**Homework 2**

**Exercise One**

**Assignment Details:**

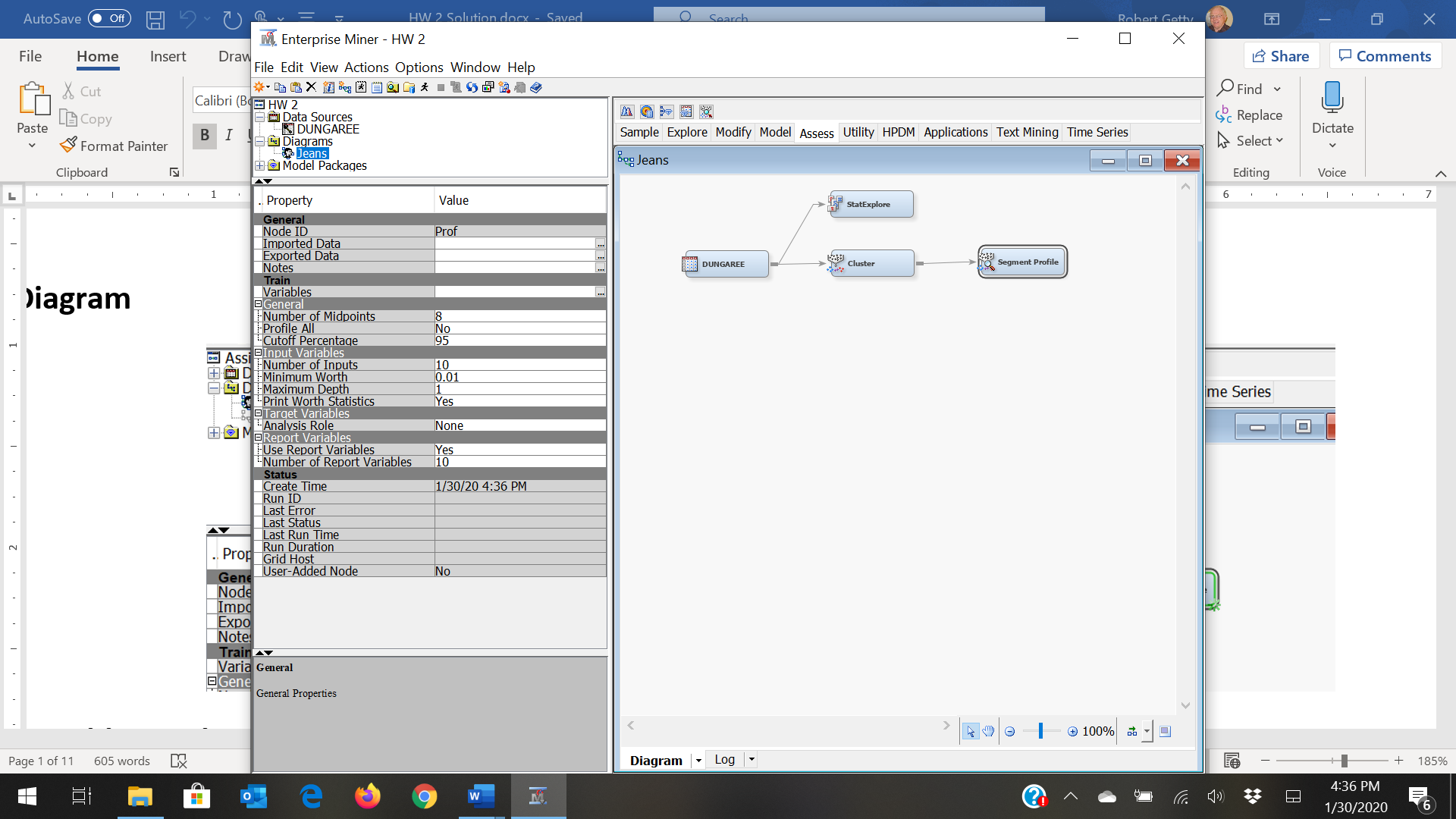
The DUNGAREE data set gives the number of pairs of four different types of dungarees sold at stores over a specific time period. Each row represents an individual store. There are six columns in the data set. One column is the store identification number, and the remaining columns contain the number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

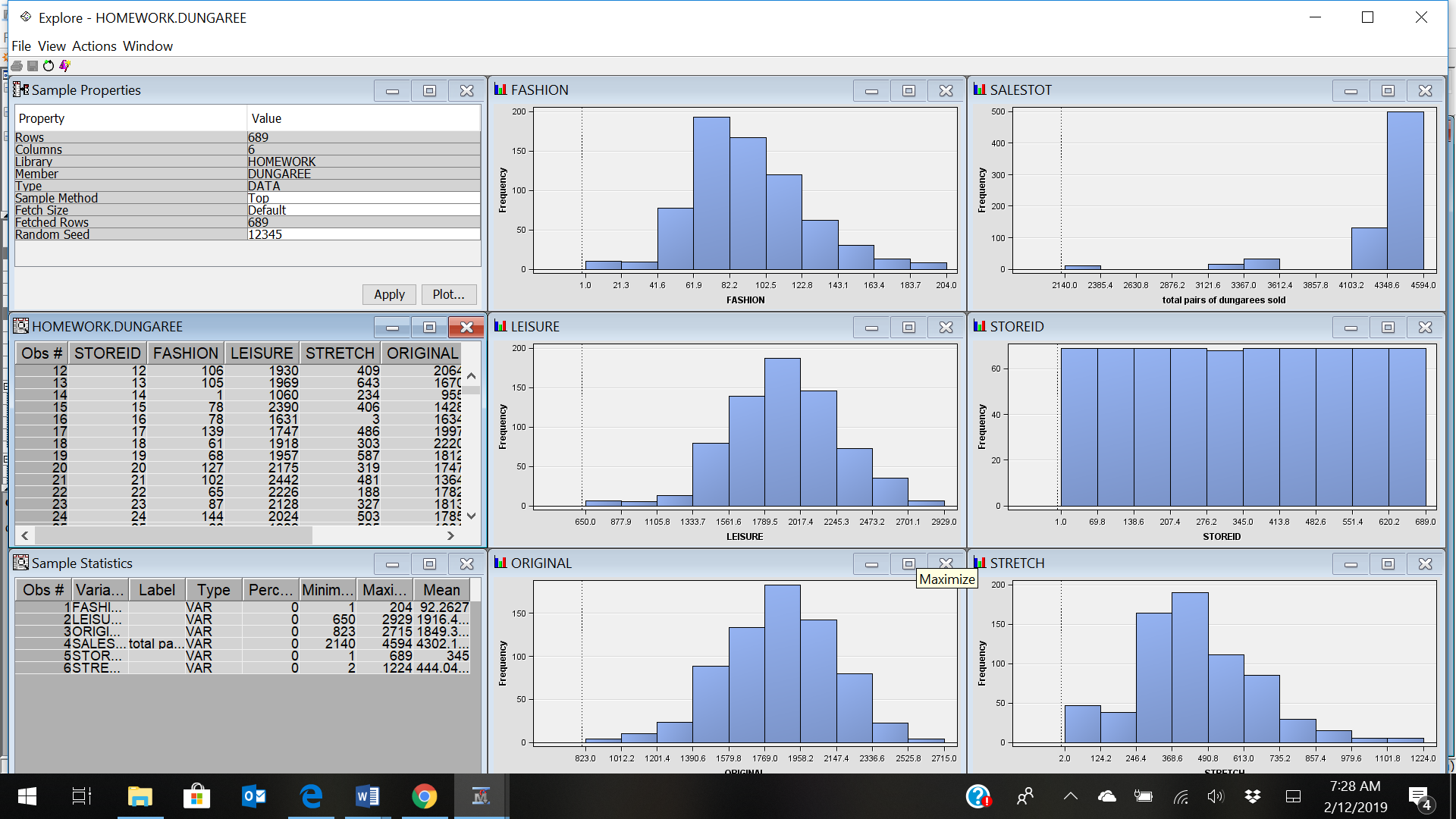
Table 1. Variable Settings and Description

