Kharazmi University

Faculty of Engineering

Industrial Engineering Department

Customer segmentation using K-mean algorithm and Self Organizing Map (SOM)

By:

Mahdi Keshavarz

Course:

Engineering data analysis

June 2020

Table of Contents

Summary

In this research, by implementing data mining techniques on a database consisting of sales transactions and data related to returned orders of a company, the customers have been divided into different segments, so planning and decision making will be more accurate.

The research has 4 steps:

- Introduction and data preparation
- k-means clustering
- SOM Clustering
- comparing results

The database consists of about 50,000 records and includes fields such as the number of purchases, discounts, profits, the type of product purchased, the geographic area of the customer, and the type of shipment. After Description and visualization, data preparation has been carried out.

In order to prepare the data, it has been tried to replace the missing data with the average amount of corresponding field and after the integration of the data, the dimensionality reduction operation is done so that both the results are more acceptable and the future calculations are done faster and as a result , a data warehouse is suitable for data mining.

There are two approaches to customer segmentation. In the first case, using k-means clustering, customers will be divided into several clusters, each representing a segment of customers. Before the implementation of the clustering operation, the statistical tendency of the data for clustering was ensured and then the optimal number of clusters was obtained using the Elbow method.

The second approach is similar to the first approach, with the difference that before clustering, the data has been transformed into 2-dimensional data using self-organizing maps. This reduction in dimensions helps to increase the accuracy of clusters and achieve better results. At the end of the research, the results of two approaches have been compared.

1- Data collection and preparation

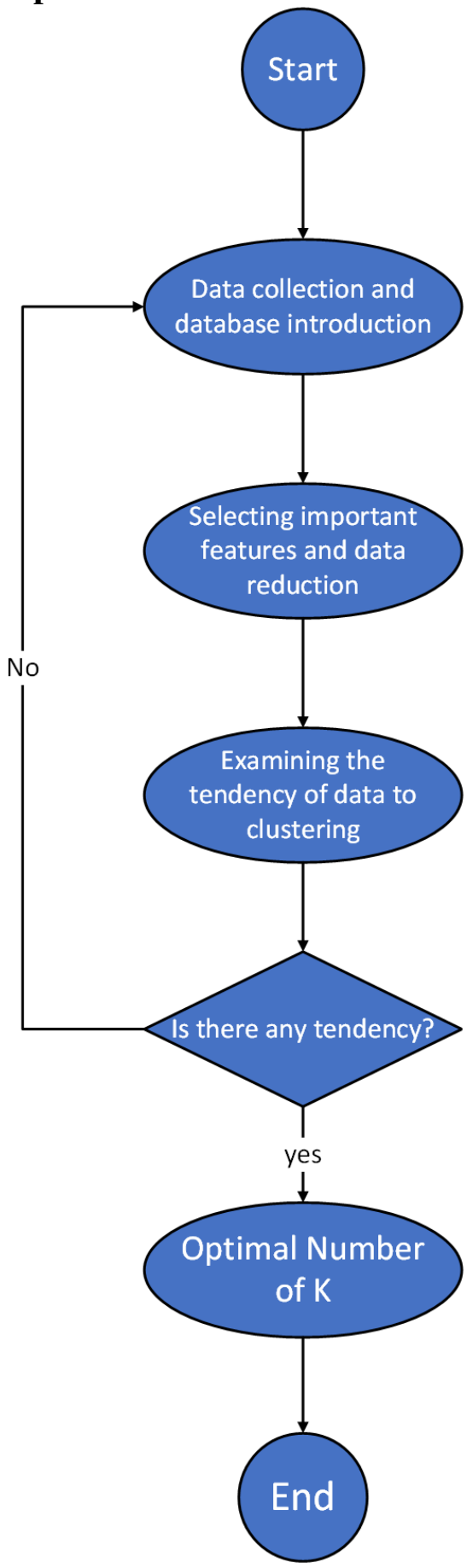


Figure 1- clustering flowchart with k-mean method

1-1- Database description

The database used shows the sales transactions of a company with more than 50 thousand records in an Excel file, where customer information, type of purchase, type of shipment, geographical area, etc. are given. There are 3 sheets in this file, and the first 2 sheets will be used. The first page is data related to orders with 51,290 records and 24 fields.

1-2- Data pre-processing

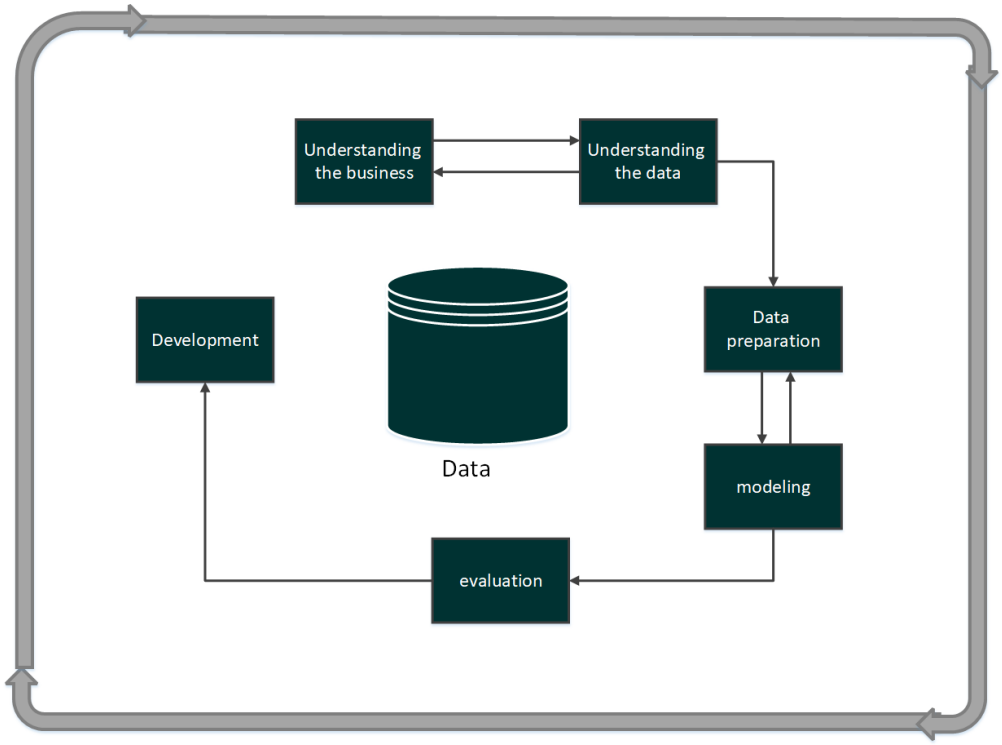


Figure 2- CRISP methodology

1-2-1- Understanding the industry

The data is related to the sales of a company whose products fall into 3 categories:

- Home Appliances
- office Equipment
- Technological equipment (copier, telephone, etc.)

Sales are done globally

1-2-2- Understanding the data

Table 1- Summary of quantitative information of database

Mean	Maximum	Minimum	Field
246.879	22638.480	0.44	Sale
3.476	14	1	Quantity
0.143	0.85	0	Discounut

28.641	8399.980	-6599.980	Profit
26.520	933.570	1	Shipping Cost

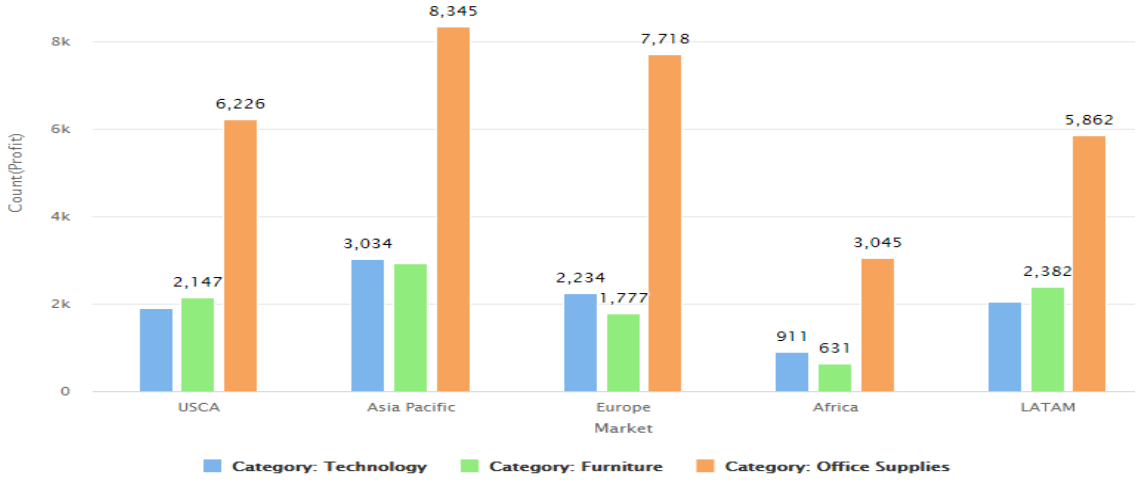


Figure 3- Profitability of markets and product categories

Figure 4 – Profitability of markets and shipping priorities

From Figure 4, we can see that there are missing data in the database.

