

# 9.1

## Segmenting Stores Using Clustering

Nitin Kalé, *University of Southern California*

Nancy Jones, *San Diego State University*

### OBJECTIVE

The objective of this exercise is to segment retail stores based on various attributes to help with sales promotions.

### ACTIVITIES

- Import and prepare data.
- Apply data mining algorithms.
- Configure predictive models.
- Create data visualizations.
- Analyze and interpret output from models.

### SOFTWARE PREREQUISITES

SAP Predictive Analytics 3.X

### DATA SET

Data file titled [Stores.csv](#)

---

## Scenario

The country manager of a retail chain (which has 150 stores) is finalizing plans for three sales promotion strategies. Data pertaining to the stores such as store location, sales turnover, store size, staff, and profit margin are stored in a CSV file. The manager wants to segment the 150 stores into three different groups based on sales turnover, profit margin, store size, and staff size so that specific strategies can be applied to each store segment. We will use clustering of retail stores data to assist the manager in developing promotion strategies.

---

# Cluster Analysis

Given a dataset, organizing it into meaningful groups is a basic and useful approach to data mining and data analysis. Clustering classifies samples into groups using a measure of association so that data points within a group are similar. Data points from different groups are not similar. Data points are multidimensional, that is they consist of several variables. Visualization is not practical for humans when datasets consist of more than three dimensions.

The input to a clustering exercise is a dataset and the number of clusters. The result of the analysis is a set of clusters. *K-means clustering* is a method of finding clusters and their centers (R) given a choice in the number of clusters (K). It is often used for market segmentation. The goal is to make the inter-cluster difference (distance) high and the intra-cluster difference (distance) low.

---

## 1. Build a segmentation analysis

1. Launch **SAP Predictive Analytics**.
2. Click **Expert Analytics** → **Expert Analytics**.
3. From the menu, choose **File** → **New**.
4. In the New Dataset window choose CSV. **Next**.
5. Search for the *Stores.csv* file provided to you.
6. Check to see if 150 rows of data have been acquired. **Create**.
7. In the **Prepare** tab, switch to the Facets view and explore the data.
8. You notice that four fields (Profit Margin, Sales Turnover, Staff Size and Store size) have been identified as Measures (which will be useful during visualization).
9. Switch to the **Predict** panel.
10. *Stores.csv* is the already added to the analysis as the data source.
11. From the **Algorithms** tab (on the right side, within Components panel), drag and drop or double click the *R-K-Means algorithm* into your analysis. See Figure 1.
12. The algorithm component is automatically connected to the data source component.
13. Hover over the R-K-Means algorithm and either click on the cog or choose **Configure Settings** (on the right).
14. In the R-K-Means properties dialog box, provide the necessary details:
  - a. In the Number of Clusters field, enter **3**.
  - b. Select all four columns to be used for cluster analysis.

- c. Retain the default values for the advanced properties.
  - d. Choose **Done**.
15. From the **Data Writers** tab, drag and drop or double click on the CSV Writer component.
16. **Configure Settings** of the CSV data writer.

