**Big Data and Predictive Analysis**

Lab 1

Assignment 2

**Cluster Analysis**

Submitted by

1. **Bank Direct Marketing Cluster Analysis**

Given the inputs, do clusters of customers exist in the bank direct marketing data set? This exercise explores the bank direct marketing data and tries to profile the resulting clusters.

Take screenshots at every stage, you might want to recheck them or paste them for several questions:

1. Create a new diagram in your project. Name the diagram **Bank Clustering**.
2. Use the **bank\_direct\_marketing** data as a data source for this clustering and profiling exercise.
3. Determine whether the model roles and measurement levels assigned to the variables are appropriate.
4. Examine the distribution of the values of these variables:

* **balance**
* **day**
* **previous**
* **duration**
* **age**
* **campaign**

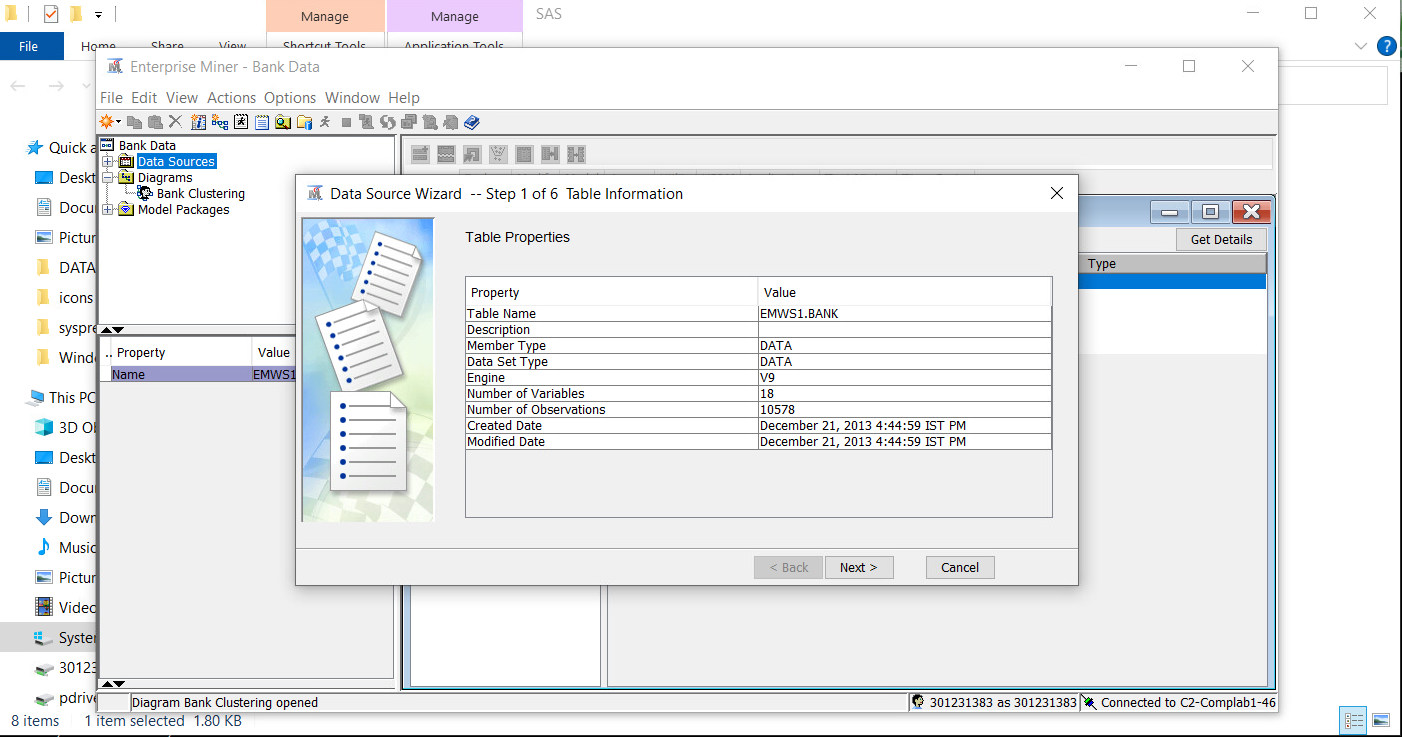
The three most heavily skewed distributions are for **balance**, **campaign**, and **previous**. Although not optimal, we could reduce the skewness of the distributions by taking the log of the variable.

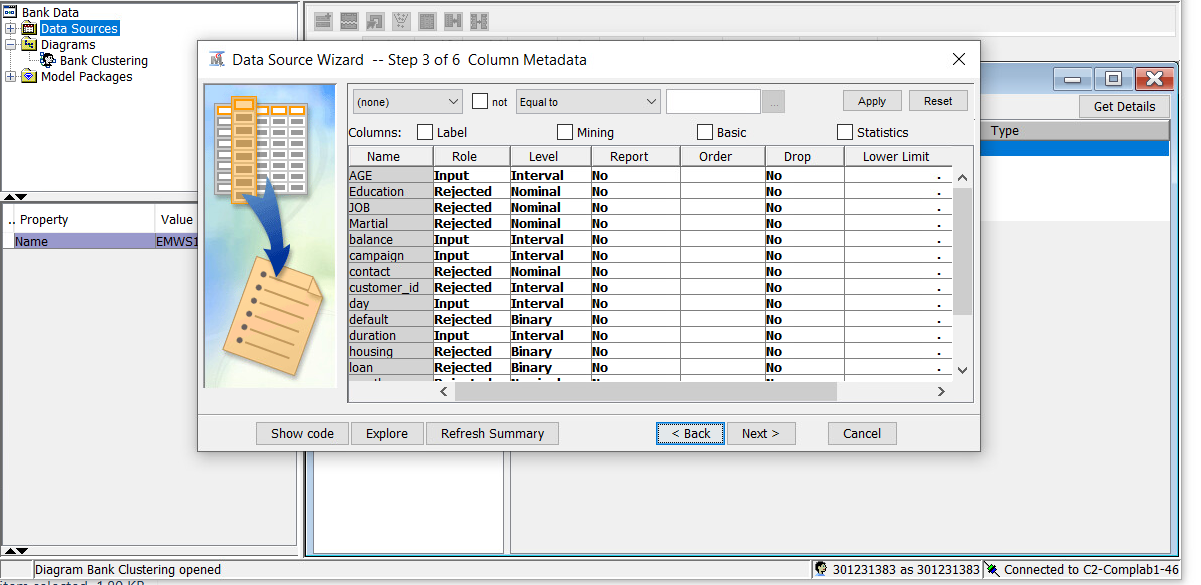
1. Drag a **Transform Variables** node onto the diagram and connect it to the **Input Data** source.
2. Apply a log transformation to the following variables:

* **balance**
* **previous**
* **campaign**

1. Connect a **Cluster** node to the **Transform Variables** node.
2. Change **Maximum Number of Clusters** to **6**.
3. Change **Use** for all the variables to **No**, except for these:

* **balance**
* **previous**
* **duration**
* **Age**
* **campaign**

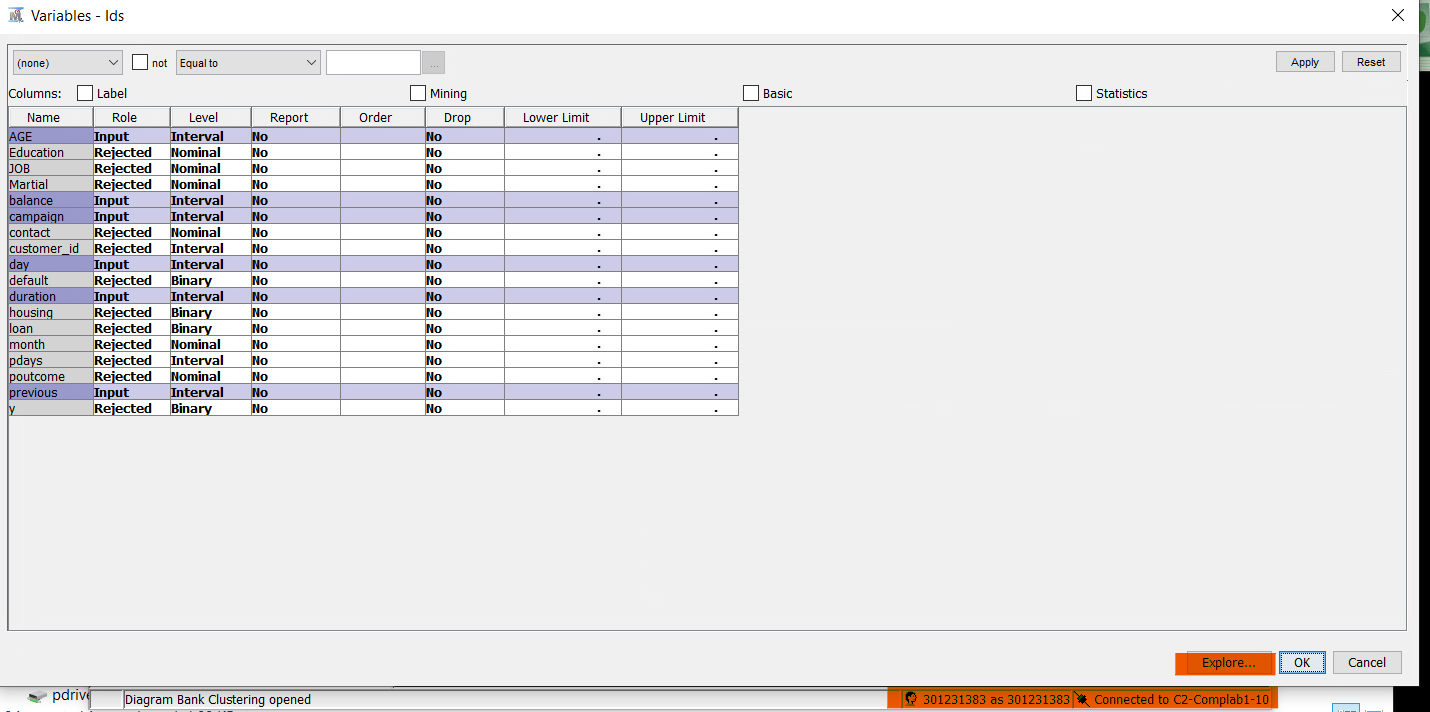


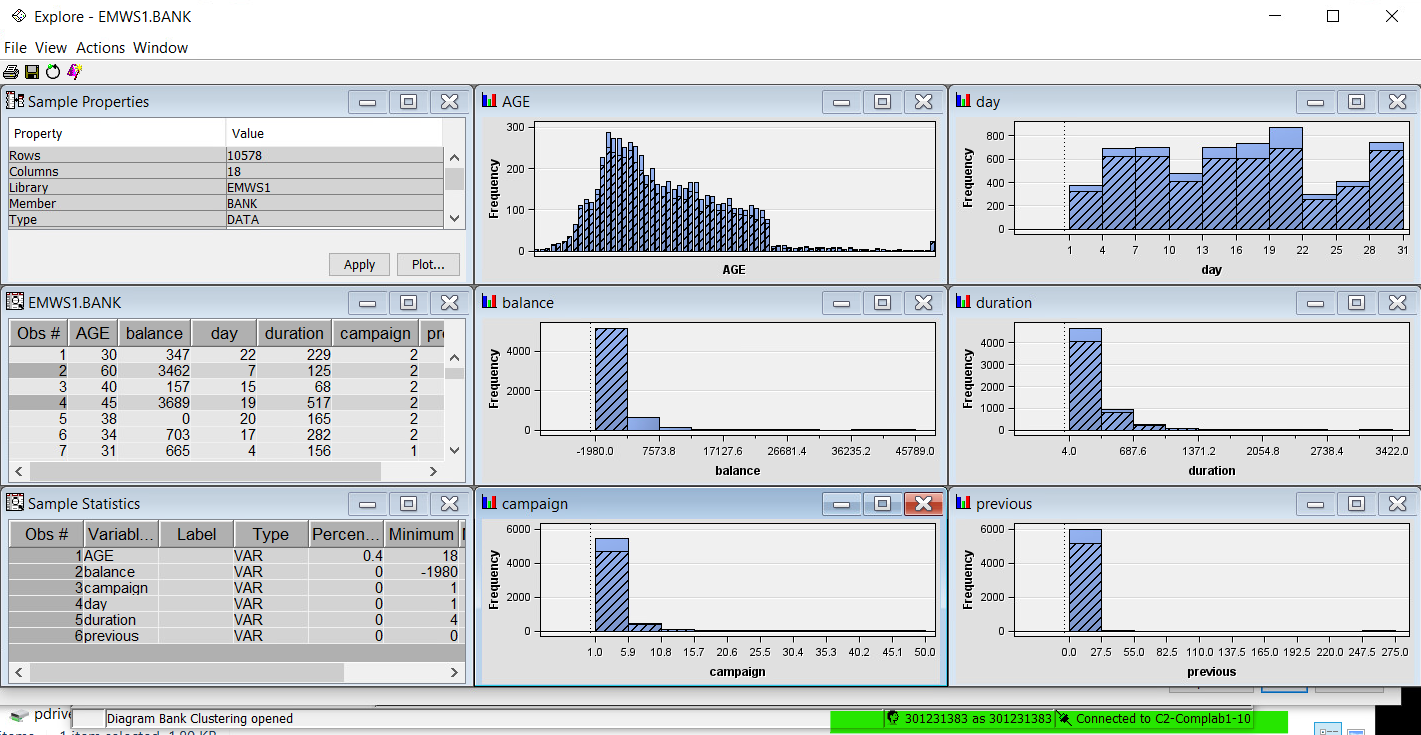


d. Examine the distribution of the values of these variables:

* **balance**
* **day**
* **previous**
* **duration**
* **age**
* **campaign**

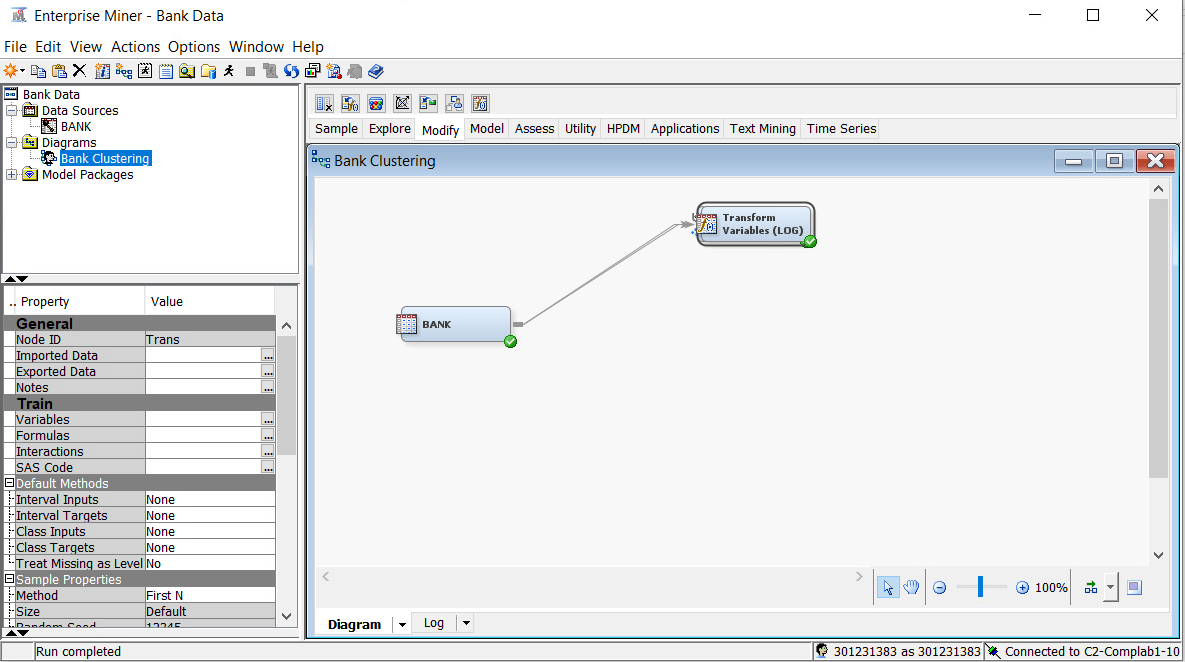
The three most heavily skewed distributions are for **balance**, **campaign**, and **previous**. Although not optimal, we could reduce the skewness of the distributions by taking the log of the variable.





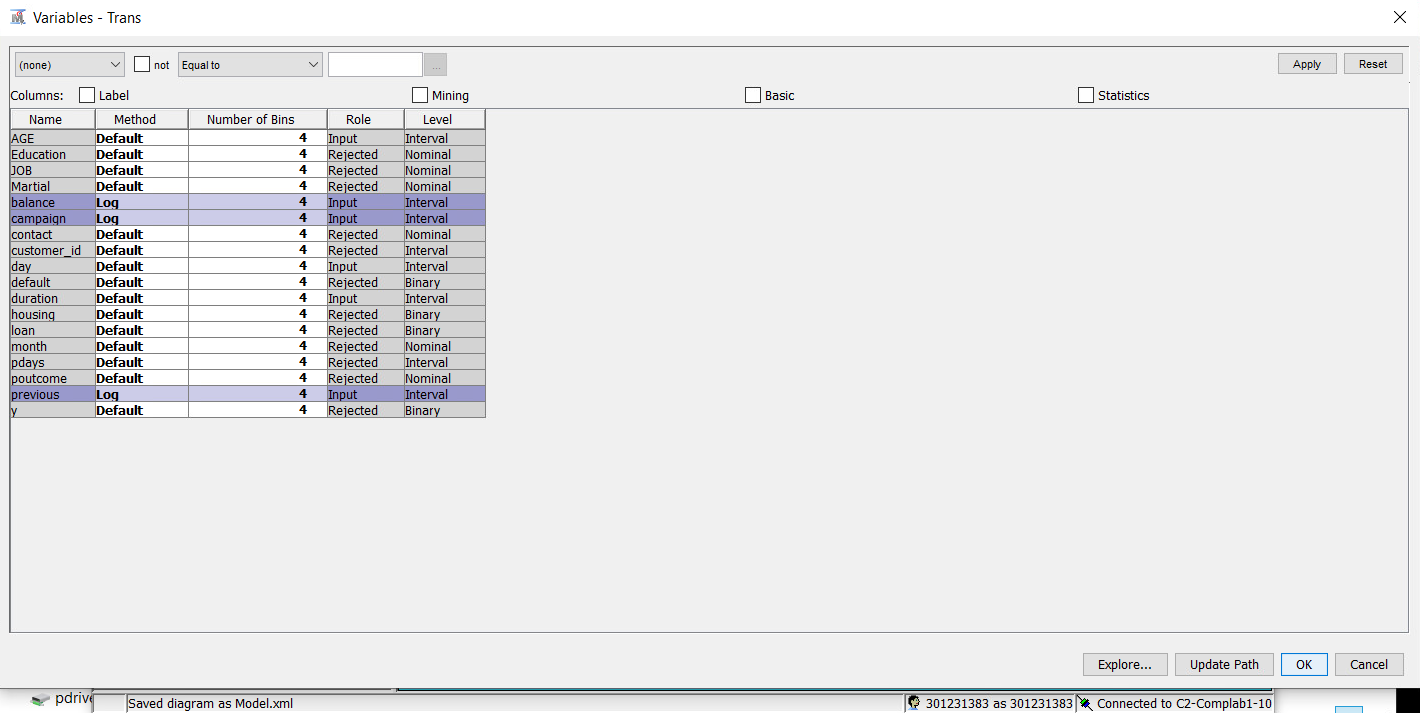
As we see the above graphs, the variable inputs for **balance, campaign and previous** are more skewed to the left.

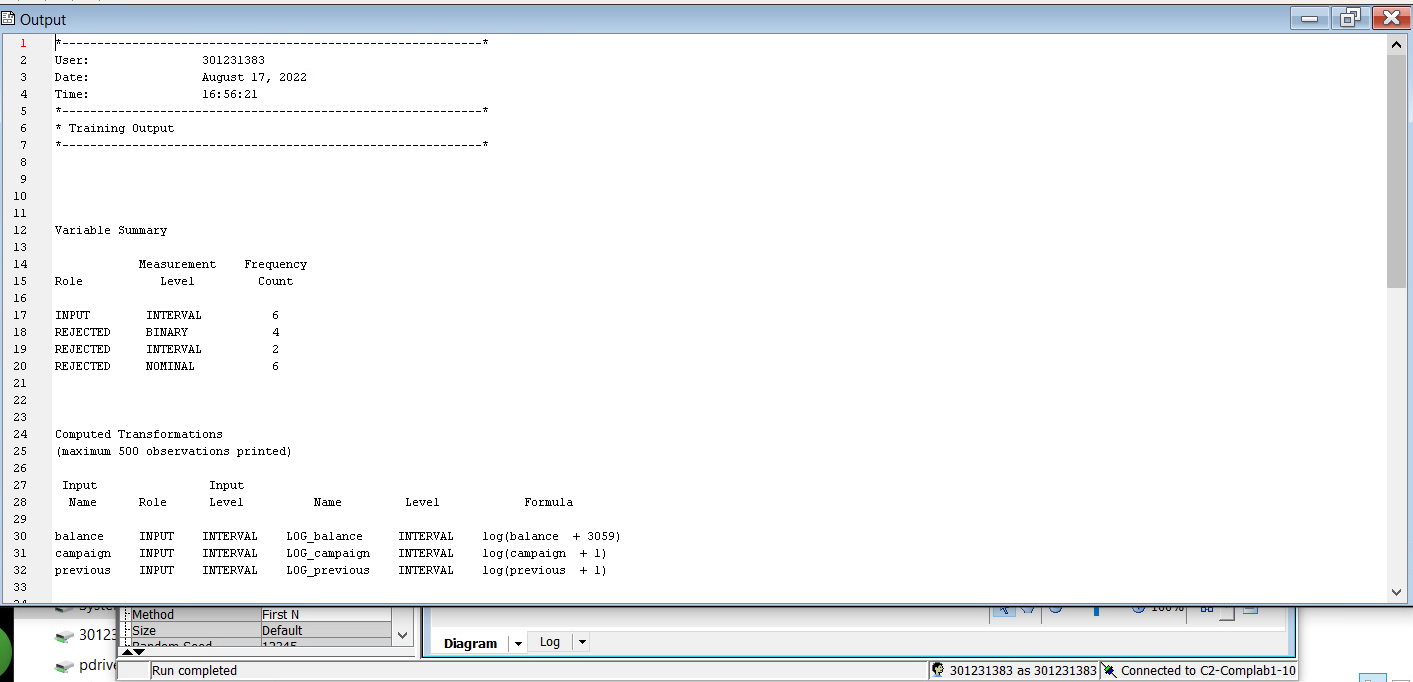
1. Drag a **Transform Variables** node onto the diagram and connect it to the **Input Data** source.



