# 

# HW 2. Cluster, DUNGAREE, PHARMACEUTICALS)

**CLUSTER ANALYSIS IN SAS ENTERPRISE MINER**

## Based on material from SAS Education

WELCOME to this SAS Enterprise Miner tutorial.

This handout introduces you to **SASEM Cluster**.

**Objectives**

* Continue to gain experience with SAS Enterprise Miner
* Learn to perform basic cluster analysis in SAS Enterprise Miner

**Instructions**

1. Submit one report per group through the UNT online learning management system
2. Clearly identify your group number and all group member names on the cover page
3. A professional quality report is expected – messy or hard-to-read reports will be penalized
4. Document your actions within SAS EM using screenshots
5. Explain your answers as clearly as possible – vague answers will be penalized

**Datasets**

* Exercise 1: Clustering Stores – dungaree.sas7bdat
* Exercise 2: Clustering Pharmaceutical Firms – Pharmaceuticals.xlsx

**Exercise 1: Clustering Stores**

**Assignment Details:**

The DUNGAREE data set gives the number of pairs of four different types of dungarees sold at stores over a specific time period. Each row represents an individual store. There are six columns in the data set. One column is the store identification number, and the remaining columns contain the number of pairs of each type of jeans sold.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Role** | **Level** | **Description** |
| STOREID | ID | Nominal | Identification number of the store |
| FASHION | Input | Interval | Number of pairs of fashion jeans sold at the store |
| LEISURE | Input | Interval | Number of pairs of leisure jeans sold at the store |
| STRETCH | Input | Interval | Number of pairs of stretch jeans sold at the store |
| ORIGINAL | Input | Interval | Number of pairs of original jeans sold at the store |
| SALESTOT | Rejected | Interval | Total number of pairs of jeans sold (the sum of FASHION, LEISURE, STRETCH, and ORIGINAL) |

Table 1. Variable Settings and Description

1. Open SAS Enterprise Miner Workstation
2. Create a new project. Name it Assignment **HW 2** and save it to your H drive or USB drive as appropriate.
3. Create a new diagram. Name it **Jeans**.
4. Import the **dungaree.sas7bdat** file into SAS Enterprise Miner
   1. Run the project startup code as discussed in class.
   2. Right click on Data Source and select Create Data Source
   3. Select SAS Table as the source and click next
   4. Click browse and select the Dungaree data set and click next
   5. Review table properties and click next
   6. Select Basic Metadata Advisor Options and click next
   7. Click next at the column metadata
   8. Select no sample data set and click next
   9. Set the role of the data set to Raw and click next
   10. Click finish
5. Drag the new **Dungaree** data source to the canvas
6. Explore the data. Determine whether the model roles and measurement levels assigned to the variables are appropriate. Examine the distribution of the variables.
   1. Are there any unusual data values?
   2. Are there any missing values that should be replaced?
7. Assign the variable STOREID the model role ID and the variable SALESTOT the model role Rejected. Make sure the remaining variables have the Input model role and the Interval measurement level. Why should the variable SALESTOT be rejected?
8. Add a Cluster node to the diagram workspace and connect it to the data source node.
9. Select the Cluster node and select Internal Standardization 🡪 Standardization. What would happen if you did not standardize your inputs?
10. Run the diagram from the Cluster node and examine the results. Does the number of clusters created seem reasonable?
11. Specify a maximum of six clusters and rerun the Cluster node. How does the number and quality of clusters compare to that previously obtained?
12. Use the Segment Profile node to summarize the nature of the clusters.

**Exercise 2: Clustering Pharmaceutical Firms**

**Assignment Details:**

1. Open SAS Enterprise Miner Workstation Open Project **HW 2**
2. Create a new diagram. Name it **PharmaIndustry**.

An equities analyst is studying the pharmaceutical industry and would like your help in exploring and understanding the financial data collected by her firm. Her main objective is to understand the structure of the pharmaceutical industry using some basic financial measures.

Financial data gathered on 21 firms in the pharmaceutical industry are available in the file **Pharmaceuticals.xls.**

For each firm, the following variables are recorded.

* 1. Market capitalization (in billions of dollars) (Mcap)
  2. Beta (Beta)
  3. Price/earnings ratio (PER)
  4. Return on equity (ROE)
  5. Return on assets (ROA)
  6. Asset turnover (AT)
  7. Leverage (Leverage)
  8. Estimated revenue growth (ERG)
  9. Net profit margin (NPM)
  10. Median recommendations (across major brokerages) (MRec)
  11. Location of firms headquarters (Location)
  12. Stock exchange on which the firm is listed (Exchange)

Use cluster analysis to explore and analyze the given dataset as follows (you need to use the File Import node as explained in Homework 1)

1. Use only the quantitative variables (1-9) to cluster the 21 firms. Use the default settings in SAS Enterprise Miner.
2. Interpret the clusters with respect to the quantitative variables that were used in forming the clusters.
3. Is there a pattern in the clusters with respect to the qualitative variables (10-12) (those not used in forming the clusters)?
4. Provide an appropriate name for each cluster using any or all of the variables in the dataset. Don’t describe the cluster, name it.
5. Do the clusters formed seem reasonable? Try different numbers of clusters and examine the results. Feel free to experiment with other criteria as needed. Explain the reasons for your selections, and identify the best clustering in your opinion (justify).