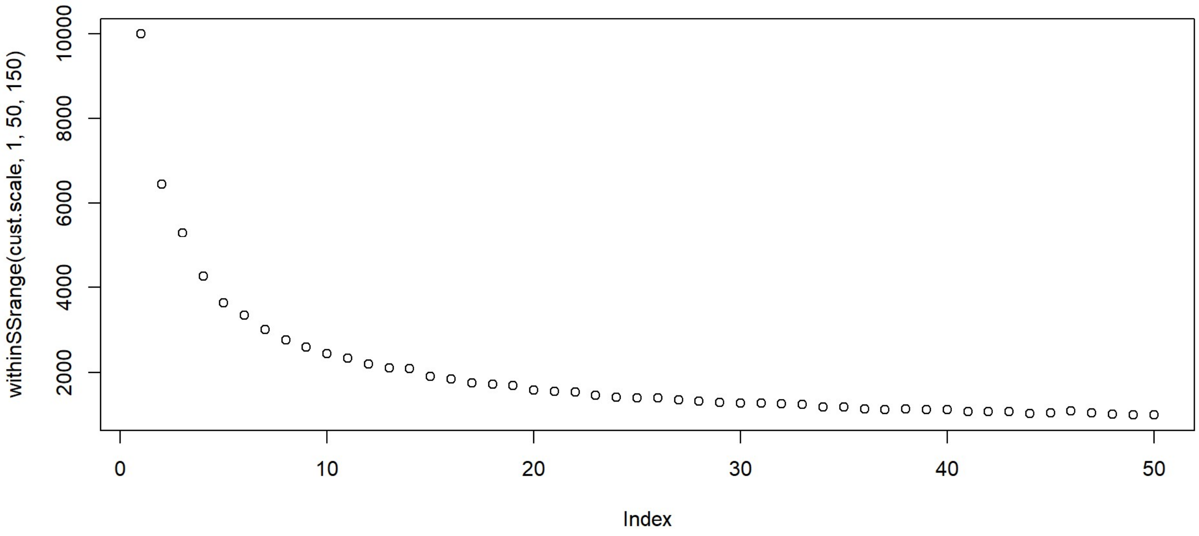
<u>Reason for choosing outliner:</u>

The NO_OF_VISIT of record of CUSTOMER_SK =1 is 369037, which is obviously higher than other records (The NO_OF_VISIT of other records in range of 5 - 290). So, we will remove this record in the analysis.

Data distribution (After outliner removal)

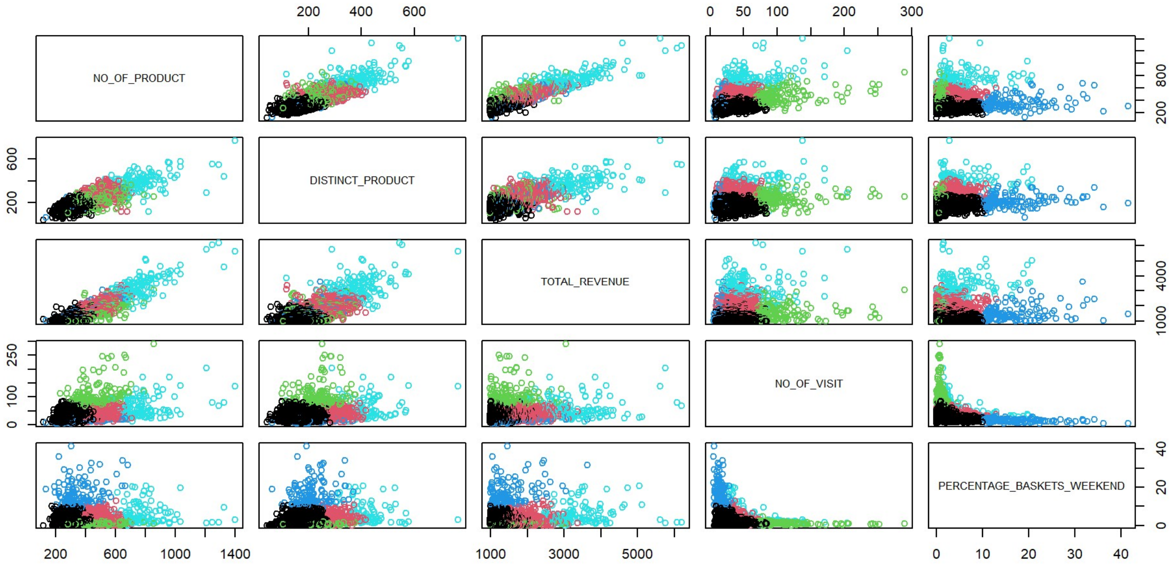


Define no. of cluster (elbow plot)