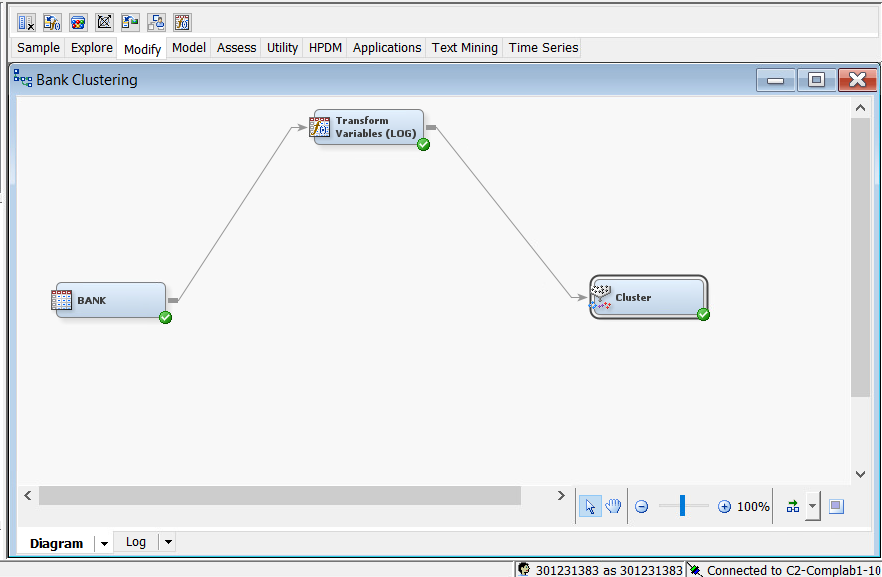
1. Apply a log transformation to the following variables:

* **balance**
* **previous**
* **campaign**

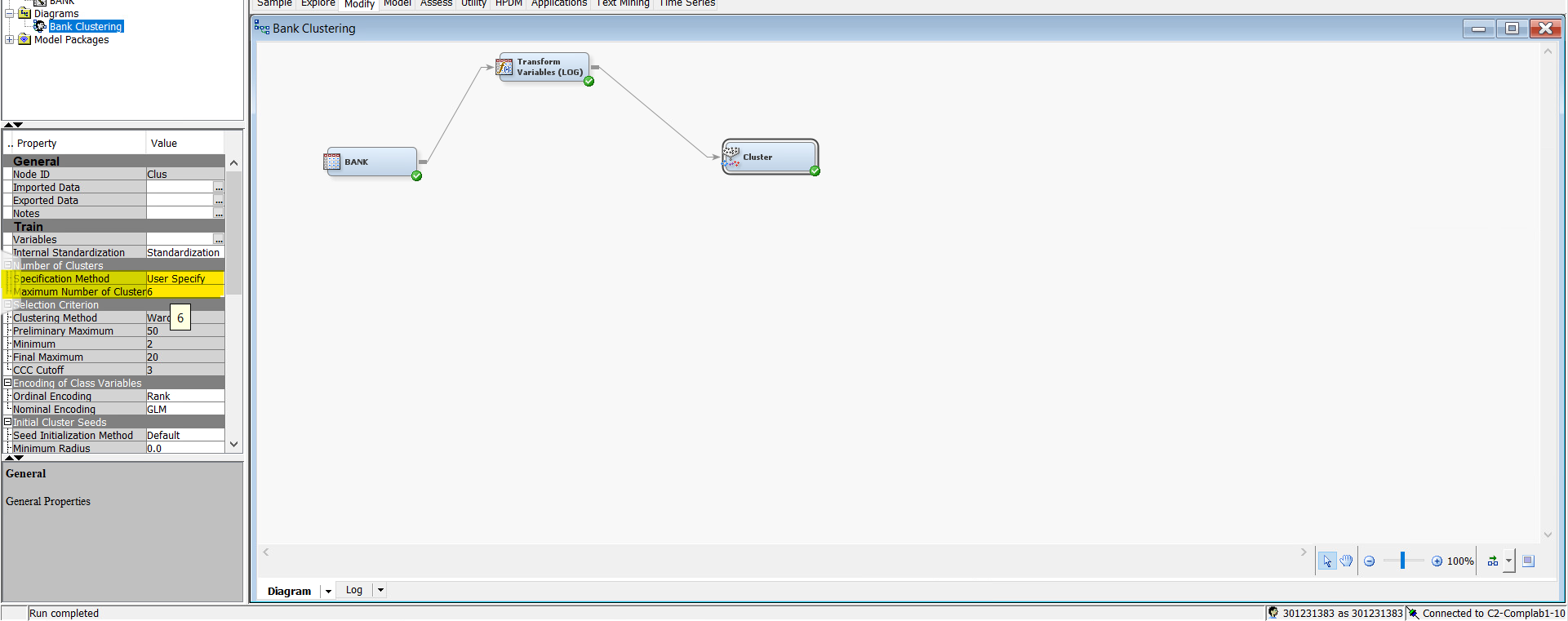




1. Connect a **Cluster** node to the **Transform Variables** node.

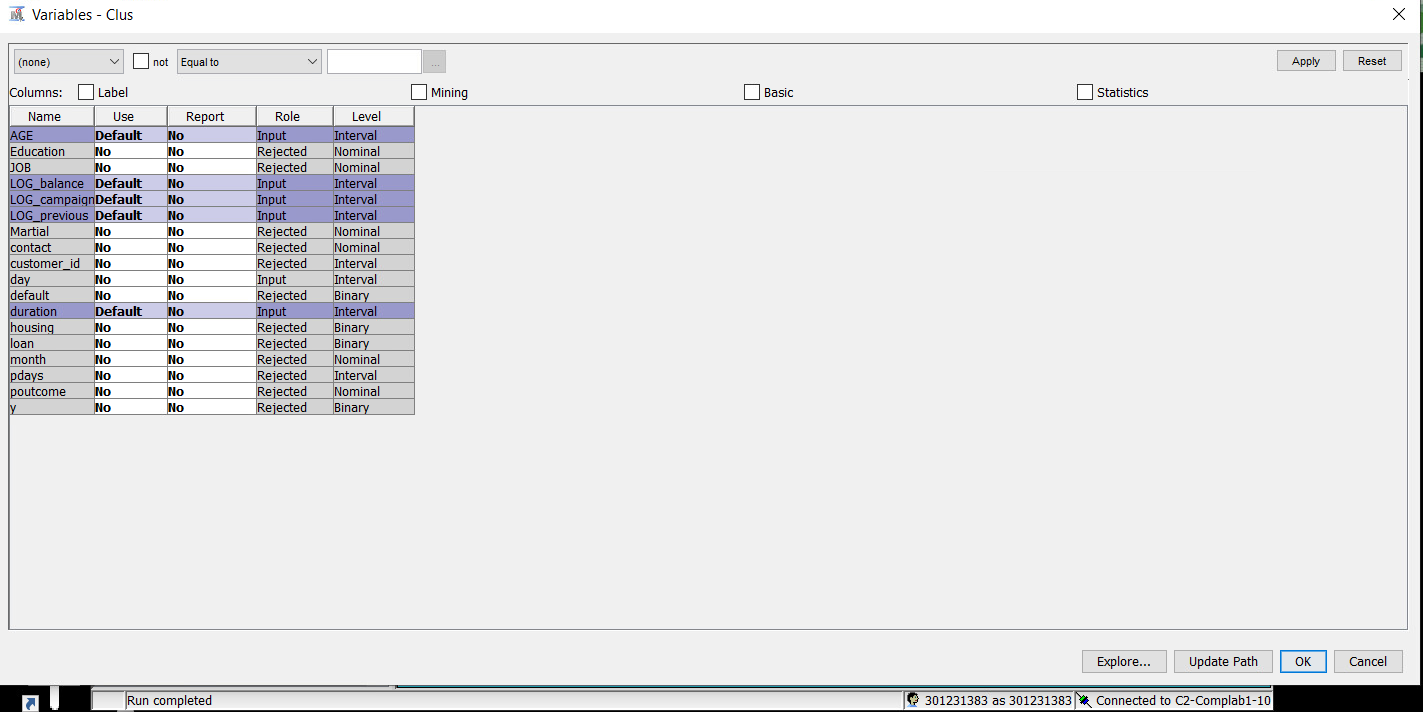


1. Change **Maximum Number of Clusters** to **6**.



1. Change **Use** for all the variables to **No**, except for these:

* **balance**
* **previous**
* **duration**
* **Age**
* **campaign**



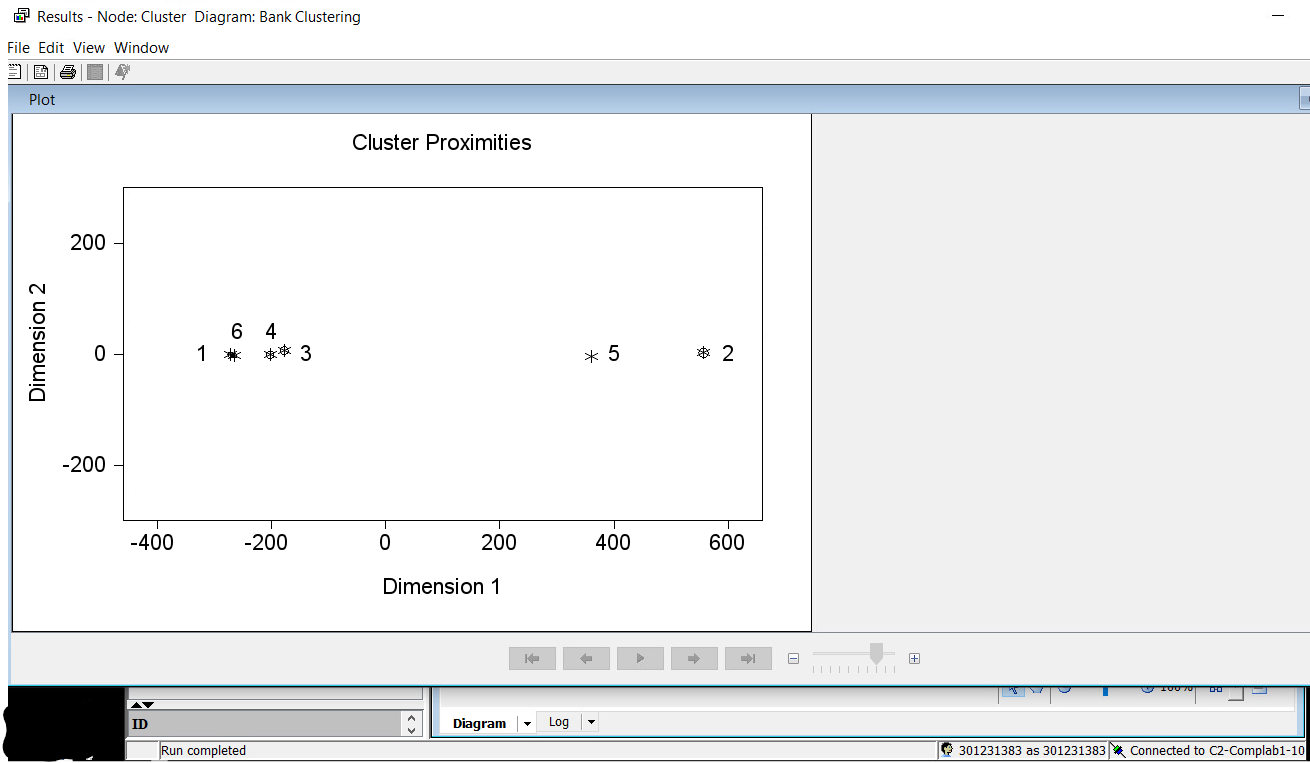
1. Run the Cluster node and view the plot of cluster proximities

SIX CLUSTERS

How many distinct clusters do you observe?

