Assignment #8 – SAS #3: Clustering and Segmentation

**NAME: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

***Instructions: Follow the steps and answer the questions below. Then email this document to your instructor.***

1. **Conducting Cluster Analysis**

The **AAEM.DUNGAREE** data set gives the number of pairs of four different types of dungarees (jeans) 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.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Model Role** | **Measurement Level** | **Description** |
| **STOREID** | Input | 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** | Input | Interval | Total number of pairs of jeans sold (the sum of FASHION, LEISURE, STRETCH, and ORIGINAL) |

You are going to cluster the data in order to see which groups of stores tend to sell certain types of jeans.

* 1. Create a new diagram in your project. Name the diagram **Jeans**.
  2. Define the data set **AAEM.DUNGAREE** as a data source. Accept the defaults.
  3. Drag the **DUNGAREE** data source to the Jeans diagram you just created.
  4. Determine whether the model roles and measurement levels assigned to the variables are appropriate. **Make sure STOREID’s Measurement Level is set to “Nominal.”**

Explore the data (right-click on the data source in the diagram and select **Edit Variables**, then click **Explore**) and examine the distribution of the variables.

* How do the SALESTOT and STOREID distributions differ from the other variables’ distributions (look at the histograms of each one)?

**ANSWER:**

* 1. Assign **STOREID** a model role of **ID** and **SALESTOT** a model role of **Rejected**. Make sure that the remaining variables have the Input model role and the Interval measurement level. Based on the variable descriptions on page 1 and your answer to part (d), why do you think that the variable **SALESTOT** should be rejected?   
       
     **ANSWER:**

* 1. Add a Cluster node to the diagram workspace and connect it to the Input Data node.
  2. Select the **Cluster** node and select **Internal Standardization** ⇨ **Standardization**. Why is it important to standardize your inputs? (hint: look at the range of the scales on the X axis of the histograms)   
       
     **ANSWER:**

* 1. Run the Cluster node and examine the results.  
       
     How many clusters are created?

**ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

What might be a problem with having so many clusters?   
  
**ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

What is the highest root mean squared standard deviation among the clusters?

Two hints and a waypoint:

* Look at the Mean Statistics window.
* **The root mean squared standard deviation means basically the same thing as the sum of squares error.**
* **Waypoint:** The lowestnon-missing root mean squared standard deviation should be .341708.  
  **ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Specify a maximum of six clusters and rerun the Cluster node.  
  
Is the root mean squared standard deviation generally **higher** or **lower** than the previous set of clusters in **(h)**?   
  
**ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
  
What does this mean? (hint: refer back to the slides regarding sum of squares error – remember that room mean squared standard deviation means the same thing as sum of squares error)

**ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Despite this, why might you prefer this set of six clusters instead of the other grouping?

**ANSWER:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

* 1. Use the Segment Profile node to summarize the nature of the clusters.

**Include a screenshot of the maximized “Profile: \_SEGMENT\_” window (similar to page 33 of the In-Class Exercise).**

Based on that diagram, how would you describe the sales of “Original” jeans in each of the six segments to the top management at the company (i.e., how do the products sold at each store segment compare to the average)?  
  
Two hints:

* Make sure you pay attention to the segment numbers – they aren’t in order!
* The X axis on each graph is the relative frequency of pairs sold

**ANSWER (put an X in the box that characterizes each segment):**

|  |  |  |
| --- | --- | --- |
|  | Lower than average | Higher than average |
| Segment 1 |  |  |
| Segment 2 |  |  |
| Segment 3 |  |  |
| Segment 4 |  |  |
| Segment 5 |  |  |
| Segment 6 |  |  |