CHAPTER 11: EXERCISE 2

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
- Publish results

SOFTWARE PREREQUISITES

- SAP Predictive Analytics 2.2

UCC PRODUCTS REQUIRED

- None

DATA SET

- Data file titled *stores_E11_2.xlsx*

SCENARIO

The country manager of a retail chain (which has 150 stores) is finalizing plans for three sales promotion strategies. Data pertaining to 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.

To build an analysis for segmentation analysis, proceed as follows:

1. Open the csv file *stores_E11_2.csv* (in Microsoft Excel) and explore its contents.
2. Close Excel

3. Launch **SAP Predictive Analytics**.
4. From the menu, choose **File → New**.
5. In the New Dataset window choose **CSV**. Next
6. Search for the *stores_E11_2.csv* file provided to you
7. Check to see if 150 rows of data have been acquired. Create.
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_E11_2.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.

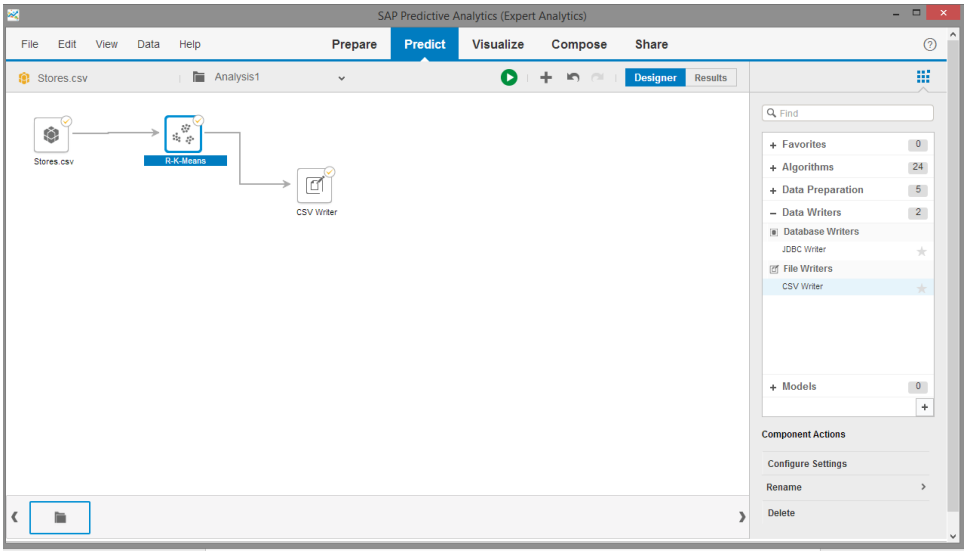

Figure 1

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

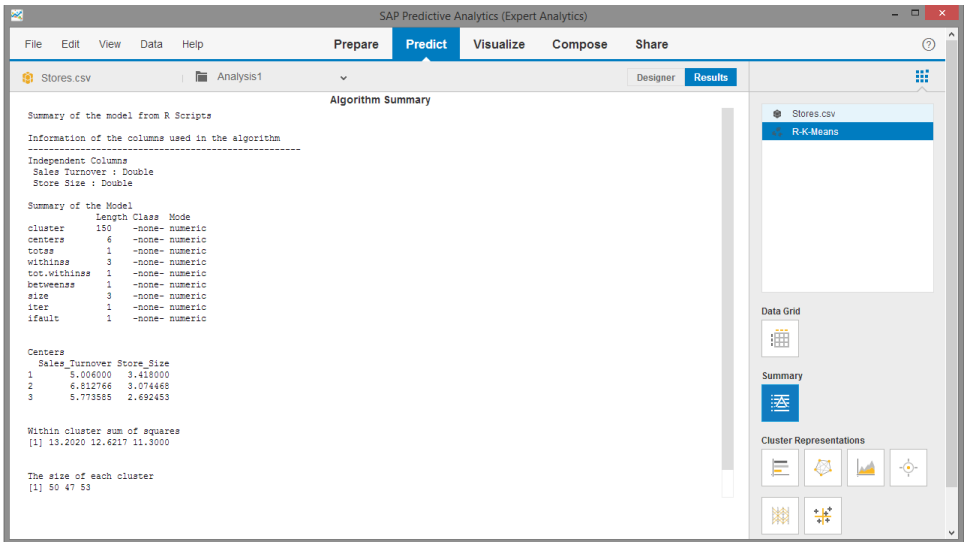

Figure 2: Cluster Results

24. 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.

25. Results visualization and interpretation

a. In the **Cluster Representations** pane, select Cluster Distribution.

i. You see a chart of cluster size vs cluster number, (Figure 3). These are the number of stores in each cluster. You can roll over the bars to see the number.

ii. Stores within a cluster are similar to each other and dissimilar to all other stores in other clusters.