A limitation on the number of clusters is one of the clustering's presumptions, so we set the maximum number at six. According to the statistics output, the segment ID contains six different clusters. This is also evident in the plot of cluster proximities below.

Cluster Proximities Plot: (**View** ⇨ **Cluster Distance** ⇨ **Plot**).

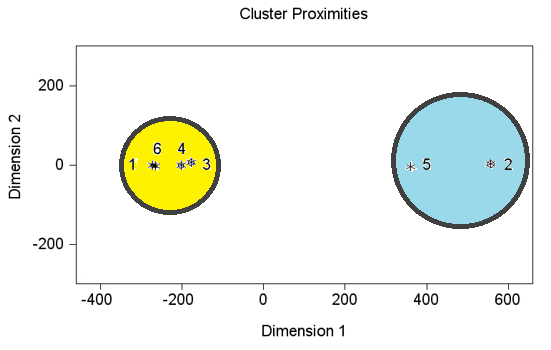


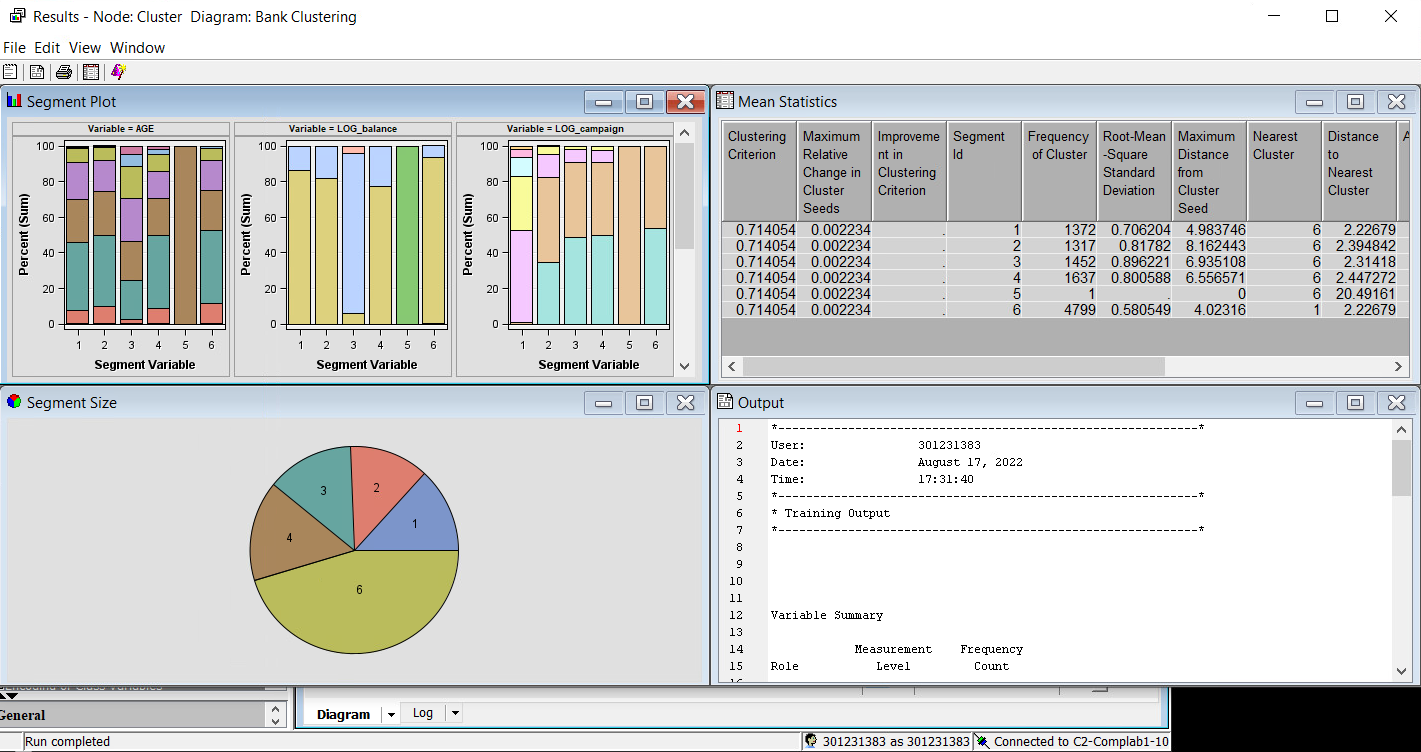
1. Are there clusters that can be combined to form fewer clusters?

The cluster proximity plot shows that cluster 6, 4, 3, 1 are in close proximities and

therefore, can be combined to one cluster. Cluster 5 and 2 can be combined as 1 cluster. In

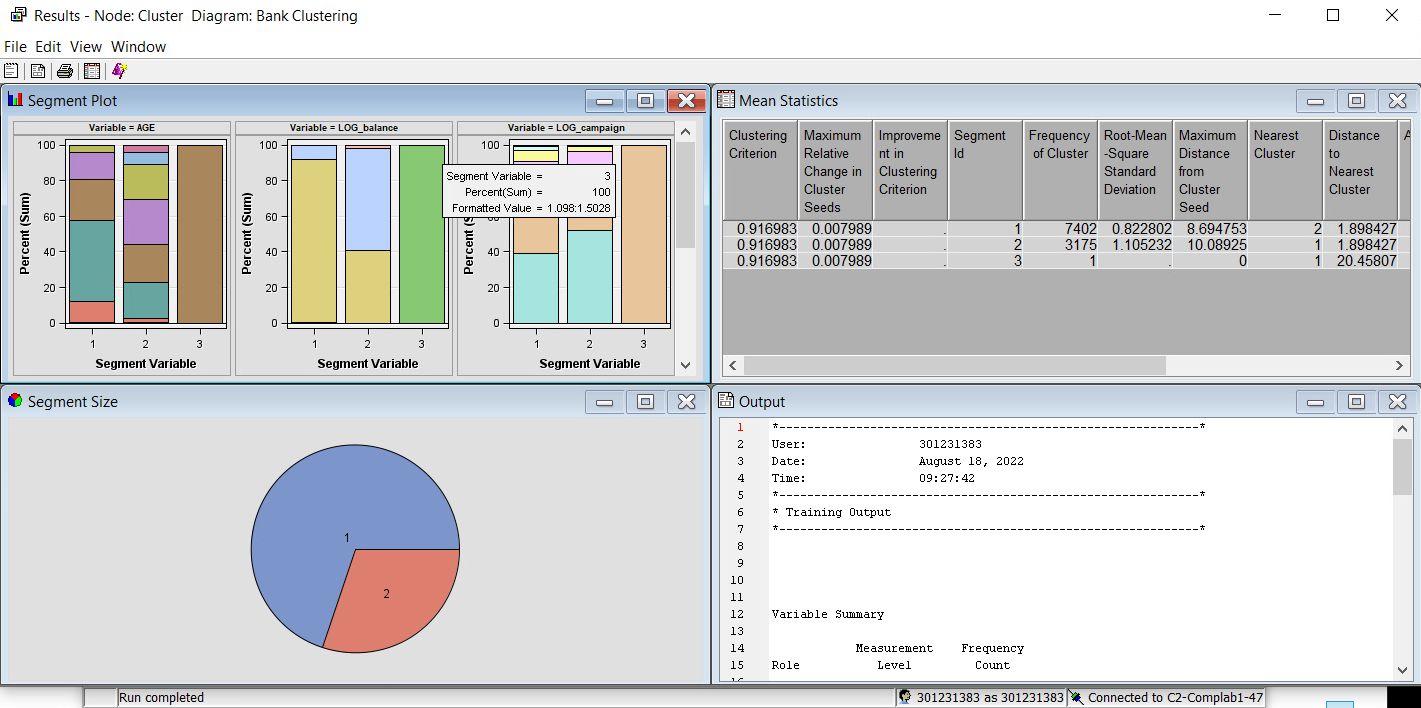
total, we can have 2 clusters. See diagram below:



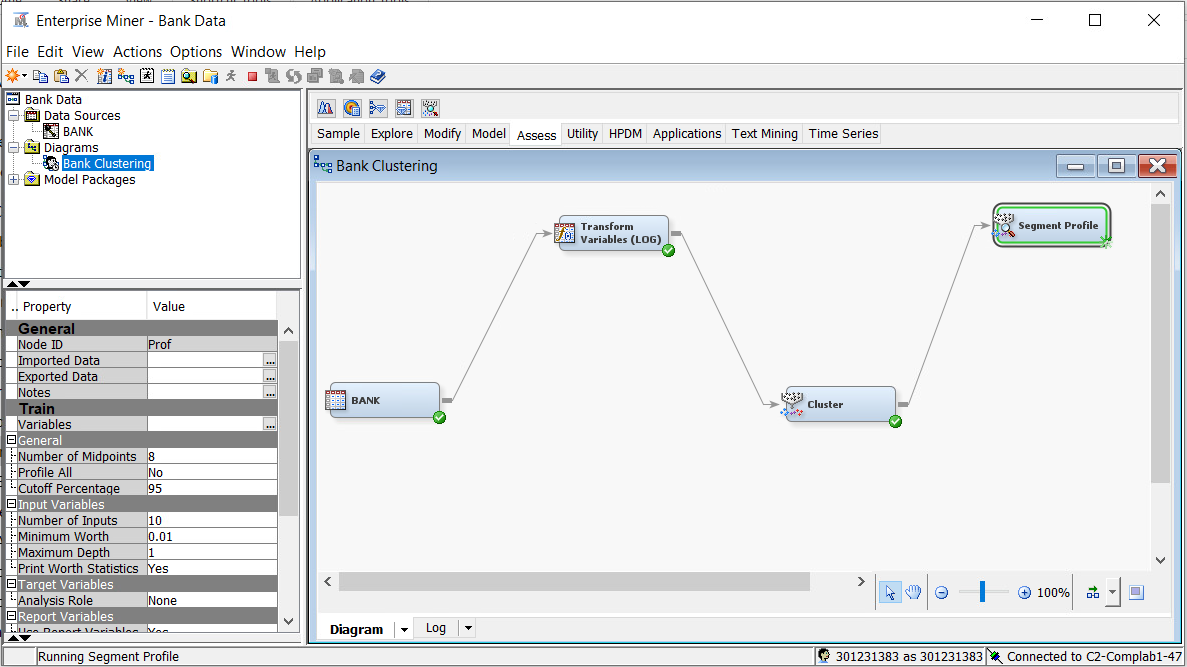


1. Change **Maximum Number of Clusters** to **3** and rerun the **Cluster** node.

We now have 2 largely independent clusters after increasing the number of clusters to 3. 7402 observations are in cluster 1 while 3175 observations are in cluster 2 (see results).