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

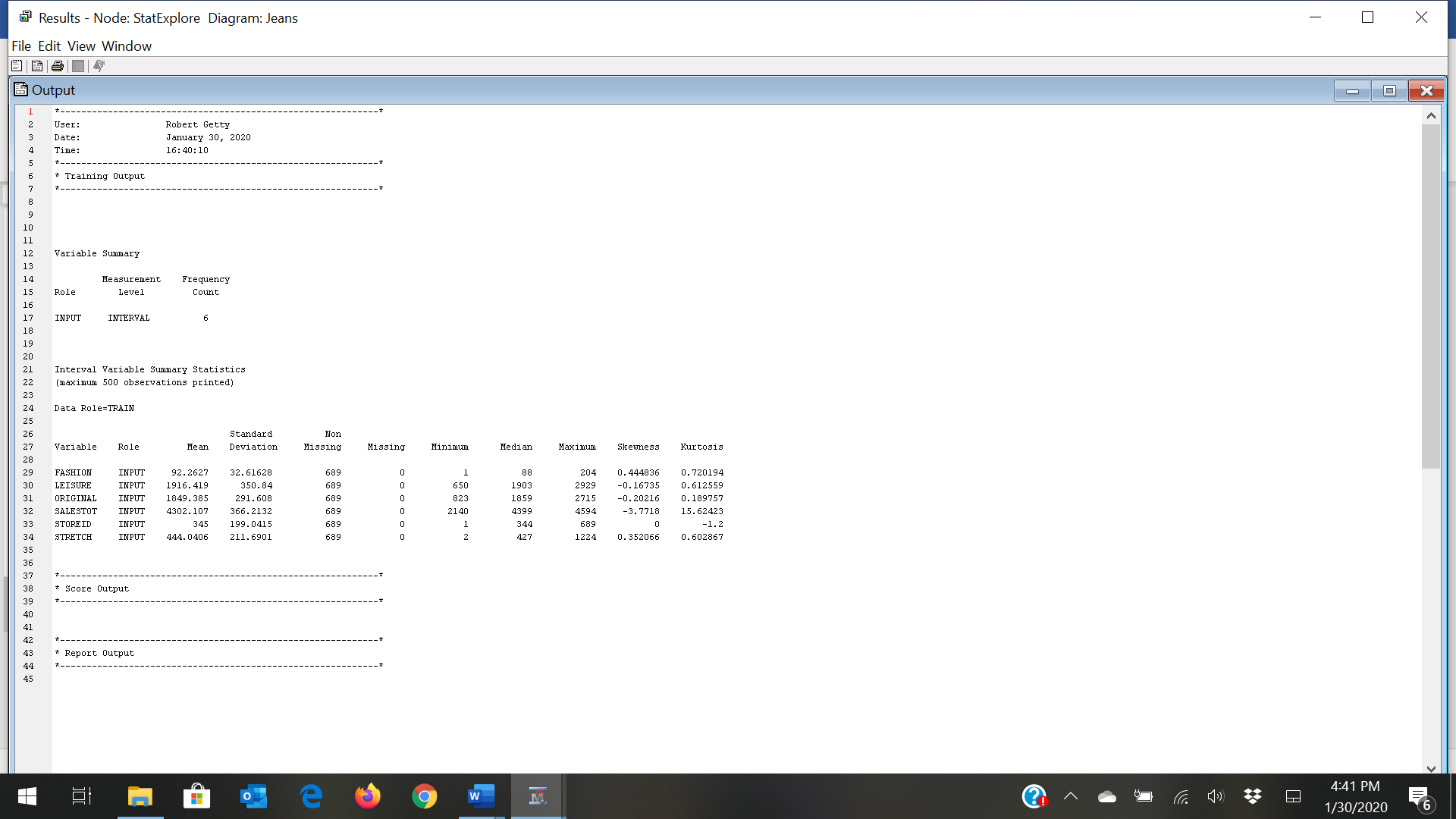
**Diagram**



**Variables Explore**



**Summary Statistics**



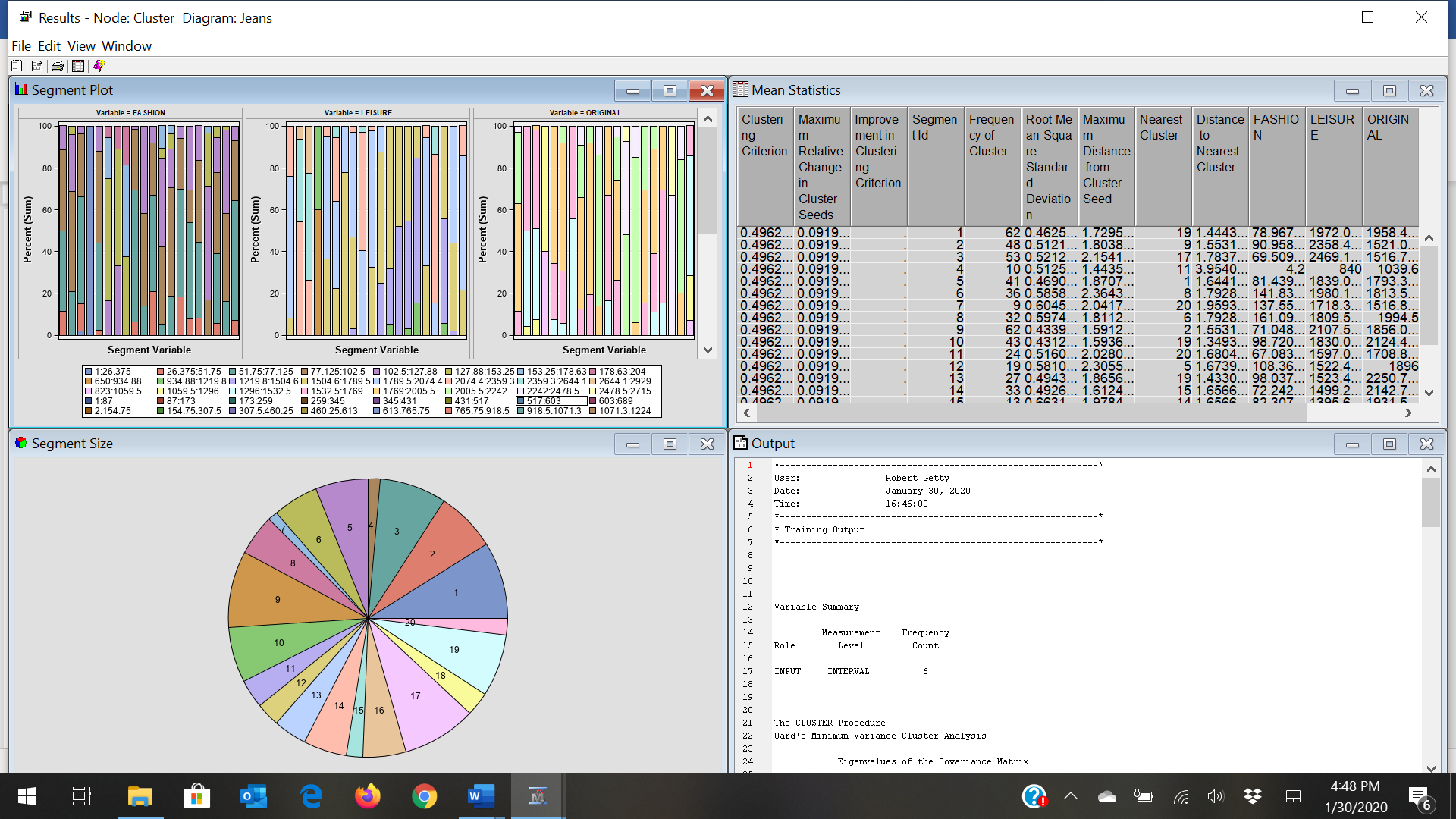
* 1. Are there any unusual data values?
  2. Are there any missing values that should be replaced?

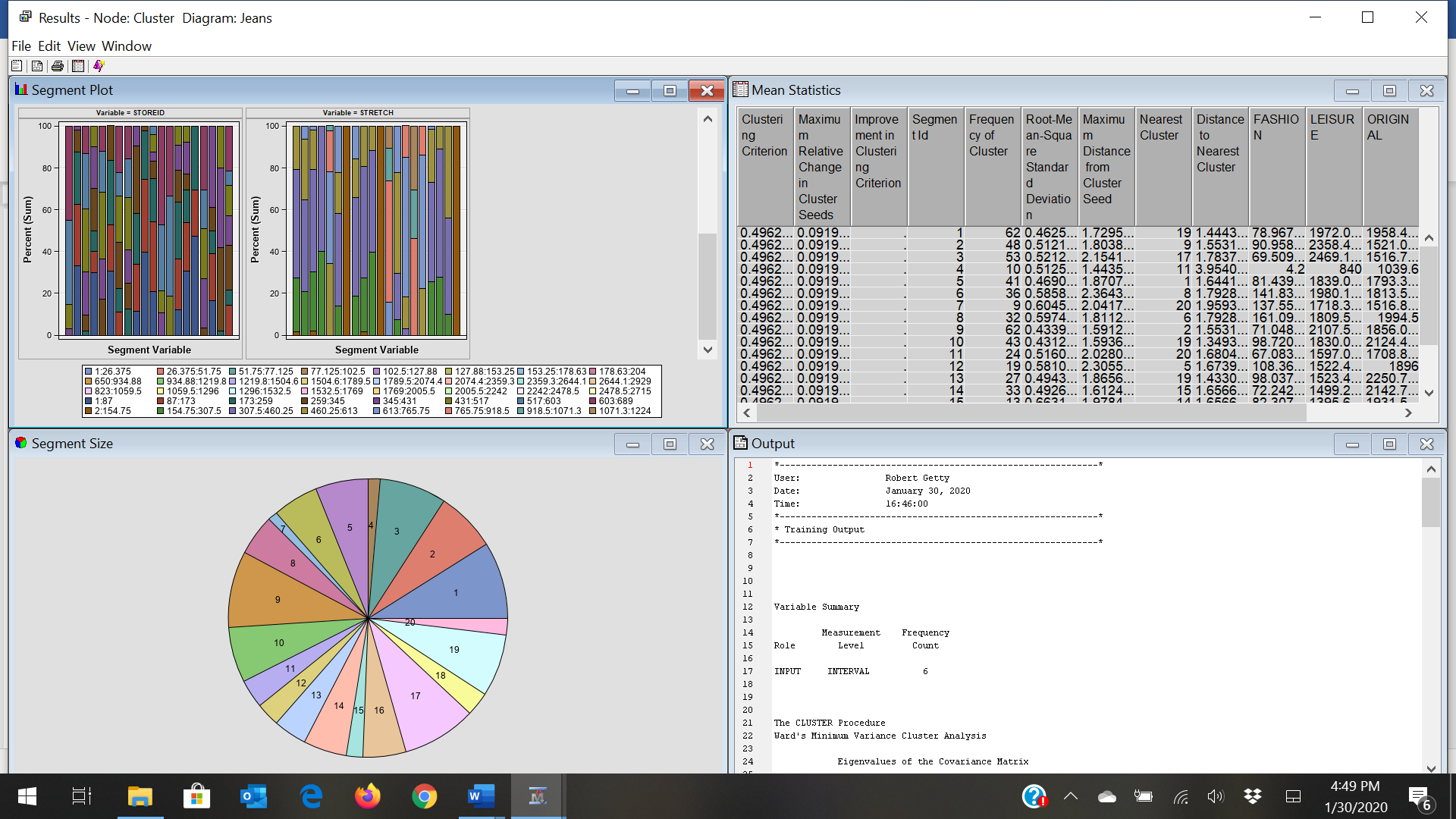
Why should the variable SALESTOT be rejected?

