MCDA5580 - Data and Text Mining

Assignment – 1

Clustering Analysis for Customers and Products of Sobeys Inc.

Submitted By:
Nikhil Bhat (A00434789)
Mehar Singh (A00434701)
Samarth Gupta (A00433096)

Contents

1. Executive Summary

One of the top food retail franchises in Canada, Sobeys Inc. wants to analyze data records of sales to provide better service to customers and increase sales to maximize profits. As part of this analysis, retail transaction records of Sobeys Inc. from 01-Jan-2015 to 14-Sept-2015 were analysed and findings were reported that may be beneficial to Sobeys Inc. in understanding the nature of its customers and products.

The following clusters were identified to study the customer behavior:

- Impulsive Buyers
- Prospective Customers
- Loyal Customers
- Active Shoppers
- Erudite Customers
- Dormant Customer

The following clusters were identified for the study of products:

- Bulk Purchases
- In Demand Products
- Incumbent Products
- Attention Seekers
- Cash Cow
- Dogs

2. Objective

Analyze the Sobeys Inc. [1] sales data to understand customer purchase patterns and product patterns to classify them in distinct segments using centroid based K-Mean clustering analysis. The segmented data for customers should be used to understand the purchase patterns of customers and recommend marketing strategies for each customer group. Additionally, the product clusters should help in identifying similar characteristics in products and recommend strategies to boost the sales of the products for each cluster.

3. About the data

The Sobeys data stored in 'dataset01' database is used:

Server: http://dev.cs.smu.ca/phpmyadmin/

Database Type: MySQL Database

Tables Used:

- **Sales219:** Contains history of transactions of Sobeys Inc. for RETAIL_OUTLET_LOCATION_SK=219 in Canada
- **Members:** Contains the data about the organization's customers
- **Items:** Contains information of all products in Sobeys Inc inventory

4. Customer Analysis

4.1 Features

For the purpose of customer data analysis, the top 2000 revenue generating customers data from the table 'sales219' is stored into a temporary table and then joined with 'members' table to get customer details and store in the final table 'CustomerCluster'.

Following is schema of the 'CustomerCluster' table for our analysis:

S.No.	Column Name	Measure	Description
1.	CUSTOMER_SK	NA	Unique identifier for each customer
2.	NumProducts	SUM	The total quantity of products bought by the customer
3.	DistinctProducts	COUNT DISTINCT	The total distinct products bought by the customer
4.	Revenue	SUM	Total amount spent by the customer
5.	TotalNoOfVisits	COUNT DISTINCT	The total number of unique transactions carried out by the customer
6.	Recency	AVG	The difference between last transaction date+1day in the database and the last day customer carried out the last transaction [2]
7.	AvgSpend	AVG	The average amount spent by the customer for each transaction
8.	CITY_NM	NA	City for each customer

4.2 Descriptive Analysis

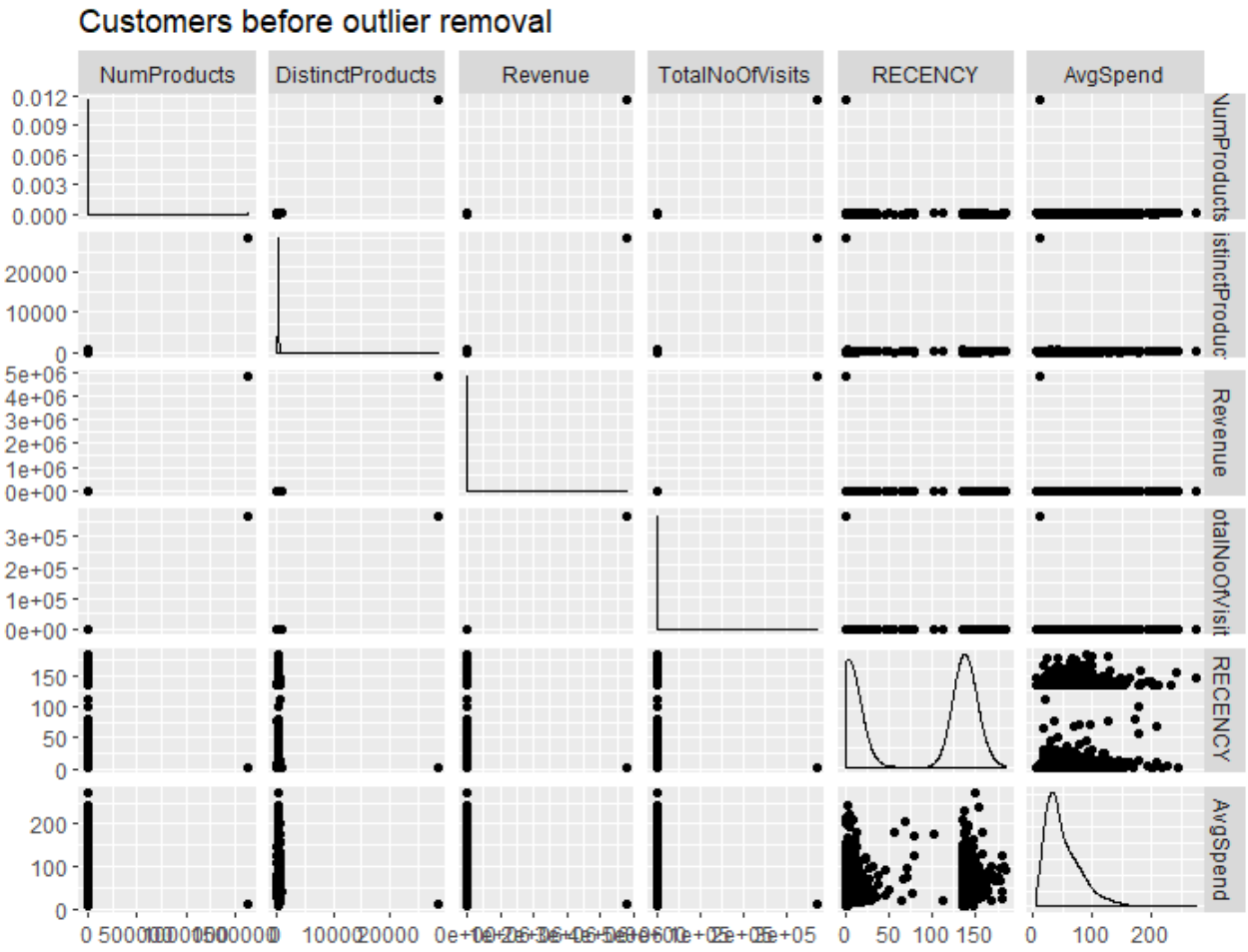
To get an idea of the data to be analysed and understand the range and distribution of the variables we take a summary of the top 2000 customers in 'CustomerCluster' table.

Summary of Top 2000 Customers:

NumProducts	DistinctProducts	Revenue	TotalNoOfVisits	Recency	AvgSpend
Min. : 214.0	Min. : 43.0	Min. : 978	Min. : 5.0	Min. : 1.00	Min. : 4.861
1st Qu.: 337.0	1st Qu.: 169.0	1st Qu.: 1124	1st Qu.: 22.0	1st Qu.: 3.00	1st Qu.: 28.977
Median : 406.0	Median : 207.0	Median : 1331	Median : 34.0	Median :134.00	Median : 42.633
Mean : 1277.8	Mean : 234.1	Mean : 3960	Mean : 225.2	Mean : 73.73	Mean : 52.274
3rd Qu.: 536.2	3rd Qu.: 256.0	3rd Qu.: 1758	3rd Qu.: 50.0	3rd Qu.:136.00	3rd Qu.: 67.379
Max. :1626564.0	Max. :28468.0	Max. :4804678	Max. :369037.0	Max. :185.00	Max. :273.919

4.3 Data Cleansing and Outliers Removal

For the purpose of data cleaning we first plot the different variables from the extracted data of CustomerCluster table to identify the outliers.