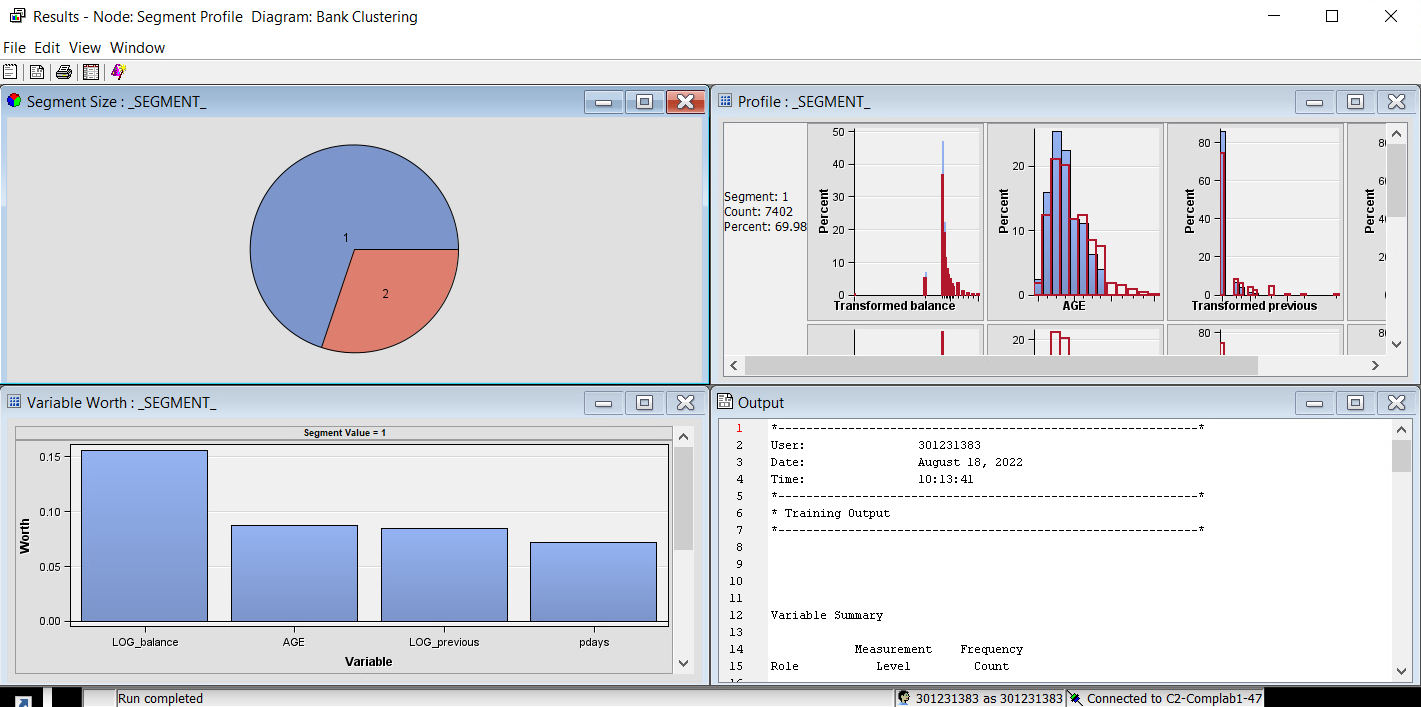


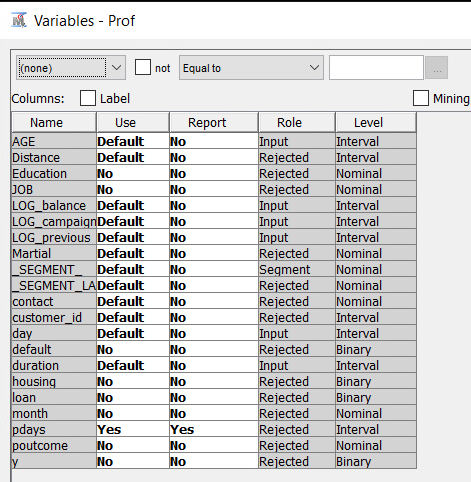
1. Drag a **Segment Profile** node onto the diagram workspace and connect it to the **Cluster** node.



1. Run the **Segment Profile** node.

Based on the cluster chart, there are different segment size of each cluster. As we see that the cluster 1 has more obervations as compared to cluster 2.



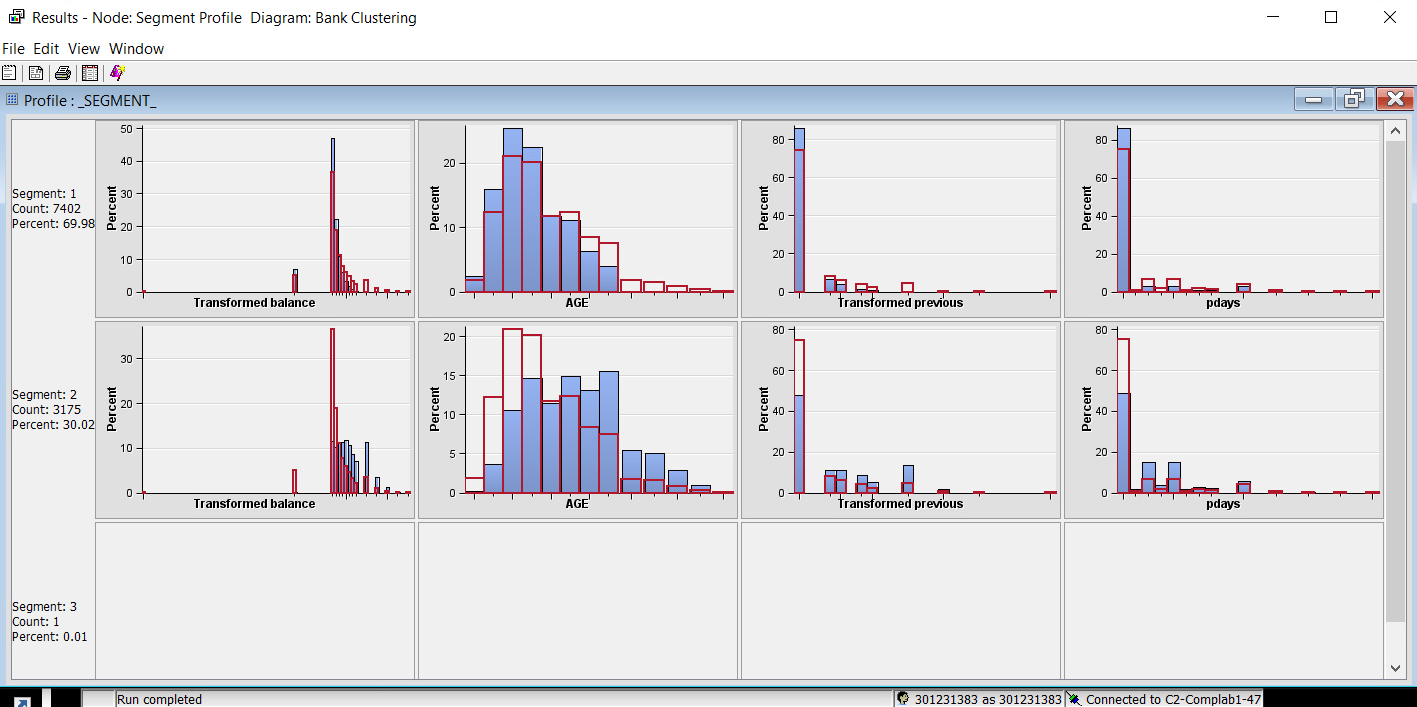


1. Inspect the profile results.

We can profile each segment formed by the cluster analysis on the graph, as well as the variable distribution inside each segment.



In the analysis, there are two clusters (segments). In each of the clusters, the variables balance, age, previous, and pdays' are significant. We examine at the profile section, as shown below, to determine the amount of its significance.



1. How would you label Clusters 1 and 2?

The Average Age for customers in cluster 1 is 37 years and 49 years for cluster 2.

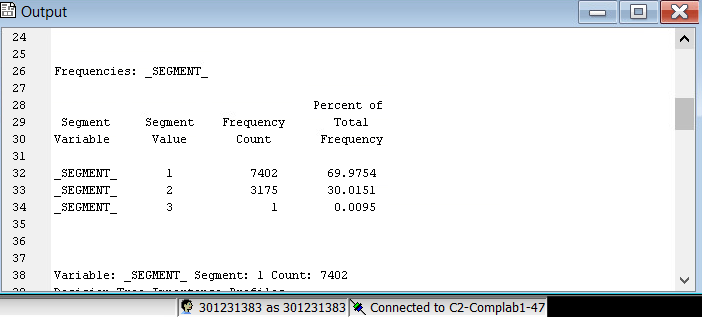
Cluster 1: Millennials & Affluent

Cluster 2: Mature & Wealthy

1. What should you do about Cluster 3 containing only one observation?

Cluster 3: Cluster 3 could not provide any useful clusters for the analysis because it only has one count or observation (0.0095).

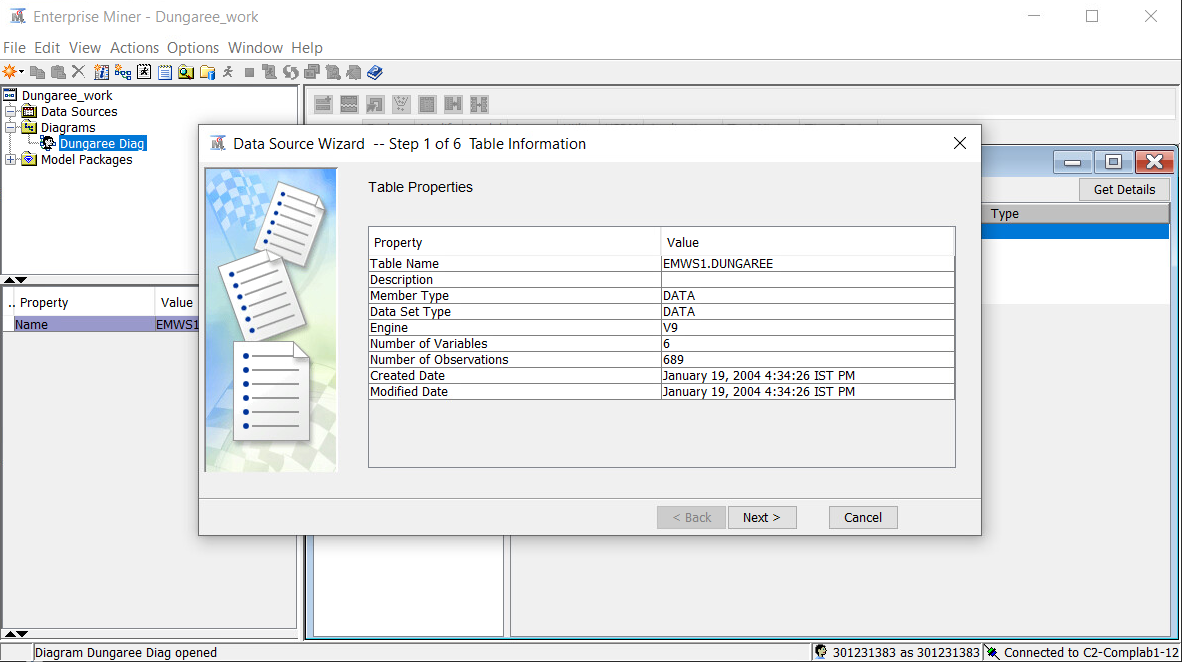
Explanation: Customers cannot be divided into this cluster or any other cluster using any other variable, hence it is irrelevant and won't be included in the analysis.