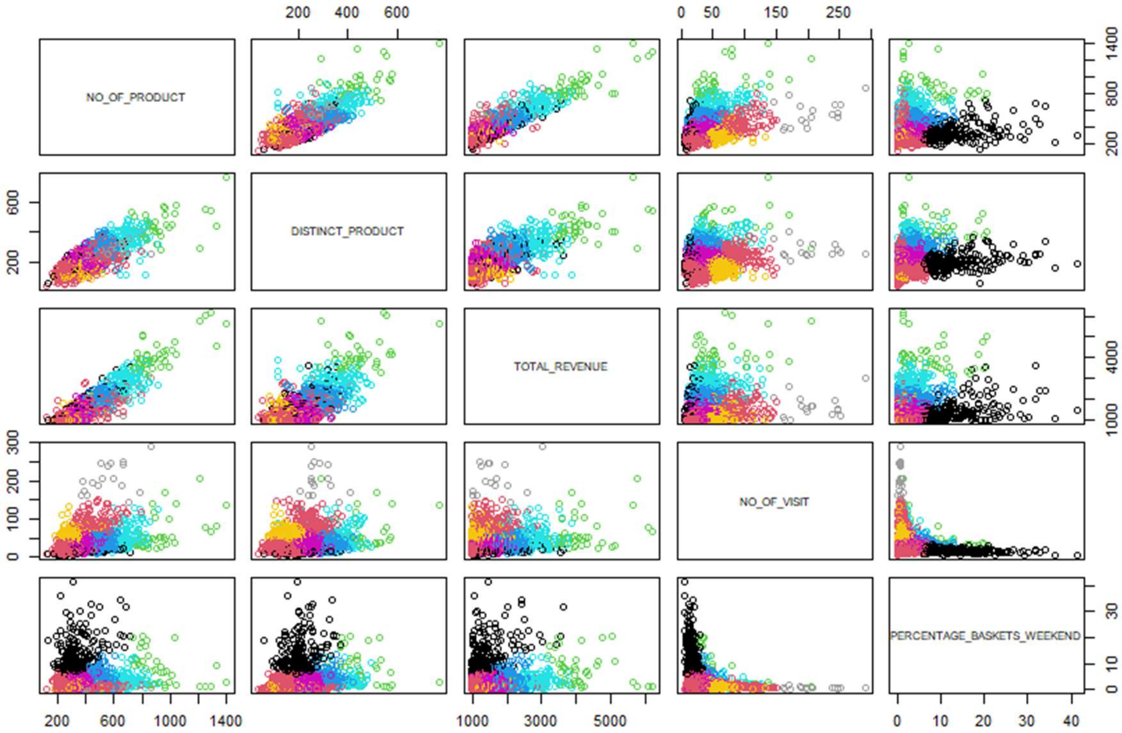


With reference to the slope of the elbow plot the number of clusters should be near 5 to 10. So, try to plot K means with clusters = 5 and 10.