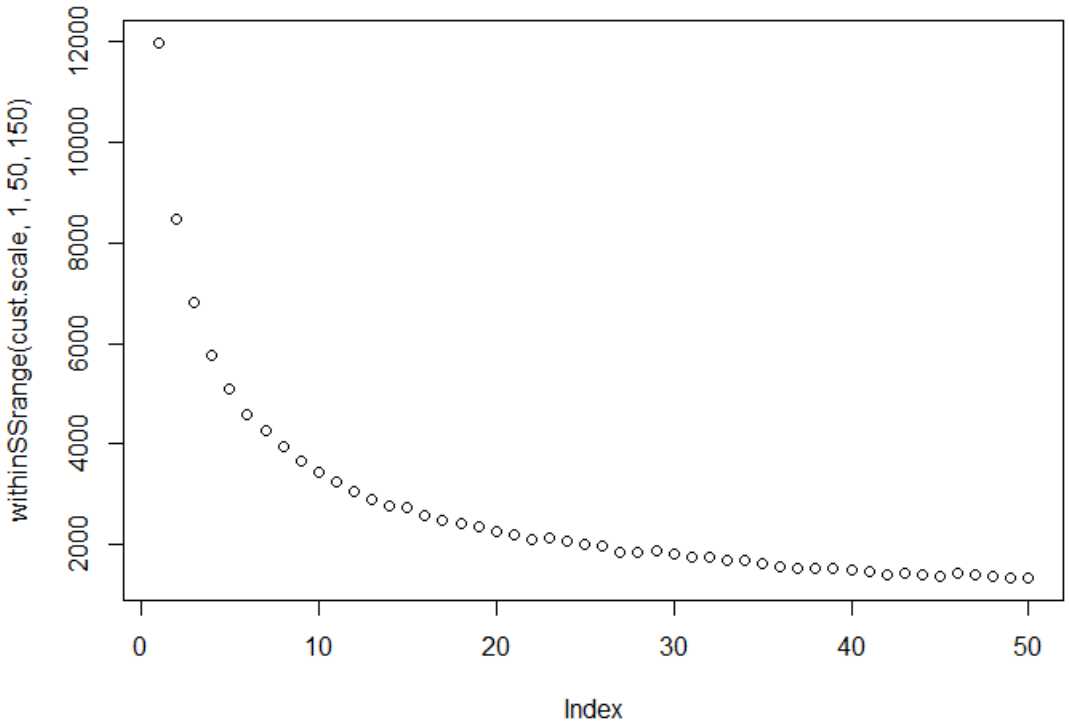
We find that Customer_SK = 1 is an outlier with very high total revenue as compared to other customers in the same table. Also, we find that this customer is of a customer type – 'NMr' and hence we consider this customer as an outlier to avoid biasing of other data points and ignore the data of this customer for the purpose of analysis.

4.4 Data Normalizing

The cleaned data has been re-scaled to a uniform scale for bringing data to a normalized format for the purpose of cluster analysis. Normalizing helps in reducing the impact of outliers and helps to compare the observations against the mean.

4.5 Selecting number of clusters

The normalized data is then passed through a K-Means Algorithm to find the number of clusters by plotting an elbow curve.



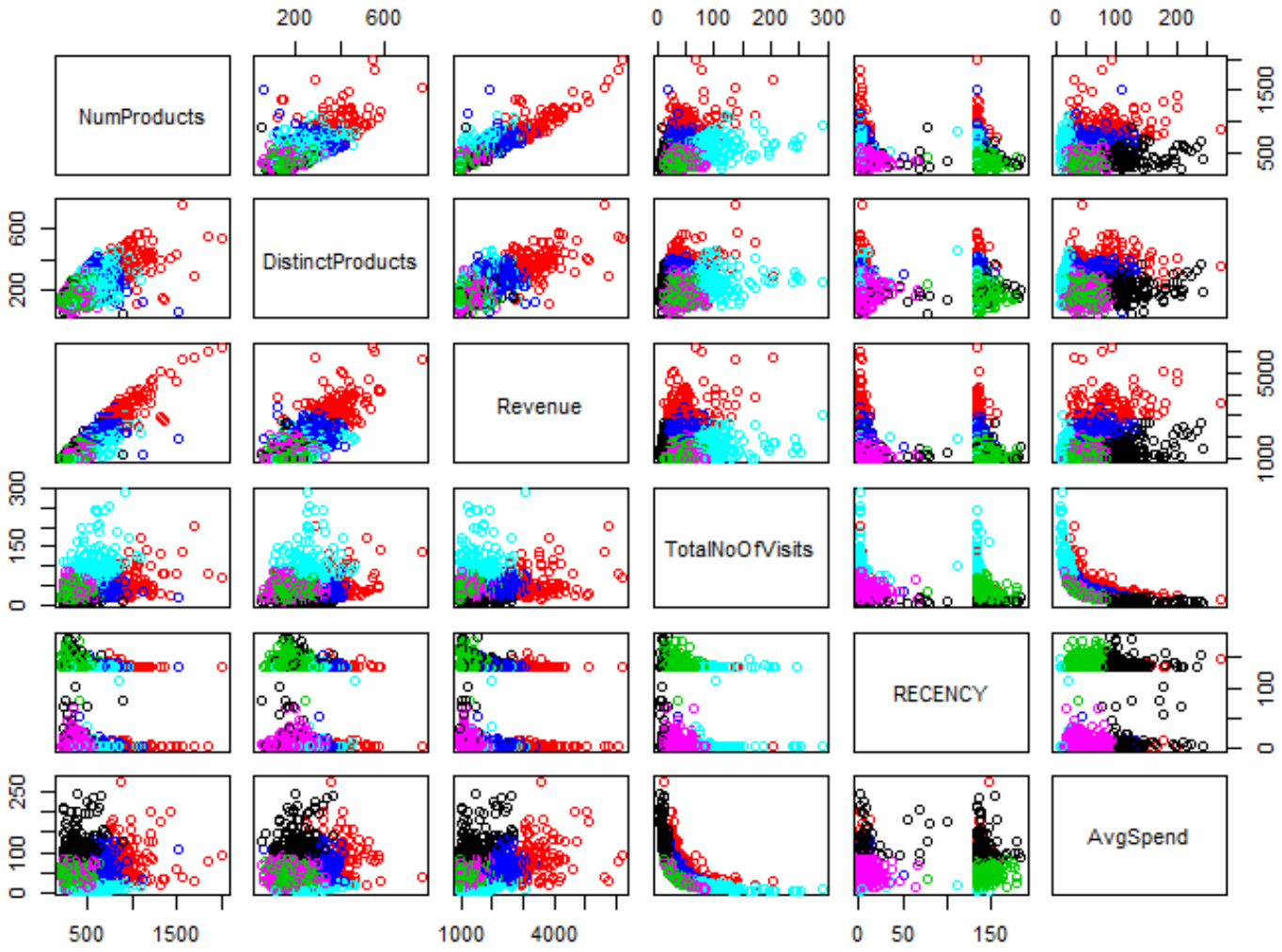
To get a more detailed view of the plotted curve for identification of the best K value, a table with WCSS values and number of clusters is generated:

```
[1] "K  : WCSS"
[1] " "
[1] "1"      ":"     "11988"
[1] "2"        ":"       "8463.359"
[1] "3"        ":"       "6812.157"
[1] "4"        ":"       "5755.768"
[1] "5"        ":"       "5089.287"
[1] "6"        ":"       "4581.246"
[1] "7"        ":"       "4262.629"
[1] "8"        ":"       "3940.889"
[1] "9"        ":"       "3661.289"
[1] "10"       ":"       "3425.922"
[1] "11"       ":"       "3236.055"
[1] "12"      ":"      "3063.93"
[1] "13"       ":"       "2887.304"
[1] "14"       ":"       "2771.486"
[1] "15"       ":"       "2725.644"
[1] "16"       ":"       "2579.126"
[1] "17"     ":"      "2493.1"
[1] "18"       ":"       "2408.909"
[1] "19"       ":"       "2365.788"
[1] "20"       ":"       "2266.843"
[1] "21"       ":"       "2207.831"
[1] "22"       ":"       "2105.958"
[1] "23"       ":"       "2122.209"
[1] "24"       ":"       "2061.986"
[1] "25"       ":"       "2001.659"
[1] "26"       ":"       "1961.208"
[1] "27"       ":"       "1851.302"
[1] "28"       ":"       "1841.793"
[1] "29"       ":"       "1865.031"
[1] "30"       ":"       "1805.529"
[1] "31"      ":"      "1751.85"
[1] "32"       ":"       "1749.193"
[1] "33"       ":"       "1678.957"
[1] "34"       ":"       "1695.355"
[1] "35"       ":"       "1627.514"
[1] "36"      ":"      "1565.21"
[1] "37"       ":"       "1523.686"
[1] "38"       ":"       "1527.559"
[1] "39"       ":"       "1514.499"
[1] "40"       ":"       "1510.537"
[1] "41"       ":"       "1461.611"
[1] "42"       ":"       "1413.709"
[1] "43"       ":"       "1431.335"
[1] "44"       ":"       "1396.484"
[1] "45"       ":"       "1366.401"
[1] "46"       ":"       "1416.411"
[1] "47"       ":"       "1396.452"
[1] "48"       ":"       "1357.331"
[1] "49"       ":"       "1330.393"
[1] "50"       ":"       "1332.781"
```