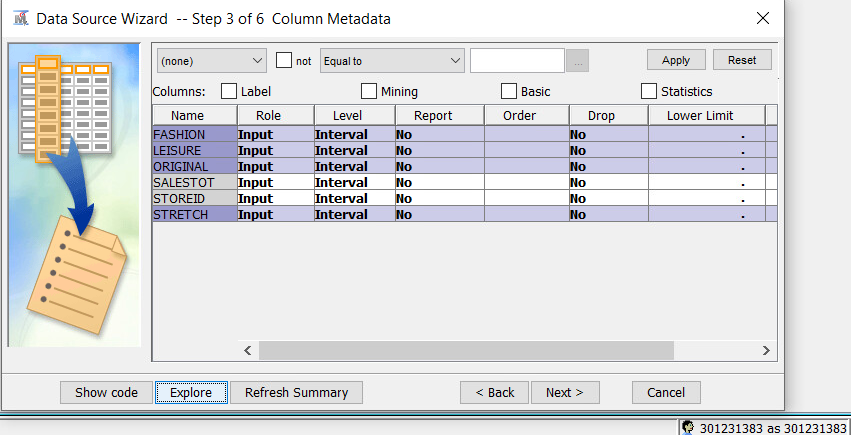


1. **Conducting Cluster Analysis**

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.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Model Role** | **Measurement 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**) |





**Answers:**

c. Examine the distribution of the variables.



* Are there any unusual data values?

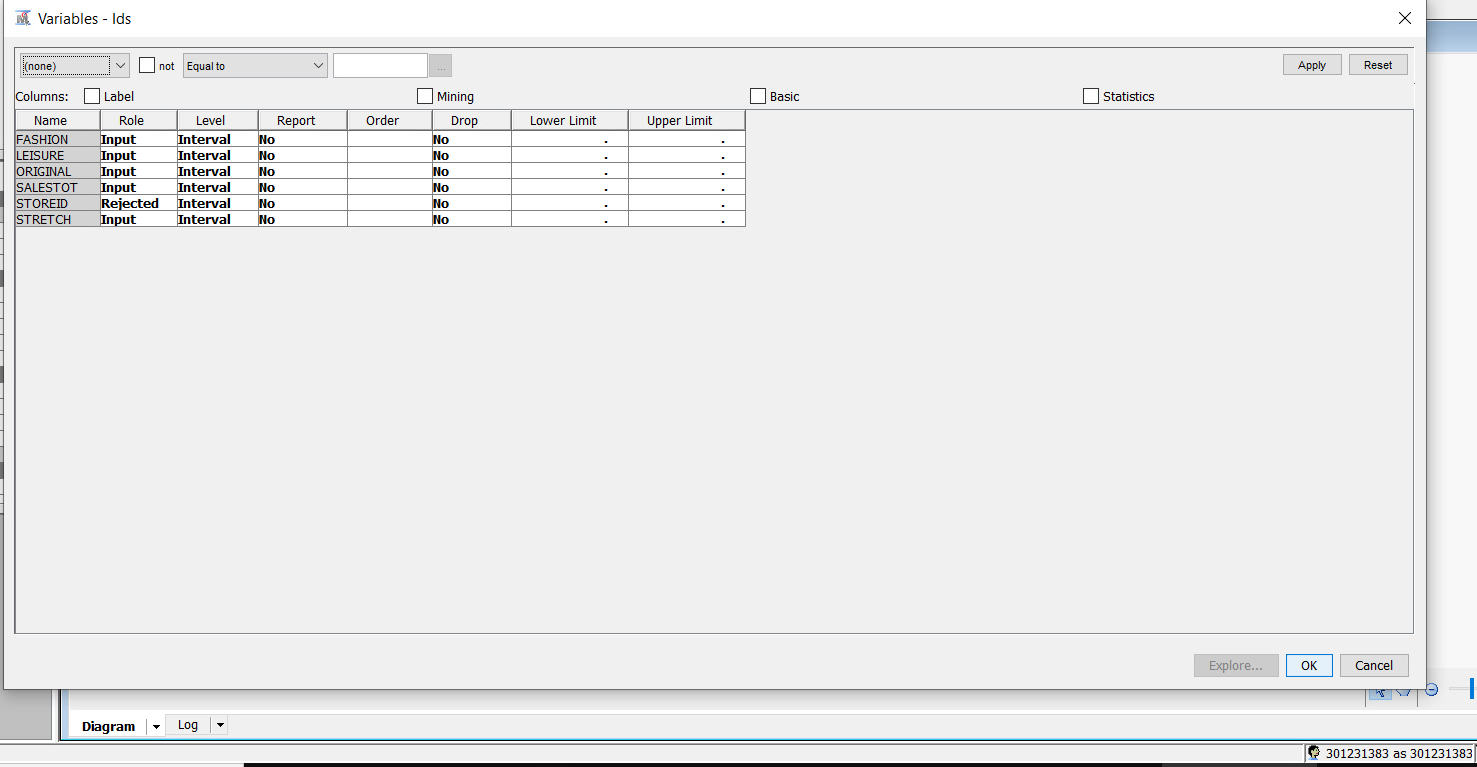
There do not appear to be any unusual data values.

* Are there missing values that should be replaced?

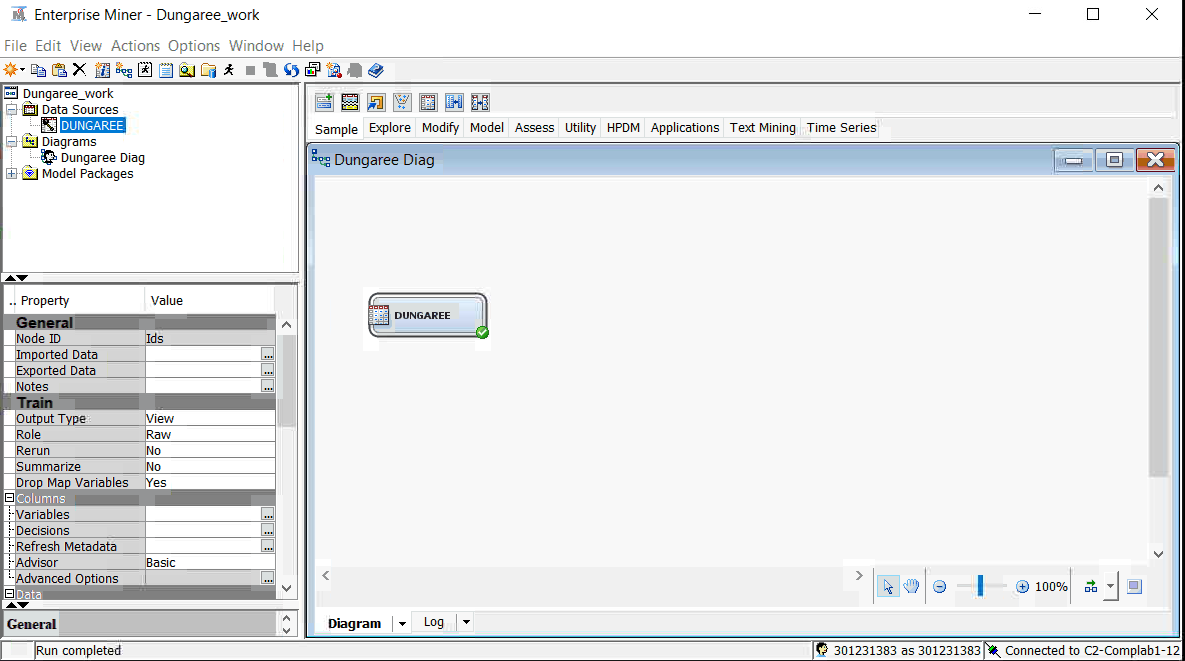
There do not appear to be any missing values.

* 1. Assign the variable **STOREID** the model role **ID** and the variable **SALESTOT** the model role **Rejected**. Make sure that the remaining variables have the **Input** model role and the **Interval** measurement level. Why should the variable **SALESTOT** be rejected?

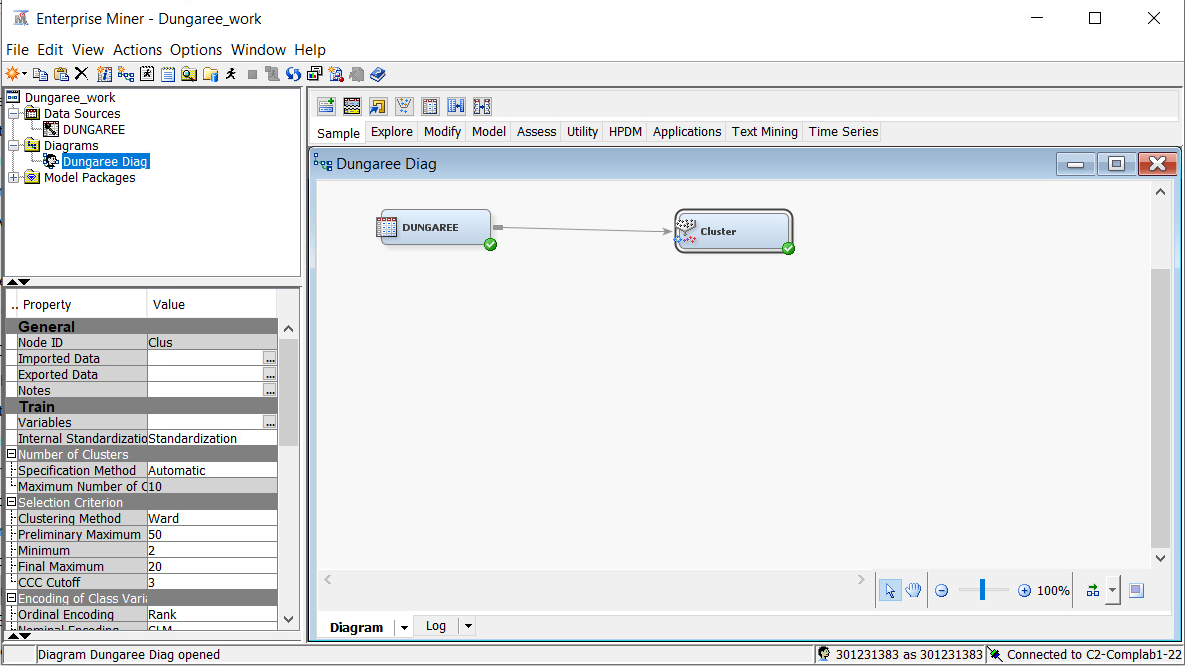
The variable SALESTOT should be rejected for the reason that it is the sum of the other input variables in the data set. Therefore, it should not be considered as an independent input value.