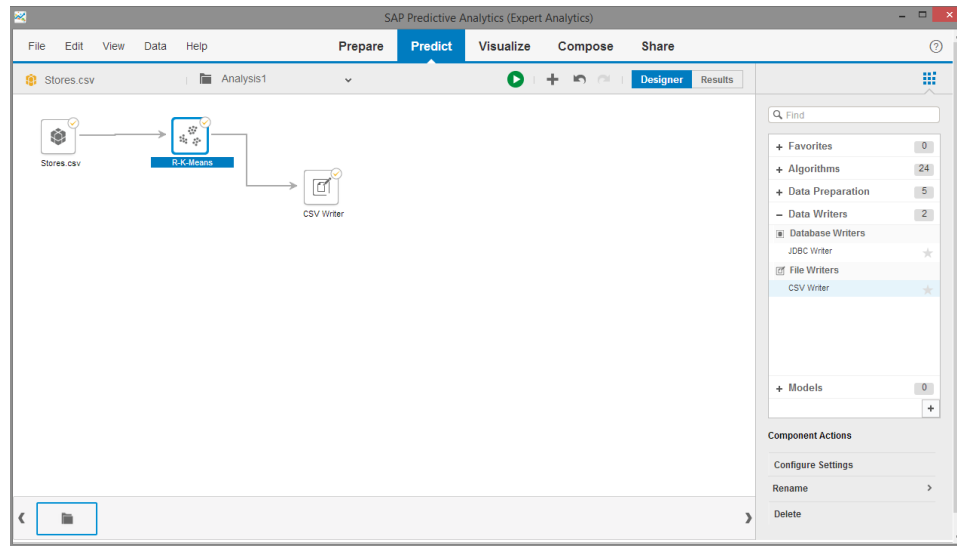


Figure 1

- a. In the CSV Writer Configure Settings, type a CSV file name to store the result (use Browse).
  - b. Chose **Done**.
17. Click to **Run** to run the analysis
18. You should receive a succeeded message. **OK**.
19. You are now in the **Results** Grid view. See Figure 2.
20. You see the name of the store, sales turnover, store size, staff size, profit margin and cluster number data. There should be three clusters numbered 1, 2 and 3.
21. Switch to the **Summary** view (Figure 3).

SAP Predictive Analytics (Expert Analytics)

File Edit View Data Help Prepare Predict Visualize Compose Share

Stores.csv Analysis1 Designer Results

ABC	Store	Sales Turn	Store Size	Staff Size	Profit Ma.	ClusterK..
	New York City	5.10	3.50	1.40	0.20	1
	Los Angeles	4.90	3.00	1.40	0.20	1
	Chicago	4.70	3.20	1.30	0.20	1
	Houston	4.60	3.10	1.50	0.20	1
	Philadelphia	5.00	3.60	1.40	0.20	1
	Phoenix	5.40	3.90	1.70	0.40	1
	San Antonio	4.60	3.40	1.40	0.30	1
	San Diego	5.00	3.40	1.50	0.20	1
	Dallas	4.40	2.90	1.40	0.20	1
	San Jose	4.90	3.10	1.50	0.10	1
	Jacksonville	5.40	3.70	1.50	0.20	1
	Indianapolis	4.80	3.40	1.60	0.20	1
	San Francisco	4.80	3.00	1.40	0.10	1
	Austin	4.30	3.00	1.10	0.10	1
	Columbus	5.80	4.00	1.20	0.20	1
	Fort Worth	5.70	4.40	1.50	0.40	1
	Charlotte	5.40	3.90	1.30	0.40	1
	Detroit	5.10	3.60	1.40	0.30	1
	El Paso	5.70	3.80	1.70	0.30	1
	Memphis	5.10	3.80	1.50	0.30	1
	Baltimore	5.40	3.40	1.70	0.20	1
	Boston	5.10	3.70	1.50	0.40	1
	Seattle	4.60	3.60	1.00	0.20	1
	Washington	5.10	3.30	1.70	0.50	1
	Nashville	4.80	3.40	1.90	0.20	1
	Denver	5.00	3.00	1.60	0.20	1
	Louisville	5.00	3.40	1.60	0.40	1

Stores.csv Showing: 150/150 Rows - 5/5 Columns Never Refreshed

Figure 2: Results Grid View

SAP Predictive Analytics (Expert Analytics)

File Edit View Data Help Prepare Predict Visualize Compose Share

Stores.csv Analysis1 Designer Results

### Algorithm Summary

Summary of the model from R Scripts

Information of the columns used in the algorithm

Independent Columns

Sales Turnover : Double

Store Size : Double

Summary of the Model

	Length	Class	Mode
cluster	150	-none-	numeric
centers	6	-none-	numeric
tot##	1	-none-	numeric
withinss	3	-none-	numeric
tot.withinss	1	-none-	numeric
betweenss	1	-none-	numeric
size	3	-none-	numeric
iter	1	-none-	numeric
ifault	1	-none-	numeric