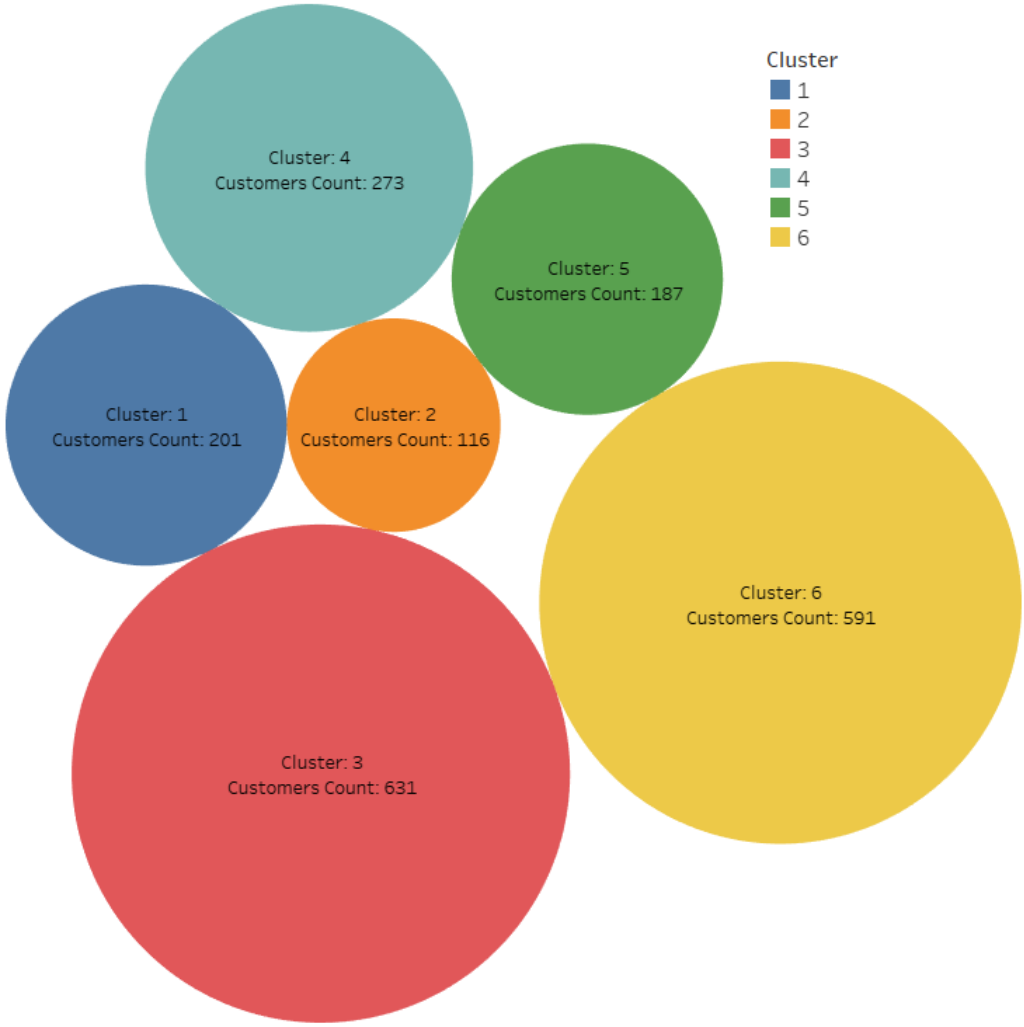
Based on the above elbow curve and generated table we see a significant drop in the WCSS values till K=6. Hence for the purpose of cluster analysis we use K=6 as the drop post this value is gradual.

4.6 Data De-normalizing and Cluster generation

After deriving the number of clusters, we again de-normalize the data to get actual centroids of the clusters and the cluster information is appended to the cleaned customer data.



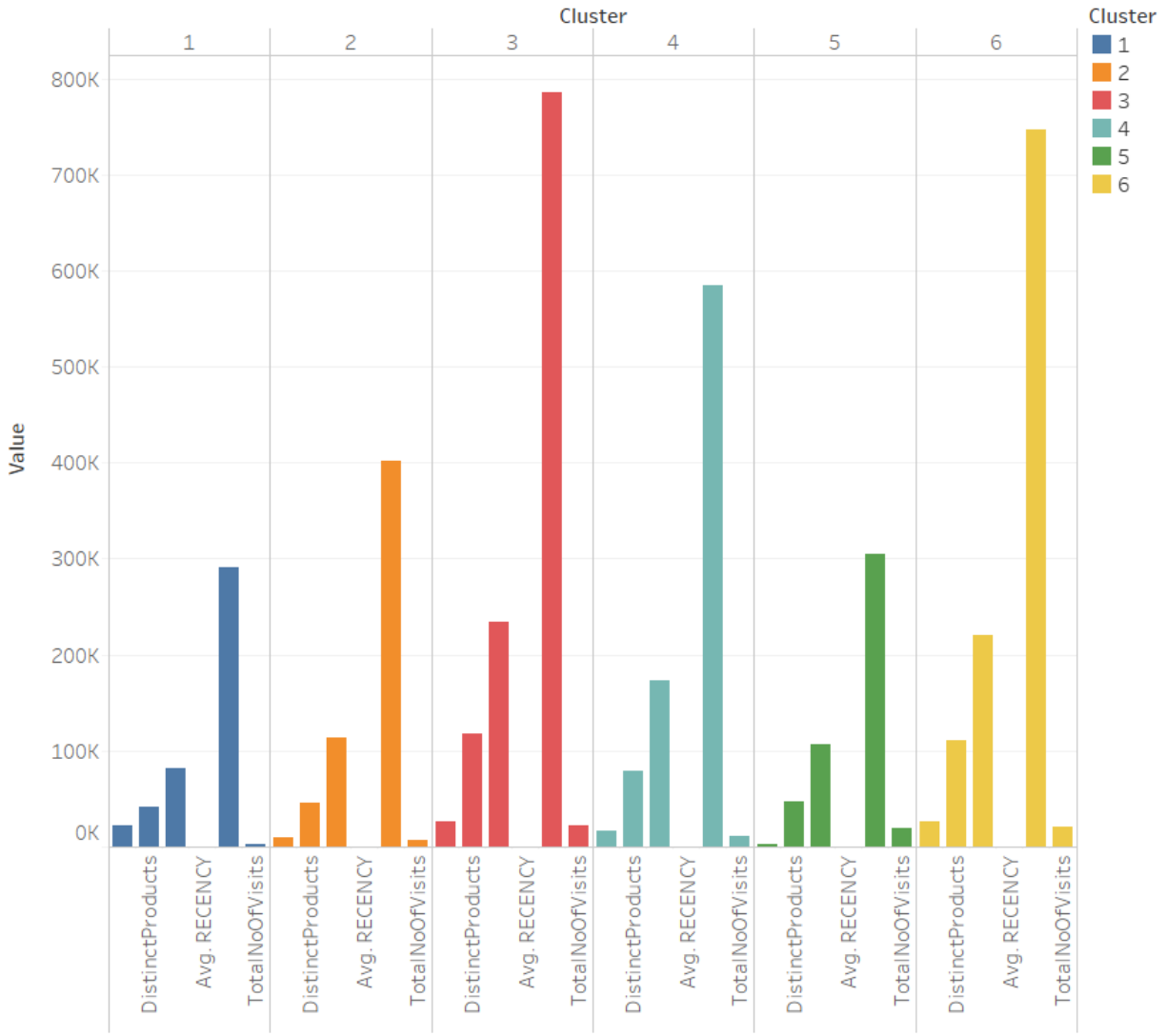
The following bubble chart shows how customers are distributed across the different clusters:



4.7 Data Analysis

4.7.1 General Analysis

On plotting selected attributes and segmenting them into clusters we get the following bar graphs. On a first visual glance it is quite evident that cluster 3 has the highest values in all significant attributes thus highlighting those customers who are most loyal and bring maximum revenue to Sobeys Inc.

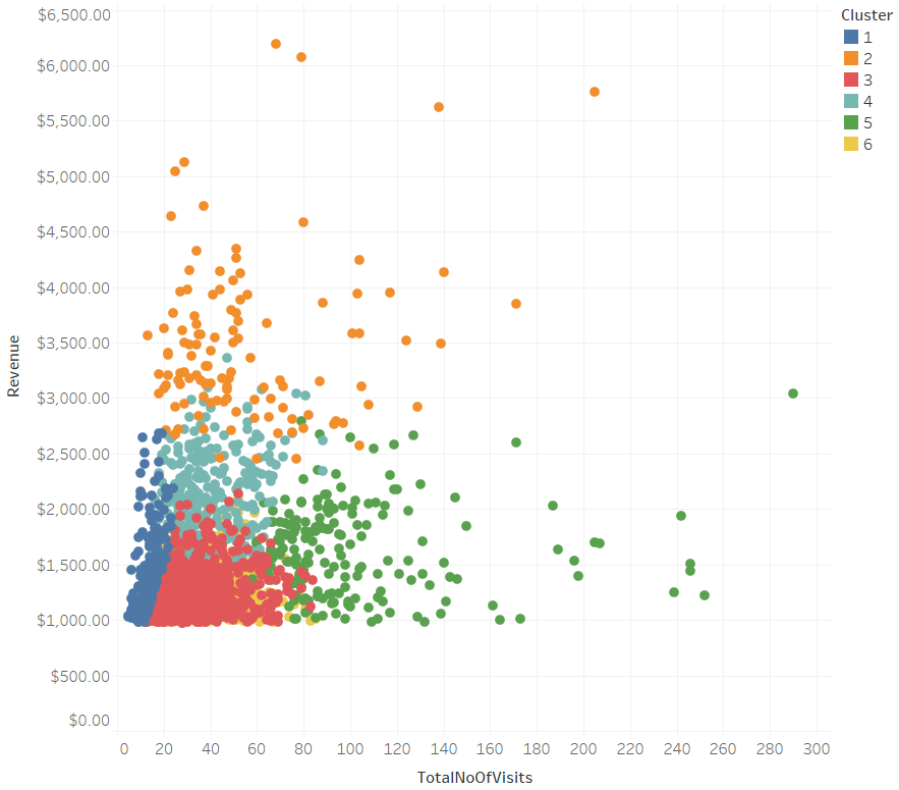


4.7.2 Average Spending vs Number of visits

A plot of total number of visits and average spending of each customer in all the six clusters is shown below. From the scatterplot, we can see that cluster 5 has high number of visits but the least average spending indicating that these customers are bargain hunters or erudite customers.