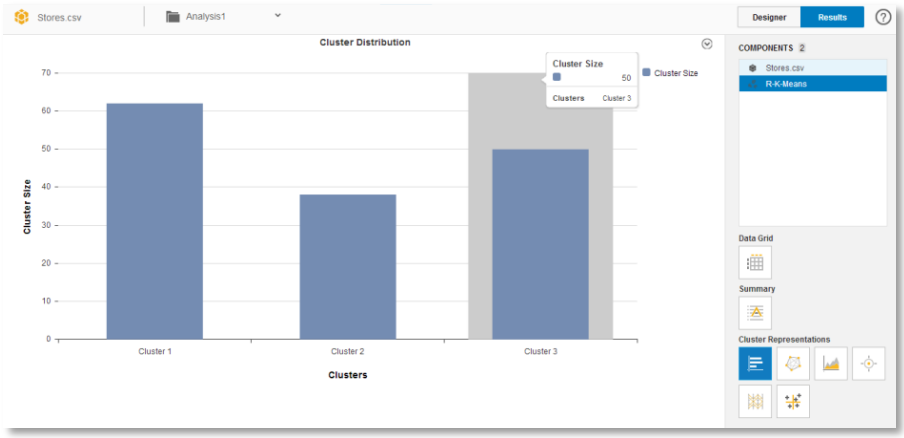

Figure 3: Cluster Distribution

b. In the **Cluster Representations** pane, select Cluster Density and Distance.

i. 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.

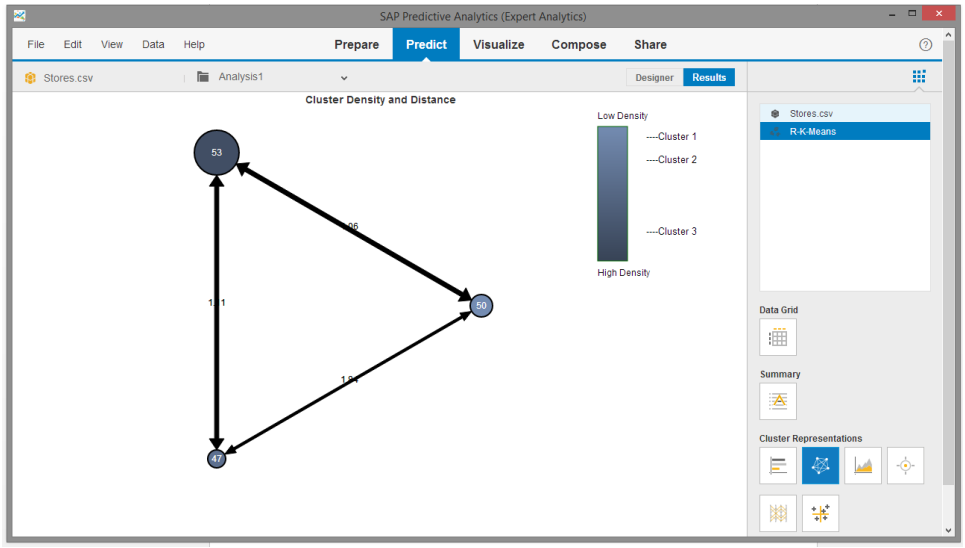

Figure 4: Cluster Density

b. In the **Cluster Representations** pane, select Feature Distribution.

   i. 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.

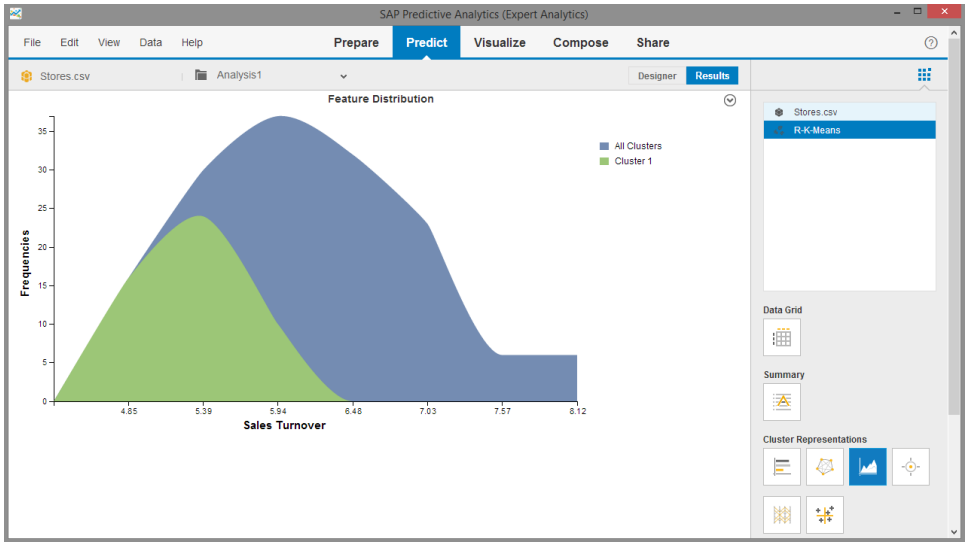

Figure 5: Feature Distribution

c. In the **Cluster Representations** pane, select Cluster Center Representation.