K means with clusters = 5

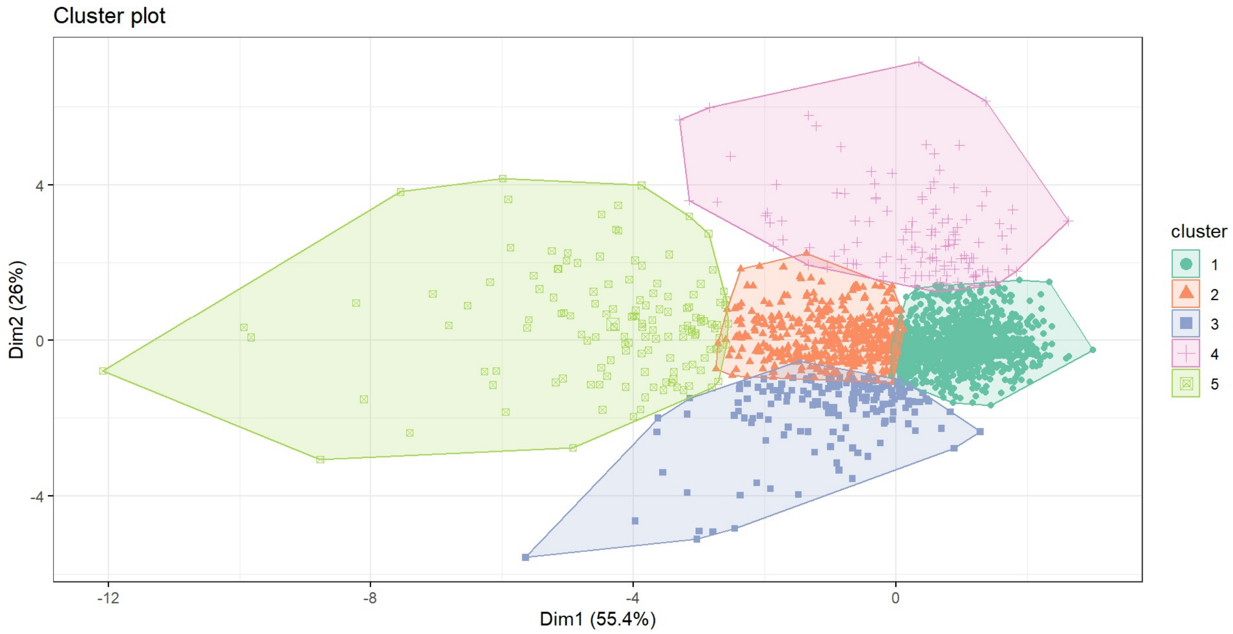


K means with clusters = 10



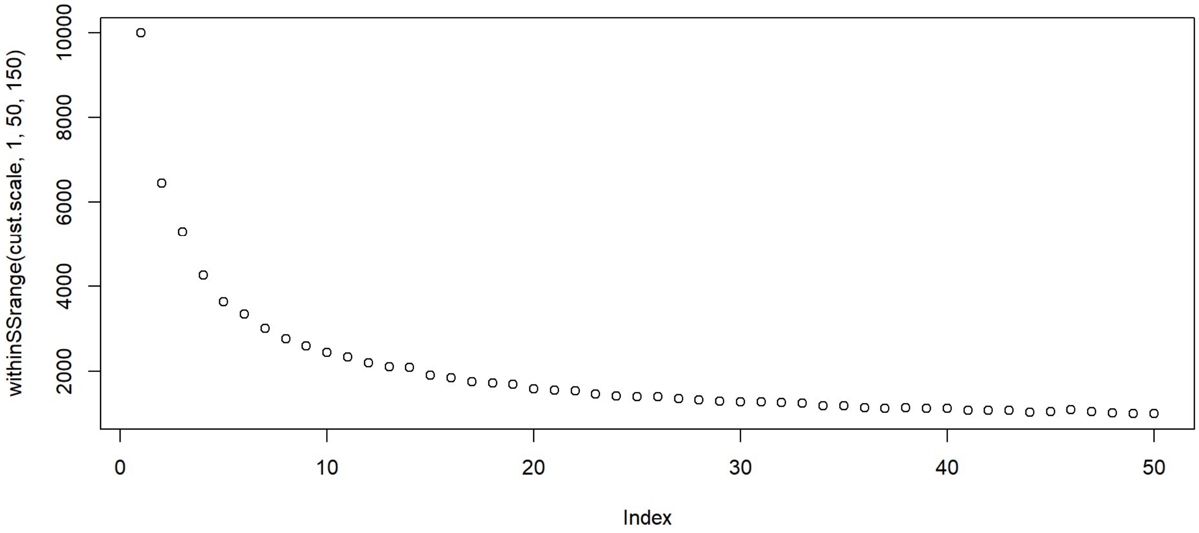
Comparing the cluster plots, when number of clusters = 10, the cluster size is small and many of them have some overlaps. So, we will choose the number of clusters = 5 for the result.

Fviz plot to visualize clusters with PCA

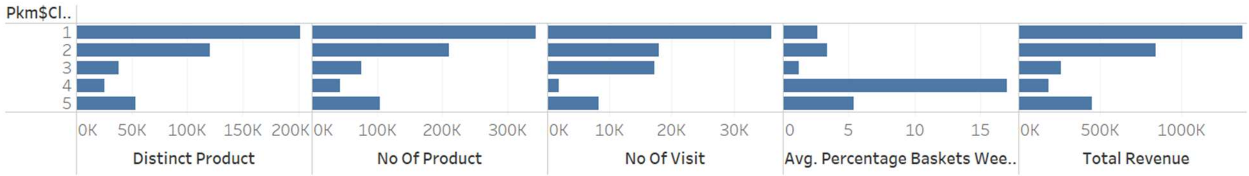


Cluster Analysis

With respect to below Elbow graph from range 1 to 50, we won't be able to see any major improvements after 5th point so we decided 5 as number of clusters to be considered which can avoid deviation of results.



Customer Cluster Summary



Customer Segment	Description	Recommendation
Cluster – 1 Elite Customers	Constantly visit the store irrespective to day Purchase large number of items based on their wish. Try to invest in different types of products. Generates large revenue to store	We can consider them to purchase membership which can help them to get discounts and bonus in return. We can make them happy to continue their shopping continuously. We can provide regular gifts during festival seasons.
Cluster – 2 Loyal Customers	They buy regularly. Involves constant purchase and generate considerable revenue	Increase the visibility of the product they purchase constantly. Increase the stocks they buy
Cluster – 3 Undecided Customers	They make high frequent visits. When compared to visits, they buy less product. They have constant dilemma to buy things	Provide discounts for the products they buy. Provide discount vouchers as encouragement to buy.
Cluster – 4 Weekend Bees	They prefer weekends when compared to weekdays. Their purchase during weekends more.	Add new and fresh stocks during weekends. Introduce weekend offers and send them email.
Cluster – 5 Need-Based Customers	They purchase based on their daily need. Daily visitors Doesn't consider buying more products	Keep basic necessary products in the front.

