CONTENTS

EXECUTIVE SUMMARY

Sobeys has provided sales data for analysis of its products so that they can mitigate the problem of product surpluses and product shortages as well as strategize in a way that products with higher revenues are always made available to the customers. We have applied k-means clustering on the given data to retrieve clusters based on number of unique transactions. Further we have analyzed products in each cluster to determine products with higher and lower revenues in order to resolve the issues mentioned earlier. Recommendation is provided in "RECOMMENDATION" section of this report.

INTRODUCTION

Sobeys is a reputed food retailer in Canada. They deal with a lot of products and thus, it is important for them to gain insights of those products and revenue that those products generate In this report we implement k-means clustering on the sales data provided by Sobeys to group similar products based on the transaction and revenues. We also provide the detailed analysis of clusters that are formed and we comment regarding products in each cluster.

PROBLEM STATEMENT

Analysis of products in order to mitigate the problem of product surpluses and product shortages as well as signifying the importance of products for the store with respect to their revenues.

DATA PREPARATION

ProductRevenueBasket table with columns ITEM_SK,revenue,baskets. ITEM_SK uniquely identifies an item in database whereas baskets are number of unique transactions of products bought from sales table.

```
mysql> create table productRevenueBasket as
    -> SELECT `ITEM_SK`,sum(`SELLING_RETAIL_AMT`) as revenue,
    -> count(distinct `TRANSACTION_RK`) as baskets from dataset01.sales219
    -> group by `ITEM_SK`;
Query OK, 32591 rows affected (1 min 58.17 sec)
Records: 32591  Duplicates: 0  Warnings: 0
```

Creation of productcluster table, by selecting top 2000 products according to highest revenue.

```
mysql> create table productcluster as
    -> SELECT * FROM `productRevenueBasket`
    -> ORDER BY `productRevenueBasket`.`revenue`  DESC
    -> limit 0,2000;
Query OK, 2000 rows affected (0.13 sec)
Records: 2000  Duplicates: 0  Warnings: 0

mysql> show tables;
+----------------------+
| Tables_in_m_shah     |
+----------------------+
| customer             |
| productRevenueBasket |
| productcluster       |
+----------------------+
3 rows in set (0.00 sec)
```

Reading CSV file in R

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

prod

```
##      ITEM_SK    revenue       baskets
## 1    11740941 126515.970   87545
## 2    11740923  78940.478   26762
## 3    11680016  72298.688   11766
## 4    11610106  59209.602    6200
## 5    11686823  55806.389   16244
```

…

summary(prod)

```
##    ITEM_SK              revenue          baskets
##  Min.   : 4633195      Min.   :  1327      Min.   :  137.0
##  1st Qu.:11692140    1st Qu.:  1662     1st Qu.:  362.0
##  Median :11753970   Median :  2285    Median :  555.5
##  Mean   :12619778    Mean   :  3832     Mean   :  982.5
##  3rd Qu.:13797938    3rd Qu.:  3685     3rd Qu.:  904.2
##  Max.   :15722923     Max.   :126516     Max.   :87545.0
```

//Applying kmeans

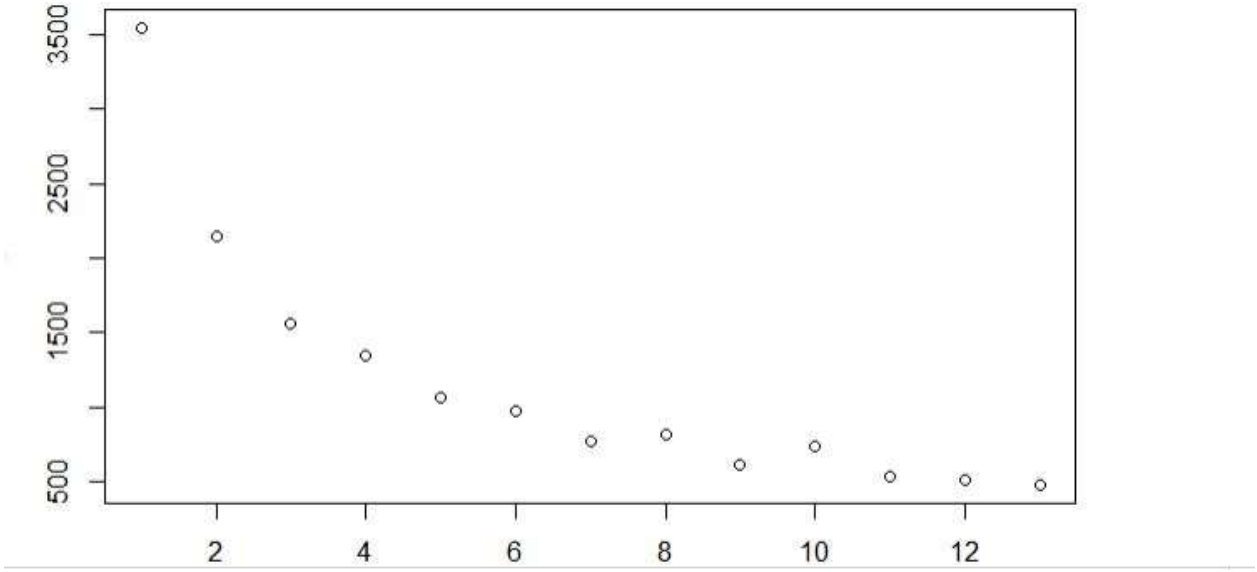
km = kmeans(prod[,2:3],5,150)