What would happen if you did not standardize your inputs?

**First Cluster Analysis**

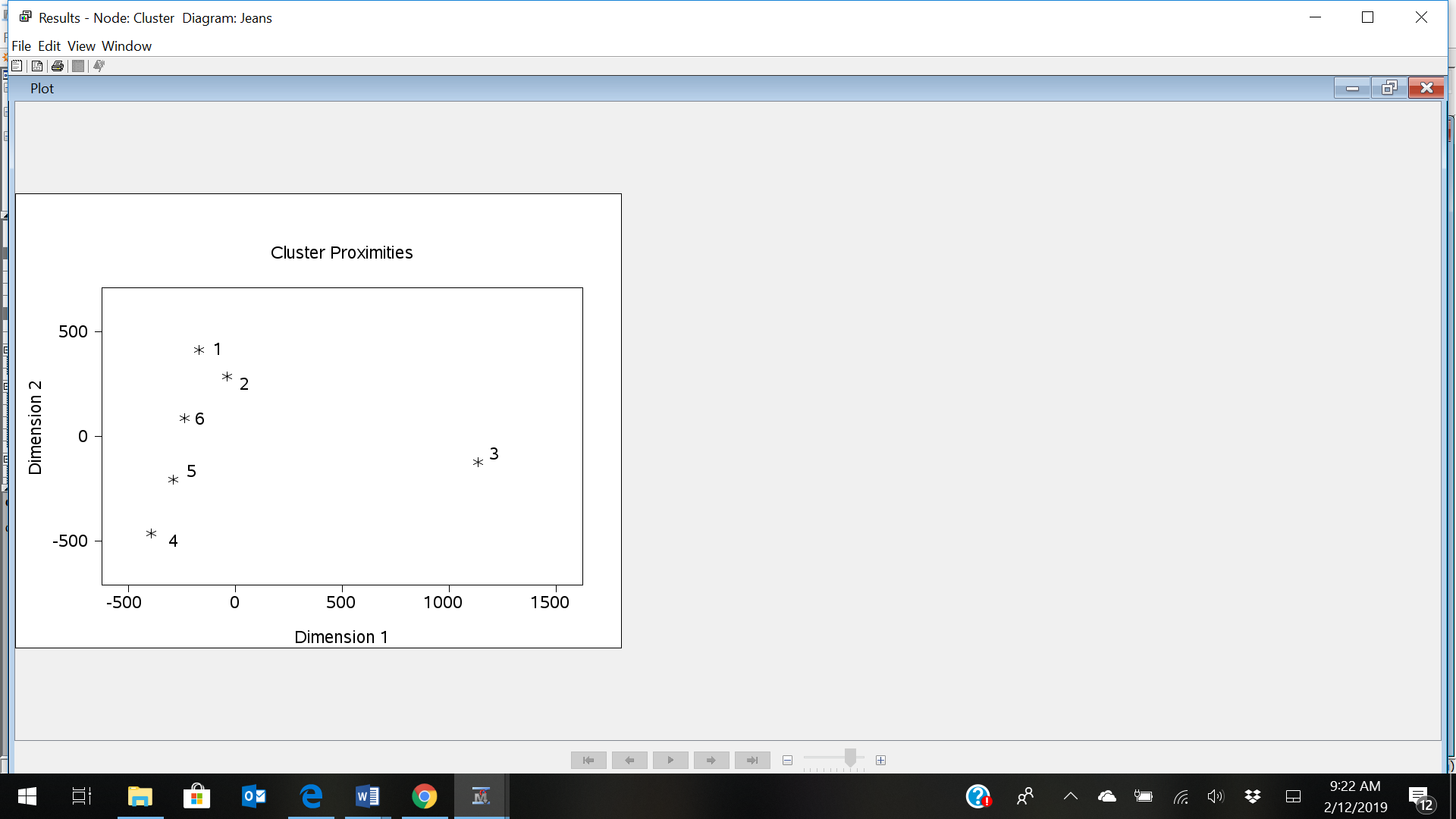
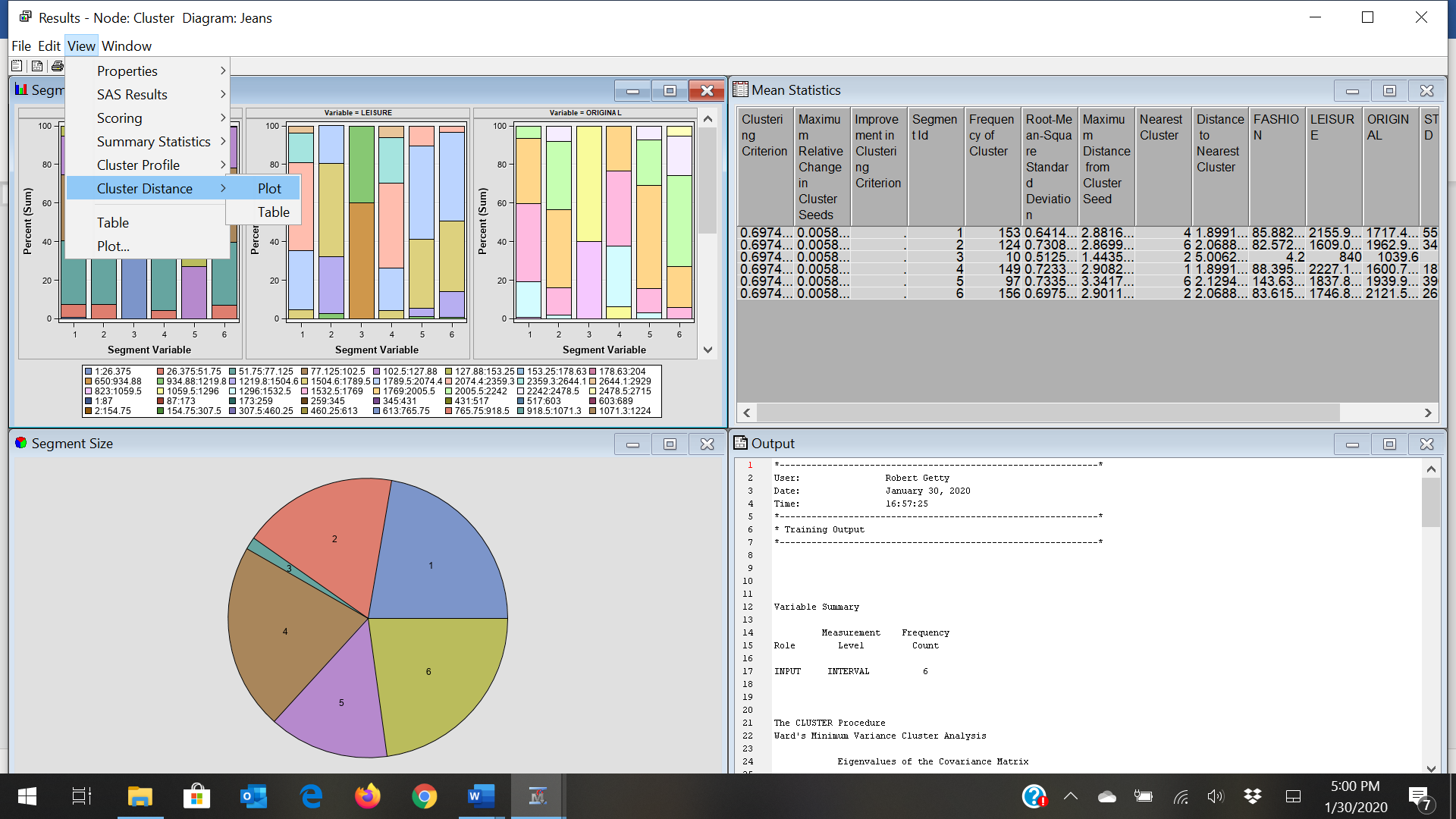
This analysis selected an unreasonable number of clusters that makes it difficult to interpret.