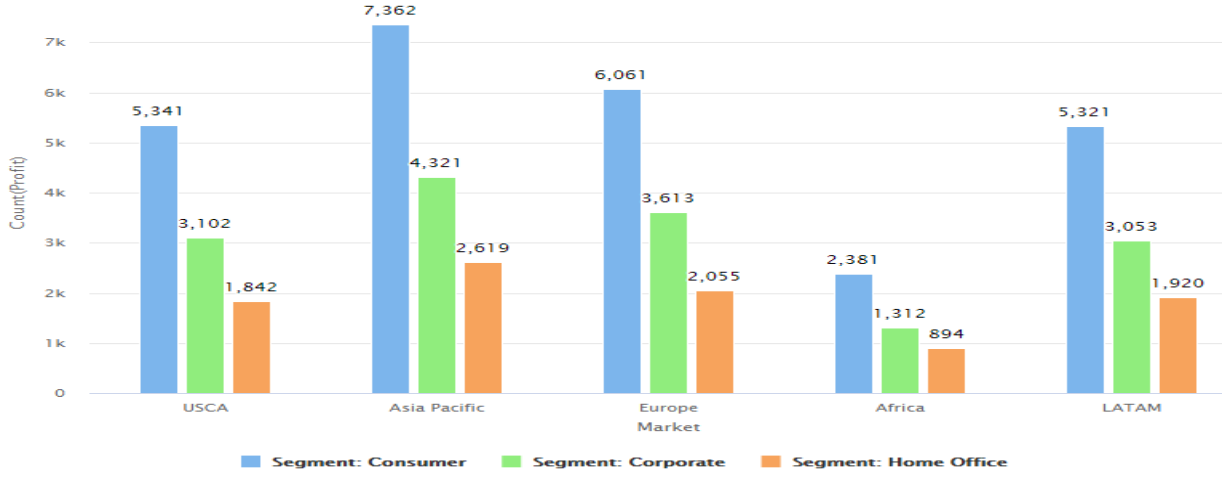


Figure 5 - Profitability of markets and segments

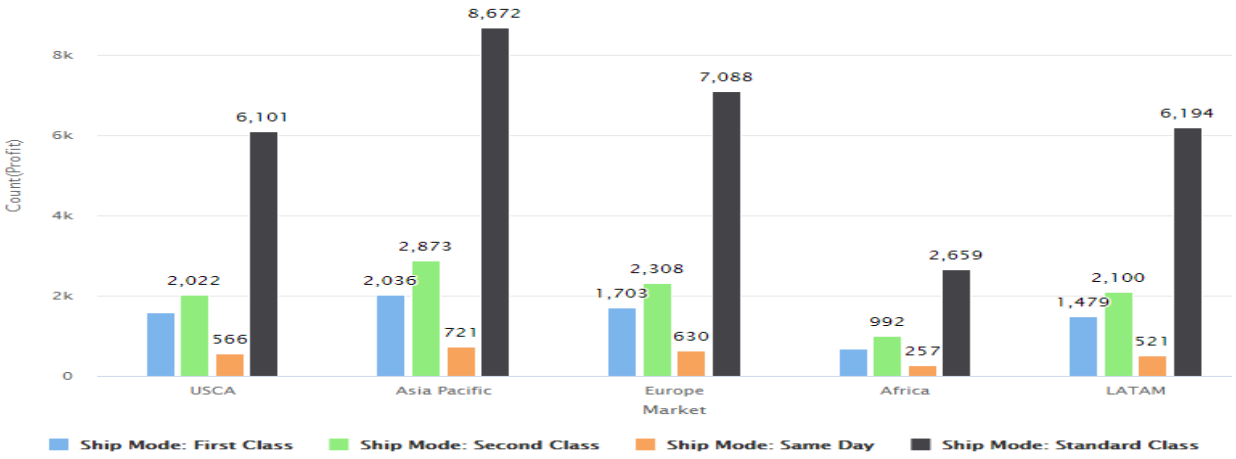


Figure 6 - Profitability of markets and ship mode

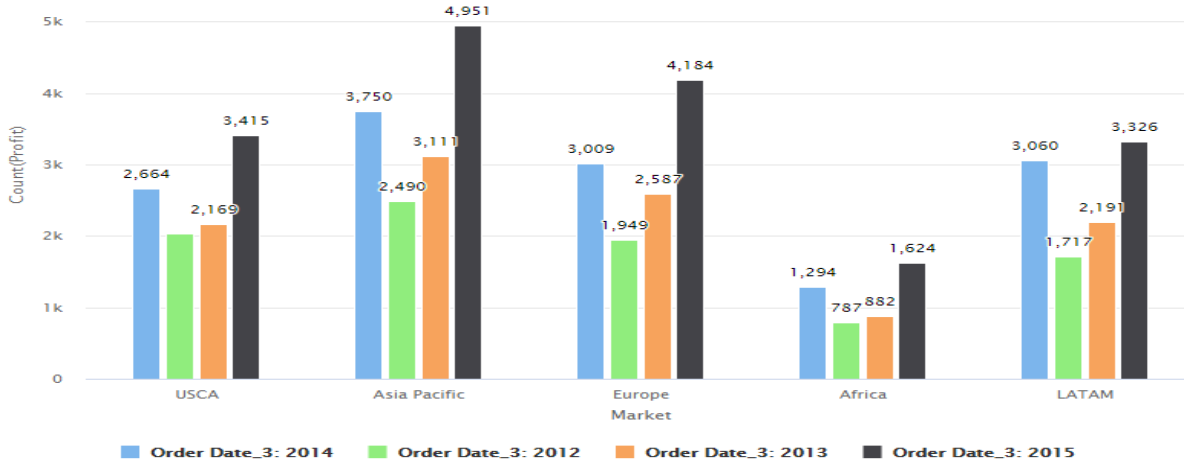


Figure 7 - Profitability of markets and order date

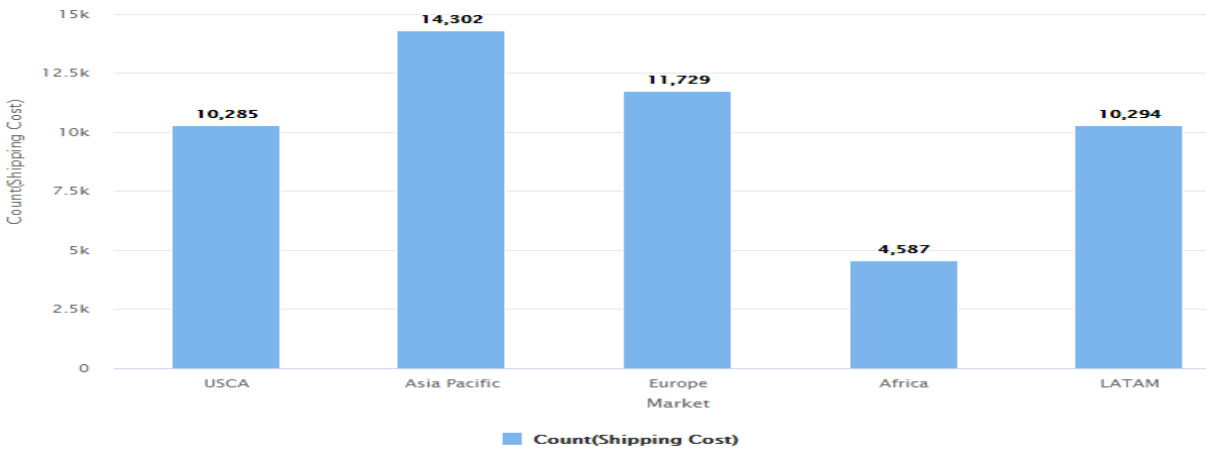


Figure 8 – Shipping cost based on different markets

1-2-3- Data preparation

There are 5 steps:

1-2-3-1- Data cleaning

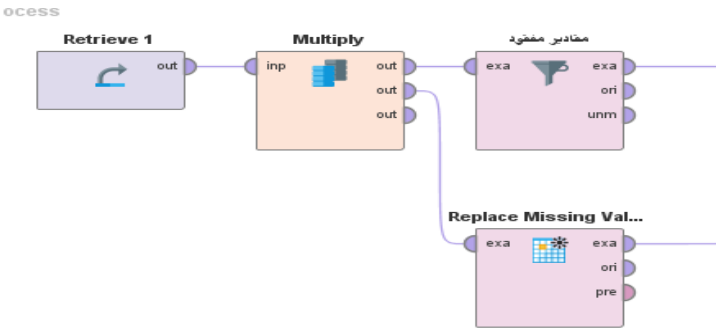


Figure 15 – Missing values treatment model

1-2-3-2- Data integration

There are 3 sheets:

- Orders
- Returns
- customers

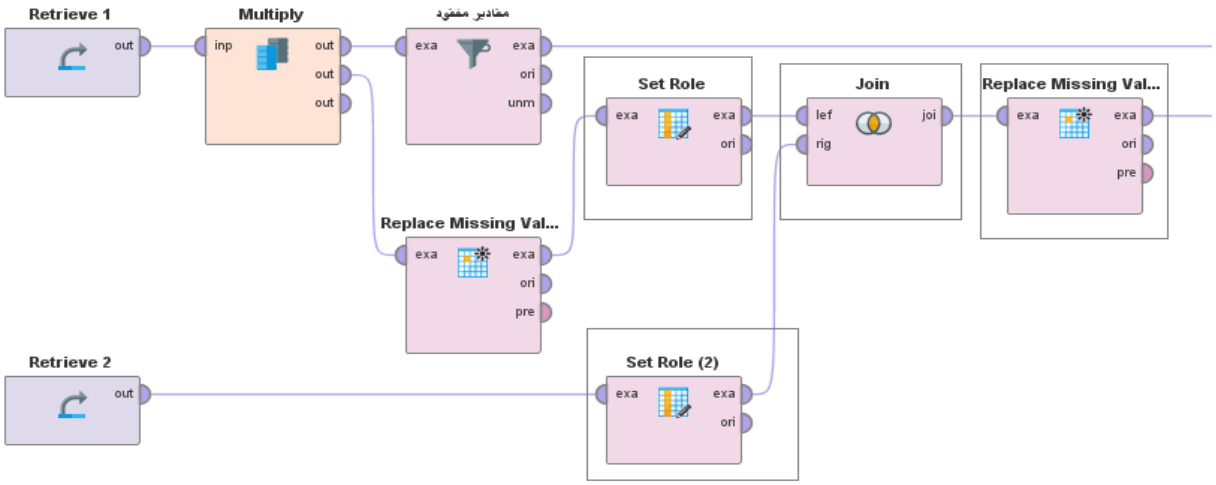


Figure 17 – Data integration model

1-2-3-3 Data transformation

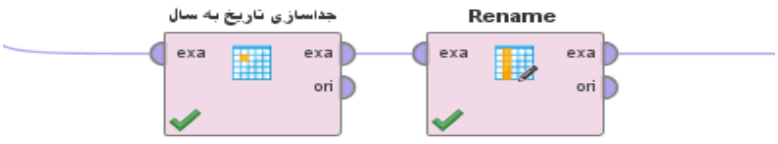


Figure 21 – Split and Rename operators

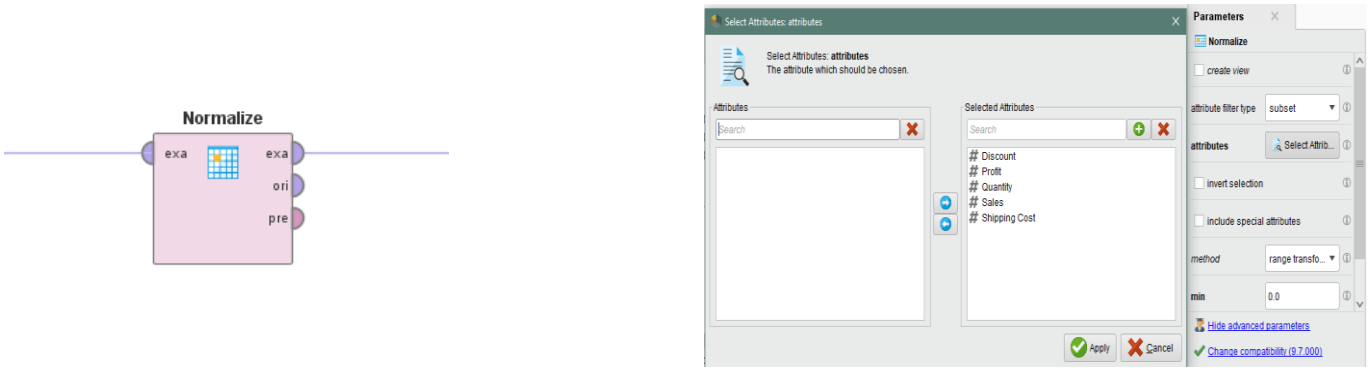


Figure 24 – Normalization model

1-2-3-4- data reduction

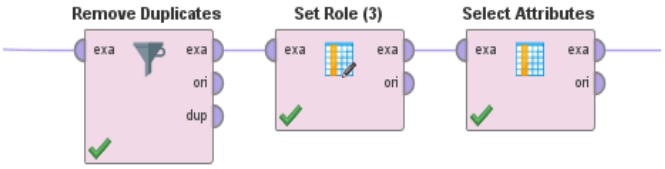


Figure 25 – Data reduction model

1-2-3-5- Data discretization

There is no need for discretization because our data is suitable for creating clusters.

1-2-4- Modeling

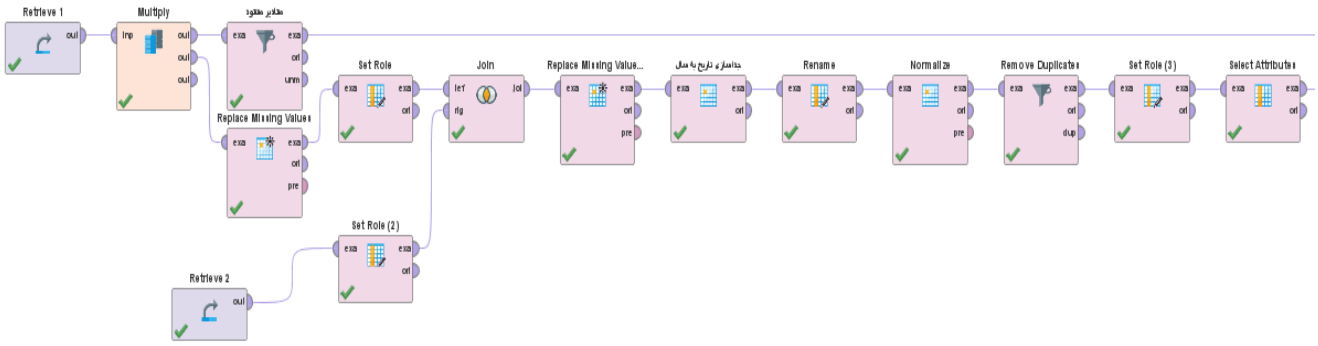


Figure 28 – Data preparation model

1-2-5- Assessment

Each of the data mining models will be evaluated after implementation.