km$centers

```
##     revenue    baskets
## 1  50106.738 11395.1818
## 2  21296.319  5736.7143
```

```
## 3   8015.049  2017.1605
## 4 126515.970 87545.0000
## 5   2356.108   578.3833
km$size
## [1]   11   49  243    1 1696
summary(prod)
##     ITEM_SK              revenue            baskets
##  Min.   : 4633195       Min.   :  1327     Min.   :  137.0
##  1st Qu.:11692140     1st Qu.:  1662     1st Qu.:  362.0
##  Median :11753970    Median :  2285    Median :  555.5
##  Mean   :12619778    Mean   :  3832     Mean   :  982.5
##  3rd Qu.:13797938    3rd Qu.:  3685    3rd Qu.:  904.2
##  Max.   :15722923     Max.   :126516    Max.   :87545.0
nprod=prod
nprod[,2]=nprod[,2]/mean(nprod[,2])
nprod[,3]=nprod[,3]/mean(nprod[,3])
summary(nprod)
##     ITEM_SK               revenue             baskets
##  Min.   : 4633195       Min.   : 0.3463     Min.   : 0.1394
##  1st Qu.:11692140     1st Qu.: 0.4336     1st Qu.: 0.3684
##  Median :11753970    Median : 0.5962    Median : 0.5654
##  Mean   :12619778    Mean   : 1.0000     Mean   : 1.0000
##  3rd Qu.:13797938    3rd Qu.: 0.9614     3rd Qu.: 0.9203
##  Max.   :15722923     Max.   :33.0121     Max.   :89.0998
km = kmeans(nprod[,2:3],5,150)
km$center
##     revenue    baskets
## 1  4.876458  4.9108976
## 2  1.873430  1.8784478
## 3 33.012095 89.0998367
## 4  0.608320  0.5488338
## 5 10.584181 13.0307048
km$size
## [1]   60  265    1 1658   16
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
}
e = withinSSrange(nprod[,2:3], 3, 15, 150)
plot(e)
```




```
nkm=kmeans(nprod[,2:3],6,150)
nkm$centers
##     revenue    baskets
## 1  1.2464561  1.2369696
## 2 33.0120950 89.0998367
## 3  0.5442913  0.4732565
## 4 11.0494030 13.5413025
## 5  5.5089761  5.6459905
## 6  2.6711876  2.7480342
nkm$size
## [1]  396    1 1433   14   43  113
realCenters = nkm$centers
realCenters[,1]=mean(prod[,2])*realCenters[,1]
realCenters[,2]=mean(prod[,3])*realCenters[,2]
realCenters
##     revenue    baskets
## 1   4776.934  1215.3838
## 2 126515.970 87545.0000
## 3   2085.949   464.9979
## 4  42345.871 13305.0000
```

```
## 5  21112.669  5547.4651
## 6  10237.093  2700.0796
```
clusteredProd=cbind(prod,nkm$cluster)

clusteredProd[1:20,]

```
##     ITEM_SK   revenue baskets nkm$cluster
## 1  11740941 126515.97   87545         2
## 2  11740923  78940.48   26762         4
## 3  11680016  72298.69   11766         4
## 4  11610106  59209.60    6200         4
## 5  11686823  55806.39   16244         4
## 6  11741143  44282.31    8602         4
## 7  11685694  43996.07    4472         5
## 8  11740964  40649.79    9287         4
## 9  12518517  39994.94    4771         5
## 10 11696675  39570.41    4207         5
## 11 11743201  38774.65   17379         4
## 12 11611881  37650.80   15657         4
## 13 11741127  33671.73    6419         5
## 14 11741816  32687.77    9866         4
## 15 11686839  31361.94    8692         5
## 16 14388093  30793.59    3395         5
## 17 11741274  29152.36   15394         4
## 18 11742966  28776.48   13197         4
## 19 11745837  28387.56   12293         4
## 20 13881134  26841.30    3239         5
```
nkm$cluster[1:20]

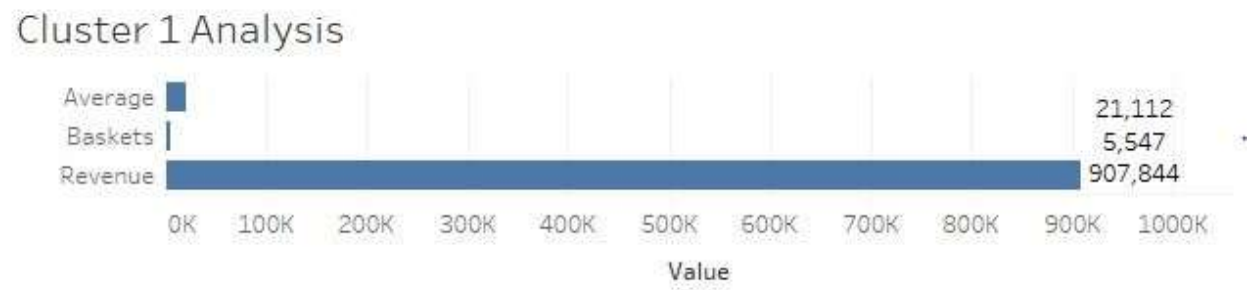
```
##  [1] 2 4 4 4 4 4 5 4 5 5 4 4 5 4 5 5 4 4 4 5
```
write(t(clusteredProd),file="clusteredProducts.csv",sep=',',ncolumns=4)

CLUSTER ANALYSIS

Cluster-1:

There are 43 products in Cluster-1. Total Revenue, Average of Revenue and Number of Baskets in Cluster-1 are as shown below.

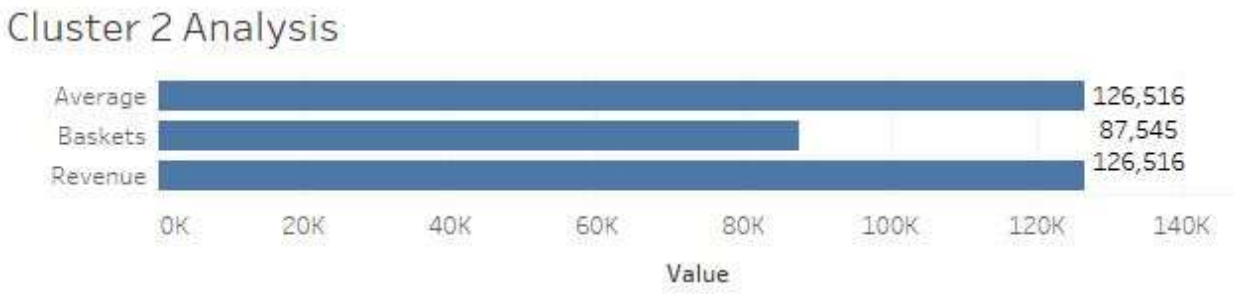
Cluster	Revenue	Average	Baskets
1	907844	21112	5547



Cluster-2:

There is 1 product in Cluster-2. Total Revenue, Average of Revenue and Number of Baskets in Cluster-2 are as shown below.