4.7.3 Revenue Vs Total number of visits

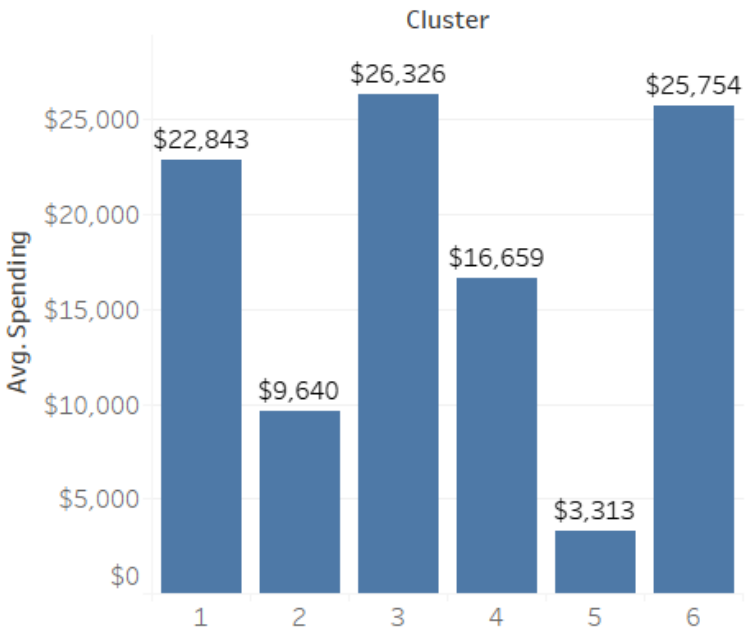
On analyzing the total number of visits and total revenue generated from the customers in each cluster, it is observed that customers belonging to cluster 1 have low number of visits and generate less revenue indicating them to be impulsive buyers.



4.7.4 Detailed Cluster Analysis

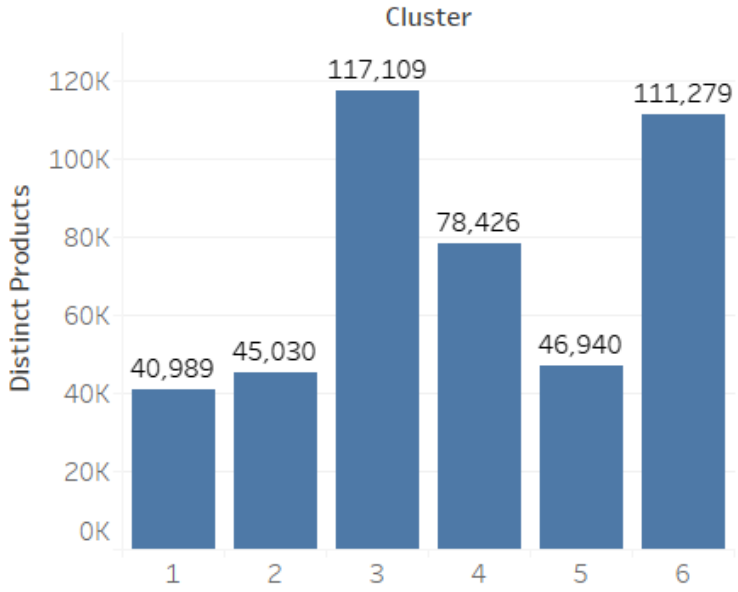
4.7.4.1 Average Spending

The following graph shows the spending nature of customers in each cluster. Customers of cluster 3 spend the most while the customers of cluster 5 spend the least.



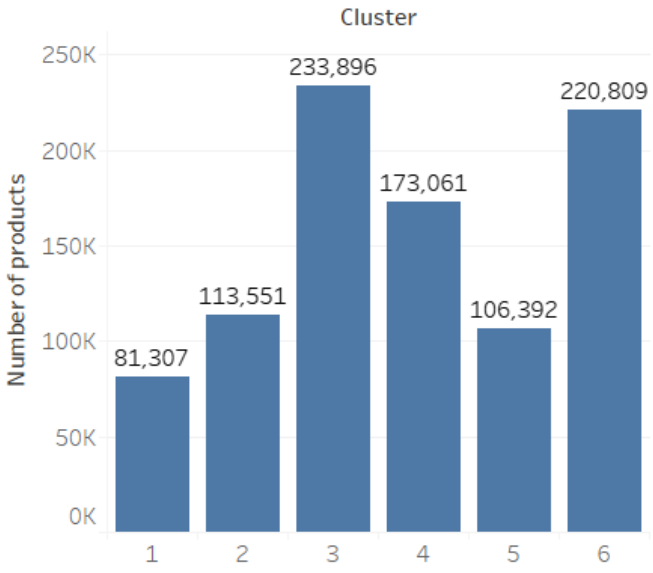
4.7.4.2 Distinct Products

This graph depicts the distinct products present in every cluster. Cluster 3 has the highest count, on the other hand, cluster 1 has the least number of distinct products.



4.7.4.3 Number of products

Cluster 3 contains the greatest number of products closely followed by cluster 6. Cluster 1 has the least count of only 81,307 products.

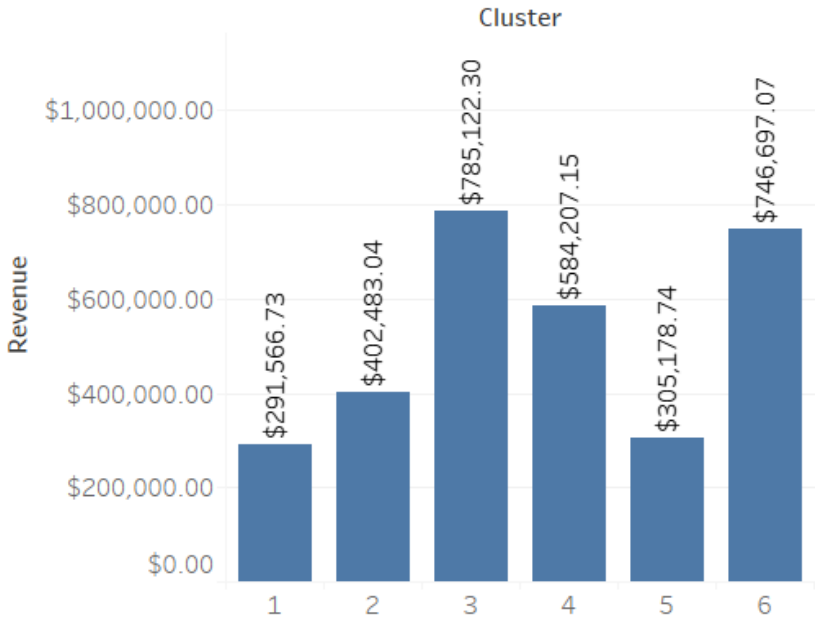


4.7.4.4 Average Recency

Customers present in cluster 3 have visited the store most recently while people in cluster 6 have the least percent of recent customers.

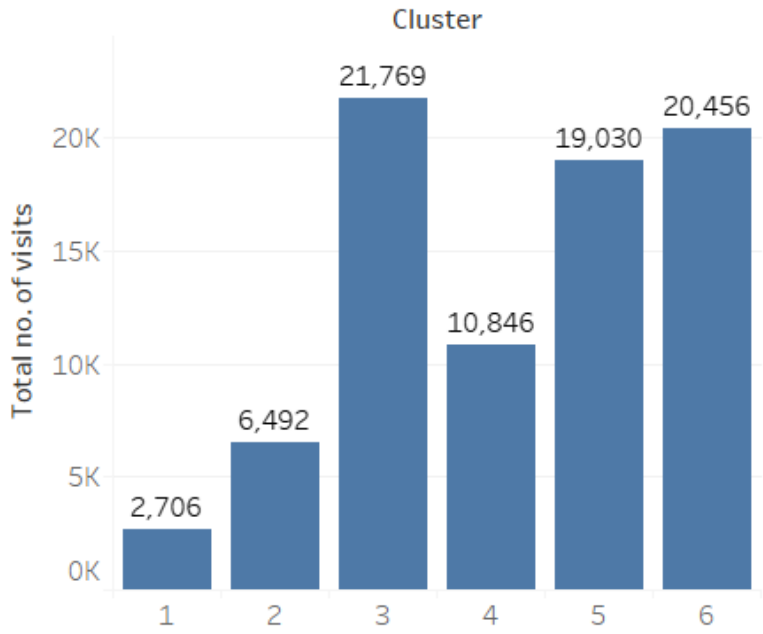
4.7.4.5 Revenue

Customers in cluster 3 generate the highest revenue, while customers in cluster 1 generate the least.