1-2-6- Implementation

Data mining models will be implemented in sections 2 and 3.

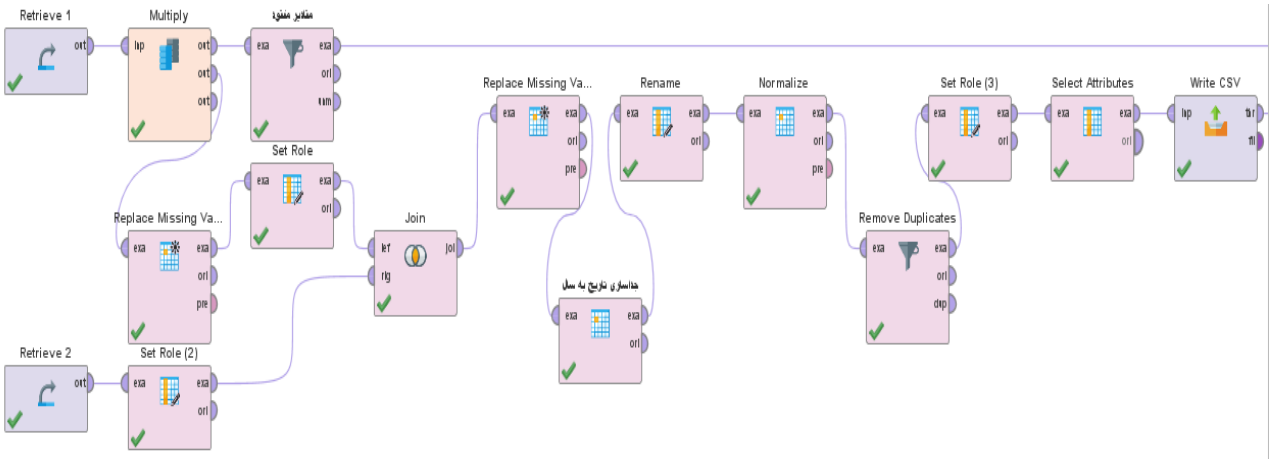


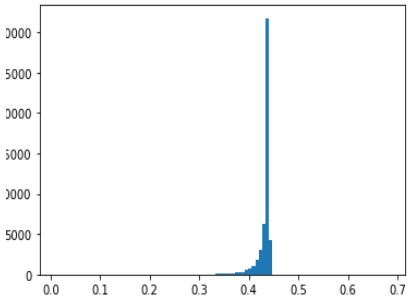
Figure 29 - The final model of step 1

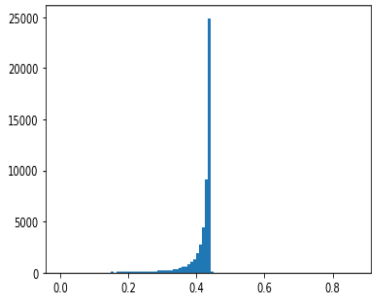
1-3- Tendency to clustering

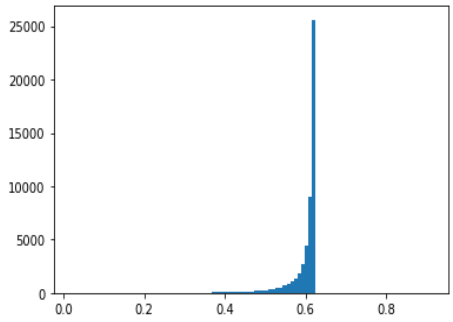
For this purpose, it is necessary to calculate the distance between all records two by two and then install a histogram from the set of resulting numbers.There are 3 numerical fields suitable for this task:

- Normalized sales
- Normalized profit
- Normal shipping cost

The existence of a peak indicates that there is some kind of correlation between the data and, .as a result, a tendency to cluster. These peaks can be seen in all 3 graphs







1-4- Optimal number of K

The ELBOW method has been used to determine the optimal number of clusters or k.

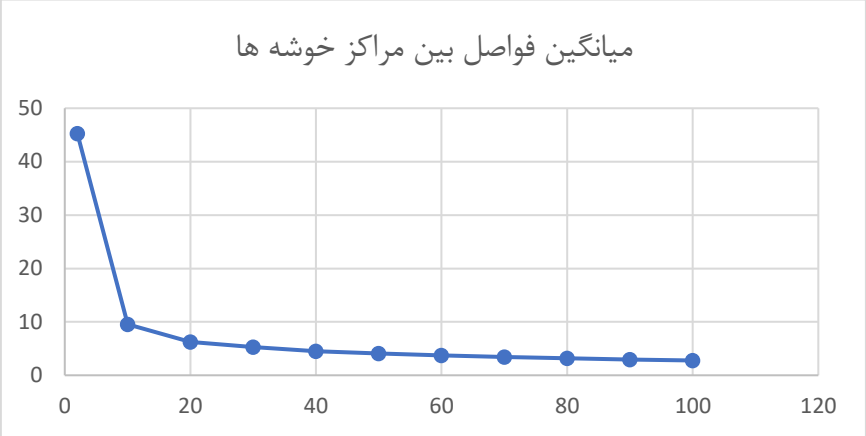


Figure 34 - The average distance between the centers of the clusters

At the point 10, there is a sudden change so 10 is the optimal number for clusters.

2- Clustering using K-means

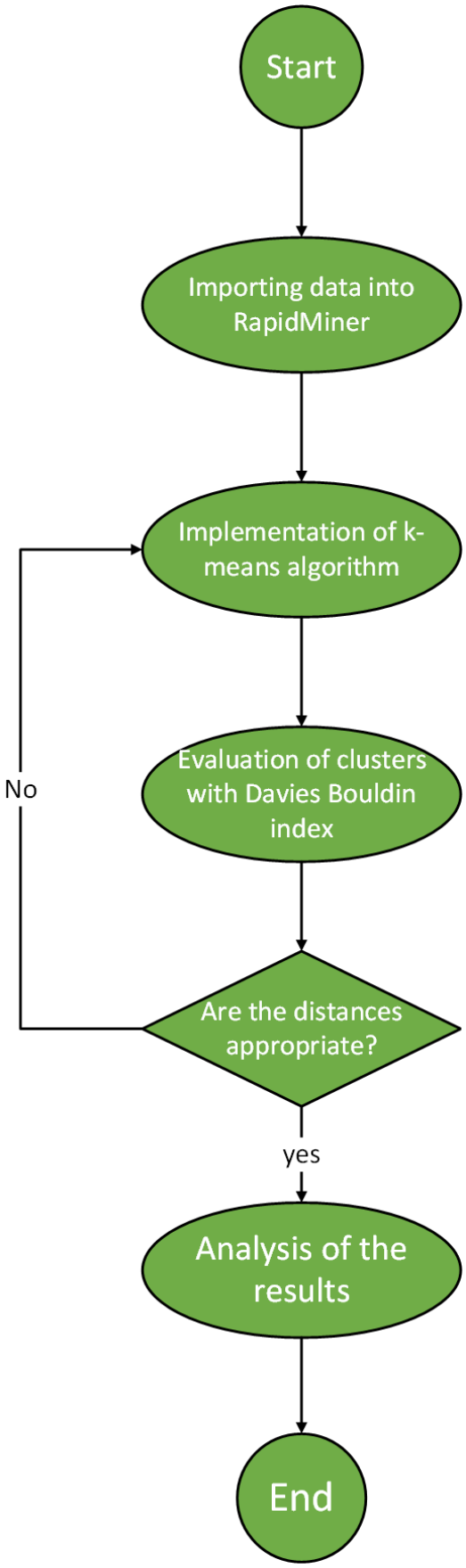


Figure 35 - Clustering flowchart with k-means

2-1- Create and run the model

There are two ways to assign numbers to qualitative variables. For ordinal qualitative variables, the following formula is used for normalization and numerical label assignment. In this formula, r represents the quality rating (for example, r=1 for the highest shipment priority or critical which has a rating of 1) and p represents the total possible states (total quality ratings).

$$z = \frac{r - 1}{p - 1}$$

Numerical labels can also be assigned to nominal qualitative variables, but since the k-means clustering algorithm is basically for quantitative variables, such an approach is not recommended.

For the binary variable of return, the number 1 represents "yes" and 0 represents "no".

In the first type of modeling, we will use only quantitative features. Then ordinal quality features and then nominal quality features are also added.

Features whose correlation is more than 0.6 are removed.

After implementation, it is necessary to validate the model.

The Cluster Model Visualizer operator will be used for better visualization.

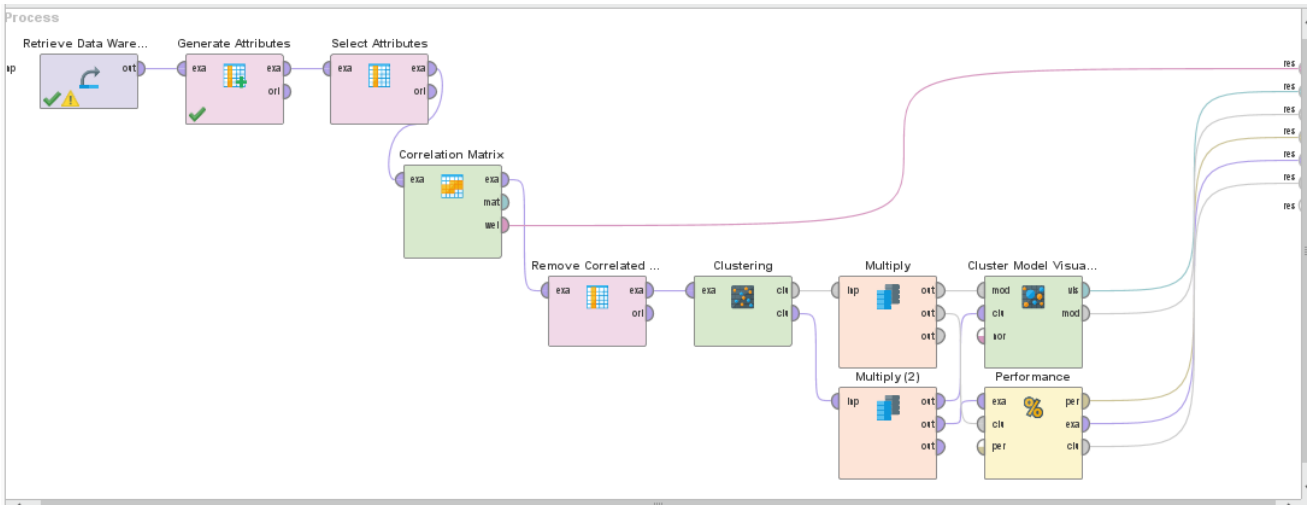


Figure 44 - k-means model

2-2- Analysis of the results

The evaluation index (Davis Bouldin) is equal to 0.704, which is a good value.

The average intra-cluster distances within the cluster is also 0.032. Therefore, the model has good validity and does not require re-implementation.

The two features of shipping date and shipping cost have been removed due to their high correlation with features of order and sales date.

Attribut...	Sales	Quantity	Discount	Profit	Shipping Cost	Order D...	Ship Date
Sales	1	0.314	-0.087	0.485	0.768	-0.003	-0.003
Quantity	0.314	1	-0.020	0.104	0.272	-0.005	-0.005
Discount	-0.087	-0.020	1	-0.317	-0.078	-0.006	-0.006
Profit	0.485	0.104	-0.317	1	0.354	0.003	0.002
Shipping...	0.768	0.272	-0.078	0.354	1	-0.003	-0.004
Order Da...	-0.003	-0.005	-0.006	0.003	-0.003	1	0.994
Ship Date	-0.003	-0.005	-0.006	0.002	-0.004	0.994	1

Figure 45 - Table of correlations

The graph below shows the correlation between the fields graphically. A redder color means more correlation.

