HOMEWORK 8
SUJATA SAHU

1.**DATA DESCRIPTION**

The data used for the project is a US superstore data from 2014-2018, The data tells about the sales of a variety of products on e-commerce websites from books, toys, clothes, shoes to food, furniture, and other household items. It basically provides a wide range of information about both product and customers as what are the discounts and profit earned on products, what is the mode of shipment a customer preferred also, tells us about the usage if it's for corporate or personal use and in what quantities it was ordered? So, this shows the broadness of the data and will give organisations a number of solutions to expand their business by helping them in demand forecast , giving a clear picture on the overall profit analysis and many more. But in this project we will be focusing on the prediction of profit of multiple products in the store by using the target variable and major independent variable and will display the conclusions using statistics and necessary visualisations.

The data consist of 21 columns(including Row ID) with almost 10000 rows. Out of which 5 are numeric variables and the rest are categorical variables and character variables.

Numeric variables

Profit - The total Profit earned on a product in a particular order

Discount - Discount provided on order

Quantity - Number of pieces ordered

Sales - Total price of the order

Postal code - The code of the area from which the item is ordered

Ship Date - The date on which the item shipped

Order Date - The date on which the item ordered

Categorical variables

Ship Mode - Mode of shipment

Segment - product used in which segment (Corporate/ Home Office/Consumer)

Country - From the country ordered

City - From the city ordered

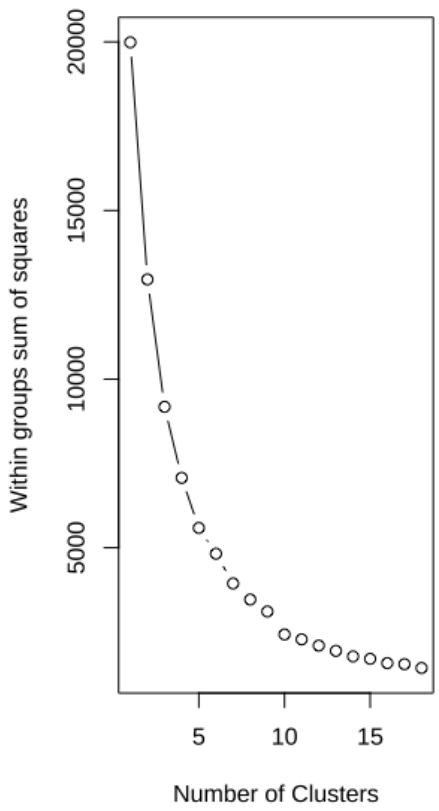
State - From the state ordered
Region - From the Region ordered
Category - product category(Office Supplies, Furniture, Technology etc)
Sub- Category - Product sub category(paper, Art, Storage etc)
Order Id - ID of the order
Customer Id - ID of the customer
product ID - ID of the product
Customer name - Name of the Customer
Product Name - Name of the product

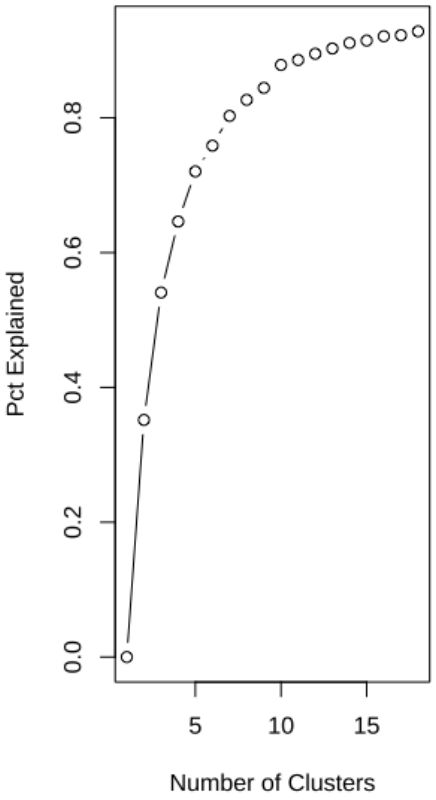
The data is extracted from Kaggle and the link to the data is
https://www.kaggle.com/datasets/juhi1994/superstore