4.7.4.6 Total number of visits

Customers in cluster 3 visited the greatest number of times followed by customers in cluster 6. Customers in cluster 1 on the other hand, made least number of visits.



4.8 Customer Profiling

Category	Description	Recommendation
Cluster 1 – Impulsive Shoppers	• High average spending • Low revenue • Low number of distinct products • Lowest number of visits	Provide deals on high priced items periodically to grab these customers
Cluster 2 – Prospective Customers	• Average recency • Low revenue • Average number of visits • Low average spending	Provide weekly fliers and advertise frequently to attract customers to the store
Cluster 3 – Loyal Customers	• Highest average spending • Highest revenue • High number of visits and recency • Highest number of distinct products	Maintain current strategy as these customers are loyal to Sobeys brand
Cluster 4 – Active Shoppers	• Average spending per visit • Average revenue • Average number of visits and recency • Average quantity of products purchased	Provide combo deals to increase number of purchased products thereby increasing revenue and average spending per visit

Category	Description	Recommendation
Cluster 5 – Erudite Customers	• Lowest spending per visit • Average revenue • High number of visits but low recency	Price matching or providing better discounts than competitors to attract Deal-Hunters
Cluster 6 – Dormant Customers	• High average spending • High revenue • High number of visits but lowest recency • High number of distinct products purchased	Targeted advertising and providing customer specific discounted coupons to bring back lost customers

4.9 Top Buyers

The top 10 customers for each cluster based on the revenue generated are listed below:

Rank	Cluster 1 Impulsive Shoppers	Cluster 2 Prospective Customers	Cluster 3 Loyal Customers	Cluster 4 Active Shoppers	Cluster 5 Erudite Customers	Cluster 6 Dormant Customers
1	59459473	37143973	23717387	59793790	21584436	61809069
2	39314414	60434066	20802419	21701323	59920039	59830917
3	21272846	64947261	51032169	22301160	61373866	39622051
4	21452032	64593270	22288093	54013758	59427703	60613178
5	61813890	28313518	37057354	59504665	22114791	58833348
6	62319212	61056436	21232652	36164600	22114425	58872897
7	40035819	59564026	22114381	22287906	59003863	62512800
8	21015957	59427755	21384489	38988153	22114125	64151004
9	40173104	40180526	40174216	60084673	34920290	58755459
10	39307935	61044591	31919946	59910413	61786992	60299979

5. Products Analysis

5.1 Features

For the purpose of products data analysis, the top 2000 revenue generating products from the table ‘sales219’ is stored into a temporary table and then joined with ‘items’ table to get product details and store in the final table ‘ProductsCluster’.

Following is schema of the ‘ProductsCluster’ table for our analysis:

S.No.	Column Name	Measure	Description
1.	ITEM_SK	NA	Unique identifier for the products
2.	Revenue	SUM	Total revenue for that item
3.	TransactionCount	COUNT DISTINCT	The total number of transactions in which the item was found. Transactions carried out by the same customer on the same day will be considered as separate transactions
4.	CustomerCount	COUNT DISTINCT	Number of distinct customers for each product
5.	AvgPrice	AVG	The average price of the product. This compensates for variations in price where a special price was calculated using coupons
6.	Category	NA	The category to which the product belongs
7.	Department	NA	The department to which the products belong
8.	ItemDesc	NA	The description of the products

5.2 Descriptive Analysis

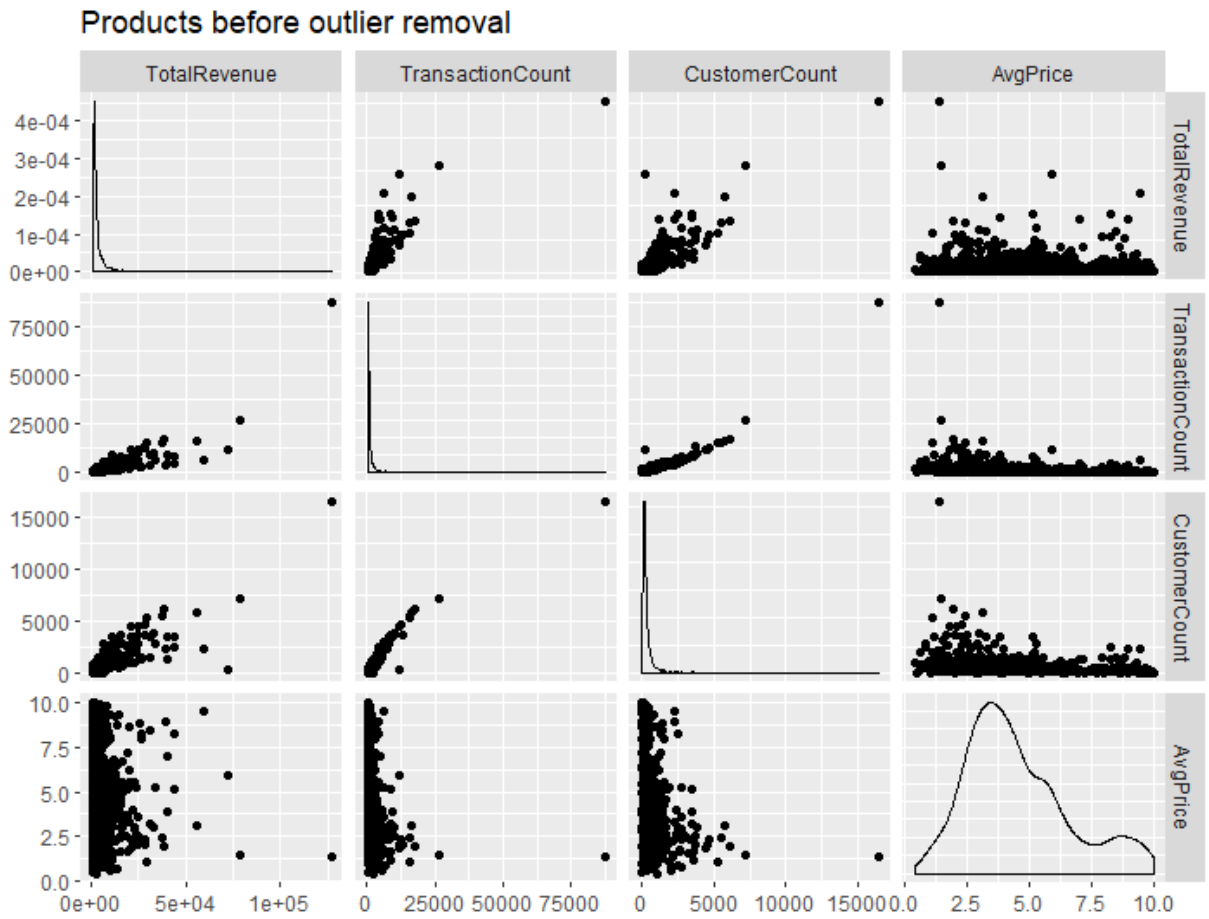
To get an idea of the data to be analysed and understand the range and distribution of the variables we take a summary of the top 2000 products in 'ProductsCluster' table.

Summary of Top 2000 products:

```
  TotalRevenue     TransactionCount  CustomerCount       AvgPrice            Category          Department
Min.    :  1327   Min.    :  137.0   Min.    :    5.0   Min.    : 0.4206   Milk Produc:  90   Grocery     :470
1st Qu.:  1662   1st Qu.:  362.0   1st Qu.:  168.0   1st Qu.: 3.0240   Other Fruit:  82   Produce     :367
Median :  2285   Median :  555.5   Median :  249.0   Median : 4.0942   Refrigerate:  74   Dairy       :319
Mean   :  3832   Mean   :  982.5   Mean   :  402.8   Mean   : 4.5170   Bread Produ:  72   Meat        :215
3rd Qu.:  3685   3rd Qu.:  904.2   3rd Qu.:  400.0   3rd Qu.: 5.6612   Yogurt      :  71   Home Meal Re:117
Max.   :126516   Max.   :87545.0   Max.   :16445.0   Max.   :10.0000   Cooking Veg:  66   Deli        :116
                                                                       (Other)    :1545   (Other)     :396
```

5.3 Data Cleansing and Outliers Removal

For the purpose of data cleaning we first plot the different variables from the extracted data of 'ProductsCluster' table to identify the outliers.