Figure 46 – Correlation between fileds

The number of records in each cluster is as follows:

Cluster 0: 2236 items
Cluster 1: 10095 items
Cluster 2: 7080 items
Cluster 3: 7990 items
Cluster 4: 2808 items
Cluster 5: 8706 items
Cluster 6: 3649 items
Cluster 7: 3761 items
Cluster 8: 1901 items
Cluster 9: 2986 items

Total number of items: 51212

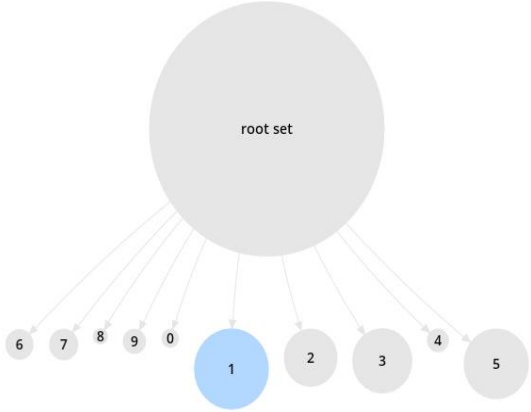


Figure 47 - Illustration of the number of members of each cluster

The centers of the clusters are as follows:

Cluster	Sales	Quantity	Discount	Profit	Order Date
Cluster 0	0.007	0.185	0.614	0.435	2013
Cluster 1	0.008	0.103	0.049	0.443	2015
Cluster 2	0.012	0.195	0.055	0.444	2012
Cluster 3	0.008	0.106	0.048	0.443	2014
Cluster 4	0.007	0.175	0.607	0.435	2014
Cluster 5	0.012	0.192	0.051	0.444	2013
Cluster 6	0.006	0.172	0.607	0.435	2015
Cluster 7	0.022	0.435	0.065	0.447	2015
Cluster 8	0.008	0.180	0.620	0.435	2012
Cluster 9	0.022	0.435	0.063	0.447	2014

Figure 48 – Cluster centers

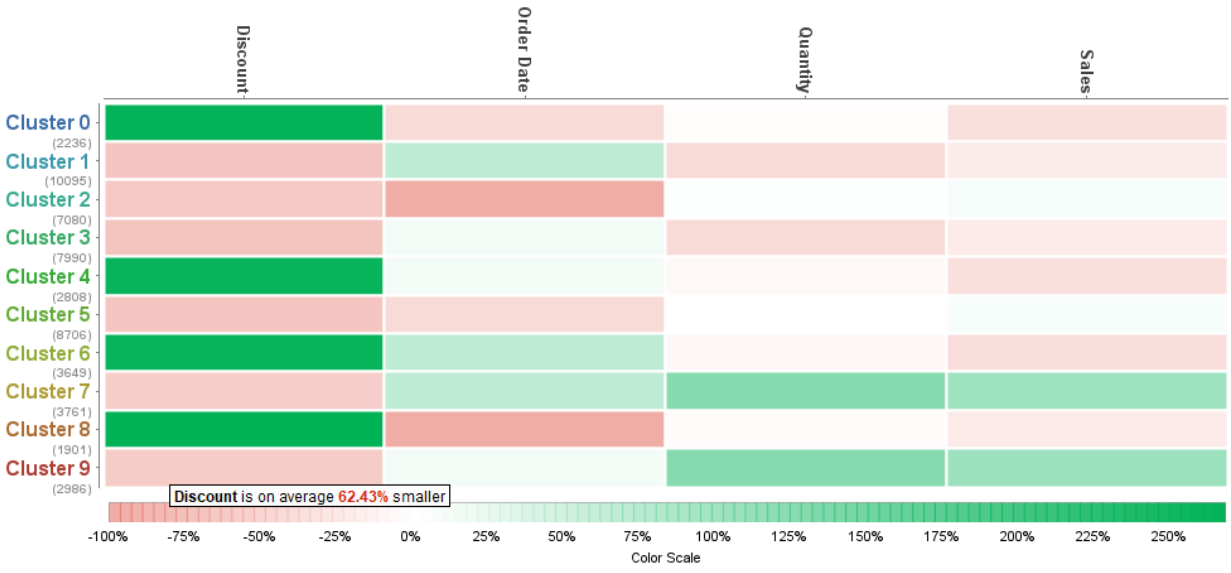


Figure 49 – Heat map of clusters and fields

The general information of the clusters can be seen in the figure below.

Figure 50 - Cluster information summary

There is also good statistical information that will be useful for the final conclusion.

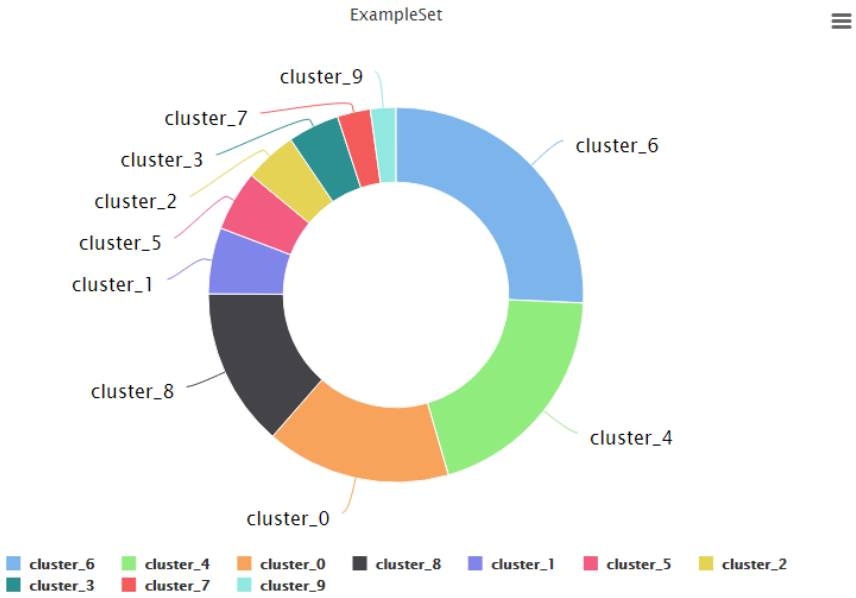


Figure 52 – Clusters based on total discount

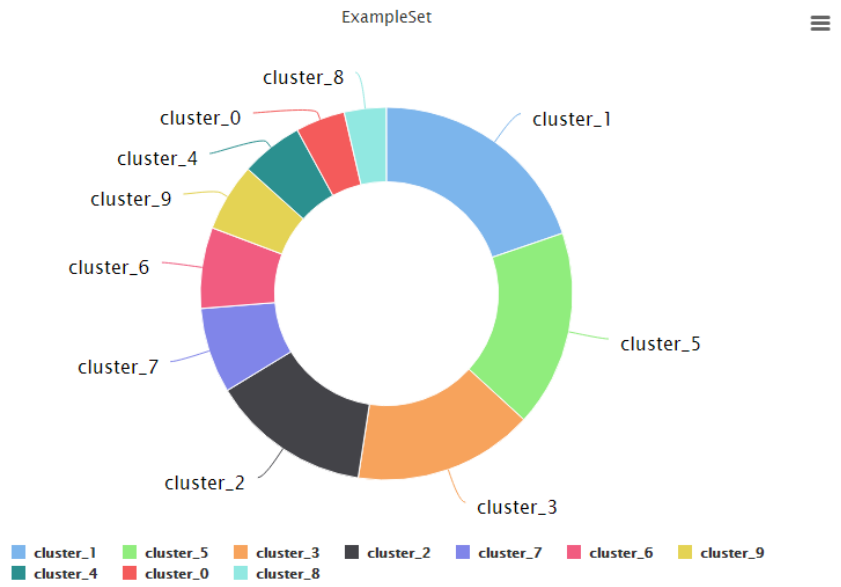


Figure 51 – Clusters based on total profitability

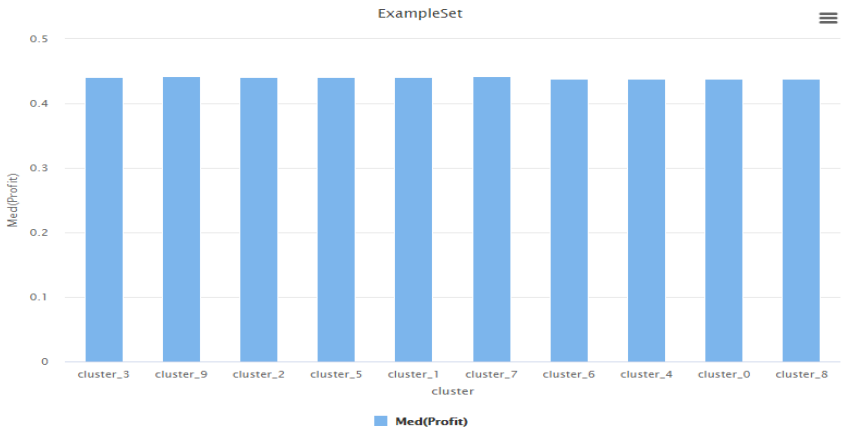


Figure 53 - Average profitability of clusters

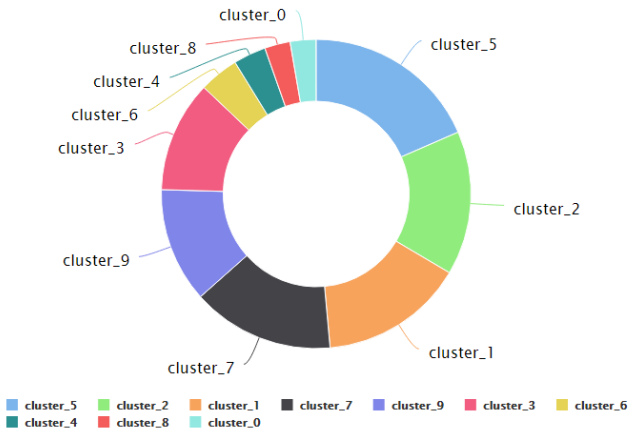


Figure 54 - Clusters based on total sales

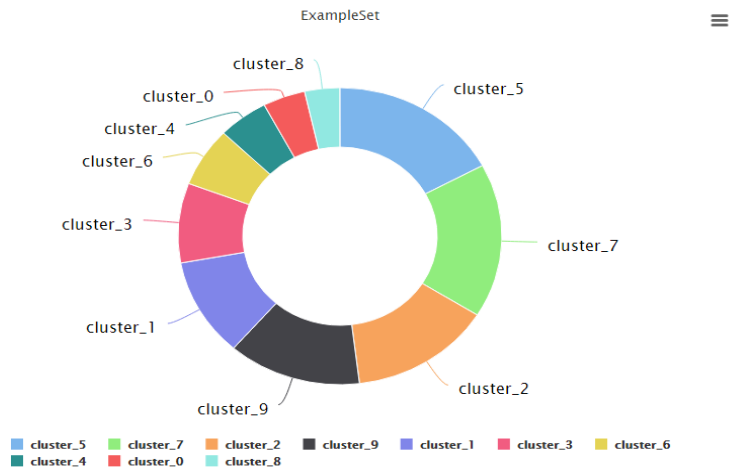


Figure 55 - Clusters based on the total number of sales

Table 2 – Ranking of clusters based on different characteristics

Average year	Number of transactions Rank	Sales revenue rank	Number of sales rank	Discount rank	Profitability rank	Cluster name
2013	9	8	5	2	7	**Cluster 0**
2015	1	5	10	9	5	**Cluster 1**
2012	4	3	3	7	3	**Cluster 2**
2014	3	6	9	10	6	**Cluster 3**
2014	8	9	7	3	8	**Cluster 4**
2013	2	4	4	8	4	**Cluster 5**
2015	6	10	8	4	9	**Cluster 6**
2015	5	1	1	5	1	**Cluster 7**
2012	10	7	6	1	10	**Cluster 8**
2014	7	2	2	6	2	**Cluster 9**