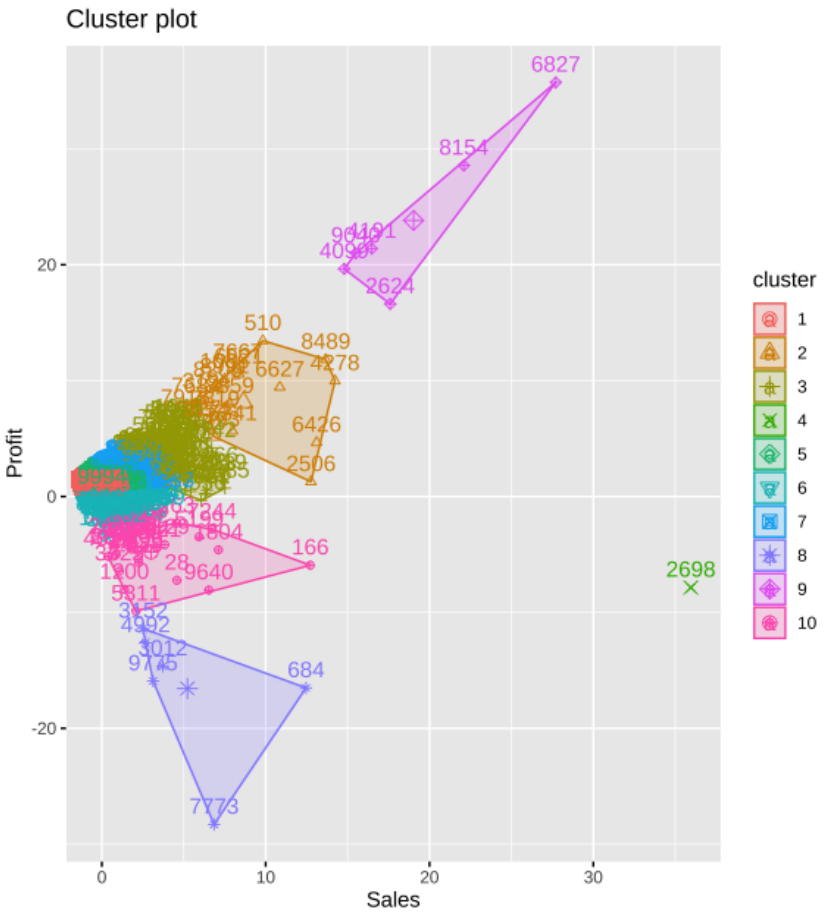
2. Performed 3 clustering models shown in R code.

k-means

K-means clustering method the aim is to find the exact group of the variables using the centroids. The value of k was determined using the elbow method.
Since my data doesn't have any missing values and k means only work for numerical variables, I created a data frame with 2 numerical variables Sales and Profit. Before performing the k-means clustering the data was scaled within sum of squares function and applied on the data frame to plot the elbow chart and found out the point at which the graph bends is at 10 so the value of k is 10.