Product Cluster Analysis

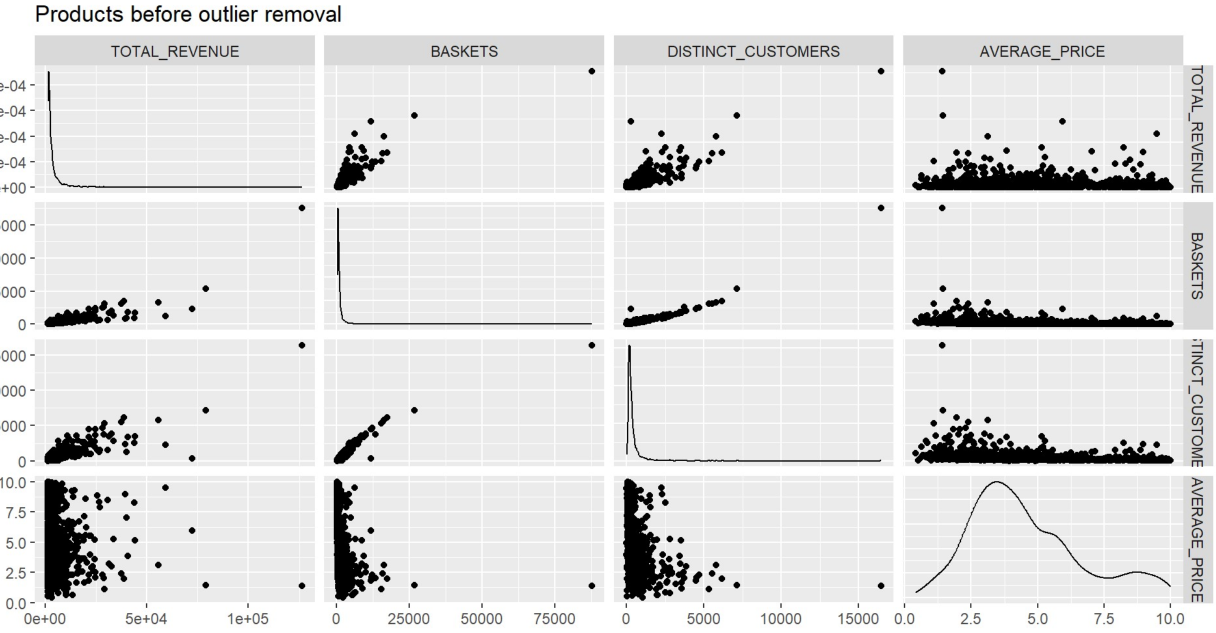
Selected Columns

Feature Name	Measurement	Description
ITEM_SK	N/A	Unique field for identifying products
TOTAL_REVENUE	COUNT	Total revenue generated by store from each product.
BASKETS	COUNT	Total number of transactions for each product.
DISTINCT_CUSTOMERS	COUNT DISTINCT	Unique customers who have purchased the product.
AVERAGE_PRICE	AVERAGE	Average sell price of the product.

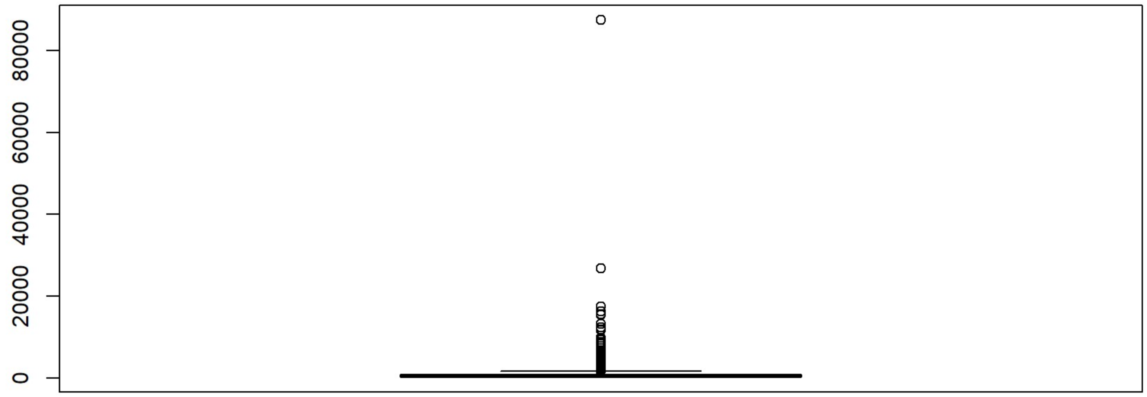
Data Cleaning

Data distribution (Before outliner removal)