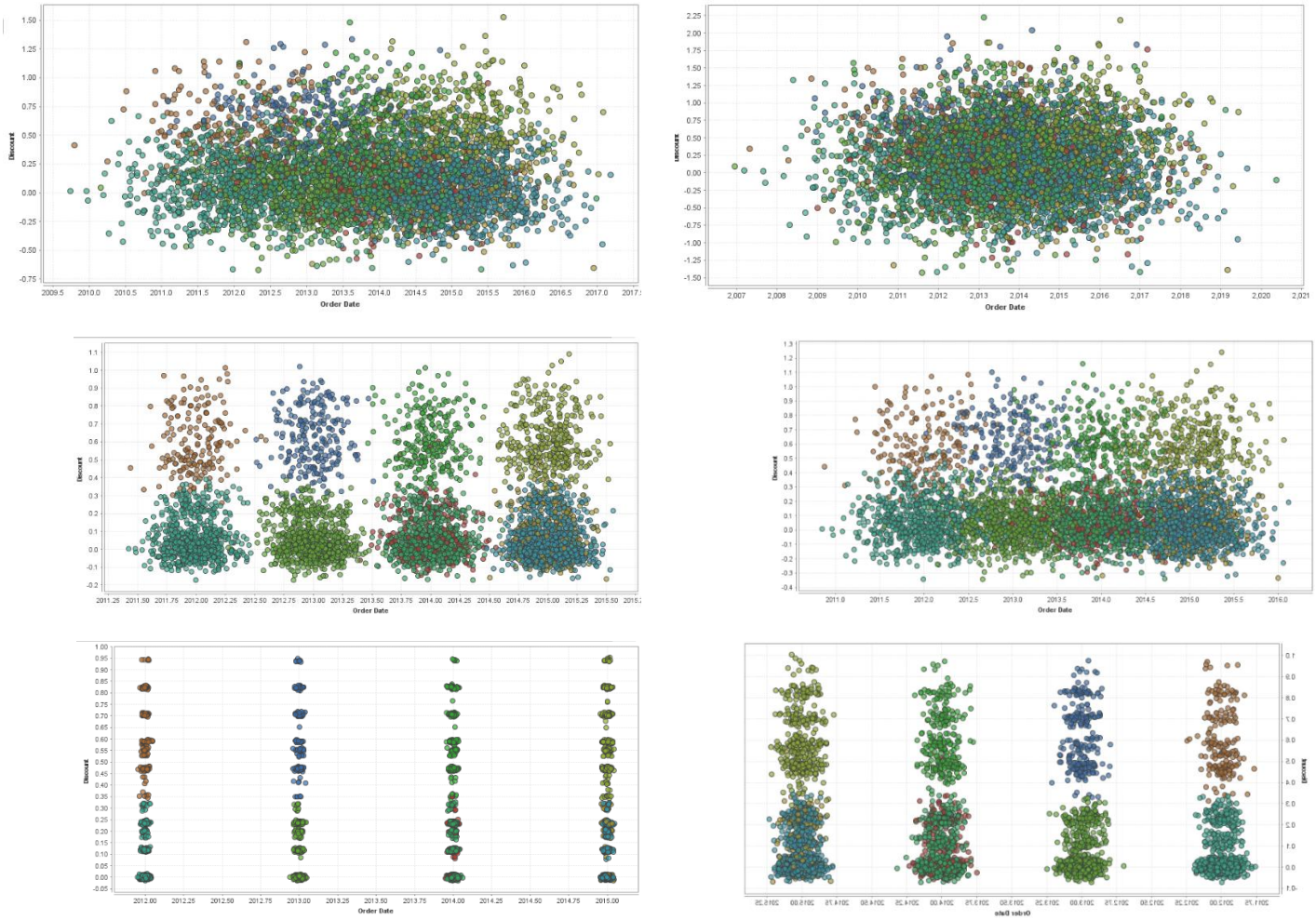


Figure 57 – Allocation of records to each cluster

Step 3: Self-Organizing Map (SOM)

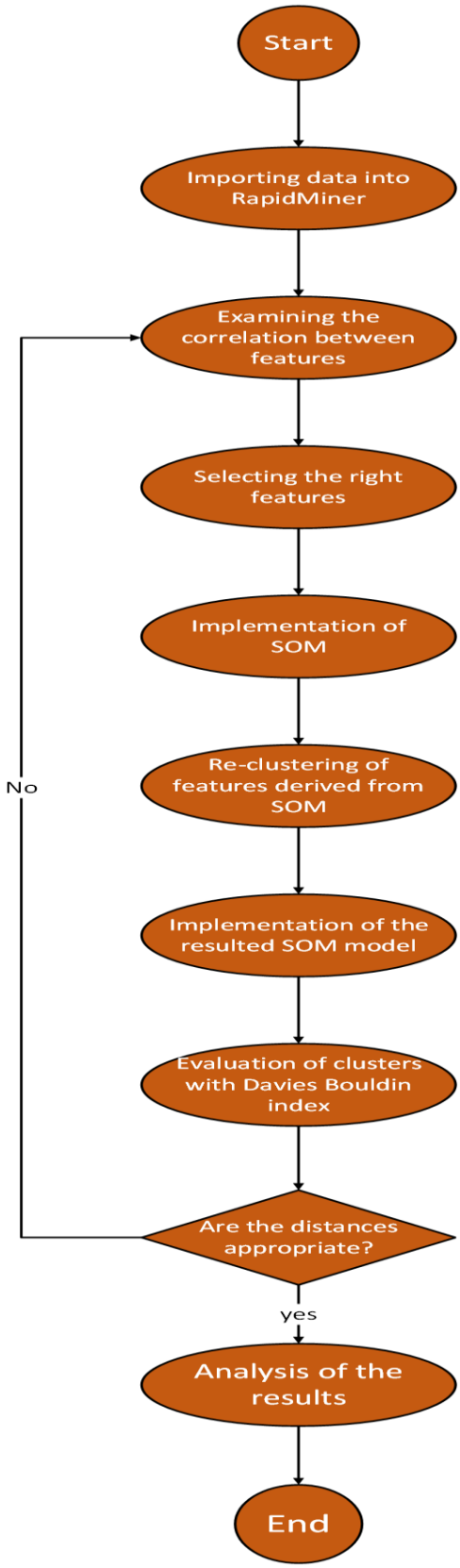


Figure 72 – Self-organizing map flowchart

3-1- Modeling

Self-organizing maps is a method to reduce the dimension of data and is placed in the category of neural networks, so qualitative variables can be used as labels for each category. In this model, there is more freedom and you can use both qualitative variables directly and using numerical symbols. Like the previous approach, different methods are tested and the best result is selected.

In the following, the correlation between the features is checked and then based on the weight of each feature, we select and display the most important ones. Also, these features and their values are selected as the input of the SOM process, and the results of the SOM, in addition to visualization, are once again clustered using the K-means method. In fact, this clustering is done on the most important features that have been selected in the previous steps. After clustering, the clusters are evaluated with the Davies-Bouldin index and the results of this evaluation are also displayed. We will also use the “Visualize Model by SOM” operator to illustrate the clusters.

The points about this model are:

A) To have different options and trial and error, the method is the same that a numerical label was determined in clustering for qualitative features.

B) To prevent the software from displaying an error message, it is necessary to select a qualitative variable each time and select it as a label in the next step (because self-organizing maps are a neural network-based method).

C) Features whose weight is 0.5 and more will be selected.

D) The number of training data of the SOM operator is 100.

E) The number of clusters is equal to 10 (also 3 is a suitable option, but first the best mode, i.e. 10 clusters and then 3 clusters will be implemented).

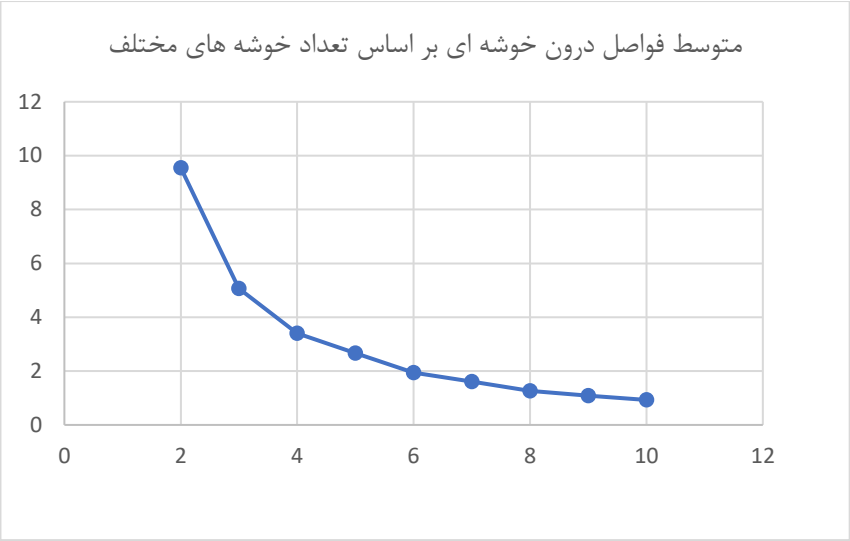


Figure 74 – Average intra-cluster distances

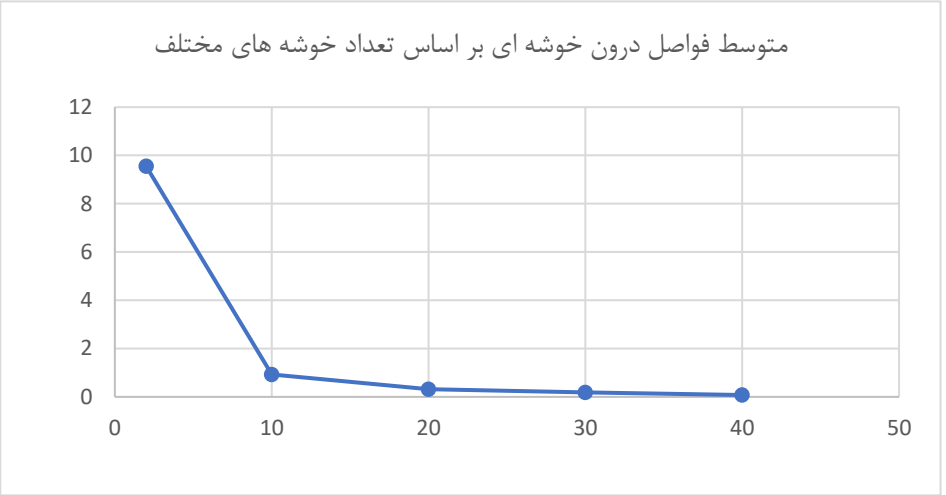


Figure 73 – Average intra-cluster distances

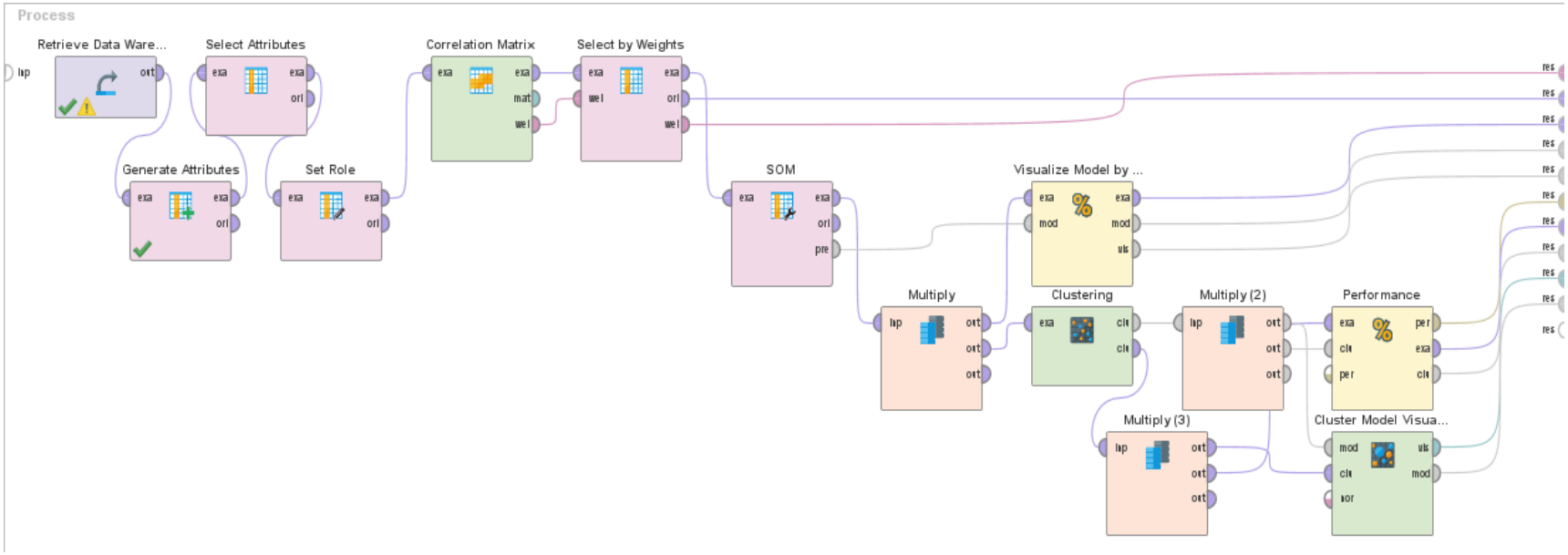


Figure 75 – SOM model

3-2- SOM model results

Davis Boldin's index for this model is 0.630 and the average intra-cluster distances is 0.883, which indicate appropriate modeling.