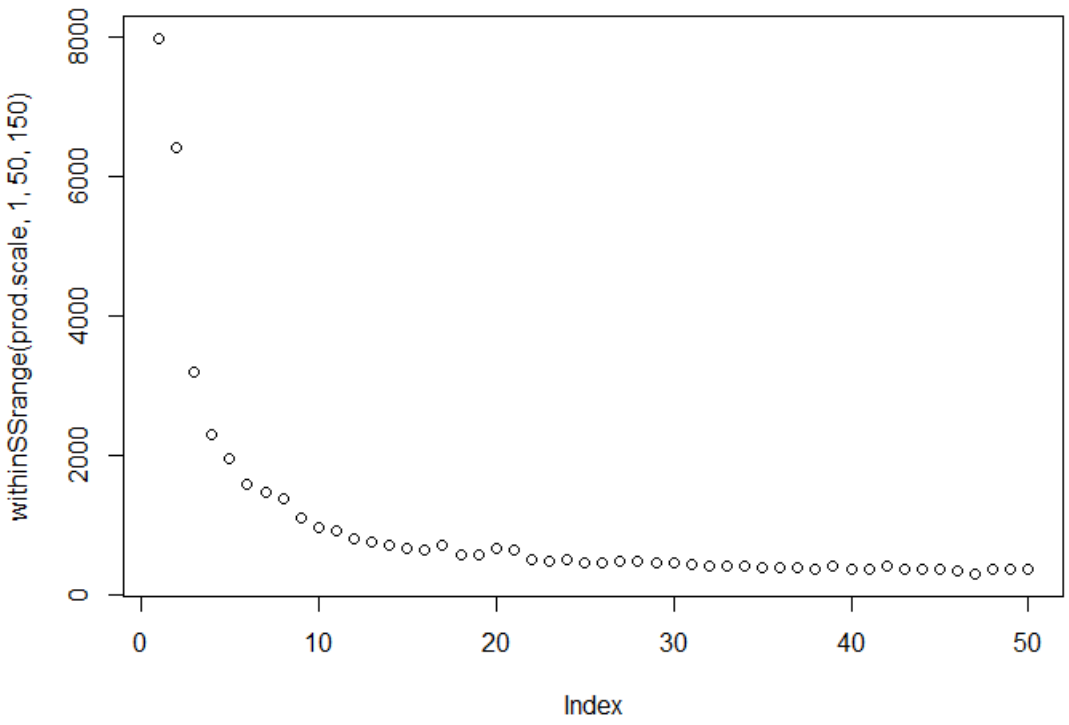
For the purpose of data cleaning, the data extracted from 'ProductsCluster' table was plotted to identify the outliers. It was observed that ITEM_SK = 11740941 (Bananas) is an outlier with a very high transaction count and revenue as compared to other items in the data.

5.4 Data Normalizing

The cleaned data has been re-scaled to a uniform scale for bringing data to a normalized format for the purpose of cluster analysis.

5.5 Selecting number of clusters

The normalized data is then passed through a K-Means Algorithm to find the number of clusters by plotting an elbow curve.



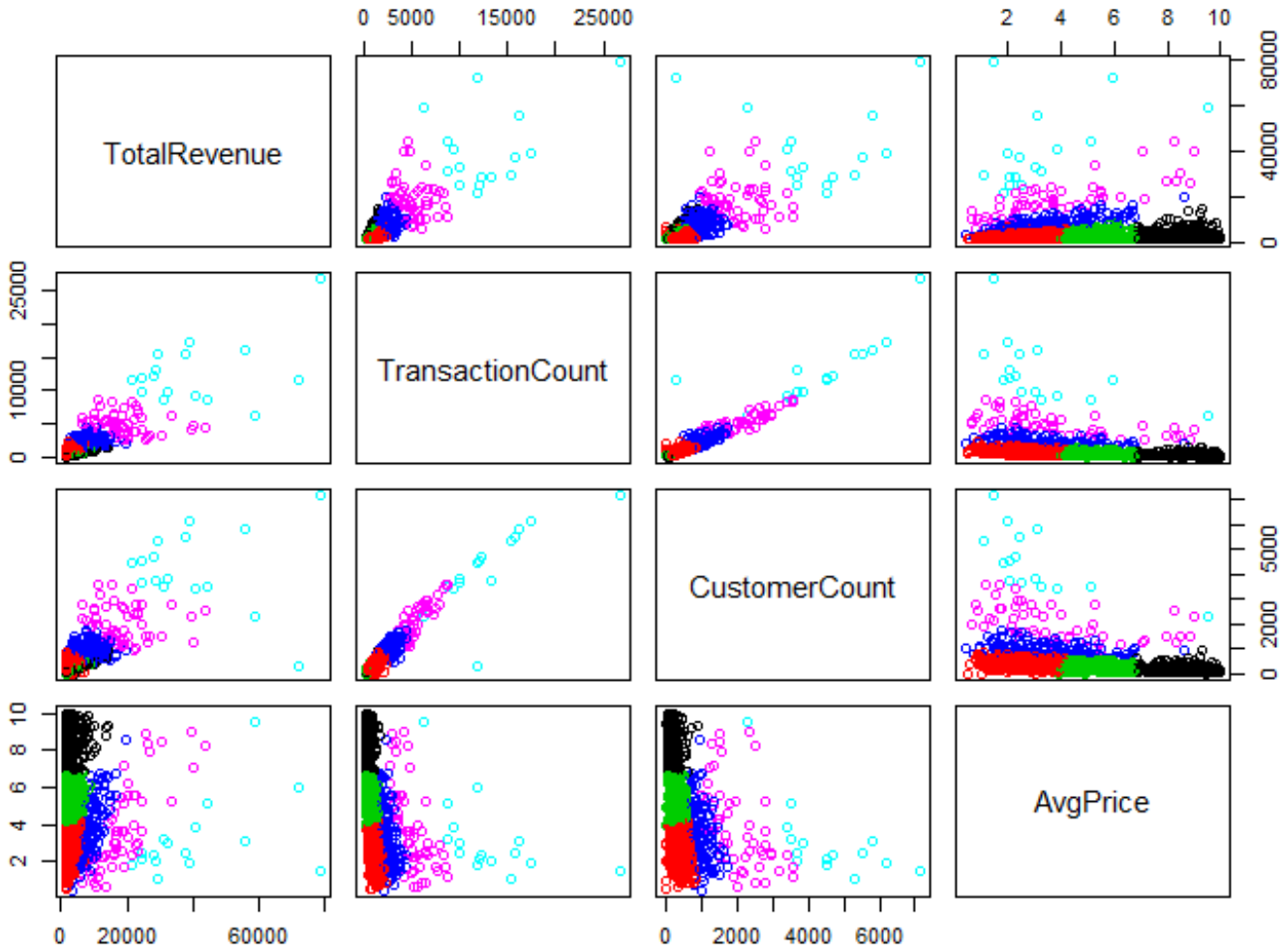
To get a more detailed view of the data in the plotted elbow curve, table showing the WCSS values and number of clusters was also generated:

```
[1] "K   : WCSS"
[1] " "
[1] "1"     ":"     "7992"
[1] "2"         ":"         "6409.778"
[1] "3"         ":"         "3195.311"
[1] "4"         ":"         "2300.168"
[1] "5"         ":"         "1949.991"
[1] "6"         ":"         "1592.037"
[1] "7"         ":"         "1473.076"
[1] "8"         ":"         "1373.535"
[1] "9"         ":"         "1096.522"
[1] "10"       ":"        "969.255"
[1] "11"       ":"        "917.233"
[1] "12"       ":"        "801.527"
[1] "13"       ":"        "768.958"
[1] "14"       ":"        "708.597"
[1] "15"       ":"        "673.186"
[1] "16"       ":"        "642.601"
[1] "17"       ":"        "711.167"
[1] "18"       ":"        "579.276"
[1] "19"       ":"        "569.108"
[1] "20"      ":"       "664.73"
[1] "21"       ":"        "654.206"
[1] "22"       ":"        "518.405"
[1] "23"       ":"        "484.691"
[1] "24"       ":"        "498.144"
[1] "25"      ":"       "456.11"
[1] "26"       ":"        "452.944"
[1] "27"       ":"        "481.328"
[1] "28"       ":"        "481.063"
[1] "29"       ":"        "463.059"
[1] "30"       ":"        "464.221"
[1] "31"       ":"        "430.712"
[1] "32"       ":"        "414.169"
[1] "33"       ":"        "427.618"
[1] "34"       ":"        "421.732"
[1] "35"       ":"        "399.854"
[1] "36"       ":"        "396.589"
[1] "37"       ":"        "389.451"
[1] "38"       ":"        "382.588"
[1] "39"       ":"        "405.965"
[1] "40"       ":"        "382.126"
[1] "41"      ":"       "374.04"
[1] "42"       ":"        "420.637"
[1] "43"      ":"       "374.63"
[1] "44"       ":"        "372.532"
[1] "45"       ":"        "369.414"
[1] "46"       ":"        "340.858"
[1] "47"       ":"        "297.627"
[1] "48"       ":"        "364.331"
[1] "49"      ":"       "362.84"
[1] "50"       ":"        "361.236"
```

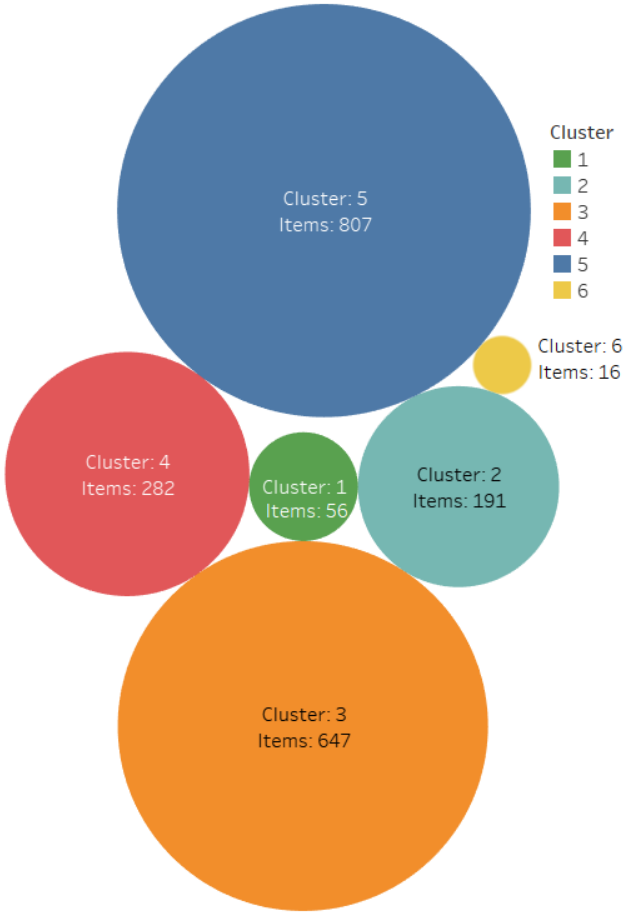
Based on the above elbow curve and generated table we see the drop in the WCSS values has reduced after K=6 (the difference in two consecutive values of WCSS has reduced). Hence for the purpose of cluster analysis K=6 clusters is used.