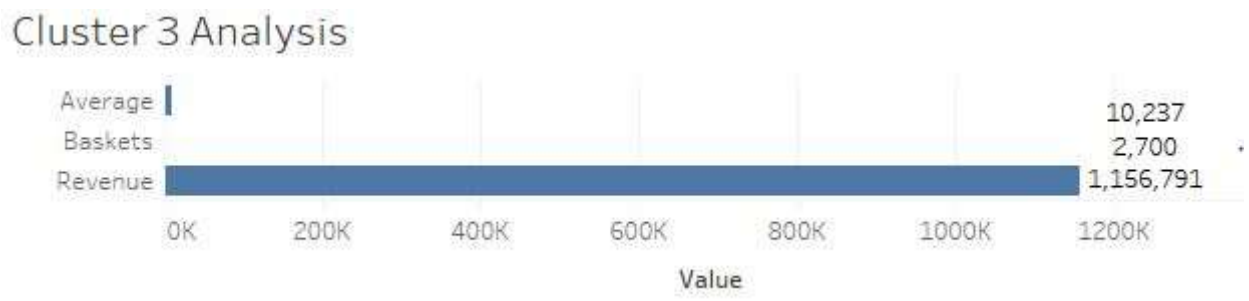
Cluster	Revenue	Average	Baskets
2	126516	126516	87545



Cluster-3 :

There are 113 products in Cluster-3. Total Revenue, Average of Revenue and Number of Baskets in Cluster-3 are as shown below.

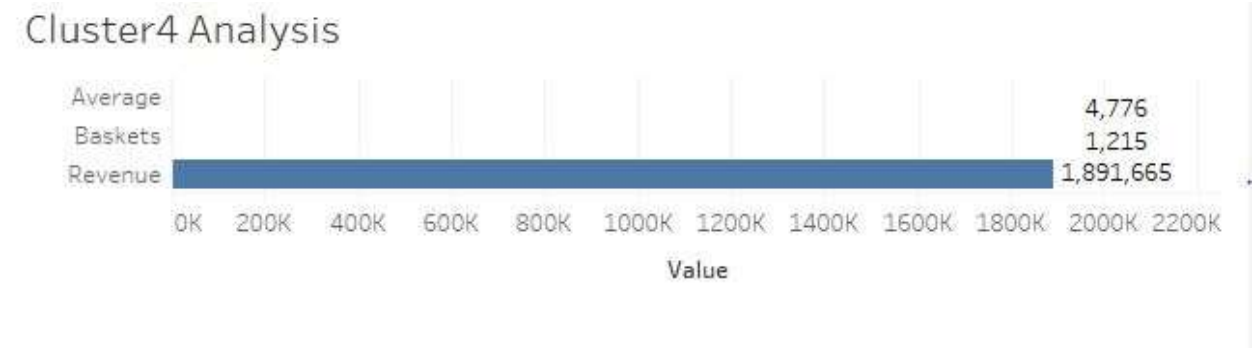
Cluster	Revenue	Average	Baskets
3	1156791	10237	2700



Cluster-4:

There are 396 products in Cluster 4. Total Revenue,Average of Revenue and number of Baskets in Cluster 4 are as shown below.

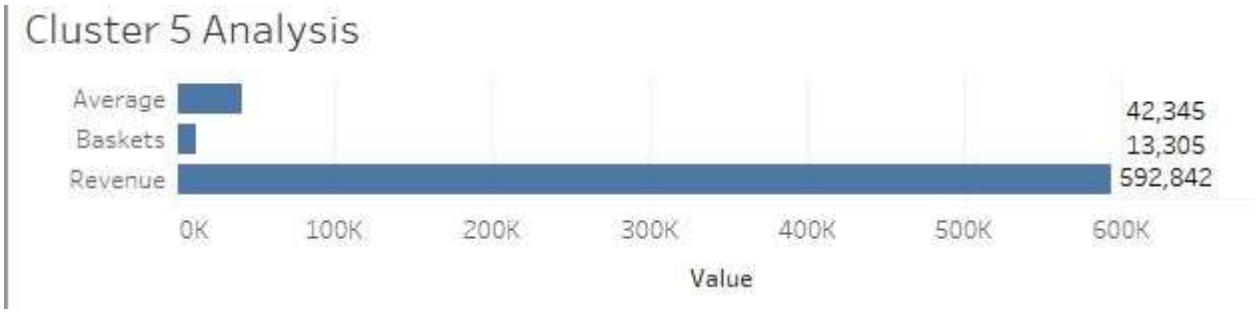
Cluster	Revenue	Average	Baskets
4	1891665	4776	1215



Cluster-5:

There are 14 products in Cluster-5. Total Revenue, Average of Revenue and Number of Baskets in Cluster-5 are as shown below.

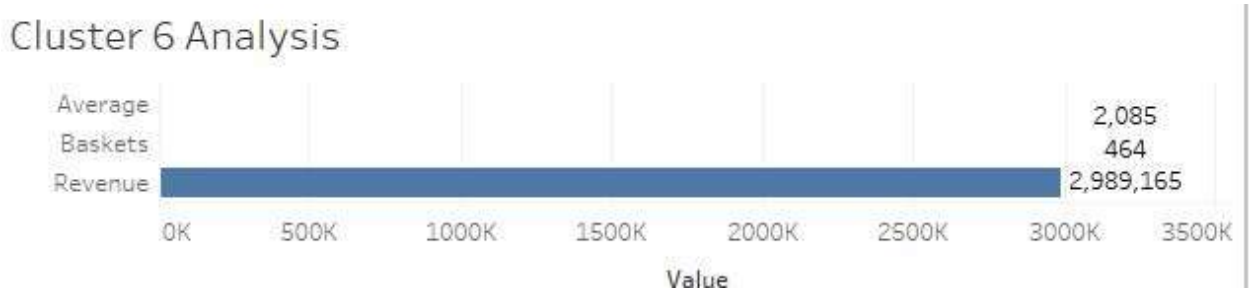
Cluster	Revenue	Average	Baskets
5	592842	42345	13305



Cluster-6:

There are 1433 products in Cluster-6. Total Revenue, Average of Revenue and Number of Baskets in Cluster-6 are as shown below.