The distribution of transactions in all 10 clusters is proportional.

Cluster 0: 6745 items	Cluster 1: 5382 items	Cluster 2: 3863 items	Cluster 3: 4697 items
Cluster 4: 3258 items	Cluster 5: 3325 items	Cluster 6: 5339 items	Cluster 7: 8293 items
Cluster 8: 4947 items	Cluster 9: 5363 items	Total number of items: 51212	

attribute	weight
Sales	0.066
Quantity	0.921
Discount	1
Profit	0.591
Shipping Cost	0.221
Order Date	0.000
Ship Date	0

Figure 77 - weight of Characteristics

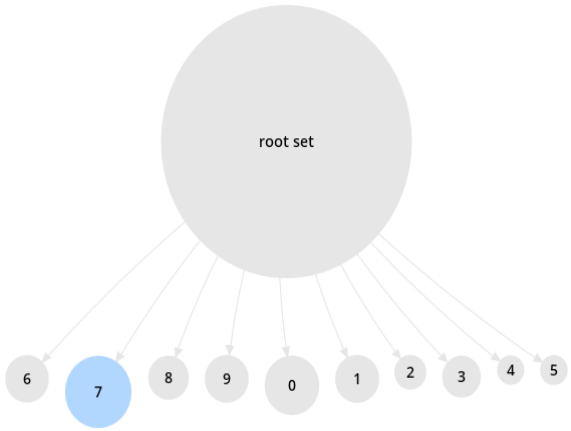


Figure 76 – The size of each cluster

Cluster	SOM_0	SOM_1
Cluster 0	2.940	8.844
Cluster 1	7.584	1.097
Cluster 2	3.245	4.098
Cluster 3	7.986	6.338
Cluster 4	5.110	7.405
Cluster 5	1.748	1.020
Cluster 6	6.188	3.999
Cluster 7	0.123	2.842
Cluster 8	1.071	7.301
Cluster 9	4.001	0.067