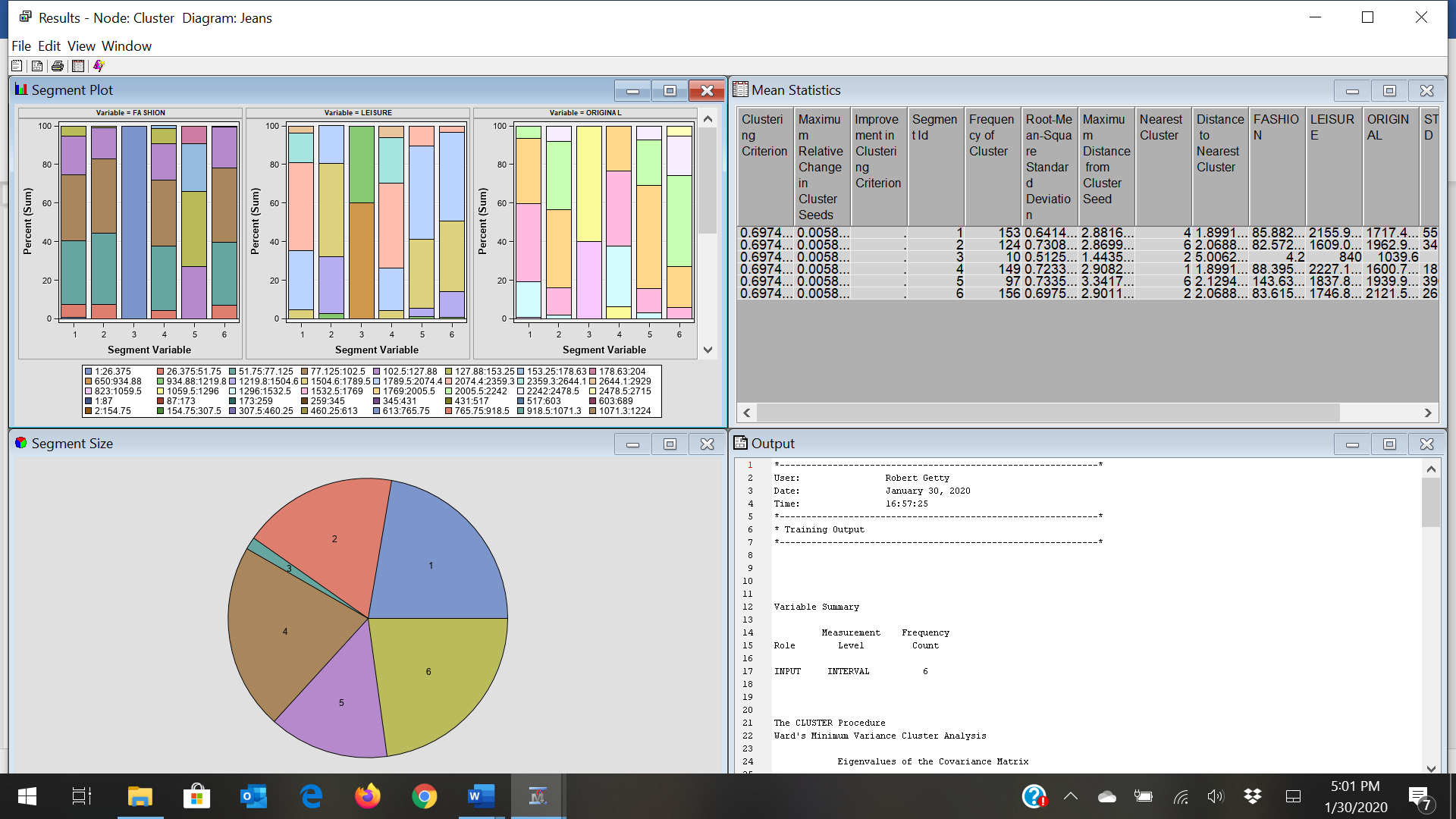


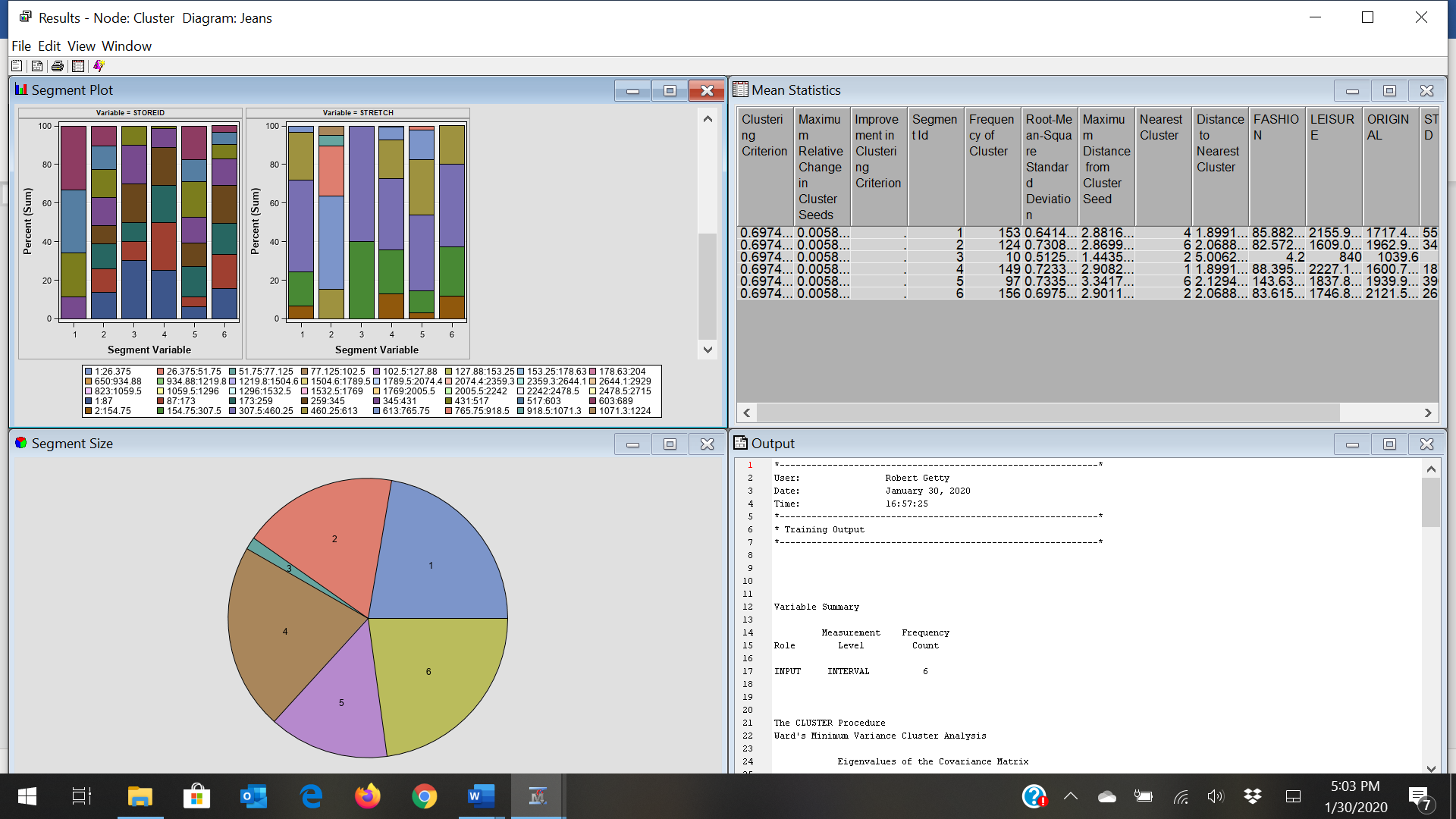


Does the number of clusters created seem reasonable?