Within groups sum of squares
5000
10000
15000
20000
5
10
15
Number of Clusters

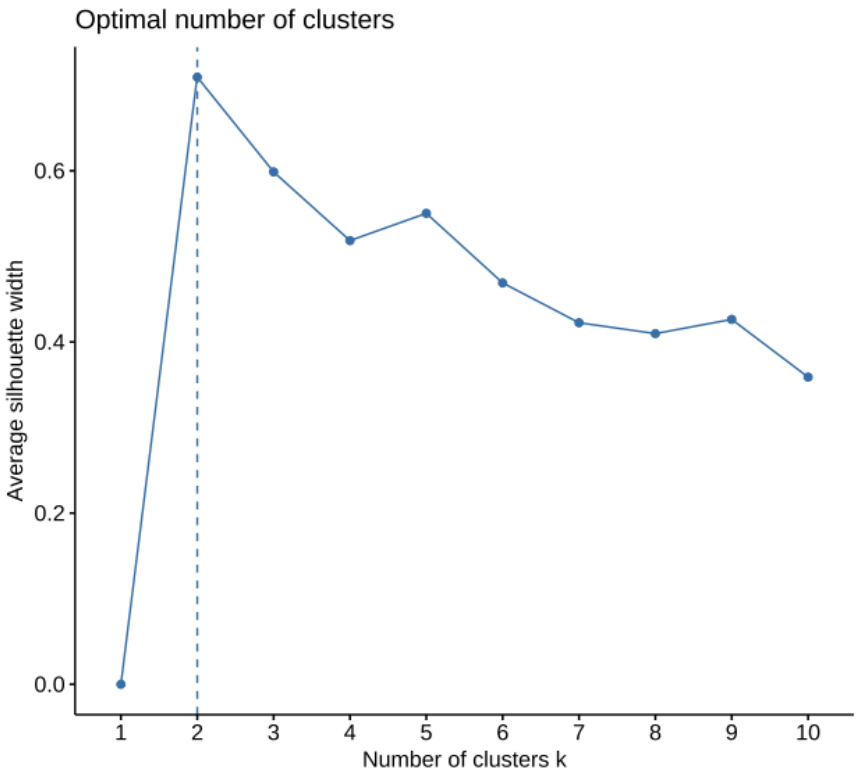

Pct Explained
0.0
0.2
0.4
0.6
0.8
5
10
15
Number of Clusters

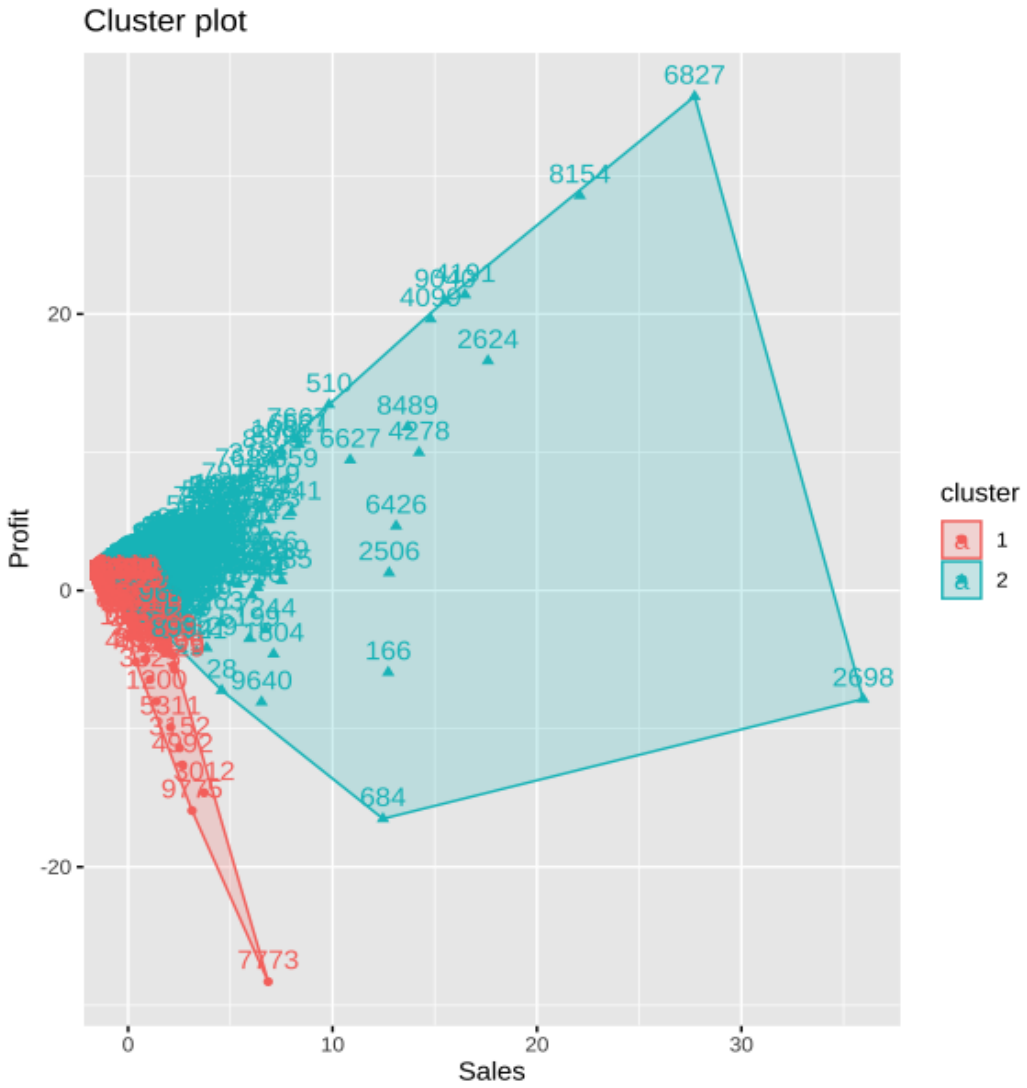


K-medoids

K-medoids clustering method also aims to divide the variables into similar groups but using cluster centres.The value of k was determined using Silhouette.