Figure 78 - The centers of the clusters

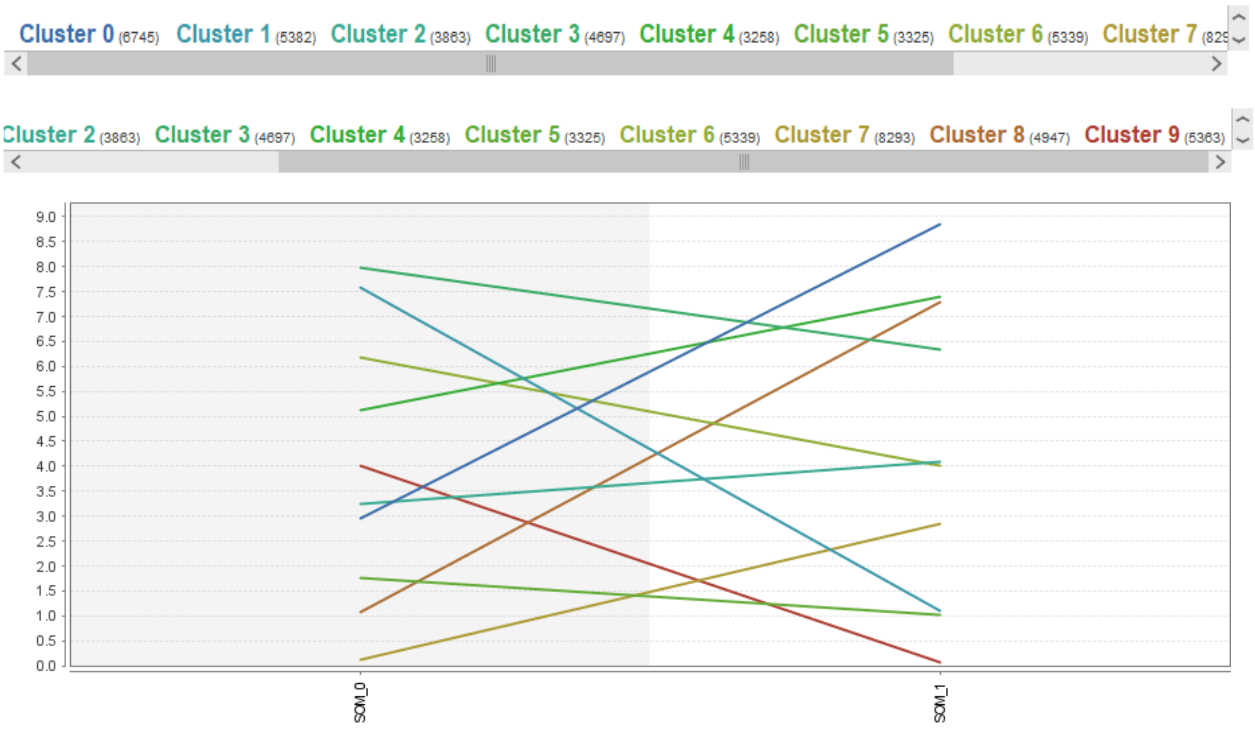


Figure 79 - Comparison of cluster centers

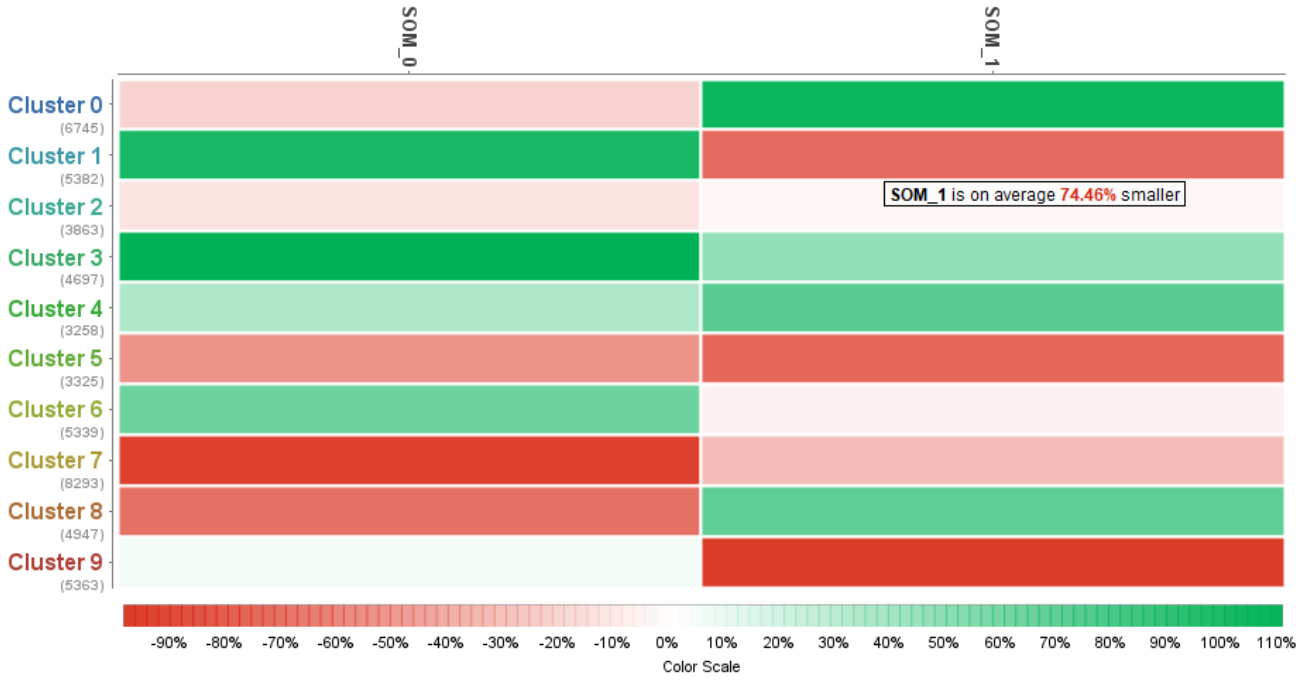


Figure 80 - Heat map of clusters

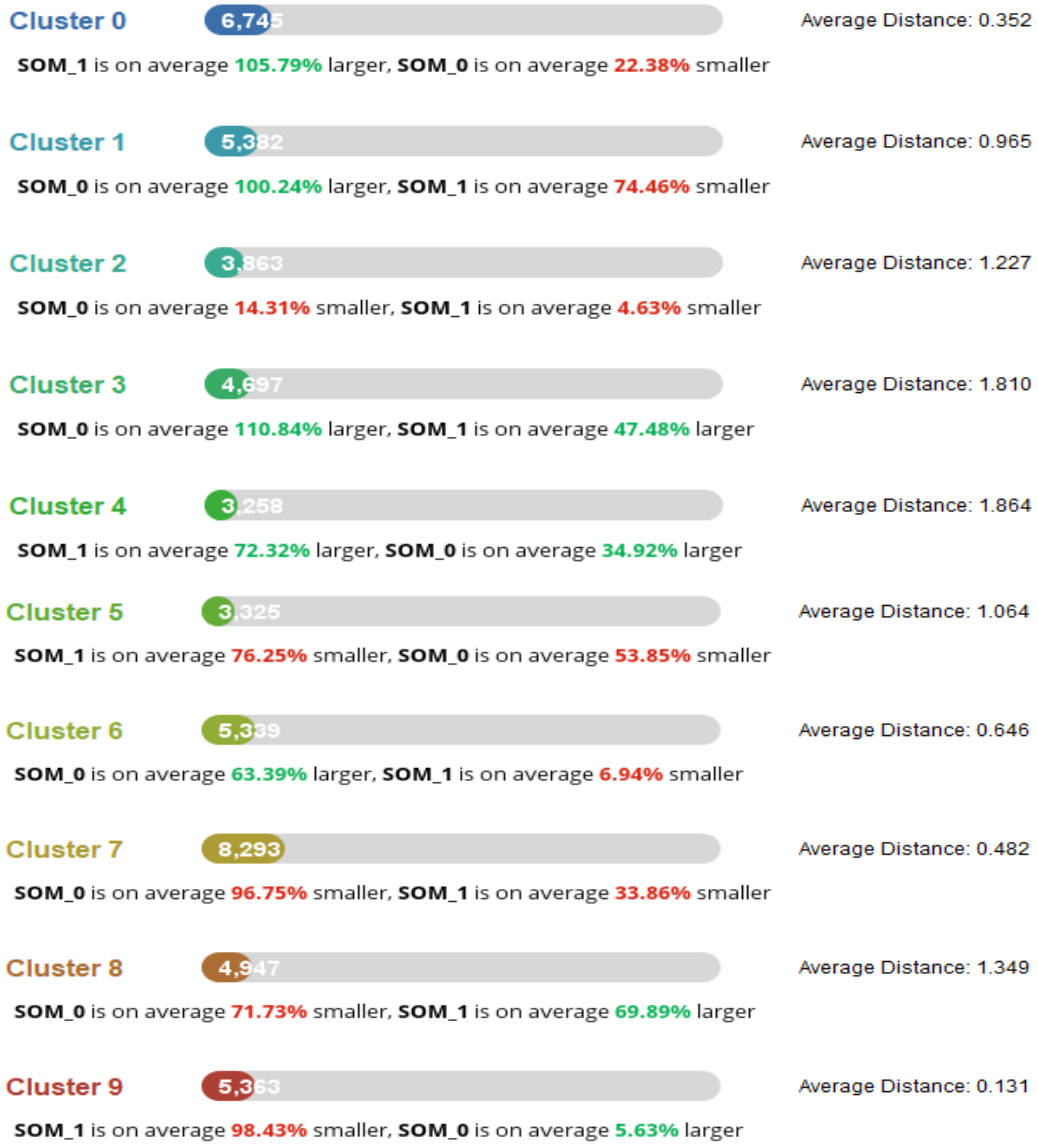


Figure 81 – Information of each cluster

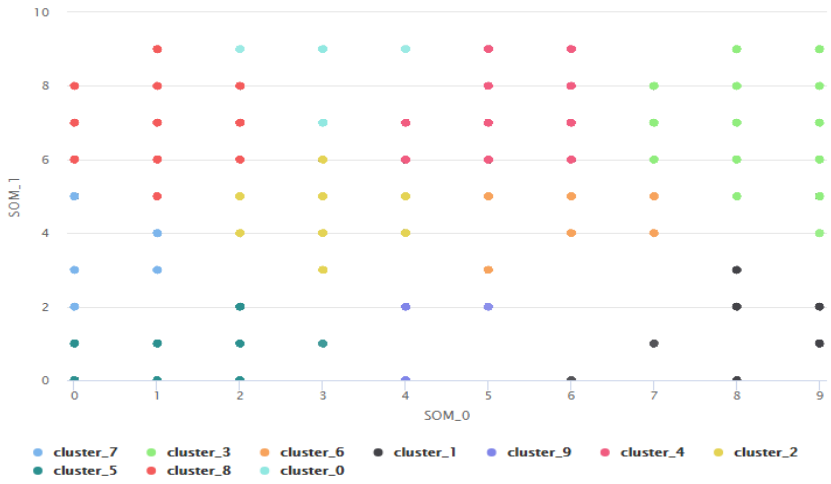


Figure 82 - Scattering of clusters

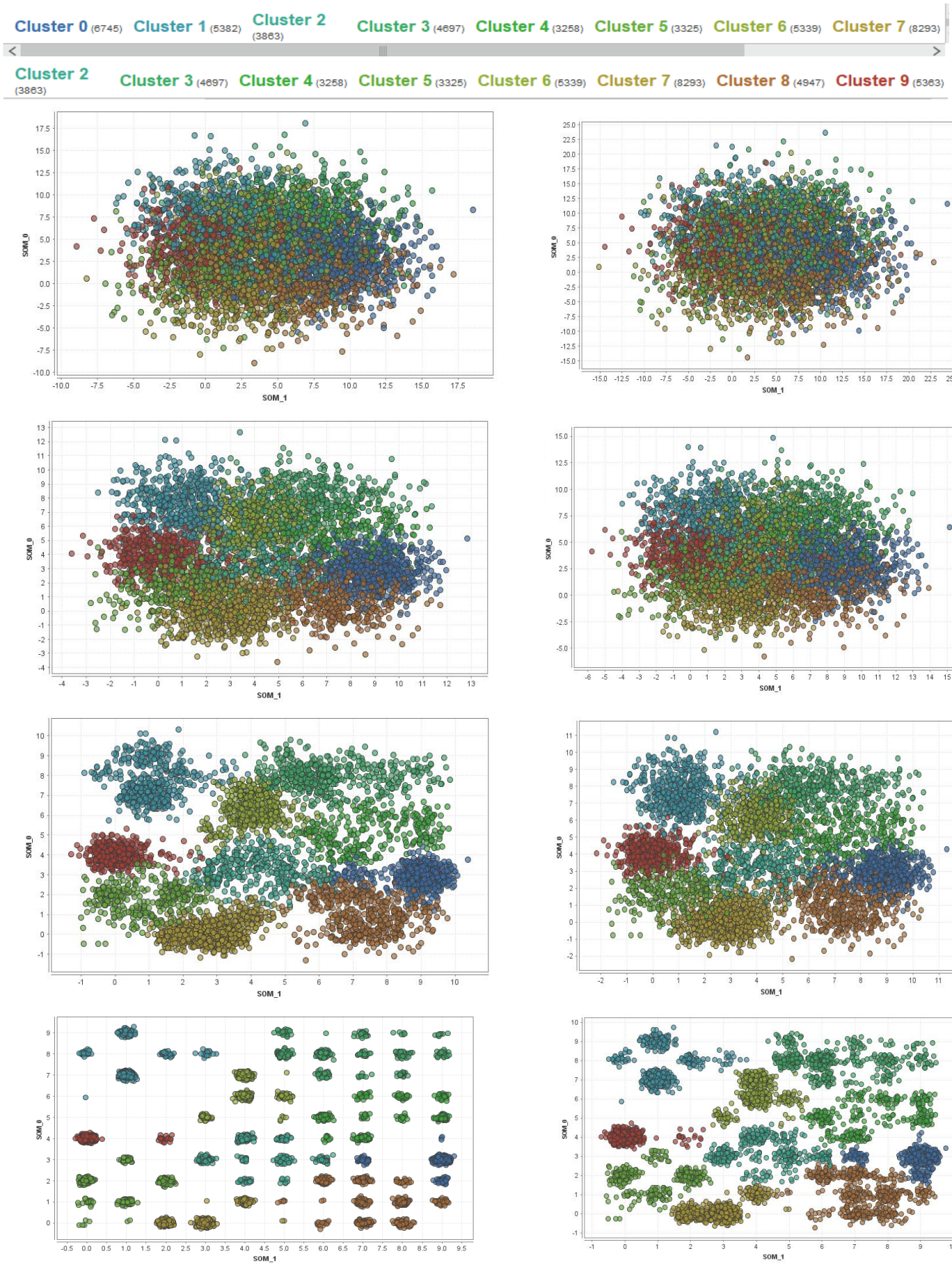


Figure 83 - Assigning records to each cluster

4 -Comparing the results of two models

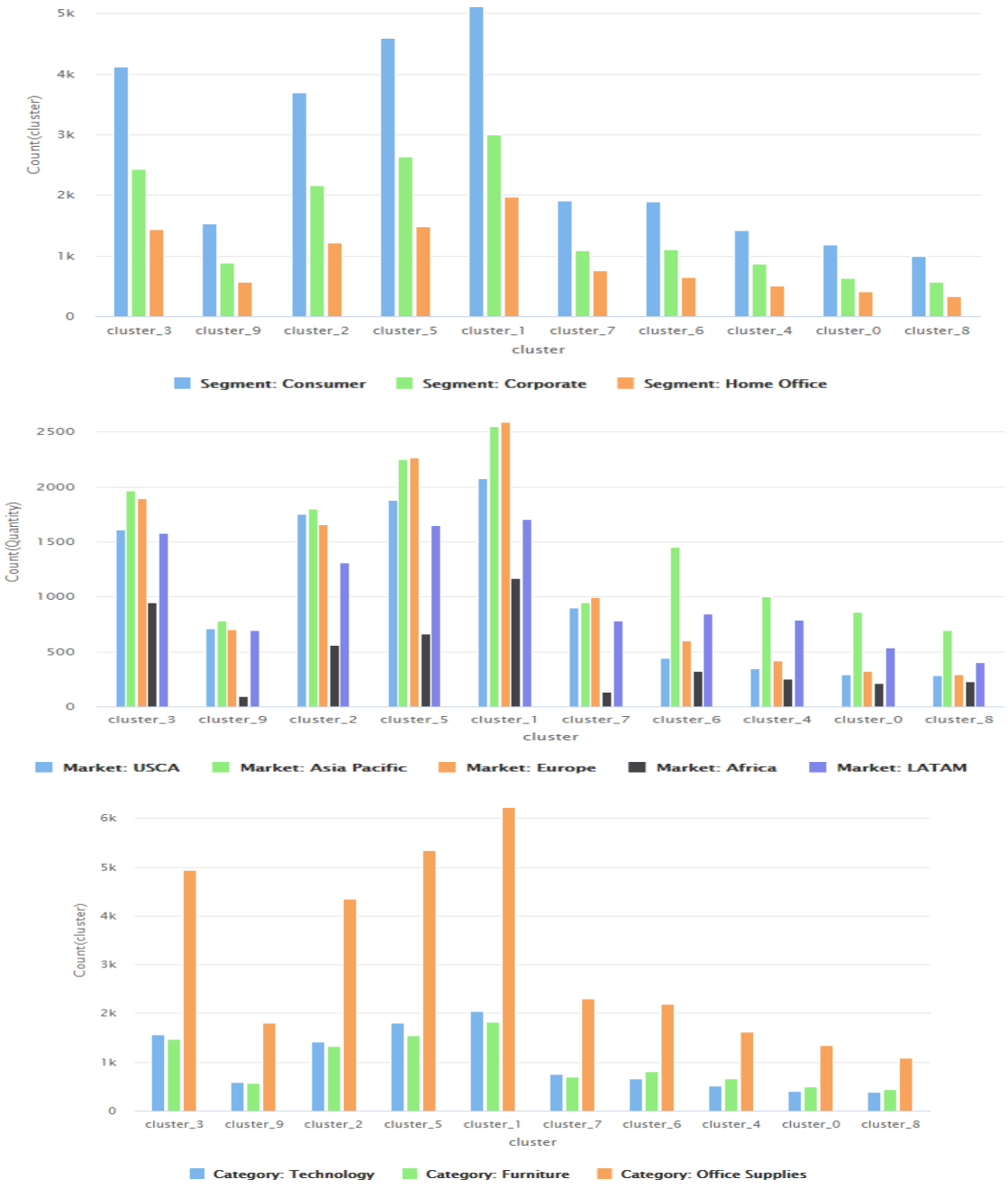
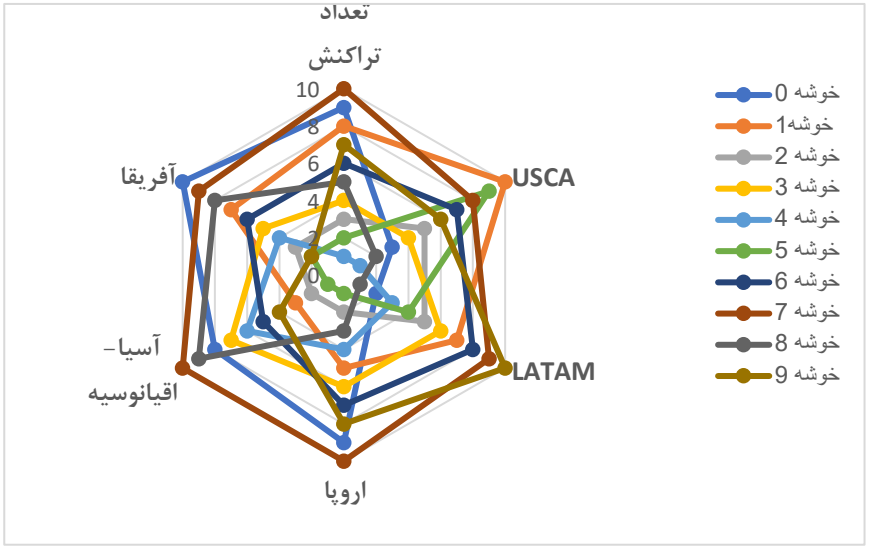