5.6 Data De-normalizing and Cluster generation

After deriving the number of clusters, we again de-normalize the data to get actual centroids of the clusters and the cluster information is appended to the cleaned customer data.



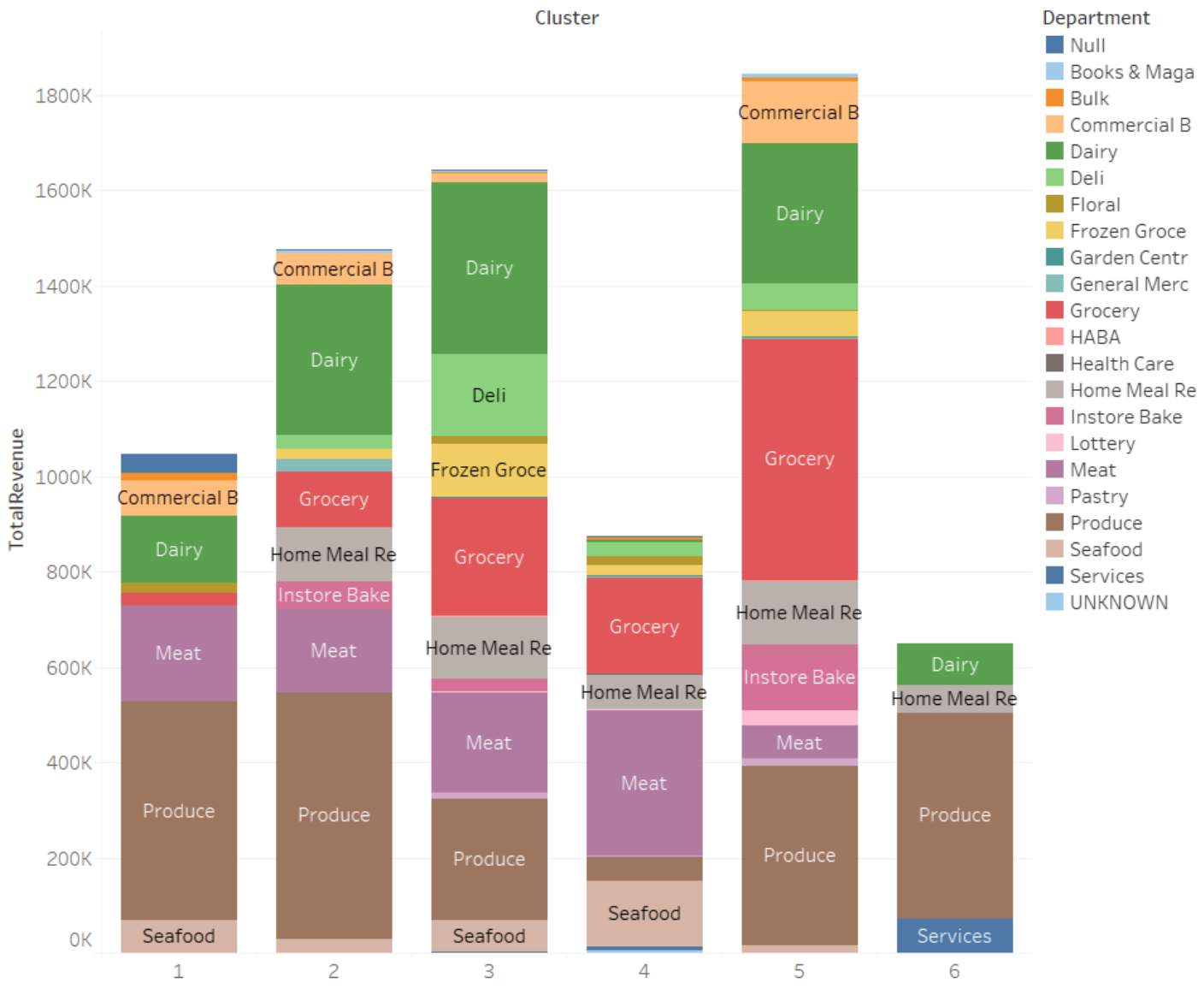
The following bubble chart shows how the products are distributed across the different clusters:



5.7 Data Analysis

5.7.1 Revenue Generation by Departments

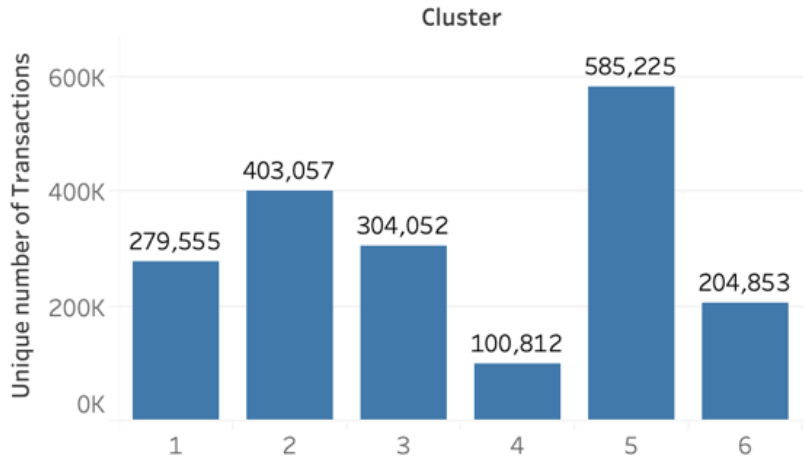
The following graph depicts the total revenue generated per cluster as well as the department of the product contained in a cluster. It can be observed that cluster 5 which contains grocery and produce as main categories produces maximum amount of revenue while cluster 6 containing produce in highest amount produces the least.



5.7.2 Detailed Cluster Analysis

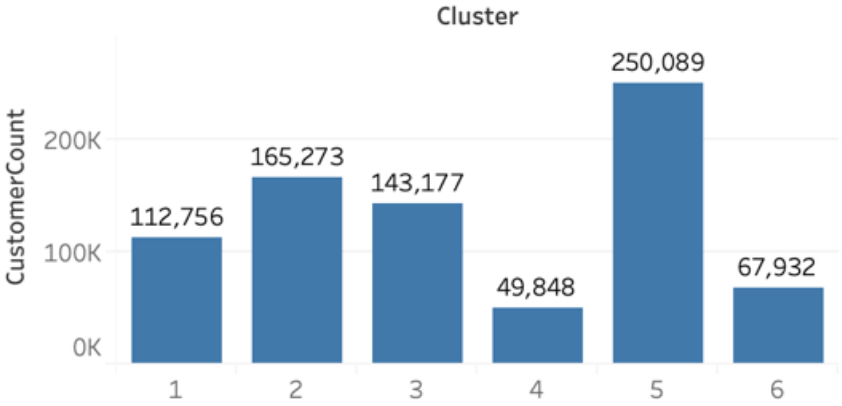
5.7.2.1 Unique Number of Transaction

It can be observed that products in cluster 5 make the greatest number of transactions which proves their loyalty. While products in cluster 4 makes the least number of transactions.



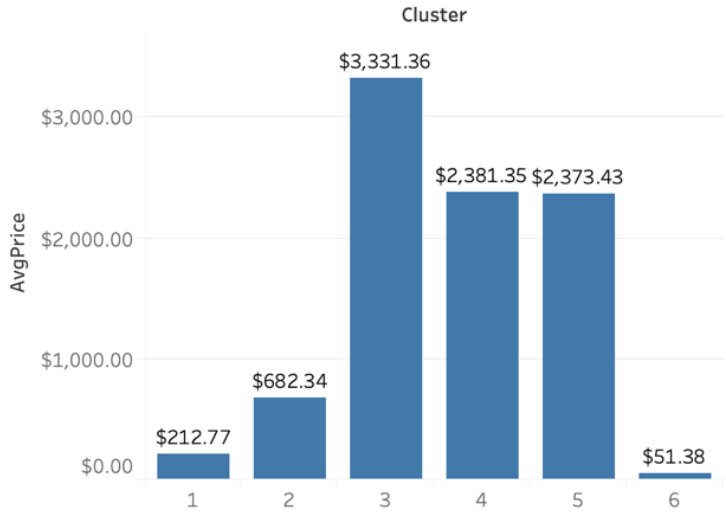
5.7.2.2 Customer Count

Cluster 5 has maximum number of products (over 250K) while cluster 4 has the least.