   i. You see a radar chart of the cluster centers (radar axes are the variables); you can change the cluster number in the Data panel

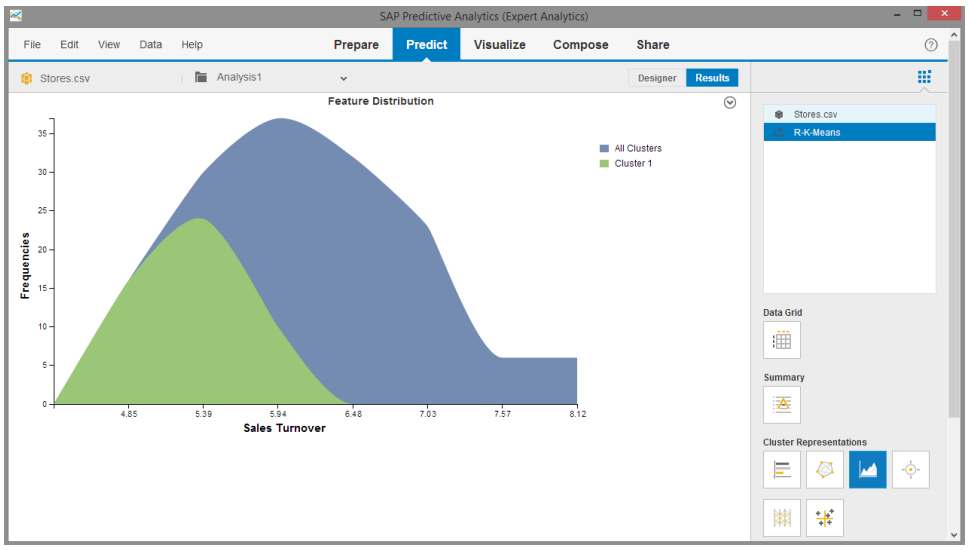

Figure 6: Radar Chart of Clusters

c. In the **Cluster Representations** pane, select Parallel Coordinate Chart.

   i. The axes are all normalized. Parallel lines between the axes imply a positive relationship between the two dimensions. Intersecting lines imply a negative relationship.

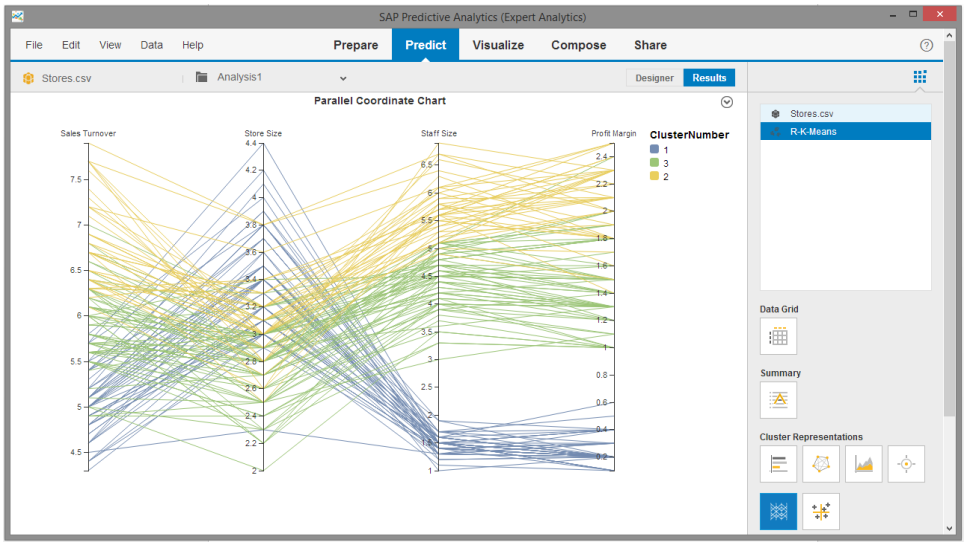

Figure 7: Parallel Coordinates Chart

d. In the **Cluster Representations** pane, select Scatter Matrix Charts.

   i. You see the scatter charts of store clusters plotted between various pairs of dimensions

Figure 8: Scatter Plots

26. 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.

27. From the **File** menu, select **Save**.

28. Enter a name for the document.

29. 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?