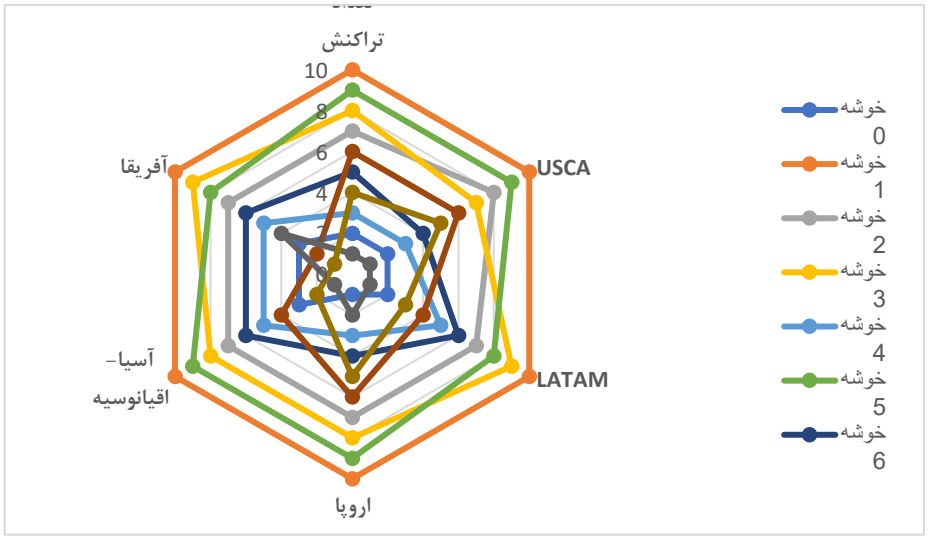
Figure 114- Distribution of nominal qualitative variables in clusters (model of step 2)

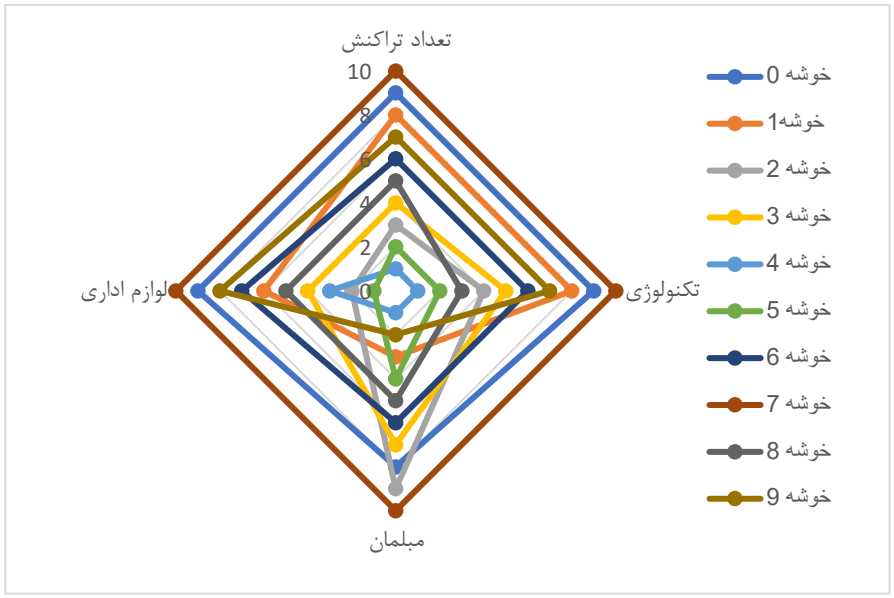
Table 11- Ranking of clusters based on different characteristics

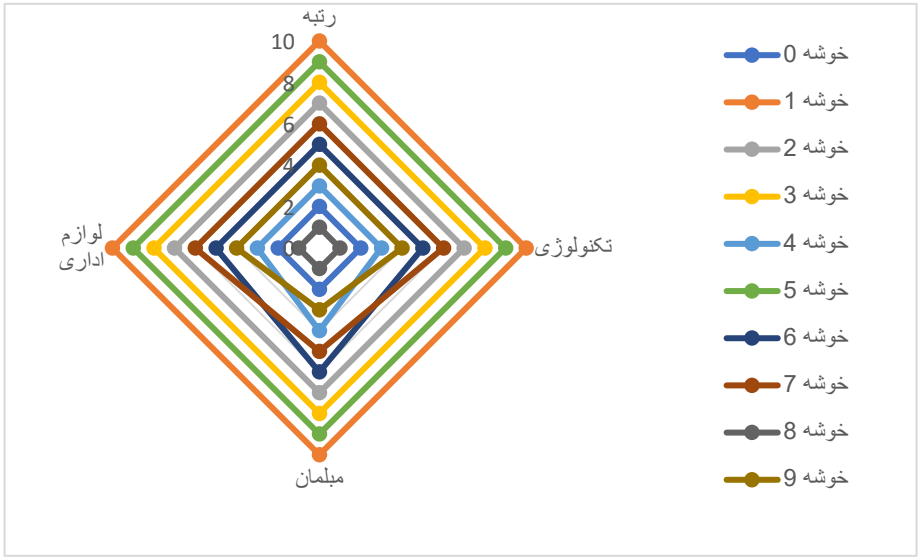
Rank in the home segment	Rank in the company segment	Rank in the consumer segment	Rank in office supplies category	Rank in furniture category	Rank in technology category	Rank in africa market	Rank in asia market	Rank in Europe market	Rank in LATAM market	Rank in USCA market	Rank in number of transactions	Cluster name
9	9	9	9	9	9	8	8	10	9	9	9	Cluster 0
1	1	1	1	1	1	1	1	1	1	1	1	Cluster 1
4	4	4	4	4	4	4	4	4	4	3	4	Cluster 2
3	3	3	3	3	3	2	3	3	2	4	3	Cluster 3
8	8	8	8	7	8	6	6	8	6	8	8	Cluster 4
2	2	2	2	2	2	3	2	2	3	2	2	Cluster 5
6	6	6	6	5	6	5	5	7	5	7	6	Cluster 6
5	5	5	5	6	5	9	7	5	7	5	5	Cluster 7
10	10	10	10	10	10	7	10	9	10	10	10	Cluster 8
7	7	7	7	8	7	10	9	6	8	6	7	Cluster 9

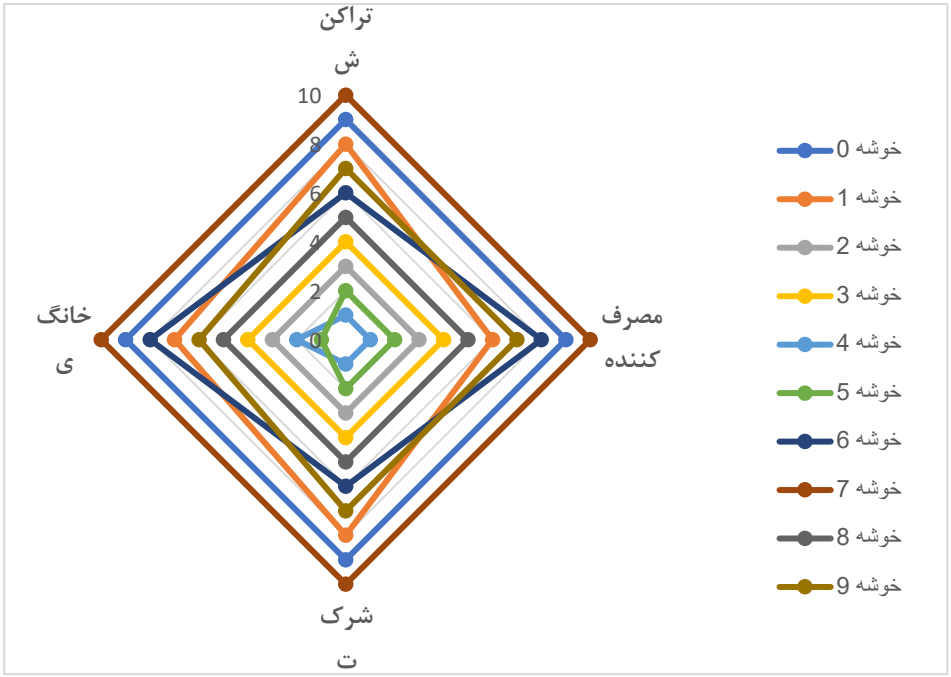
This table is drawn for step 2 model.


تعداد تراکنش
10
8
6
4
2
0
USCA
LATAM
اروپا
آسیا- اقیانوسیه
آفریقا
خوشه 0
خوشه1
خوشه 2
خوشه 3
خوشه 4
خوشه 5
خوشه 6
خوشه 7
خوشه 8
خوشه 9


تراکنش
10
8
6
4
2
0
USCA
LATAM
اروپا
آسیا- اقیانوسیه
آفریقا
خوشه 0
خوشه 1
خوشه 2
خوشه 3
خوشه 4
خوشه 5
خوشه 6


تعداد تراکنش
10
8
6
4
2
0
تکنولوژی
مبلمان
لوازم اداری
خوشه 0
خوشه1
خوشه 2
خوشه 3
خوشه 4
خوشه 5
خوشه 6
خوشه 7
خوشه 8
خوشه 9


رتبه
10
8
6
4
2
0
تکنولوژی
مبلمان
لوازم اداری
خوشه 0
خوشه 1
خوشه 2
خوشه 3
خوشه 4
خوشه 5
خوشه 6
خوشه 7
خوشه 8
خوشه 9



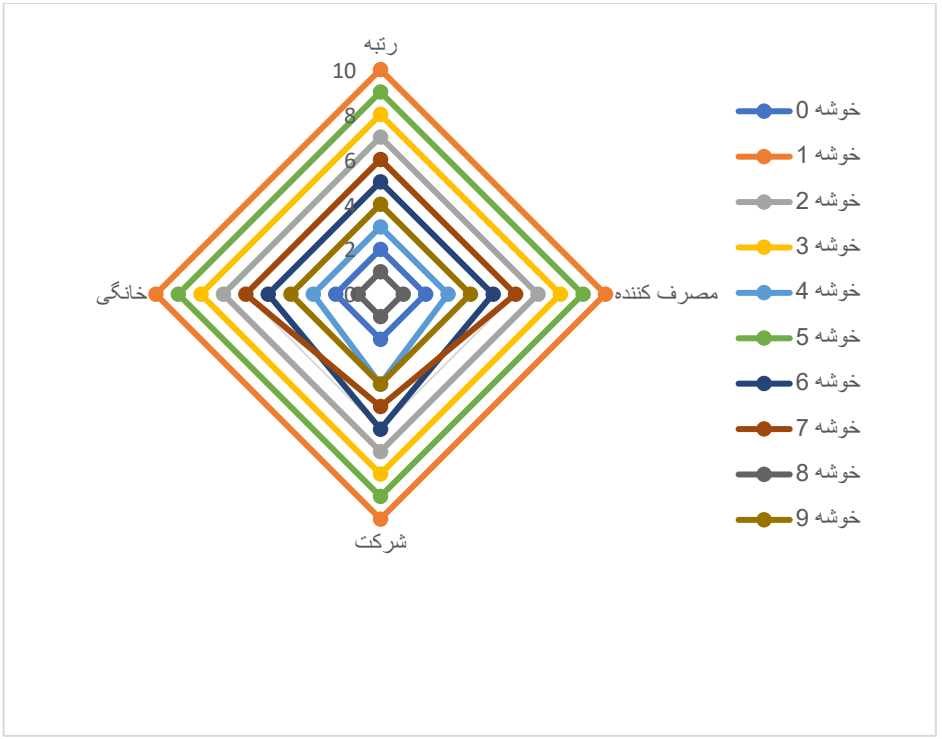


Figure 115 - Comparison of the results of step 2 and 3

The graphs look similar, but to be sure, the rank distance of similar clusters (in terms of cluster size) was calculated in both step 3 and 4 models and the result is 0.0090. This small number shows the closeness of the results, but in step 3, due to the dimension reduction process before clustering and removing the features that had little weight, there are more accurate results.

Finally, the results of both models are reliable and both are almost the same. Customer segmentation based on the results shows that customer clusters are located in a close distance in terms of average profitability. Also, there is a slight difference between different clusters in product categories and customer segments. for example, if a cluster is the second cluster in the office supplies category, it is most likely also the second cluster in the furniture category. This is different for markets, and clusters have different behaviors towards different markets, each of which was explained in detail in the results section of steps 2 and 3.

Attachments

Table 12- Guide to attached files

Description	Format	File name
Initial database	xlsx	global_superstore_2016
Database first sheet RapidMiner Input	csv	edited - global_superstore_2016-sheet 1
Database Second sheet RapidMiner Input	csv	edited - global_superstore_2016-sheet 2
Python source code to calculate the tendency of data to clustering	py	11
Edited Excel file suitable for Python	xlsx	Edited_global_superstore_2016
Centers of clusters and calculating the distance between the rank of clusters of steps 2 and 3	xlsx	Centroids