**Second Cluster Analysis**







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

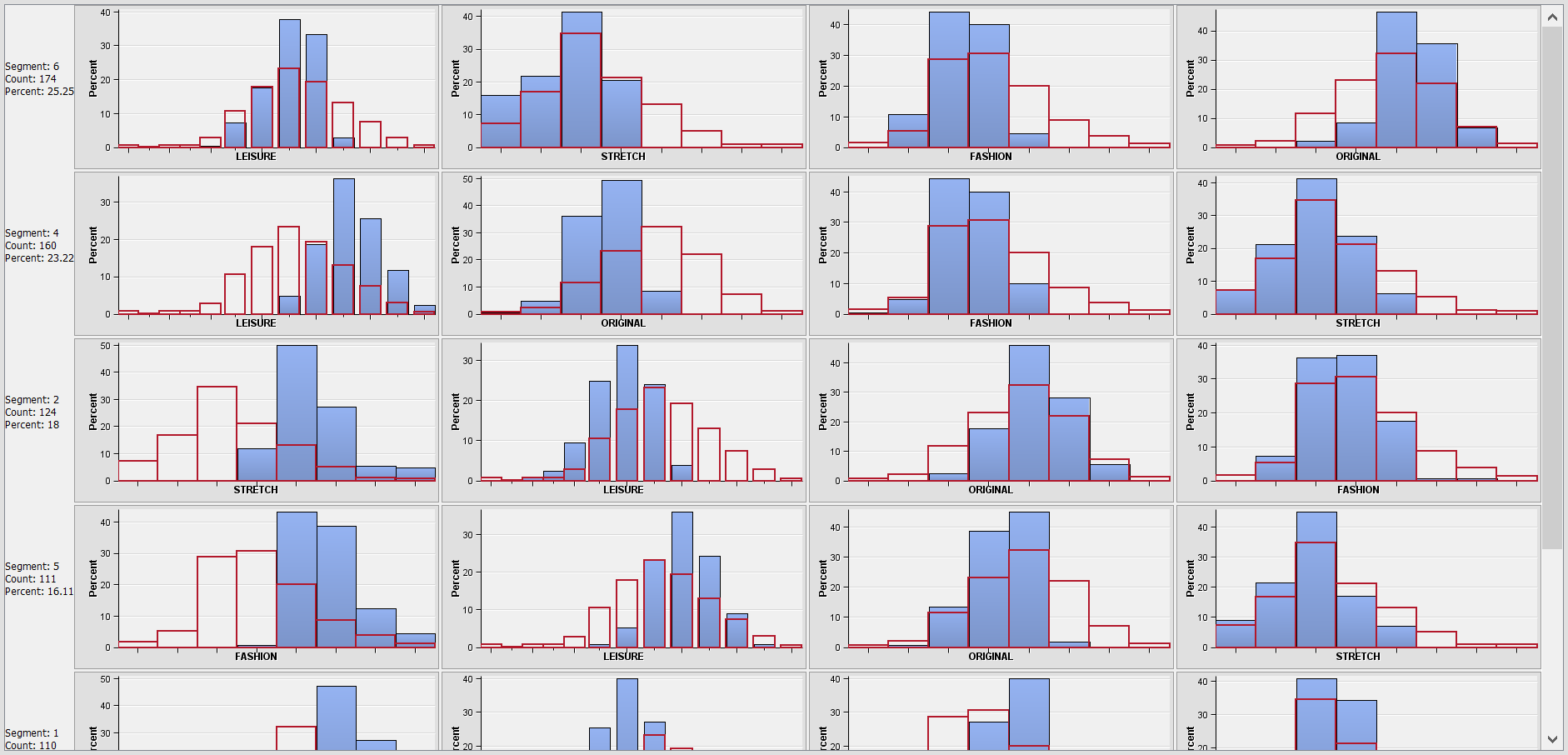
**Overview of the Segment Profile Node**

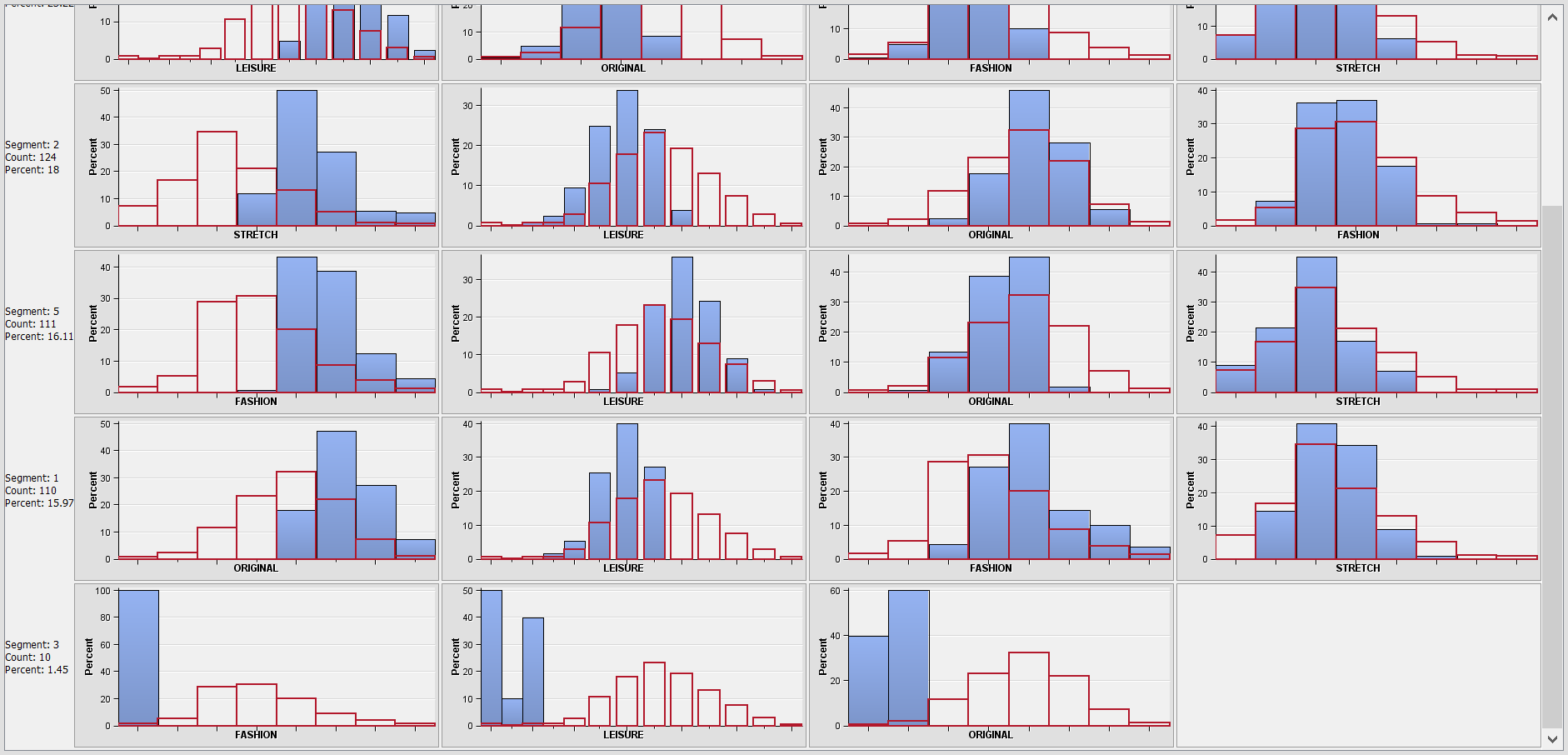
In the Enterprise Miner SEMMA data mining process, the Segment Profile node belongs to the Assess group. The Segment Profile node enables you to examine segmented or clustered data and identify factors that differentiate data segments from the population. The node generates various reports that aid in exploring and comparing the distribution of these factors within the segments and population.

The key variables for a specific segment are determined by creating a binary pseudo target variable based on the membership of the segment. Two methods can be used to determine the differentiating variables. In the default method, each of the interval and class target variables are ranked based on their logworth value. The value in turn, is based on the pseudo target. The interval variables are binned to identify their maximum logworth. The second method builds a decision tree of specified depth. Variables are selected based on their total differentiating capabilities. This second method permits a degree of interaction between the variables.