Performed k-medoids using pam function Using nbclust function and method -Silhouette, the value of k determined was 2

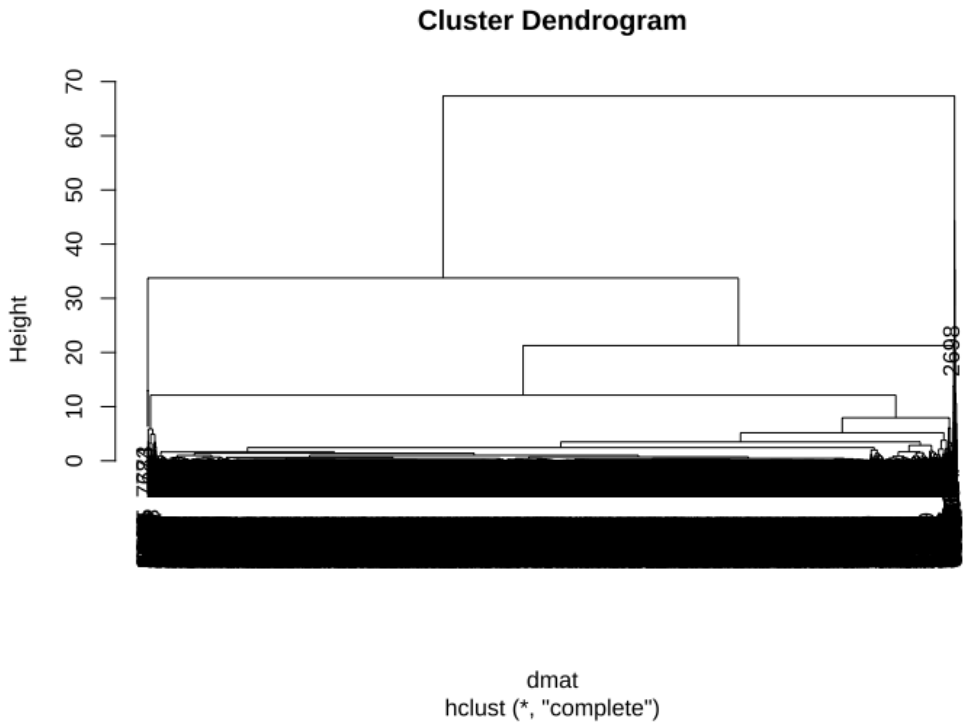

Optimal number of clusters
Average silhouette width
0.0
0.2
0.4
0.6
1
2
3
4
5
6
7
8
9
10
Number of clusters k



Hierarchical clustering

Since my data set consist of huge amount of data therefore the plot is unclear.