Centers

	Sales_Turnover	Store_Size
1	5.006000	3.412000
2	6.812766	3.074468
3	5.773555	2.692453

Within cluster sum of squares

[1] 13.2020 12.6217 11.3000

The size of each cluster

[1] 50 47 53

Figure 3: Results Summary

22. You can see the *center coordinates* of the three clusters; also the size of each cluster which is the number of stores in each cluster.

## 2. Results visualization and interpretation

1. In the **Cluster Representations** pane, select Cluster Distribution.
  - a. You see a chart of cluster size vs cluster number, (Figure 4). These are the number of stores in each cluster. You can roll over the bars to see the number.

- b. Stores within a cluster are similar to each other and dissimilar to stores in other clusters.

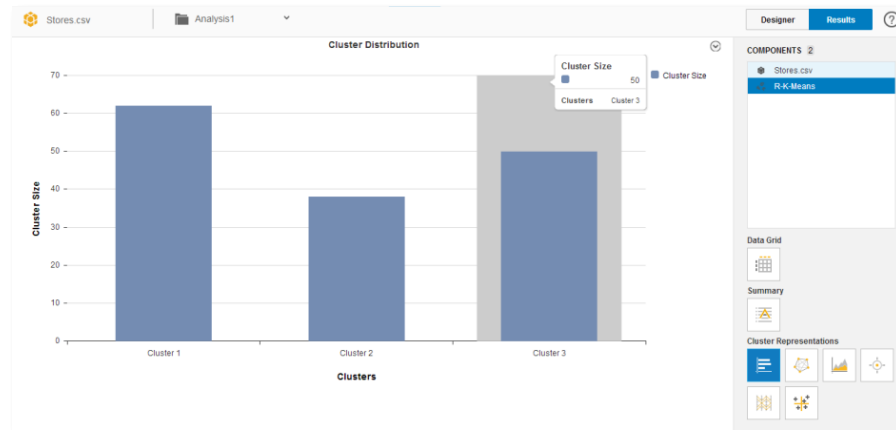


Figure 4: Cluster Distribution

2. In the **Cluster Representations** pane, select Cluster Density and Distance.
  - a. You see that cluster 1 has the lowest/weakest density and cluster 3 has the highest. Low density clusters imply clusters of noise, outliers, or other loosely associated data. The distance shows how dissimilar the clusters are.