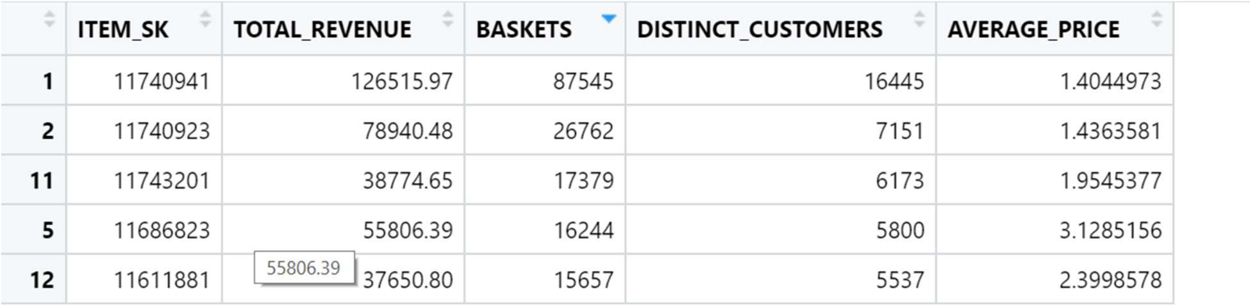
There is one clear outlier which can be seen in second column with more than 75000 baskets. There is one more with more than 25000 baskets.



Box Plot for field "Baskets"



Outliner: ITEM_SK = 11740941 and 11740923 (First two rows from below)

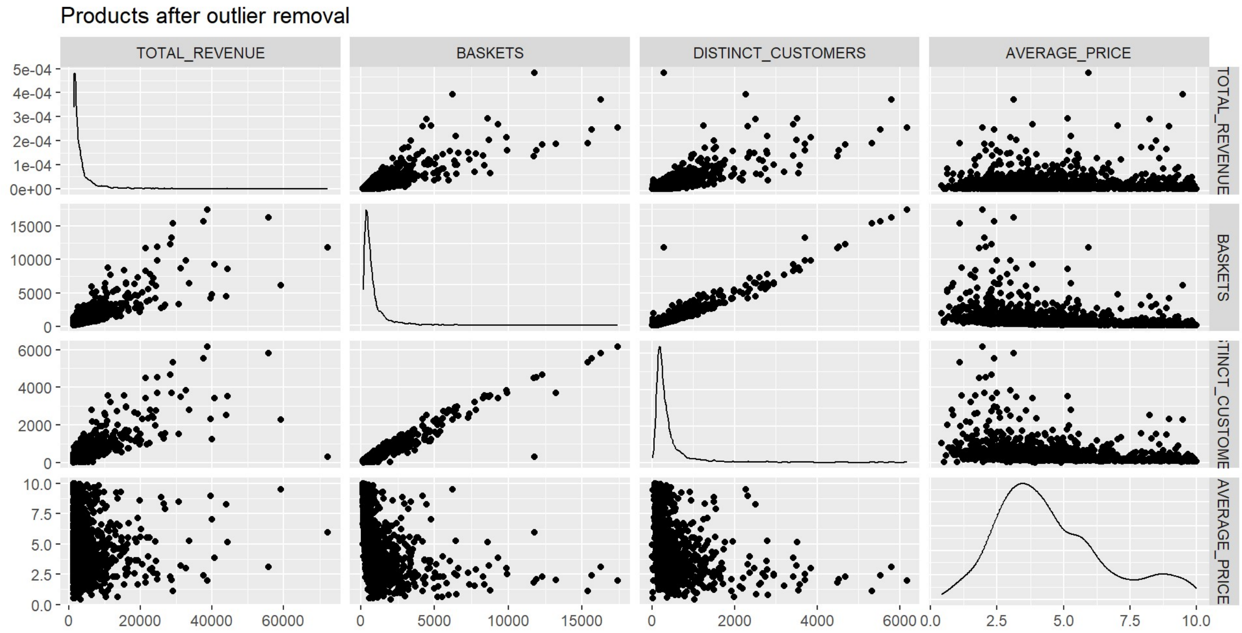


	ITEM_SK	TOTAL_REVENUE	BASKETS	DISTINCT_CUSTOMERS	AVERAGE_PRICE
1	11740941	126515.97	87545	16445	1.4044973
2	11740923	78940.48	26762	7151	1.4363581
11	11743201	38774.65	17379	6173	1.9545377
5	11686823	55806.39	16244	5800	3.1285156
12	11611881	37650.80	15657	5537	2.3998578