Hierarchical clustering is good for small amount of data



Comparing and contrasting

Details:

(a) As for conventional k-means, the algorithm takes every point and clusters by calculating centroids which change all the time. It is computationally costly.
(b) Regular K-means, although worse, are generally better than hierarchical, even after refinement.
(c) Hierarchical is good for small amounts of data whereas k-means and bisecting k-means are generally good for large amounts of data.
(d) K- medoids in many ways are similar to K-means, But In contrast to the k-means algorithm, k-medoids algorithm chooses points as centres that belong to the dataset. Because of this it is sensitive to outliers.
(e) Number of clusters

Initially while computing the k means i considered 18 clusters but the elbow method shows that the data can be divided into 10 clusters whereas in k medoids case the cluster the Silhouette method shows the value of k as 2

(f) Size

Size of K-medoids

K-medoids divides the data into two clusters in 75% and 25%

```
> usDataKMed$clusinfo
     size max_diss   av_diss diameter  separation
[1,] 8471 29.10403 0.1685204 29.50150 0.006811613
[2,] 1523 44.52387 1.3057111 54.42746 0.006811613
```

Size of k-means

```
> usDataKM$size
 [1] 7864   20  107    1 1357  234  369    6    6   30
```

Kmeans shows uneven cluster size and distributed the clusters into very small and very large groups

(g) K-medoids

```
> usDataKMed$medoids
          Sales      Profit
[1,] -0.3153944 -0.08822371
[2,]  0.6901458  0.34722559
```

K-means centroid

```
> usDataKM$centers
        Sales       Profit
1  -0.2703251  -0.08915445
2   8.6419983   8.12283117
3   3.9096791   2.76811719
4  35.9547504  -7.85338705
5   0.3720809   0.20096567
6   0.7312967  -1.16538332
7   1.5907511   0.94903138
8   5.2249027 -16.56510400
9  19.0303381  23.82311671
10  3.0050089  -4.78108604
```