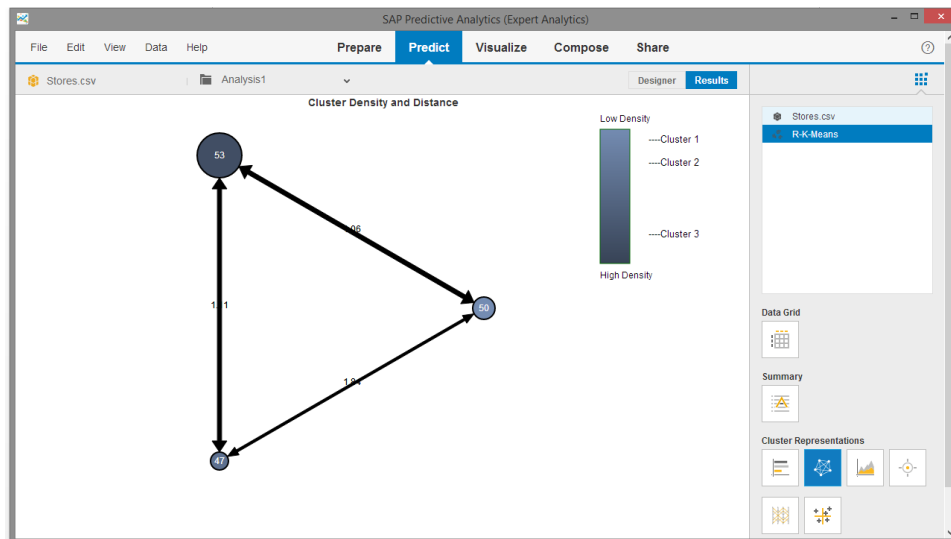


Figure 5: Cluster Density

3. In the **Cluster Representations** pane, select Feature Distribution.
  - a. The graph lets you compare the distribution of the variable in a particular cluster against the entire dataset. You can change the Measure being displayed and the cluster number in the Data panel on the right side.

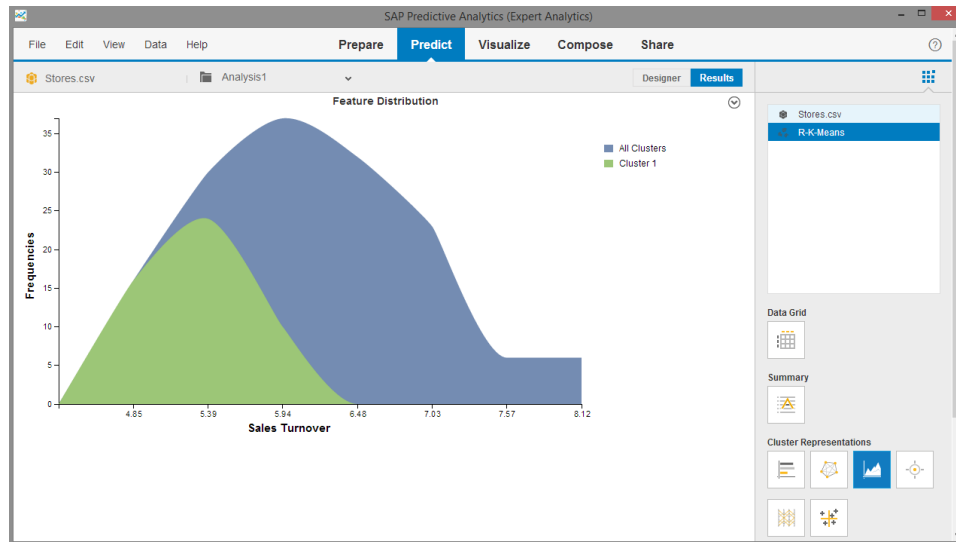


Figure 6: Feature Distribution

4. In the **Cluster Representations** pane, select Cluster Center Representation.
  - a. You see a radar chart of the cluster centers (radar axes are the variables); you can change the cluster number in the Data panel.