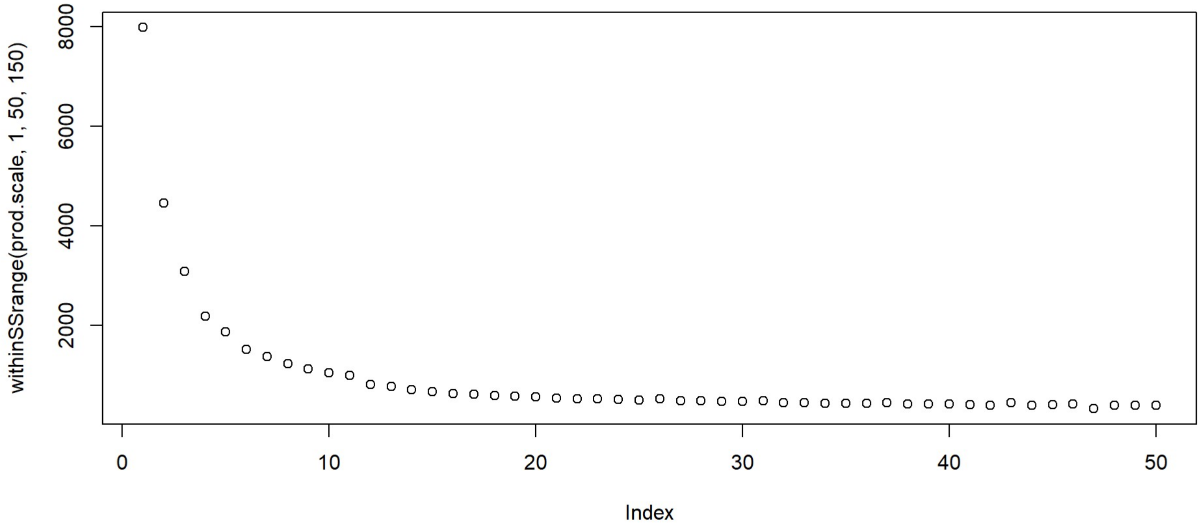
<u>Reason for choosing outliner</u>

First two data points are having 75000 and 25000 as number of baskets. Next point is around 17000 baskets. Hence removing both the points as outlier.

Data distribution (After outliner removal)

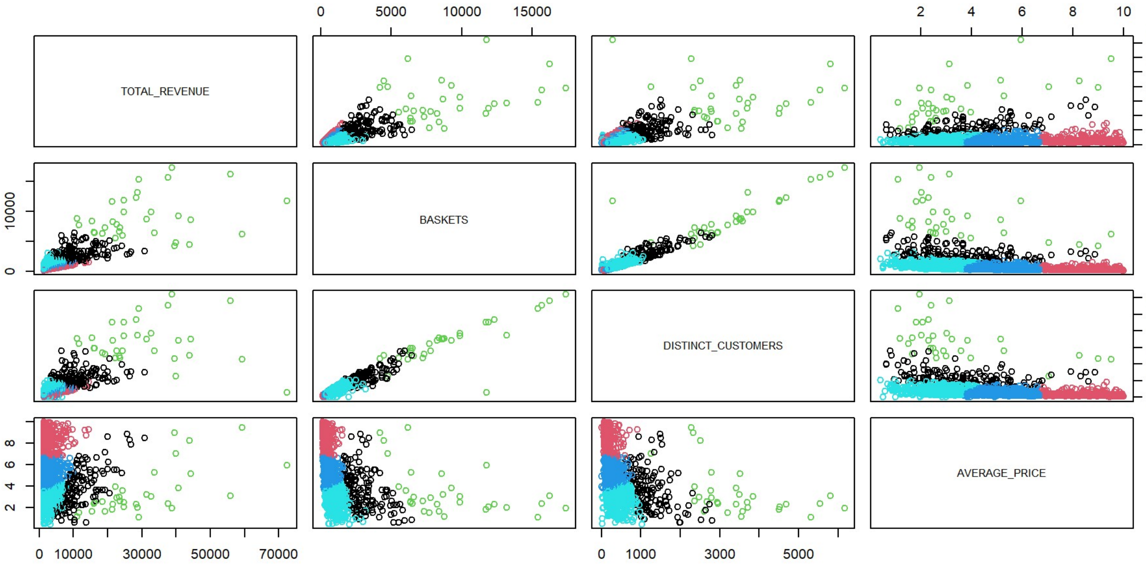


Define no. of cluster (elbow plot)