Cluster	Revenue	Average	Baskets
6	2989165	2085	464

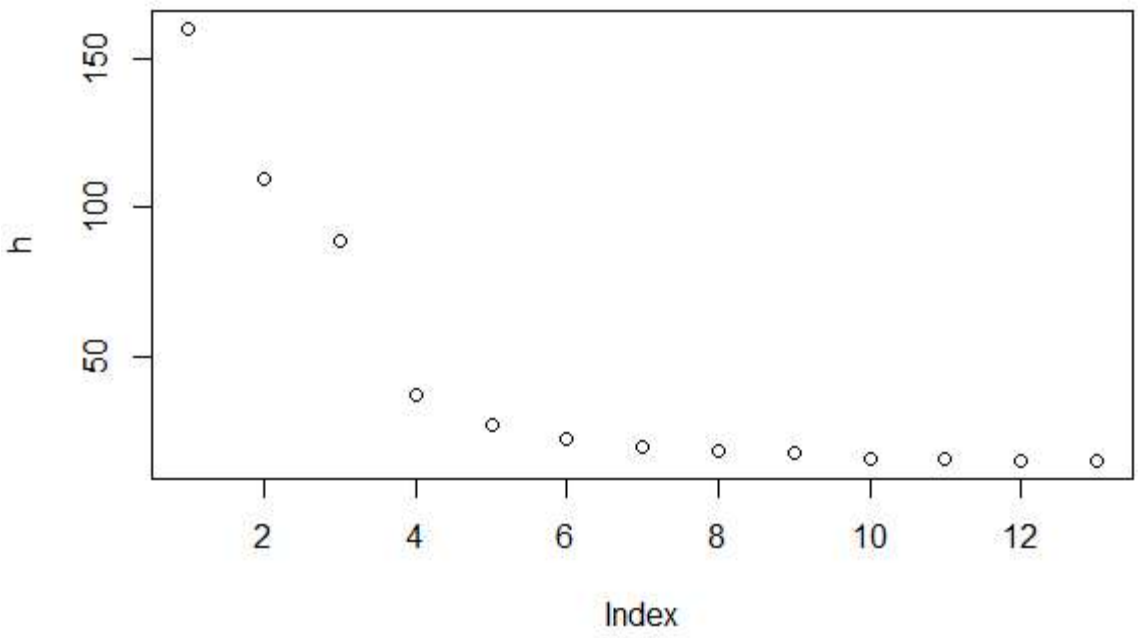


Internal clustering on cluster-6:

Choosing the best k-value for k-means clustering.

```
> withinSSrange(nprods[,3], 3, 15, 150)
 [1] 159.90411 109.44498  88.59989  37.02715  26.80764  22.01224  19.77159  18.28574
 [9]  17.35882  15.67043  15.23466  15.03699  14.89416
> h = withinSSrange(nprods[,3], 3, 15, 150)
> plot(h)
```

Scatter Plot for best k-value



Applying k-means clustering

```
> km=kmeans(nprods[,3],5,150)
> km$centers
       [,1]
1 2.9644534
2 5.5880695
3 1.7977361
4 1.0257121
5 0.5653973

> km$size
[1]  42  17  94 256 557
```

We get 5 clusters with above mentioned sizes.

For 5 clusters:

1. In R, when 5 clusters of products are formed using 150 iterations, we can observe product clusters with respective sizes 1433,14,43,396 and 113 and within sum of square percentage is 90.1%
2. For 5 clusters the minimum revenue by products in store is 1327 and maximum revenue is 126516

3. The minimum number of baskets is 137.0 and maximum number of baskets is 87545.0 As the difference between minimum and maximum number of baskets is very large this does not give a clear picture for analysing sales of products.
4. After this, when normalisation is applied, to analyse what should be the right number of clusters, in order to perform better analysis and receive good results. We notice that when number of clusters is increased to 6 the withinss tends to remain the same.
So we decide number of clusters to be 6.
1. When k-means is again applied, clusters of size 1,1433,14,13,396 are formed and withinss of 92.1.%
2. Then cluster binding is performed to analyse that different products belong to which corresponding cluster.