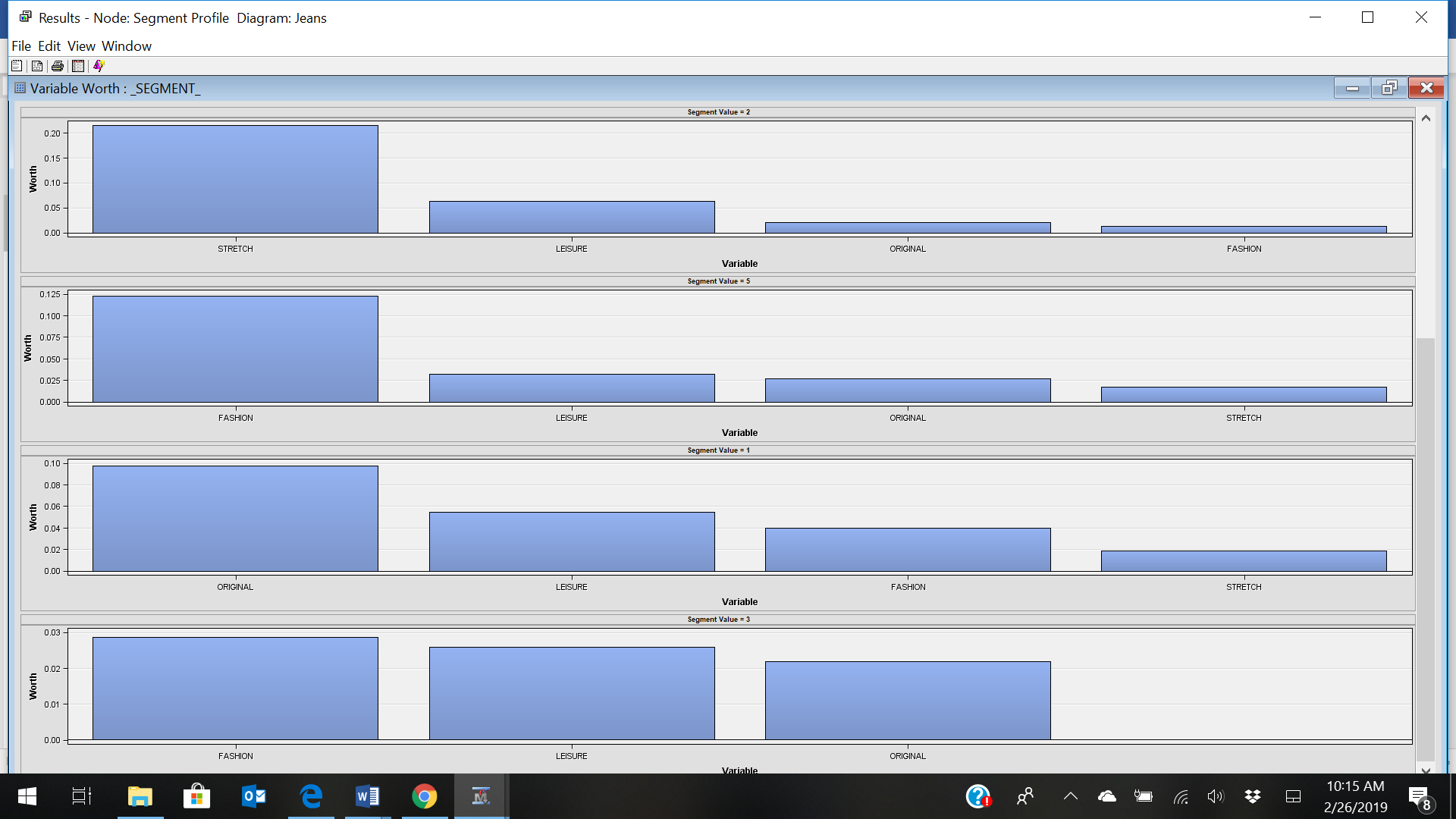
<https://www.youtube.com/watch?v=Cm0fDWHpbwg>

Since the measurement levels for all the variables is set as Interval, the results of Segment Profile plot is displayed as histogram. The blue shaded region represents with-in the segment distribution and the red outlined histogram represents the population distribution. The height of the histogram is scaled by the percentage of segment population. The output shows the relative difference between segment and population for all the segments from 1 to 6. Each segment shows profile plots, for all the four variables LEISURE, STRETCH, FASHION and ORIGINAL. The results of the node reveal that the four variables mark importance in generating all the segments except segment 3. Stretch the dungarees do not have importance in generating segment 3. Hence this segment might not be effective for the sales of stretch dungarees.