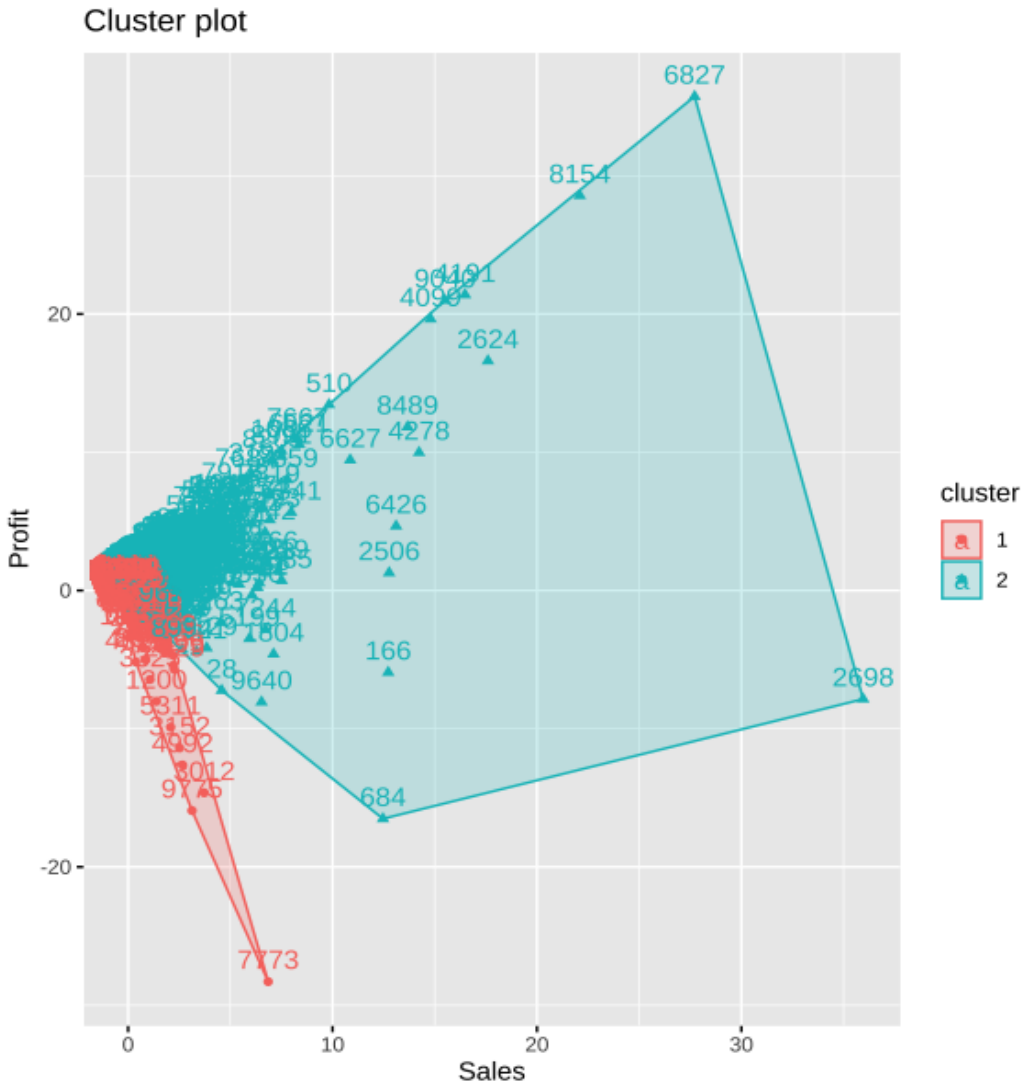
Interpreting K- Medoids Clustering

K medoids

In the k-medoids method, each cluster is represented by a selected object within the cluster. The selected objects are named medoids and correspond to the most centrally located points within the cluster.The PAM algorithm requires the user to know the data and to indicate the appropriate number of clusters to be produced. This is estimated using the function *fviz_nbclust*

Checking the cluster data I can see that my first object belongs to cluster 1 and my second object belongs to cluster 2 similarly below the output shows the objects belonging to cluster 1 or cluster 2. We can see the k medoids are robust to the outliers.A high average silhouette width indicates a and average clustering.

```
> usDataKMed$clustering
   [1] 1 2 1 2 1 1 1 2 1 1 2 2 1 2 1 1 2 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1
  [35] 1 2 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2
  [69] 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [103] 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1
 [137] 1 1 1 1 1 1 1 1 1 2 2 1 2 1 2 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 2 1 2 1 1
 [171] 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1
 [205] 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 1 2 1 1
 [239] 1 1 1 1 1 1 2 1 1 2 1 1 1 2 1 2 1 1 1 2 1 1 1 1 2 2 1 1 1 1 1 1 2 1
 [273] 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1
 [307] 1 1 1 1 2 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 2 1
 [341] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1 2 1 1 1 2
 [375] 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1
 [409] 2 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 2 1 2 1 1 1 2 1 1 2 1 1 2 2 1 1 1 1
 [443] 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1
 [477] 1 1 1 1 2 1 1 1 1 2 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2
 [511] 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1 1 2 1 1 2 1 1 1 2 1 1 1 1 1 1 2 2 2 1
 [545] 1 1 1 1 1 1 1 2 1 1 2 1 1 2 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1
 [579] 1 1 1 2 1 1 1 2 1 2 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1
 [613] 1 2 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 2 1 1 1 1
 [647] 1 1 1 1 1 2 1 1 1 1 1 2 1 2 2 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [681] 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1
 [715] 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1
 [749] 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 2 1 2 1 1 1 1 1 1 1 2 1 1 1 1
 [783] 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 2 1 1 1 2 1
 [817] 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [851] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1
 [885] 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 2 1 1
 [919] 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 2 1 1 2 2 2 1 1 1
 [953] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 2 1 2 2 1 1
 [987] 1 1 2 1 2 1 1 1 2 2 1 1 1 1
```


Cluster plot
6827
8154
2624
510
8489
6627
4278
6426
2506
166
2698
684
9640
7773
20
0
-20
Profit
0
10
20
30
Sales
cluster
1
2