* 1. Drag the **DUNGAREE** data source to the diagram workspace.



* 1. Add a **Cluster** node to the diagram workspace and connect the **Input Data** node to it.



* 1. What would happen if you did not standardize your inputs?

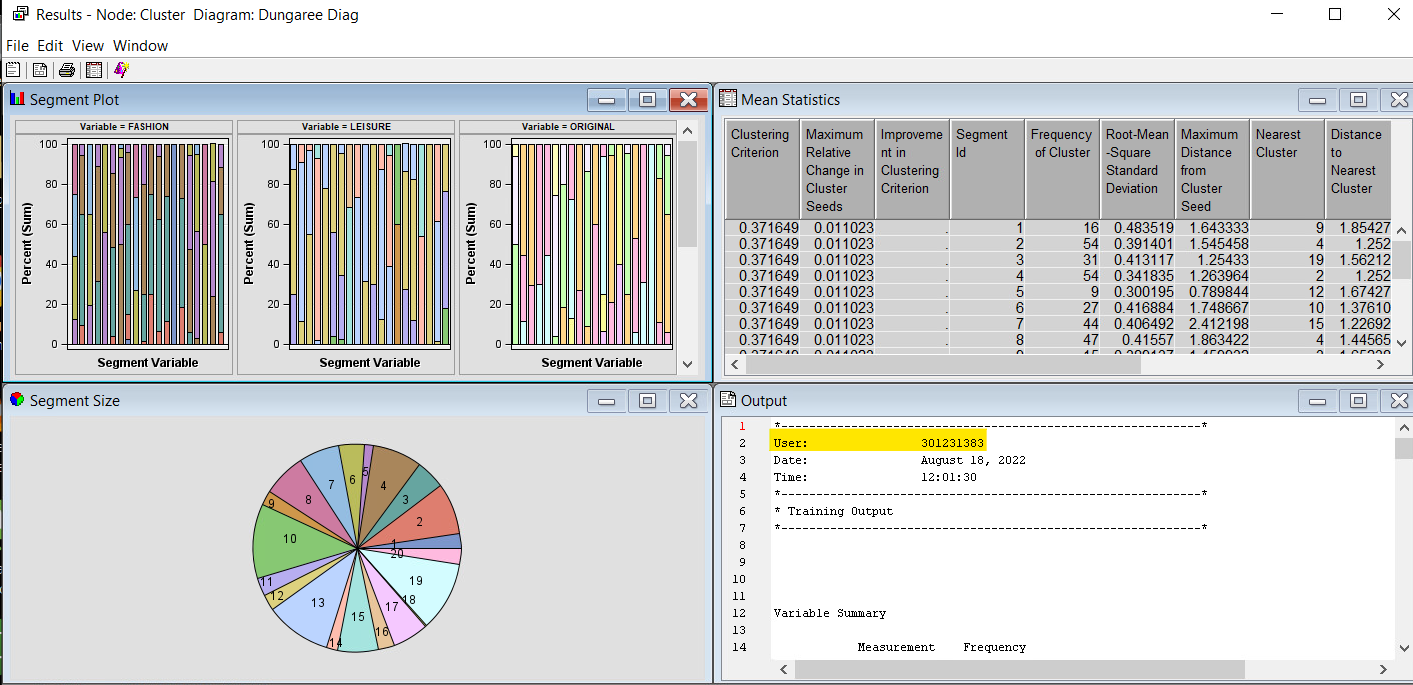
Not all of the data's variables will have the same scale.

Without standardisation, variables having a wide range will dominate those with a narrower range. For instance, Leisure and Original will be more important than Stretch and Fashion.

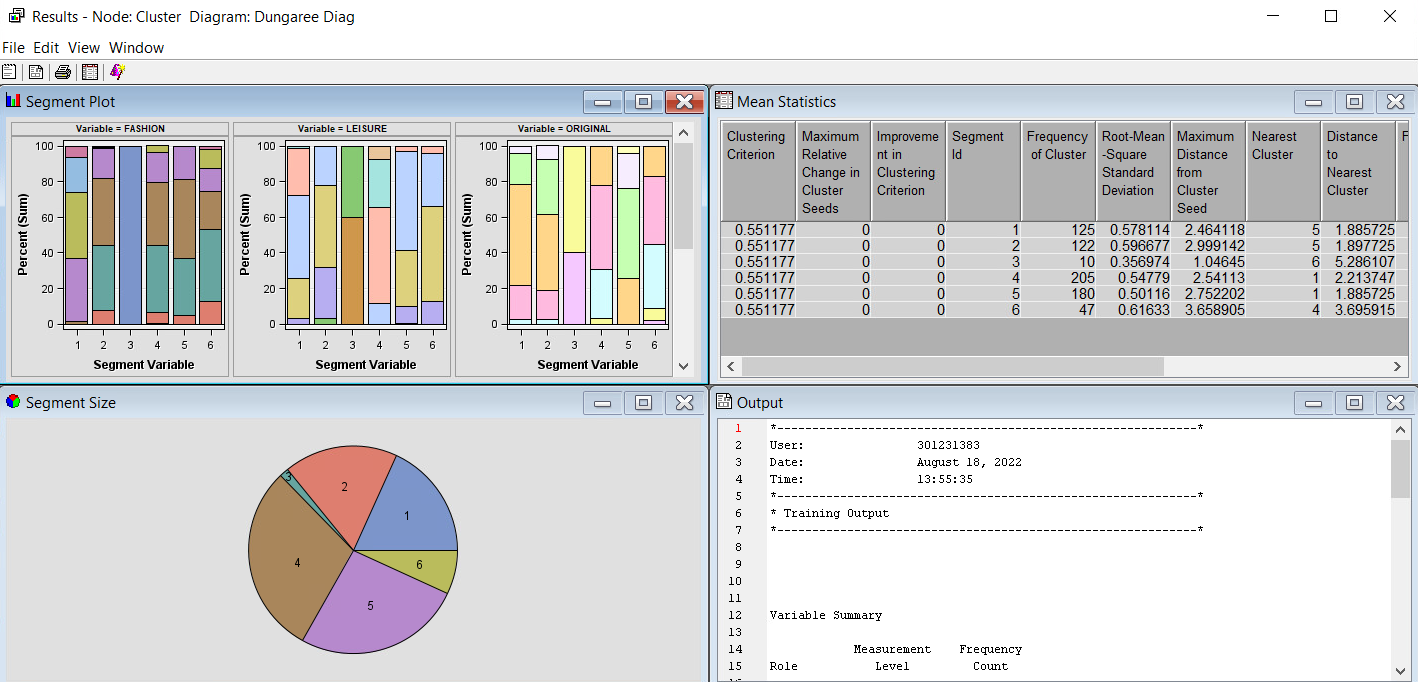
* 1. Run the diagram from the Cluster node and examine the results. Does the number of clusters that are created seem reasonable?

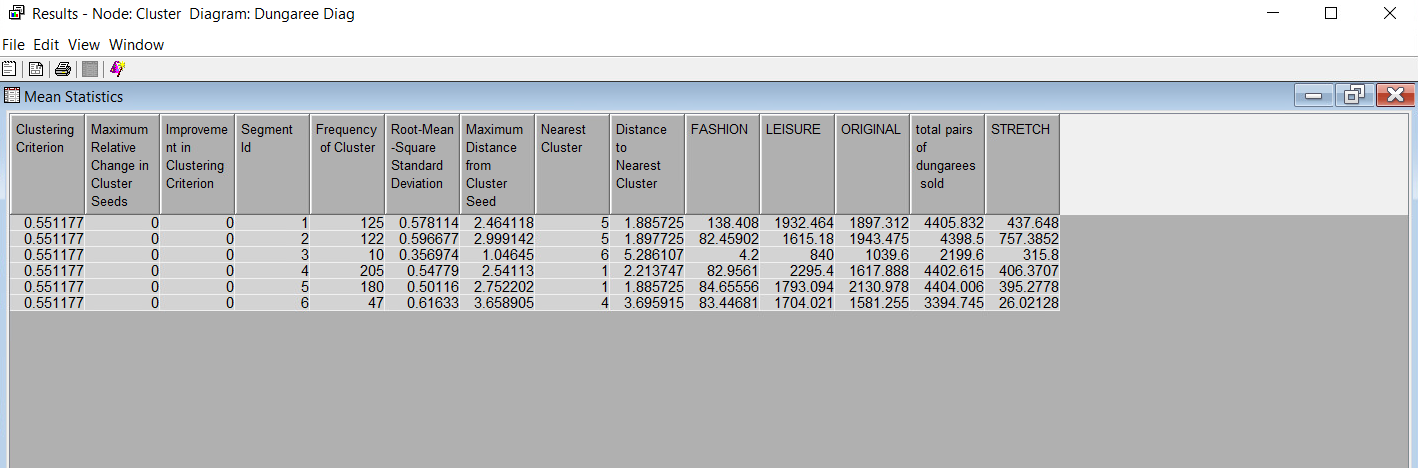
20 clusters in total were generated automatically. There are too many clusters here, thus it doesn't make sense. The presumptions of a finite number of clusters are not met. Additionally, there are 689 total observations, therefore 20 clusters is a large amount for the study.

For the SAS programme to execute a manageable number of segments, it is critical to provide the required number of clusters.



Specify a maximum of six clusters and rerun the Cluster node.