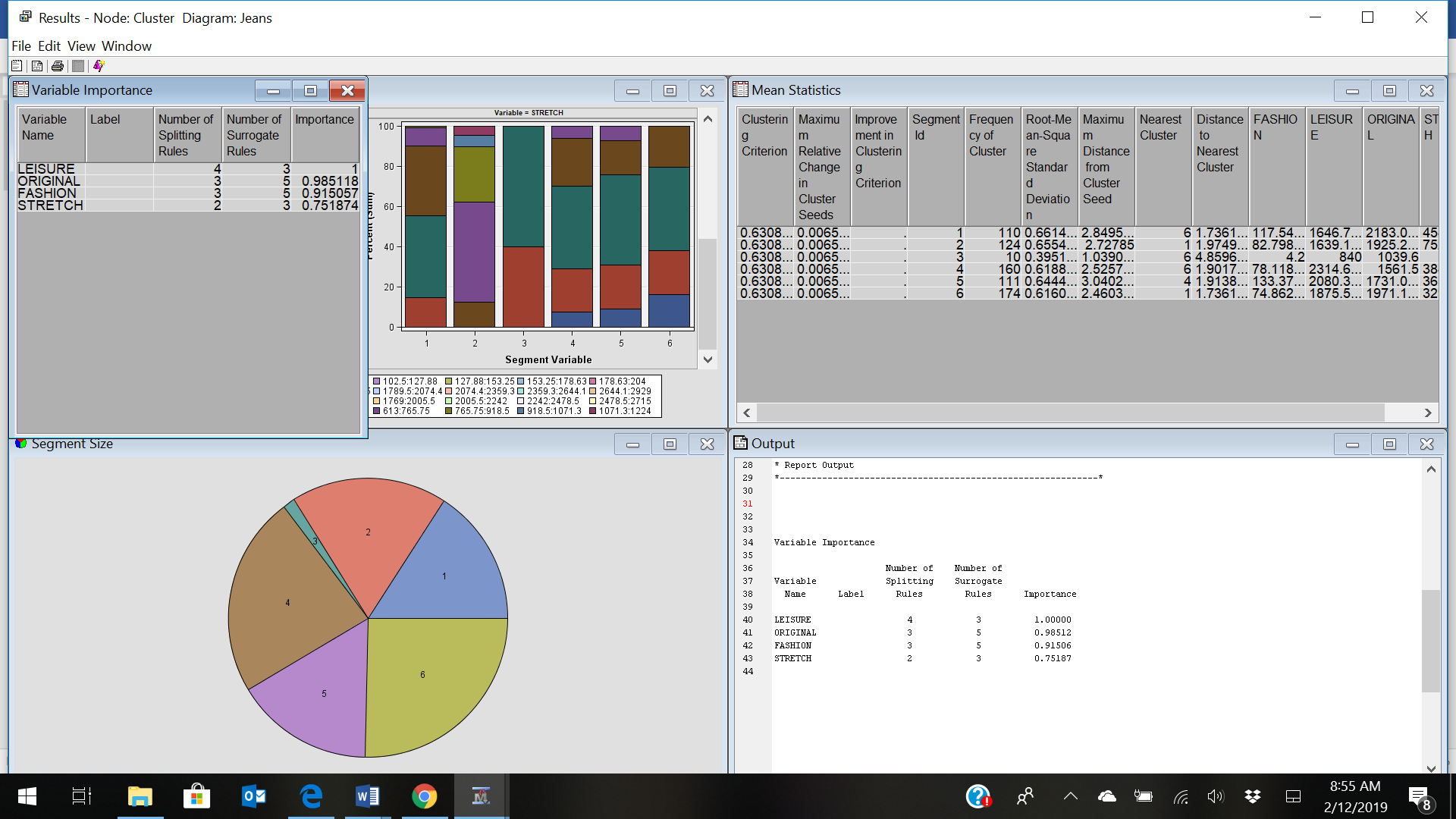






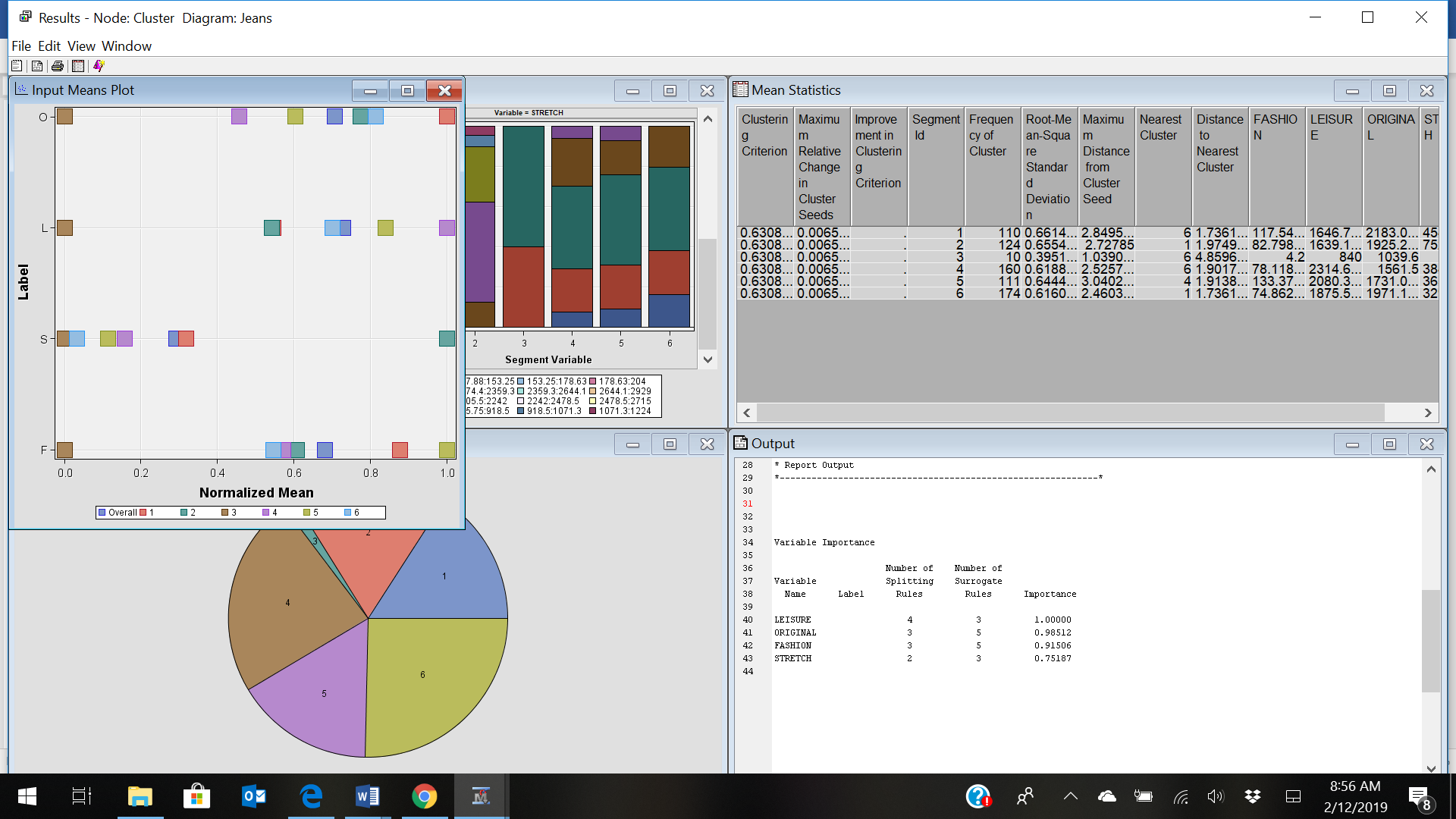


Variable Importance



**Cluster Plots**

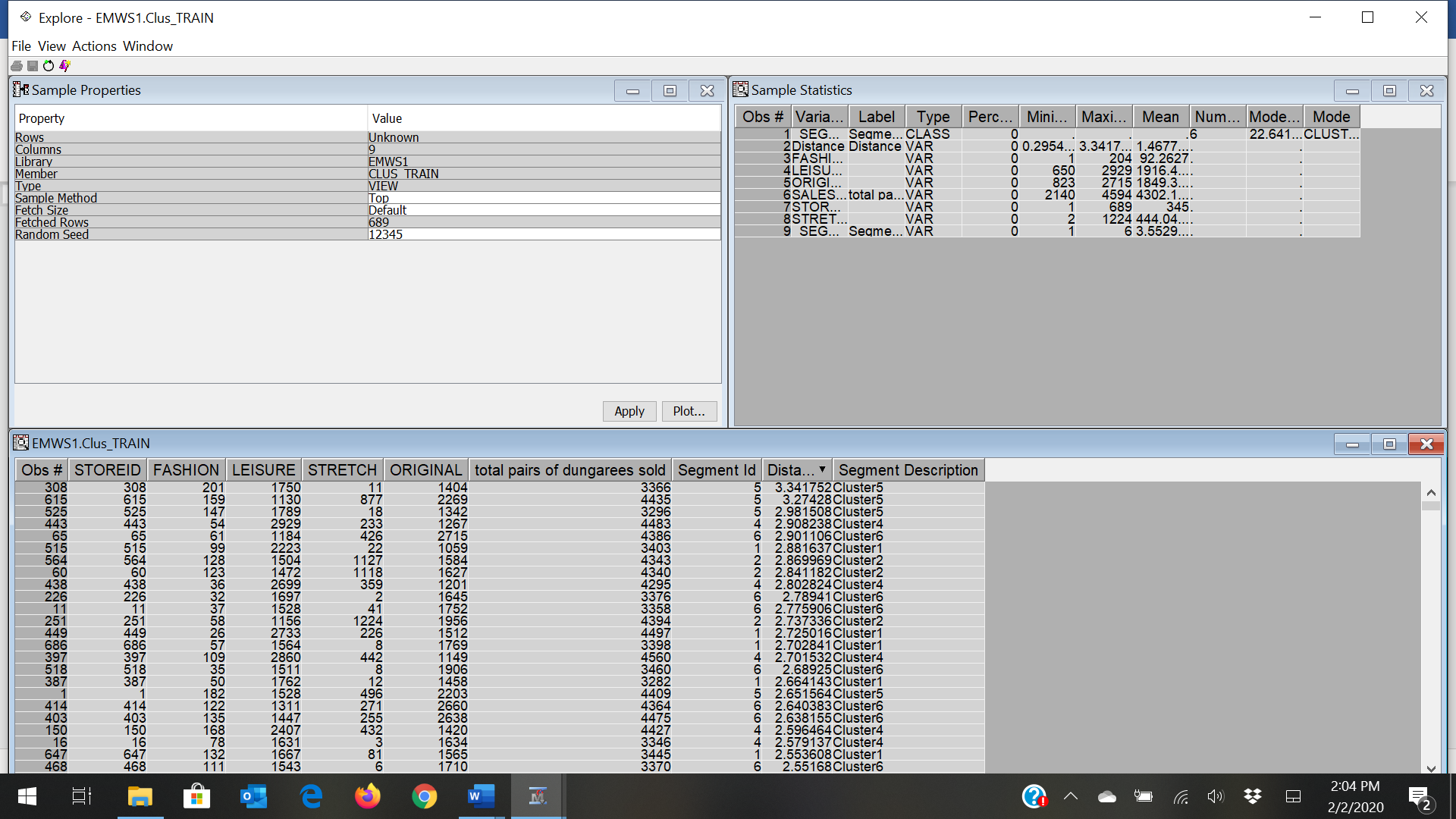
Input Means Plot



***Examining the Clusters***

Select the **Cluster** node in your process flow diagram. Click the ellipsis button next to the **Exported Data** property. The Exported Data — Cluster window appears. Click **TRAIN** and click **Explore**.

The Clus\_TRAIN window contains the entire input data set and three additional columns that are appended to the end. Scroll to the right end of the window to locate the **Segment ID**, **Distance**, and **Segment Description** columns. The **Segment ID** and **Segment Description** columns display what cluster each observation belongs to.



Next, create a plot that compares the number of hits for each cluster.

On the main menu, select **Actions** **Plot**.

In the Select a Chart Type window, click **Box**. Click **Next**

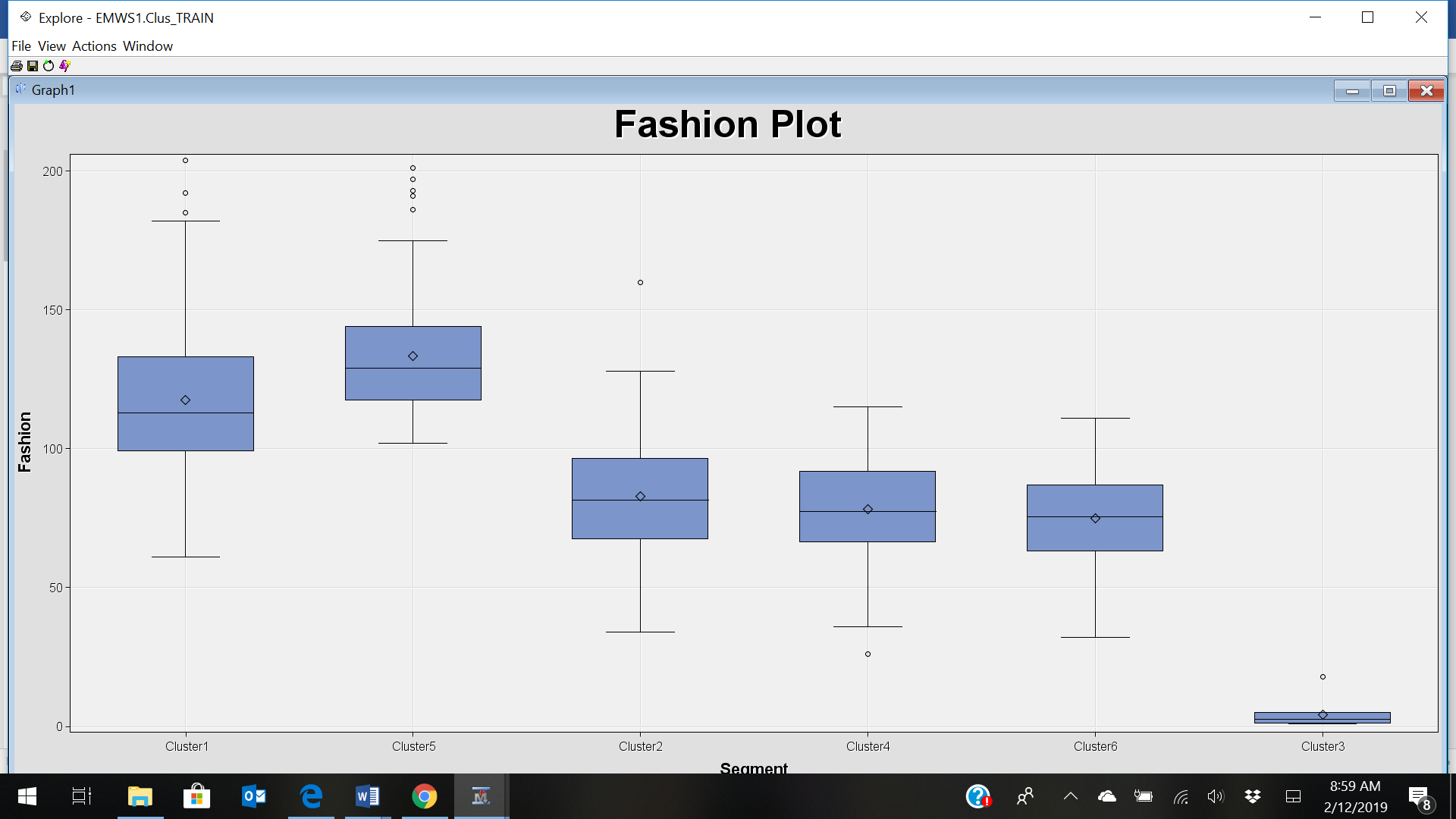
In the Select Chart Roles window, find the \_SEGMENT\_LABEL\_ variable. Set the **Role** for \_SEGMENT\_LABEL\_ to **X**. Next, find the FASHION variable. Set the **Role** for FASHION to **Y**. Click **Next**.