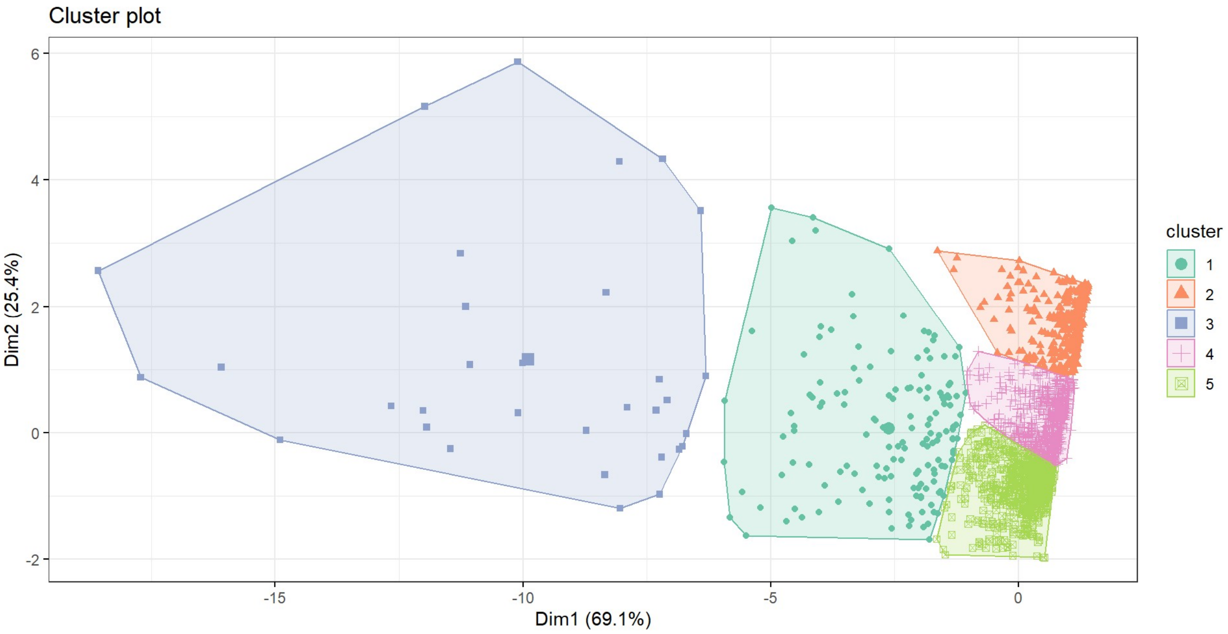


With reference to the slope of the elbow plot the number of clusters should be 5 or 6 since error is reducing till those points. Keeping it 5 since difference in errors between 5 and 6 clusters is not significant and 5 clusters will be easier to visualize and analyze.

K means with clusters = 5

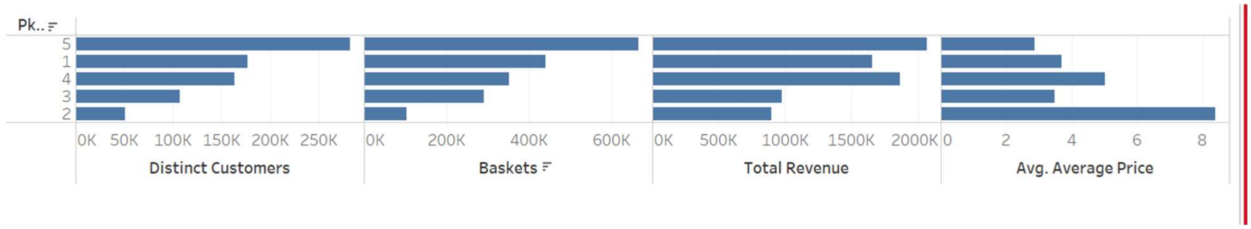


Fviz plot to visualize clusters with PCA



Cluster Analysis

Product Cluster Summary



Customer Segment	Description	Recommendation
Cluster – 1 Premium Staples	Many customers are buying it frequently. Revenue generation is good.	Continue to monitor.
Cluster – 2 Premium Products	Average price is higher. Less transactions but decent revenue.	Increase the visibility of the product.
Cluster – 3 Problem Area	Low prices but still not many customers are buying it.	Explore further to see if individual products within this cluster are loss- making.
Cluster – 4 Semi Premium	Prices are slightly on higher side. Many customers buy it frequently. Revenue generation is good.	Continue to monitor.
Cluster – 5 Staples	Low price but many customers buy it frequently. Best revenue generator.	Increase the price a little bit to see the response. May be?

Conclusion

The analysis had met the objective stated. 5 customers group and 5 product group are defined. The “Premium Staples”,” Semi Premium”,” Staples” product group and “Elite” customer group are the most beneficial group for revenue generation, continuous monitoring is needed to ensure they are doing well in future as well.

The “Loyal”, “Weekend Bees”, and “Need-Based” are potential customer groups, we discovered performance on them seems to be positive and we had defined recommendations improve the revenue generated from them.

For product cluster “Premium”, it also had potential to improve its sales performance by increase the visibility.

Appendix

Customer Cluster Analysis

SQL

```
SELECT
CUSTOMER_SK,
count(ITEM_QTY) as NO_OF_PRODUCT,
count(distinct ITEM_SK) as DISTINCT_PRODUCT,
sum(SELLING_RETAIL_AMT) as TOTAL_REVENUE,
count(distinct TRANSACTION_RK) as NO_OF_VISIT,
COUNT(IF(WEEKDAY(`date`)>=5, 1, NULL))/count(distinct TRANSACTION_RK) as
PERCENTAGE_BASKETS_WEEKEND
FROM dataset01.sales219
GROUP BY CUSTOMER_SK
ORDER BY TOTAL_REVENUE DESC
LIMIT 2000;
```