5.7.2.3 Average Price

Products in cluster 6 have minimum prices and can be considered as cheap products while cluster 3 contains the most expensive products.



5.7.2.4 Revenue

From the below graph, the products in cluster 5 generate the maximum amount of revenue for Sobeys Inc. On the other hand, the products in cluster 6 generate the least amount of revenue and need to be given more attention.

5.8 Product Profiling

Category	Description	Recommendation
Cluster 1 – Bulk Purchases	• Generates medium revenue • Low price • Medium Customer count • Average number of transactions	Since they are bought in bulk, the offers on these products should be increased
Cluster 2 – In Demand Product	• Generates high revenue • Low Price	Increase stock of the product to meet customer demands

Category	Description	Recommendation
	• High Customer Count • Frequently bought	
Cluster 3 – Incumbent products	• Generates high revenue • Highest price • Average customer count • Average number of transactions	The products in this cluster can be sold as combo deals with related products to boost sales
Cluster 4 – Attention Seekers	• Generates low revenue • Medium price • Lowest customer count • Rarely bought	Advertise these products in different mediums to grab consumer attention
Cluster 5 – Cash Cow [3]	• Generates highest revenue • Medium price • Highest customer count • Most frequently bought	Maintain good stock of these products as they are in good demand
Cluster 6 – Dogs [3]	• Generates lowest revenue • Lowest price • Low customer count • Medium transactions	More deals should be put in order to make the customers use these products Only optimum stock should be kept in store to avoid risk of loss of revenue

5.9 Top Sellers

The top 10 products for each cluster based on the number of transactions carried out for each product are listed below:

Rank	Cluster 1 Bulk Purchases	Cluster 2 In Demand Products	Cluster 3 Incumbent products	Cluster 4 Attention Seekers	Cluster 5 Cash Cow	Cluster 6 Dogs
1	Farmers Mi	Farmers Mi	Liberte Yo	Celebratio	Dempster B	Avocada Ha
2	Bens Holsu	CocoCola C	BlkDiamd C	NewWorld D	Scotsbrn M	Comp Eggs
3	Strawberri	Dempster B	Scotsbrn 1	Prime Chic	Farmers Mi	Red Seedle
4	Bens Villa	Big8 Sprin	Sabra Humm	Sens Bnls	Artisan Ba	English Cu
5	Comp Baby	Crispy Chi	Comp Chees	MLPrime Ch	Danone Oik	Red Cluste
6	Onions Gre	Farmers 1%	Kraft Pean	Prime Chck	Lays Chips	Lemons Lar
7	Lean Groun	MapleLea B	HaagnDaz I	Extra Lean	Scratch EP	Red Bell S
8	Comp Carro	Farmers Sk	CrackBar C	Mina Chick	Mini Chick	Hothouse T
9	Comp Mushr	Farmers 10	Danone Oik	New World S	Liberte Yo	On Line Lo
10	Farmers 1%	BlkDiamd C	Crispy Tor	Pork Loin	Pita Brd L	Green Pepp

Appendix A (References)

[1] "Sobeys," [Online]. Available: https://en.wikipedia.org/wiki/Sobeys.

[2] "RFM," [Online]. Available: https://towardsdatascience.com/who-is-your-golden-goose-cohort-analysis-50c9de5dbd31.

[3] "BCG," [Online]. Available: https://nl.wikipedia.org/wiki/BCG-matrix.

Appendix B (SQL Scripts)

1. SQL – Customers

```
SELECT MAX(Date) from dataset01.sales219;

CREATE TABLE CustomersC1 as SELECT CUSTOMER_SK, sum(ITEM_QTY) as
NumProducts, count(distinct ITEM_SK) as DistinctProducts,
SUM(SELLING_RETAIL_AMT) as Revenue, count(distinct( TRANSACTION_RK )) as
TotalNoOfVisits, DATEDIFF('2015-09-15',max(date)) as Recency,
SUM(SELLING_RETAIL_AMT)/count(distinct( TRANSACTION_RK )) as AvgSpend
FROM dataset01.sales219 GROUP BY CUSTOMER_SK ORDER BY Revenue DESC LIMIT
2000;


ALTER TABLE CustomersC1 CONVERT TO CHARACTER SET utf8 COLLATE 'utf8_general_ci';


CREATE TABLE CustomerCluster as
SELECT c.CUSTOMER_SK, c.NumProducts, c.DistinctProducts, c.Revenue, c.TotalNoOfVisits, c.Recency,
c.AvgSpend, m.CITY_NM, m.CUSTOMER_TYPE_CD from CustomersC1 c JOIN
dataset01.members m ON c.CUSTOMER_SK = m.CUSTOMER_SK;
```

2. SQL – Products

```
CREATE TABLE ProductC1 as SELECT
ITEM_SK,
sum(SELLING_RETAIL_AMT) as Revenue,
count(distinct TRANSACTION_RK) as TransactionCount,
count(distinct CUSTOMER_SK) as CustomerCount,
avg(SELLING_RETAIL_AMT/NULLIF(ITEM_QTY,0)) as AvgPrice  FROM dataset01.sales219
GROUP BY ITEM_SK
ORDER BY Revenue DESC
LIMIT 2000;


ALTER TABLE ProductC1 CONVERT TO CHARACTER SET utf8 COLLATE 'utf8_general_ci';

CREATE TABLE ProductsCluster2 as
SELECT p.ITEM_SK, p.Revenue, p.TransactionCount, p.CustomerCount, p.AvgPrice, i.PM_CAT_ENG_DESC as
Category,i.PM_DEP_ENG_DESC  as Department,i.ITEM_ENG_DESC as ItemDesc from ProductC1 p JOIN
dataset01.items i ON p.ITEM_SK = i.ITEM_SK;
```

Appendix C (R Scripts)

1. R Script – Customers

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

set.seed(5580)
cust <- read.csv("C:/CDA/Sem 2/Data Mining/Assignment1/customercluster.csv")
View(cust)

summary(cust[2:7])

ggpairs(cust[,2:7],upper=list(continuous=ggally_points), lower=list(continuous="points"),title="Customers
before outlier removal")

cust.clean <- cust[cust$CUSTOMER_SK != 1, ]
ggpairs(cust.clean[,2:7],upper=list(continuous=ggally_points),
lower=list(continuous="points"),title="Customers after removing CUSTOMER_SK=1")

summary(cust.clean[2:7])

cust.scale = scale(cust.clean[,2:7])
View(cust.scale)

withinSSrange <- function(data,low,high,maxIter)
{
  withinss = array(0, dim=c(high-low+1));
  print("K  : WCSS")
  print(" ")
  for(i in low:high)
  {
    withinss[i-low+1] <- kmeans(data, i, maxIter)$tot.withinss
    print (c(i,":",round((withinss[i-low+1]),3)))
  }
  withinss
}

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

ckm = kmeans(cust.scale, 6,150)
cust.realCenters = unscale(ckm$centers,cust.scale)
clusteredCust = cbind(cust.clean,ckm$cluster)

View(clusteredCust)

plot(clusteredCust[,2:7],col=ckm$cluster)
```

```
write.csv(clusteredCust,file="C:/CDA/Sem 2/Data
Mining/Assignment1/customercluster_output.csv",col.names = FALSE)
```

2. R Script – Products

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

set.seed(5580)
prod <- read.csv("C:/CDA/Sem 2/Data Mining/Assignment1/productscluster.csv")
View(prod)

summary(prod[2:7])
ggpairs(prod[,2:5],upper=list(continuous=ggally_points), lower=list(continuous="points"),title="Products
before outlier removal")

boxplot(prod$Baskets)


prod.clean <- prod[prod$ITEM_SK != 11740941, ]
ggpairs(prod.clean[,2:5],upper=list(continuous=ggally_points),
lower=list(continuous="points"),title="Products after removing ITEM_SK=11740941")

summary(prod.clean[2:7])
prod.scale = scale(prod.clean[,2:5])
View(prod.scale)

withinSSrange <- function(data,low,high,maxIter)
{
 withinss = array(0, dim=c(high-low+1));
 print("K  : WCSS")
 print(" ")
 for(i in low:high)
 {
  withinss[i-low+1] <- kmeans(data, i, maxIter)$tot.withinss
  print (c(i,":",round((withinss[i-low+1]),3)))
 }
 withinss
}

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

pkm = kmeans(prod.scale, 6,150)
prod.realCenters = unscale(pkm$centers,prod.scale)
clusteredProd = cbind(prod.clean,pkm$cluster)
```

```
View(clusteredProd)

plot(clusteredProd[,2:5],col=pkm$cluster)

write.csv(clusteredProd,file="C:/CDA/Sem 2/Data
Mining/Assignment1/productcluster_output.csv",col.names = FALSE)
```