REVENUE ANALYSIS OF CLUSTERS

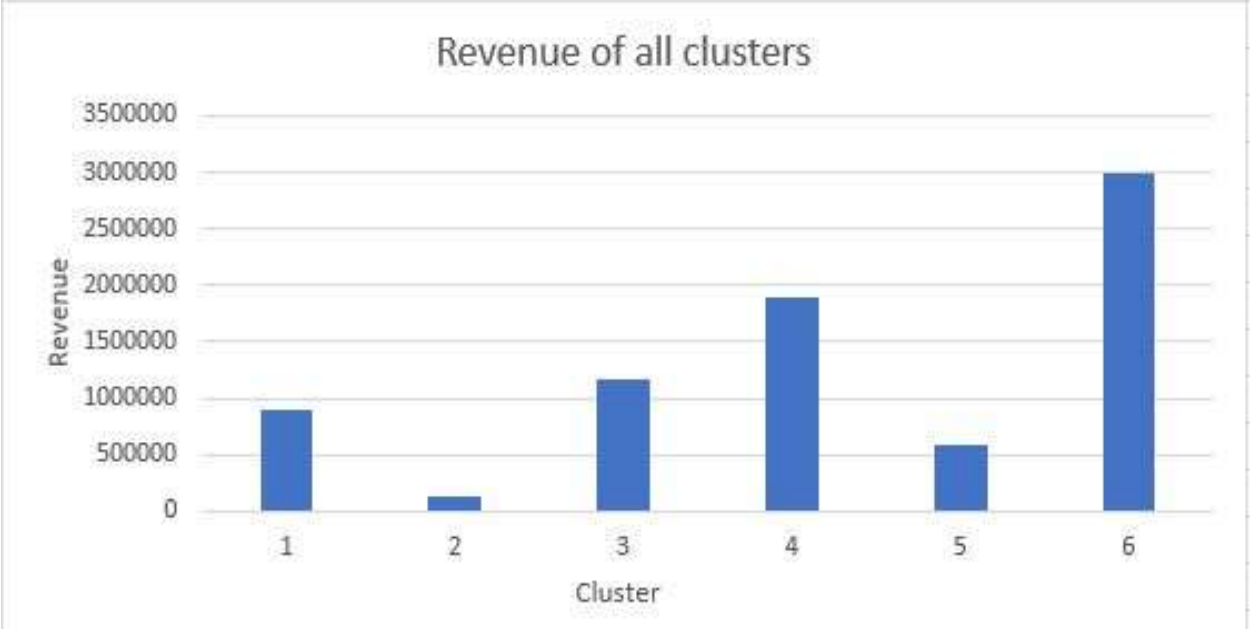
Cluster	Number of products	Revenue	Revenue %
1	43	907844	11.8
2	1	126516	1.6
3	113	1156791	15.09
4	396	1891665	24.6
5	14	592842	7.7
6	1433	2989165	38.9
	Total Revenue	7664823	

Number of products	Revenue %	Cluster
1433	38.9	6
396	24.6	4
113	15.09	3
43	11.8	1
14	7.7	5
1	1.6	2

- Revenue% of cluster-2 is 1.6 with 1 product BANANA.
- Revenue of cluster-5 is 7.7 with 14 products.Revenue could have been increased to 14% when compared to cluster 2 that has only one product and accounts to 1.6% of total revenue.But this cluster has better revenue compared to other clusters like 1 and 3.
- Reason for above statement (cluster 5 is better compared to other clusters) is that cluster 1 has 43 products which are nearly 2.5 times to products in cluster 5 but cluster 1 does not even has twice the revenue as compared to cluster 5.

- Similarly cluster 3 has nearly 2.7 times products as compared to cluster 1 but has revenue only 15% which is just 4% increase in revenue compared to cluster 1.
- Cluster 3 is better(more profitable) than cluster 4 as cluster 4 has 3 times as many as products as cluster 3 but increase in revenue is just 9 from 15% to 24%
- Cluster 6 has highest revenue but number of products in cluster 6 are 1433 which implies it has twice the sum of products in all other clusters 1,2,3,4,5 but it does not even has twice the revenue.It accounts to only 39% whereas revenue sum of all other clusters 1,2,3,4,5 account to 61% with only half of products of cluster1.

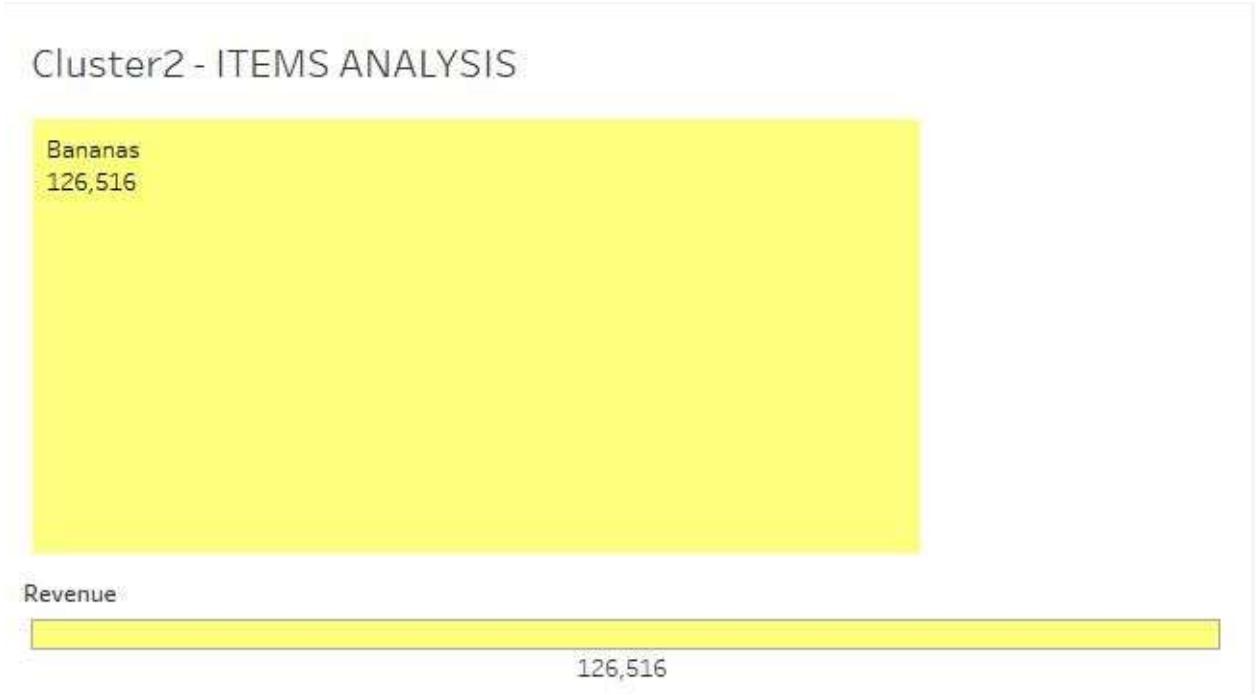
Below is the figure that shows revenue of individual cluster:



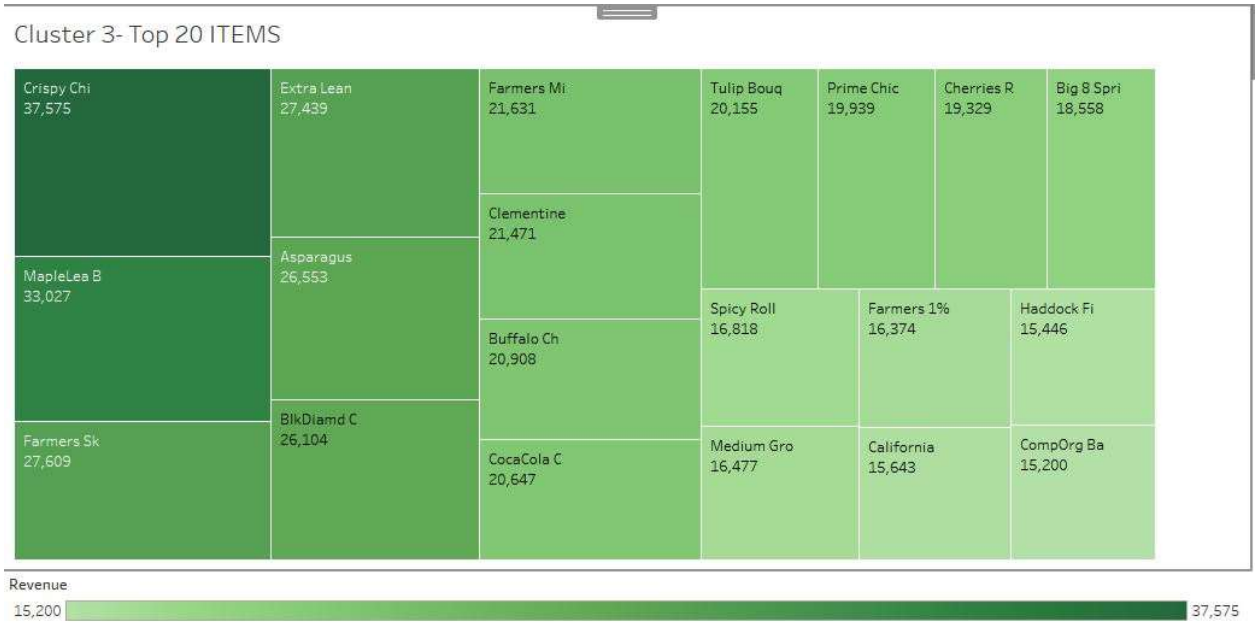
CLUSTER PROFILES

Below is the figure that shows top 20 items that contribute to maximum profit in cluster1 with item LeanGroun contributing to revenue of 63,840

Below is the figure that shows single item in cluster2 that contributes to 1.6% revenue in total revenue.

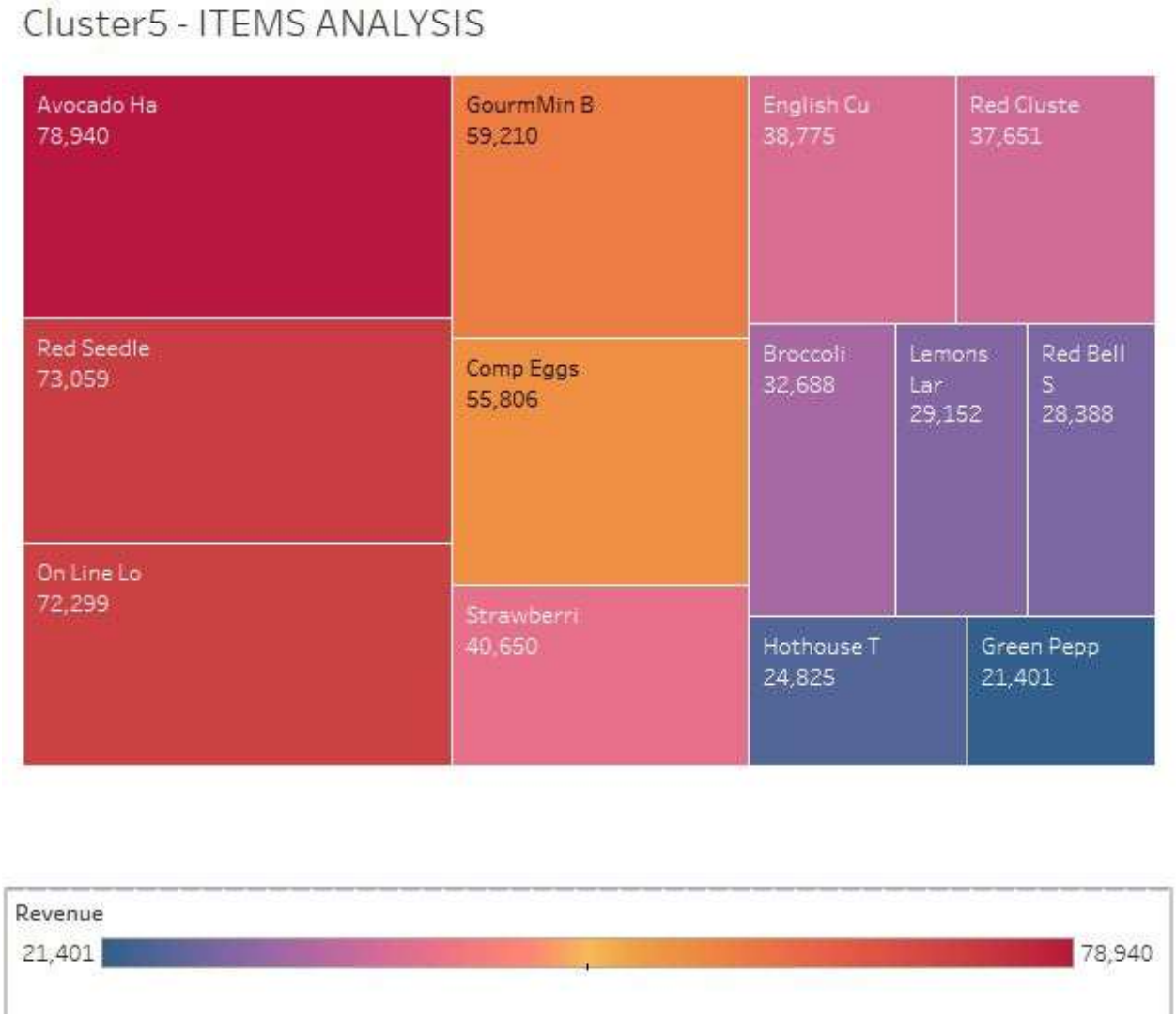


Below is the figure that shows top 20 items that contribute to maximum profit in cluster3 with item CrsipyChi contributing to revenue of 37,575.