R Script

```
library(ggplot2)
library(GGally)
library(DMwR)
library(factoextra)

set.seed(84)

cust <- read.csv("sales219_CustomerCluster_2000-2.csv")
summary(cust)

ggpairs(cust[, which(names(cust) != "CUSTOMER_SK")],
      upper = list(continuous = ggally_points),
    lower = list(continuous = "points"),
    title = "Customers before outlier removal")

boxplot(cust$DISTINCT_PRODUCT) # For Box and Whisker plot.
                # here cust is dataset and Distinct_prodcust is column

cust.clean <- cust[cust$CUSTOMER_SK != 1, ]   # Remove outliers

ggpairs(cust.clean[, which(names(cust) != "CUSTOMER_SK")],
    upper = list(continuous = ggally_points),
    lower = list(continuous = "points"),
    title = "Customers after outlier removal")

# Normalize data using scale and exclude CUSTOMER_SK column.
# -1 will remove first column that is CUSTOMER_SK and keep all other.

cust.scale = scale(cust.clean[-1])

withinSSrange <- function(data,low,high,maxIter)
{
  withinss = array(0, dim=c(high-low+1));
  for(i in low:high)
  {
    withinss[i-low+1] <- kmeans(data, i, maxIter)$tot.withinss
  }
  withinss
}
```

```
# Elbow plot to determine the optimal number of clusters between 1 and 50.

plot(withinSSrange(cust.scale,1,50,150))

# K-means using k=5 for products based on results of  elbow plot.
pkm = kmeans(cust.scale, 5, 150)

# Denormalize data by reversing scale function
cust.realCenters = unscale(pkm$centers, cust.scale)

# Bind clusers to cleansed Data
clusteredcust = cbind(cust.clean, pkm$cluster)

# Visualizing clusering results.
# Here we want all rows so we are not mentioning anything
# but we want columns only from 2 to 6
# (we don't want to visualize first column - ITEM_SK).
plot(clusteredcust[,2:6], col=pkm$cluster)



# fviz plot to visualize clusters with PCA

fviz_cluster(pkm, clusteredcust[2:6],
        palette = 'Set2',
        geom = "point",
        ellipse.type = "convex",
        ggtheme = theme_bw()
)
```

```
# Write data to csv

write.csv(clusteredcust, file='CustClusterOutput.csv')
```

Product Cluster Analysis

SQL

```
SELECT

ITEM_SK,

sum(SELLING_RETAIL_AMT) as TOTAL_REVENUE,

count(distinct TRANSACTION_RK) as BASKETS,

count(distinct CUSTOMER_SK) as DISTINCT_CUSTOMERS,

avg(SELLING_RETAIL_AMT/NULLIF(ITEM_QTY,0)) as AVERAGE_PRICE

FROM dataset01.sales219

GROUP BY ITEM_SK

ORDER BY TOTAL_REVENUE DESC

LIMIT 2000;
```

R Script

```
library(ggplot2)

library(GGally)

library(DMwR)

library(factoextra)


set.seed(84)


prod <- read.csv("productcluster.csv")

summary(prod)


ggpairs(prod[, which(names(prod) != "ITEM_SK")],
        upper = list(continuous = ggally_points),

        lower = list(continuous = "points"),

        title = "Products before outlier removal")


boxplot(prod$BASKETS) # For Box and Whisker plot.

                     # here prod is dataset and BASKETS is column


prod.clean <- prod[(prod$ITEM_SK != 11740941) &

                  (prod$ITEM_SK != 11740923), ]   # Remove outliers


ggpairs(prod.clean[, which(names(prod) != "ITEM_SK")],

        upper = list(continuous = ggally_points),

        lower = list(continuous = "points"),

        title = "Products after outlier removal")


# Normalize data using scale and exclude ITEM_SK column.

# -1 will remove first column that is ITEM_SK and keep all other.


prod.scale = scale(prod.clean[-1])


withinSSrange <- function(data,low,high,maxIter)

{

  withinss = array(0, dim=c(high-low+1));

  for(i in low:high)

  {

    withinss[i-low+1] <- kmeans(data, i, maxIter)$tot.withinss

  }

  withinss
```

```r
}

# Elbow plot to determine the optimal number of clusters between 1 and 50.

plot(withinSSrange(prod.scale,1,50,150))

# K-means using k=5 for products based on results of  elbow plot.
pkm = kmeans(prod.scale, 5, 150)

# Denormalize data by reversing scale function
prod.realCenters = unscale(pkm$centers, prod.scale)

# Bind clusers to cleansed Data
clusteredProd = cbind(prod.clean, pkm$cluster)

# Visualizing clusering results.
# Here we want all rows so we are not mentioning anything
# but we want columns only from 2 to 6
# (we don't want to visualize first column - ITEM_SK).
plot(clusteredProd[,2:5], col=pkm$cluster)



# fviz plot to visualize clusters with PCA

fviz_cluster(pkm, clusteredProd[2:5],
        palette = 'Set2',
        geom = "point",
        ellipse.type = "convex",
        ggtheme = theme_bw()
)


# Write data to csv
write.csv(clusteredProd, file='ProdClusterOutput.csv')
```