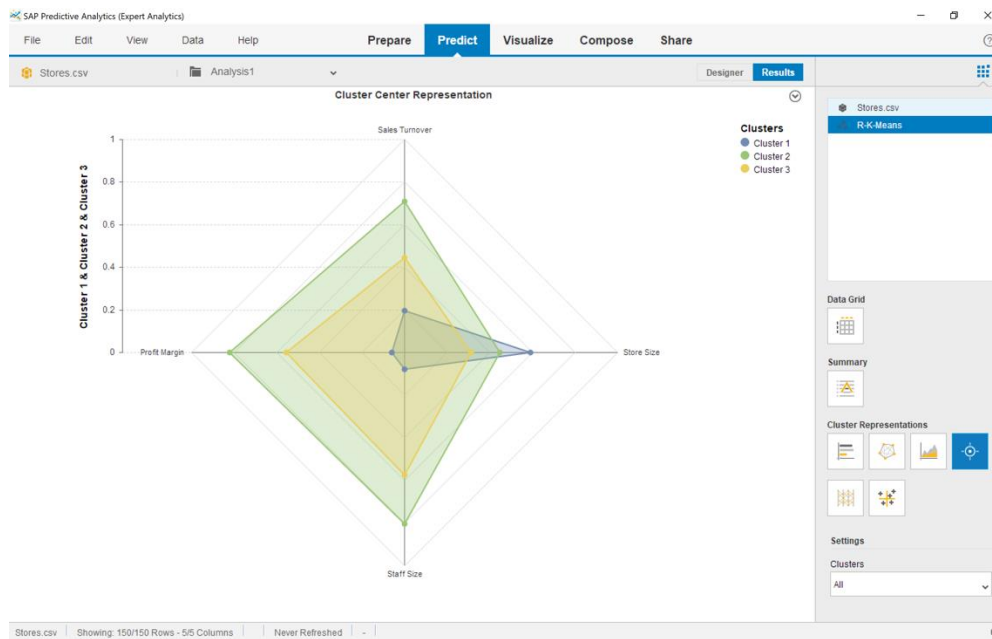


Figure 7: Radar Chart of Clusters

5. In the **Cluster Representations** pane, select Parallel Coordinate Chart.
  - a. The axes are all normalized. Parallel lines between the axes imply a positive relationship between the two dimensions. Intersecting lines imply a negative relationship.



Figure 8: Parallel Coordinates Chart

6. In the **Cluster Representations** pane, select Scatter Matrix Charts.
  - a. You see the scatter charts of store clusters plotted between various pairs of dimensions

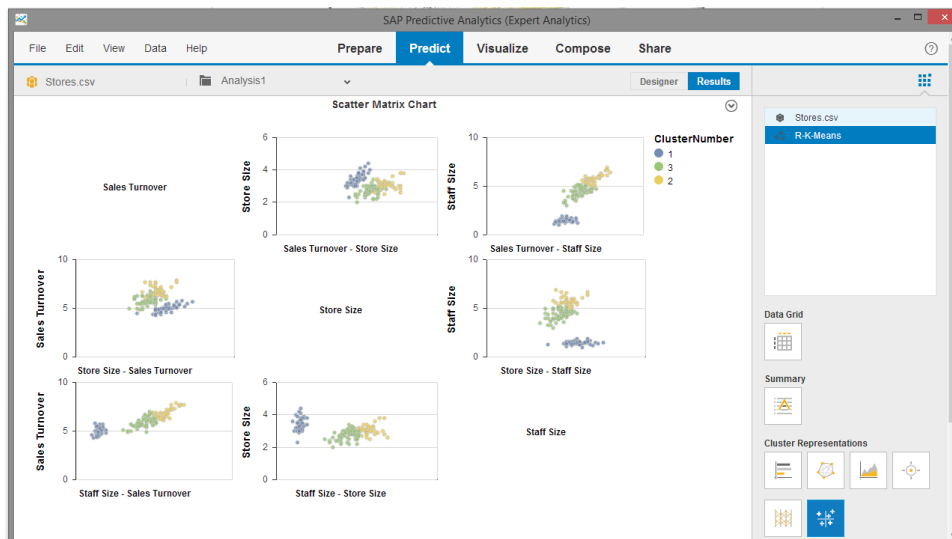


Figure 9: Scatter Plots

7. The fitted and forecast results are stored in the csv file. You can open the saved csv file and explore the three clusters that have been generated.
8. From the **File** menu, select **Save**.
9. Enter a name for the document.
10. Choose **Save**.

**Question 1: Which cluster has the most number of stores?**

**Question 2: List the name of one store in each cluster.**

**Question 3: What can the manager do with these segmentation results?**

---

## Challenge Activity 1

Choose one cluster to analyze further using visualizations. Provide a detailed description/analysis of the stores within the cluster you have chosen. Based on what you see in this cluster, what kind of marketing strategy to improve sales for the stores in the cluster do you recommend?

*Hint: On the Visualize tab of Predictive Analytics, Select Analysis 1 (R-K Means).*

---

## Challenge Activity 2

Using the Visualize tab in Predictive Analytics, compare and contrast the attributes of the stores within the three clusters.

---