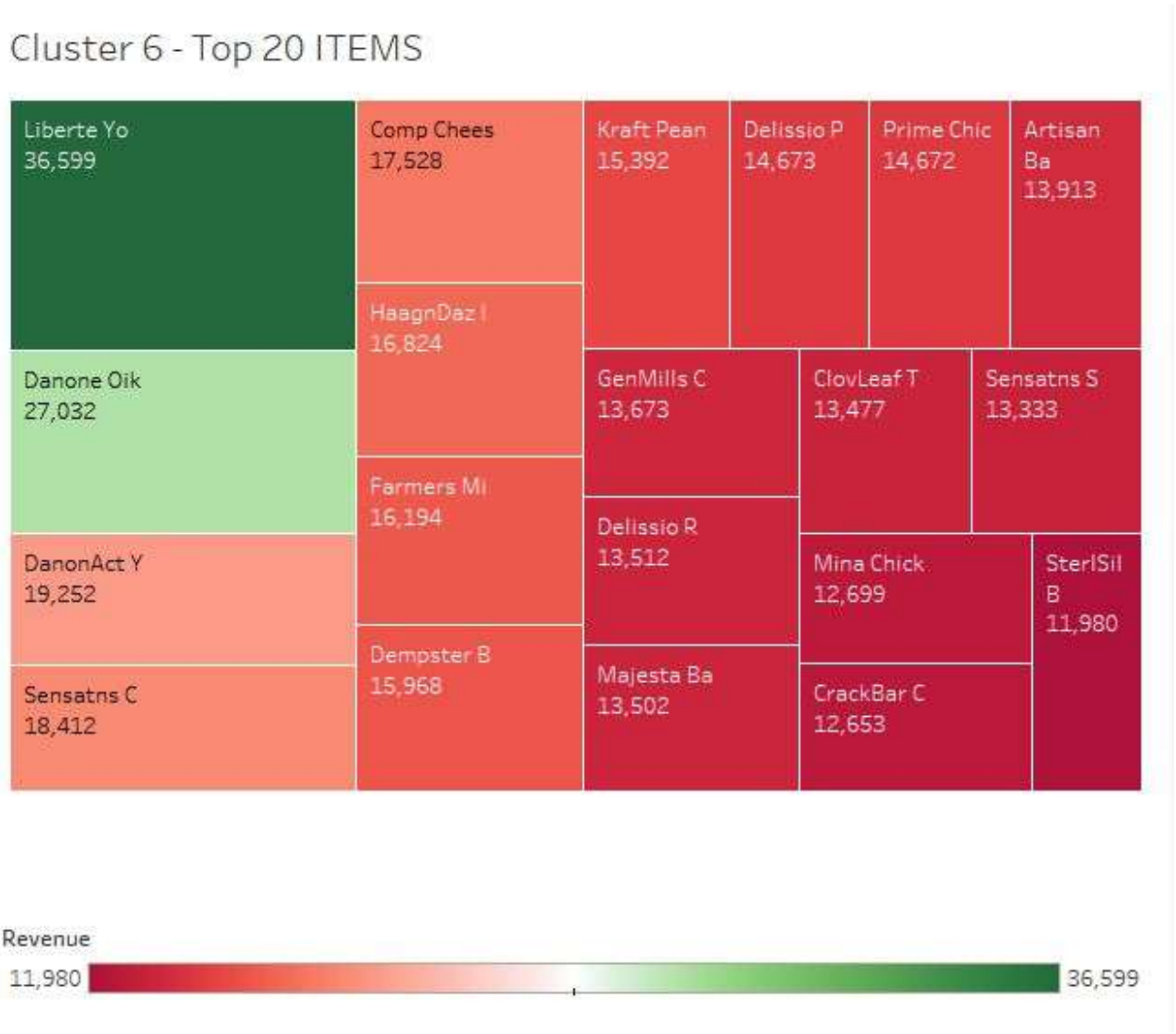


Below is the figure that shows top 20 items that contribute to maximum profit in cluster 4 with item Scotsbrn M contributing to revenue of 36,098

Below is the figure that shows all 14 items that contribute to maximum profit in cluster 5 with item Avocado Chi contributing to revenue of 78,940



Below is the figure that shows top 20 items that contribute to maximum profit in cluster 6 with item Liberto Yo contributing to revenue of 36,599

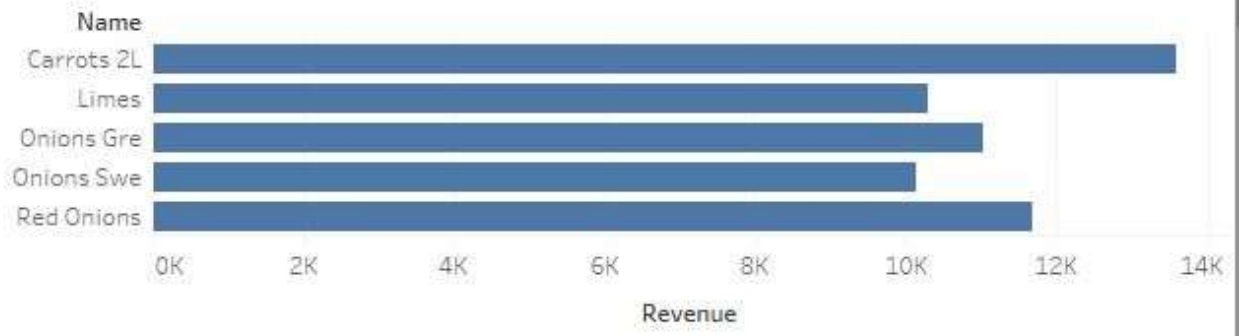


Products that contribute to lowest revenue in clusters 1,3,4,6:

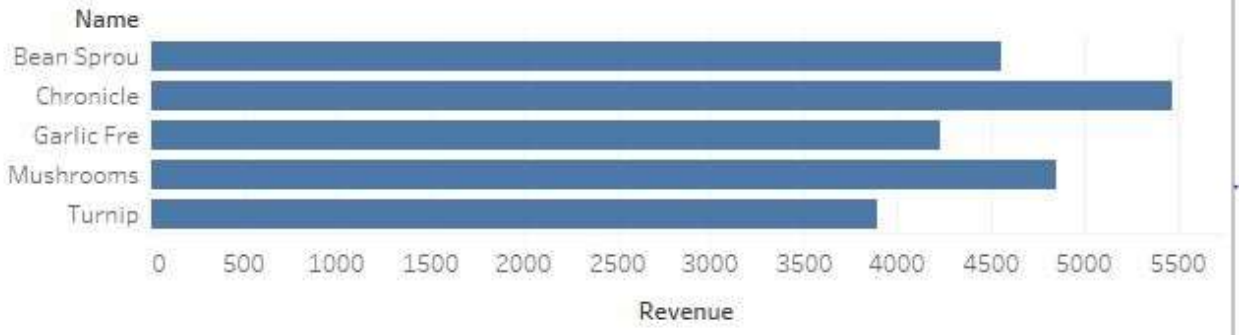
Below are the figures that show 5 products in every cluster 1,3,4,6 that contribute to lowest profits in respective clusters.

Since cluster 5 is better than all clusters lowest products in this have not been taken into account.

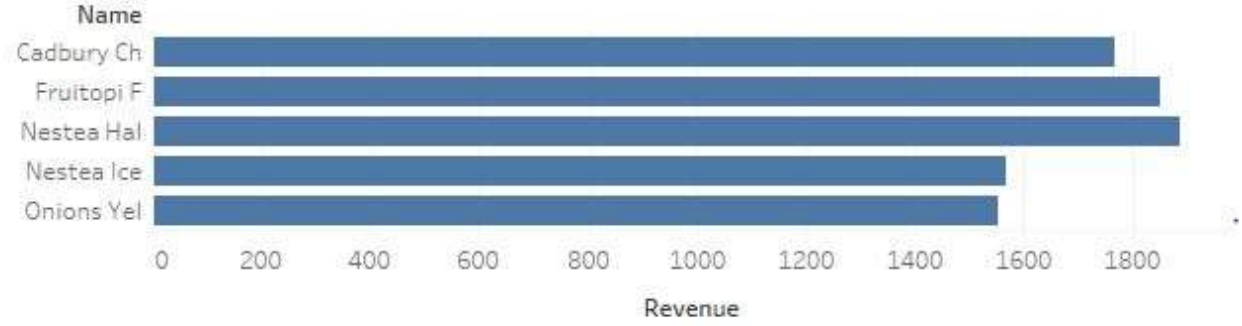
Cluster1 - Low Revenue Products

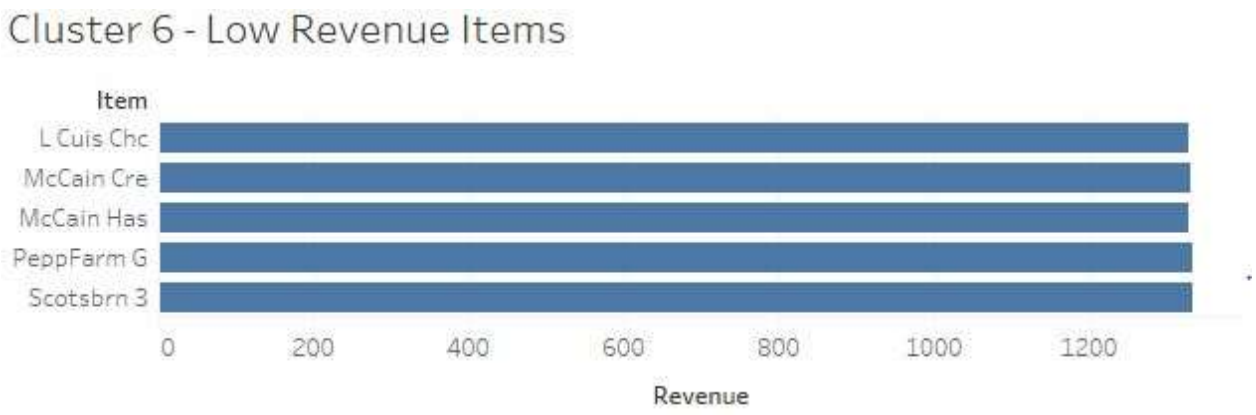


Cluster3 - Low Revenue Products



Cluster 4 - Low Revenue Items





RECOMMENDATIONS

These recommendations are provided as per REVENUE ANALYSIS OF CLUSTERS (Refer to page 9 in report)

- Cluster 5 accounts to more revenue with less number of products.(This is already analysed in “Revenue Analysis of clusters” part of report) .There should be more focus on products of this cluster.
- It is not required to add any product in cluster 2 as it can serve as a base cluster for knowing revenue of all other products.(clusters)
- Cluster 1 has to considered to be added with more products which can maximise revenue as this is better in revenue with minimum products compared to clusters 3 and 4.
- Cluster 6 has highest revenue but number of products in cluster 6 are 1433 which implies it has twice the sum of products in all other clusters 1,2,3,4,5 but it does not even has twice the revenue.It accounts to only 39% whereas revenue sum of all other clusters 1,2,3,4,5 account to 61% with only half of products of cluster1.