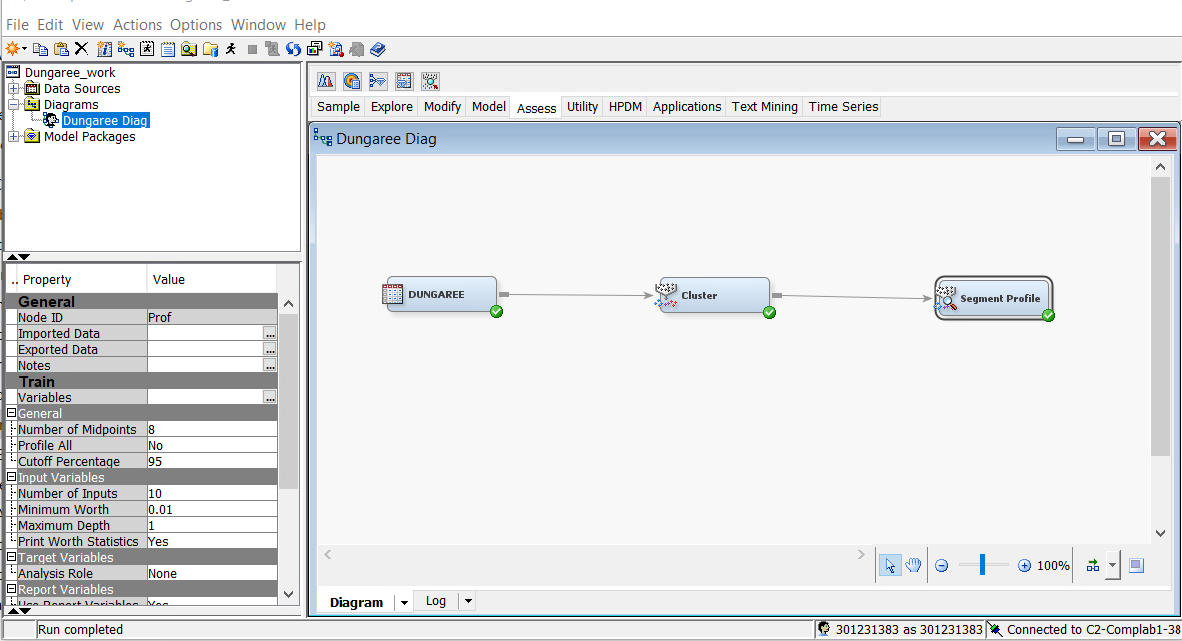


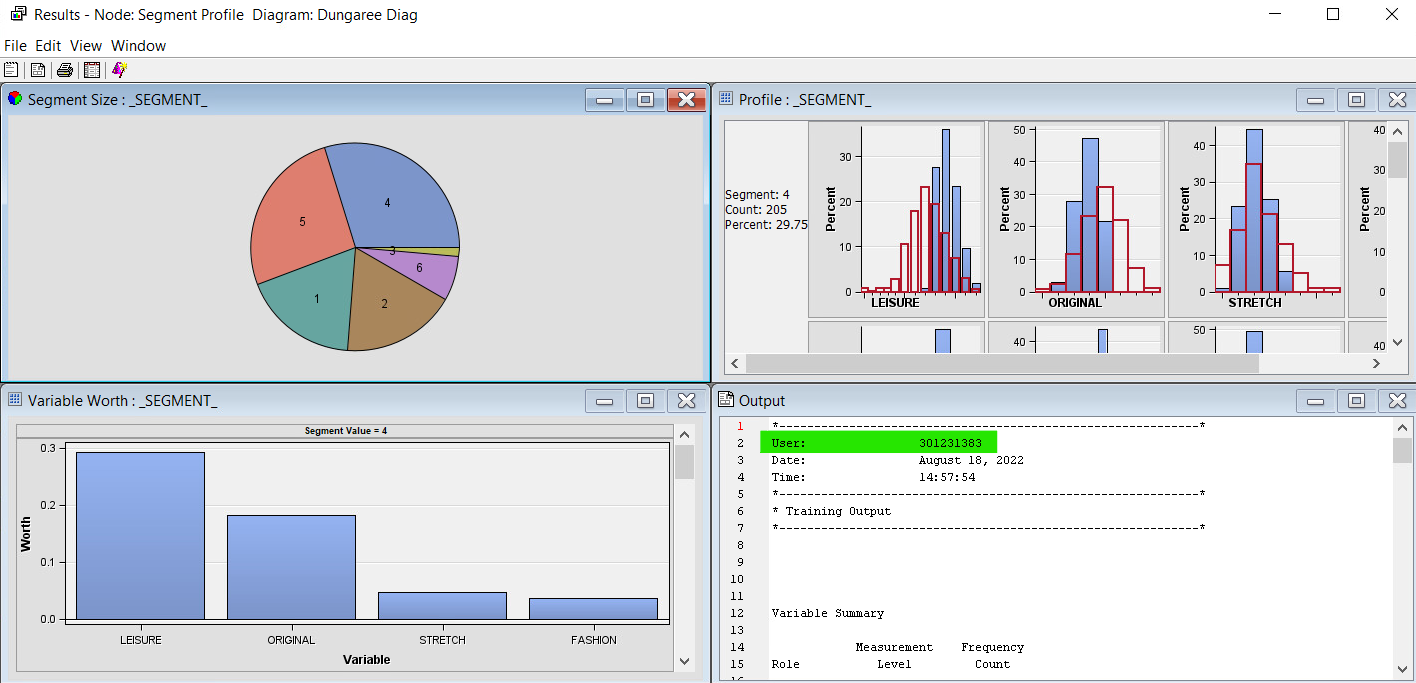


* 1. How does the number and quality of clusters compare to that obtained in part e ?

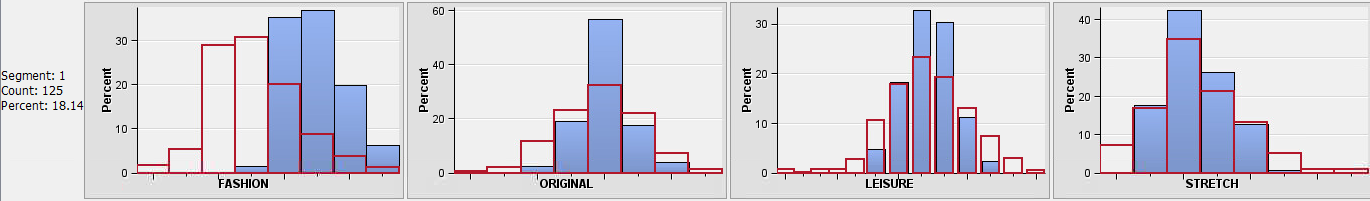
Six clusters are more appropriate for the study than the 20 that e automatically generated, making the analysis easier to interpret. It is predicated on the notion of a finite number of clusters. Except for cluster 3, five of the six clusters have similar observations or frequencies. The least amount of observations are in Cluster 3.

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

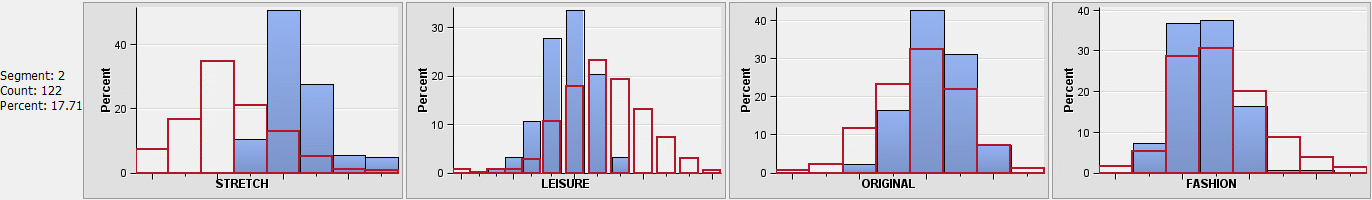




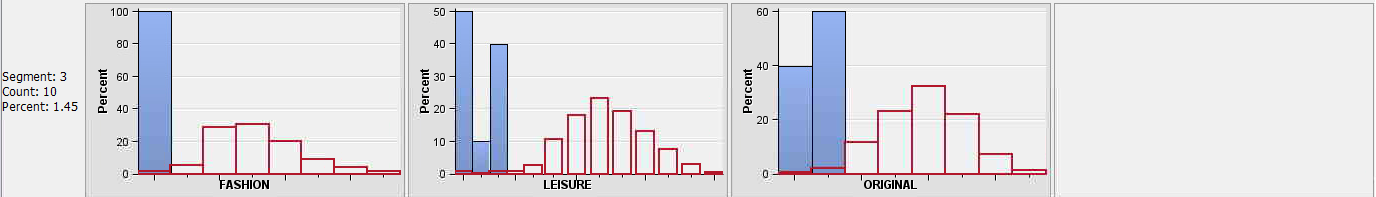
Segment 1: Segment 1 demonstrates that the retailer sells more Original jeans. Original and Leisure characteristics, which indicate that more of these types of jeans are offered in stores, set apart this sector. When evaluated to population statistics, the distribution for Leisure and Stretch is very similar to a normal distribution.



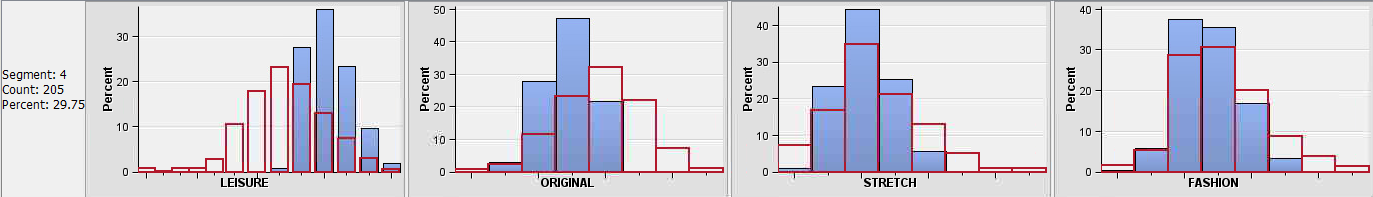
Segment 2: The retailer appears to sell more stretch jeans, as demonstrated in segment 2. Stretch and Leisure factors, which indicate that more of these types of jeans are sold in stores, set apart this sector. For the segment profile, the Original and Fashion distributions are very similar to a normal distribution.



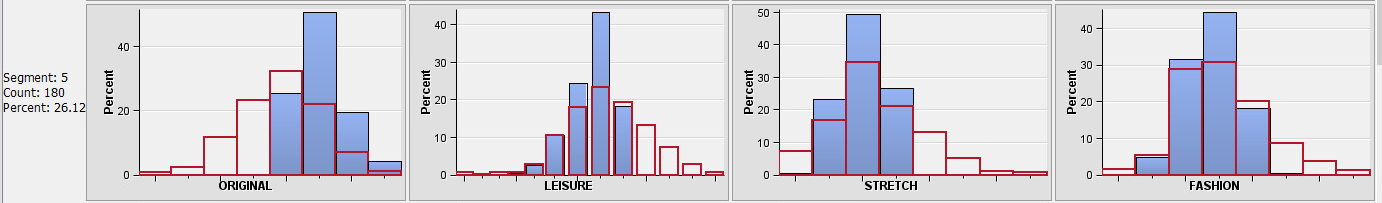
Segment 3: Stores selling all jean designs in modest quantities can be found in segment 3.



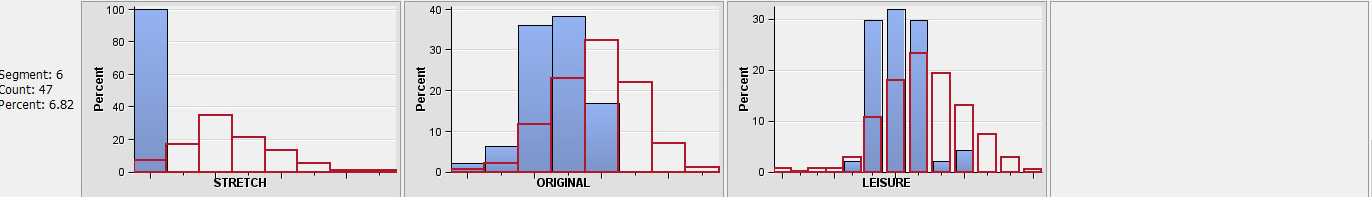
Segment 4: Retailers offering more leisure jeans than original can be found in Segment 4.



Segment 5: Retailers offering more designer jeans than original can be found in Segment 5.



Segment 6: There are stores in segment 6 that sell more original jeans than original, but fewer stretch and fashion styles.



* 1. Based on this analysis what can you conclude about the purchasing pattern of the customers?

Same as above answer.