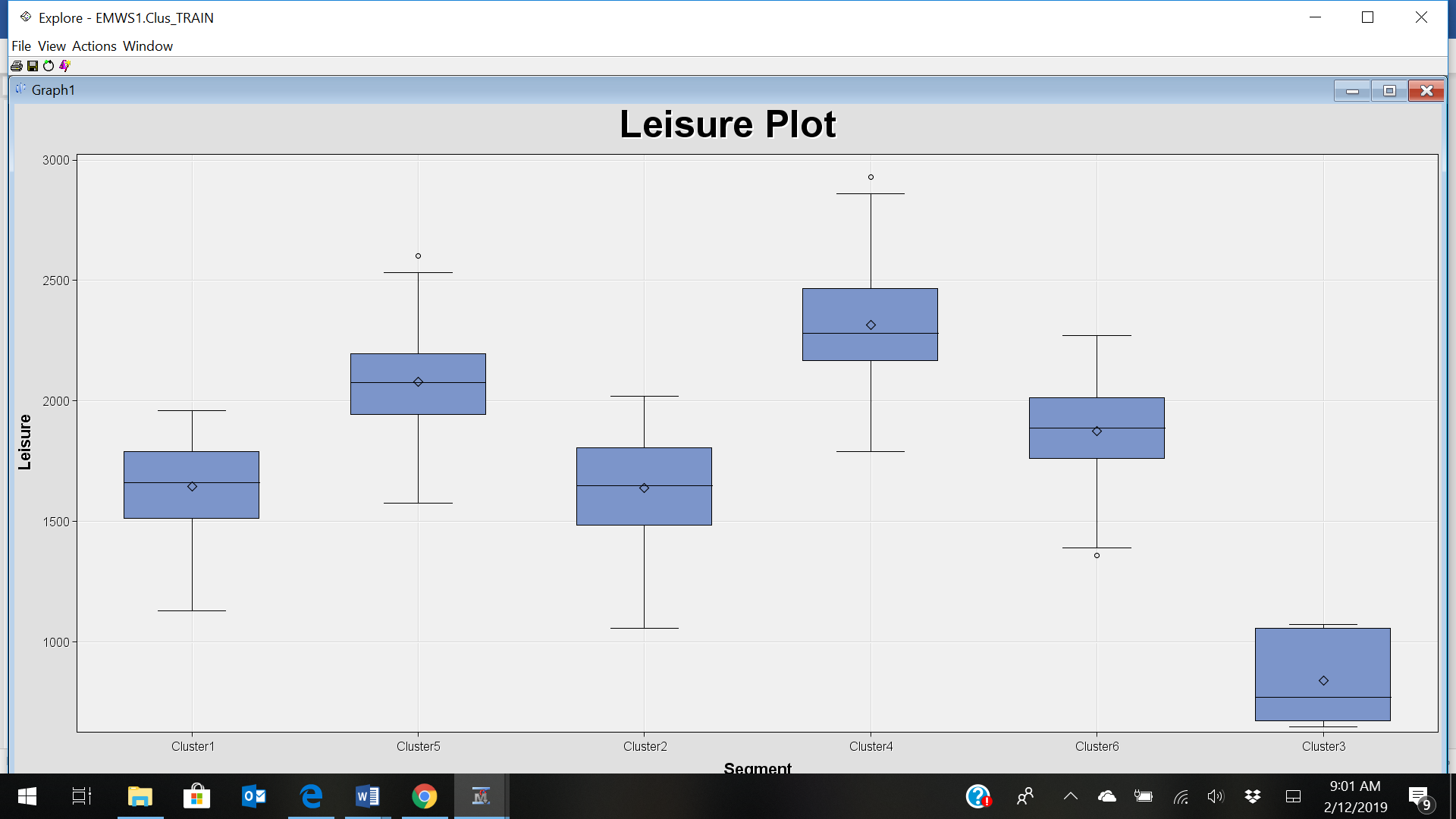
In the Data WHERE Clause window, click **Next**.

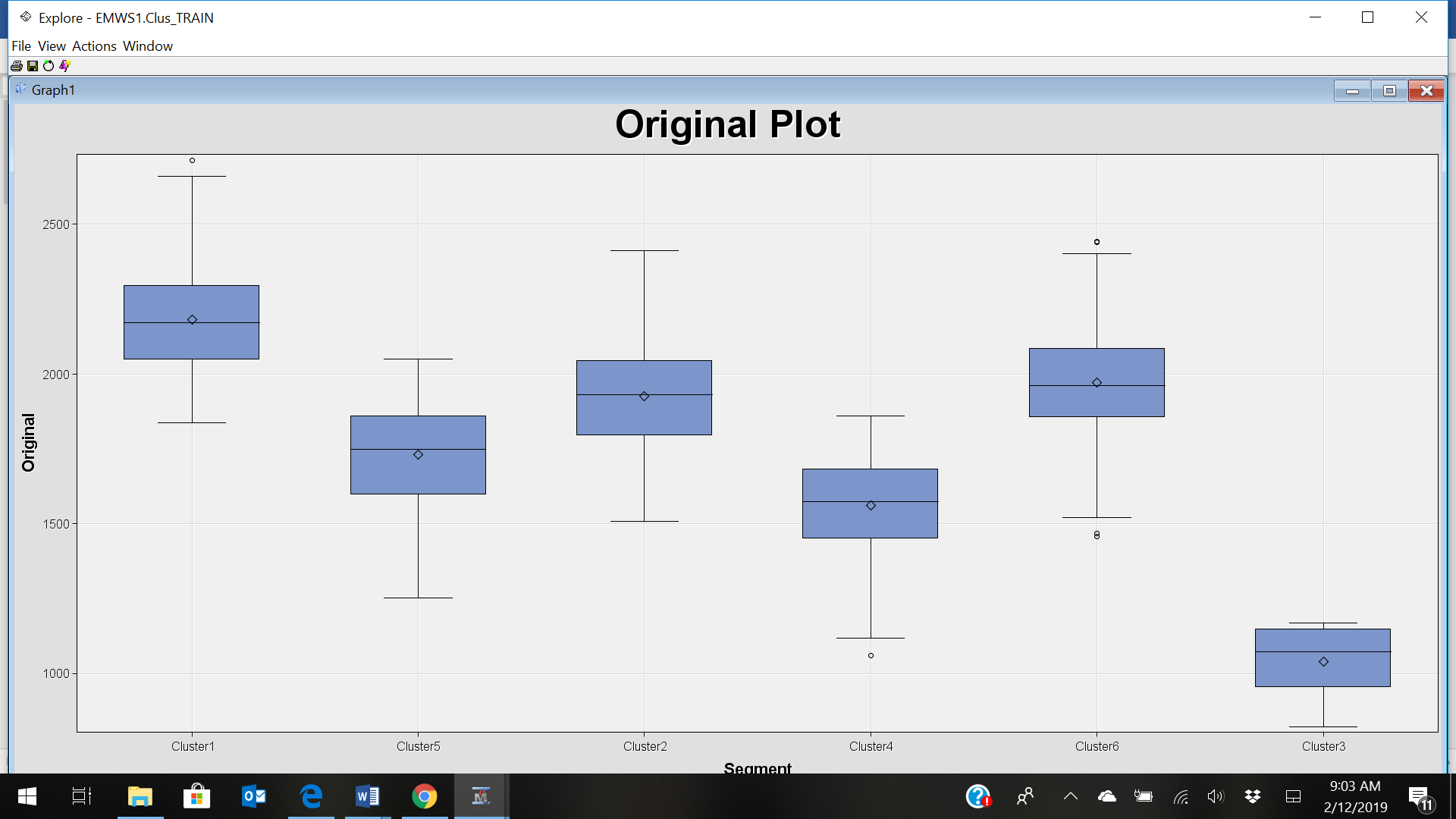
In the Chart Titles window, enter **Fashion Plot** in the **Title** field. Leave the **Footnote** field blank. Enter **Segment** for the **X Axis Label**. Enter **Fashion** for the **Y Axis Label**. Click **Next**.

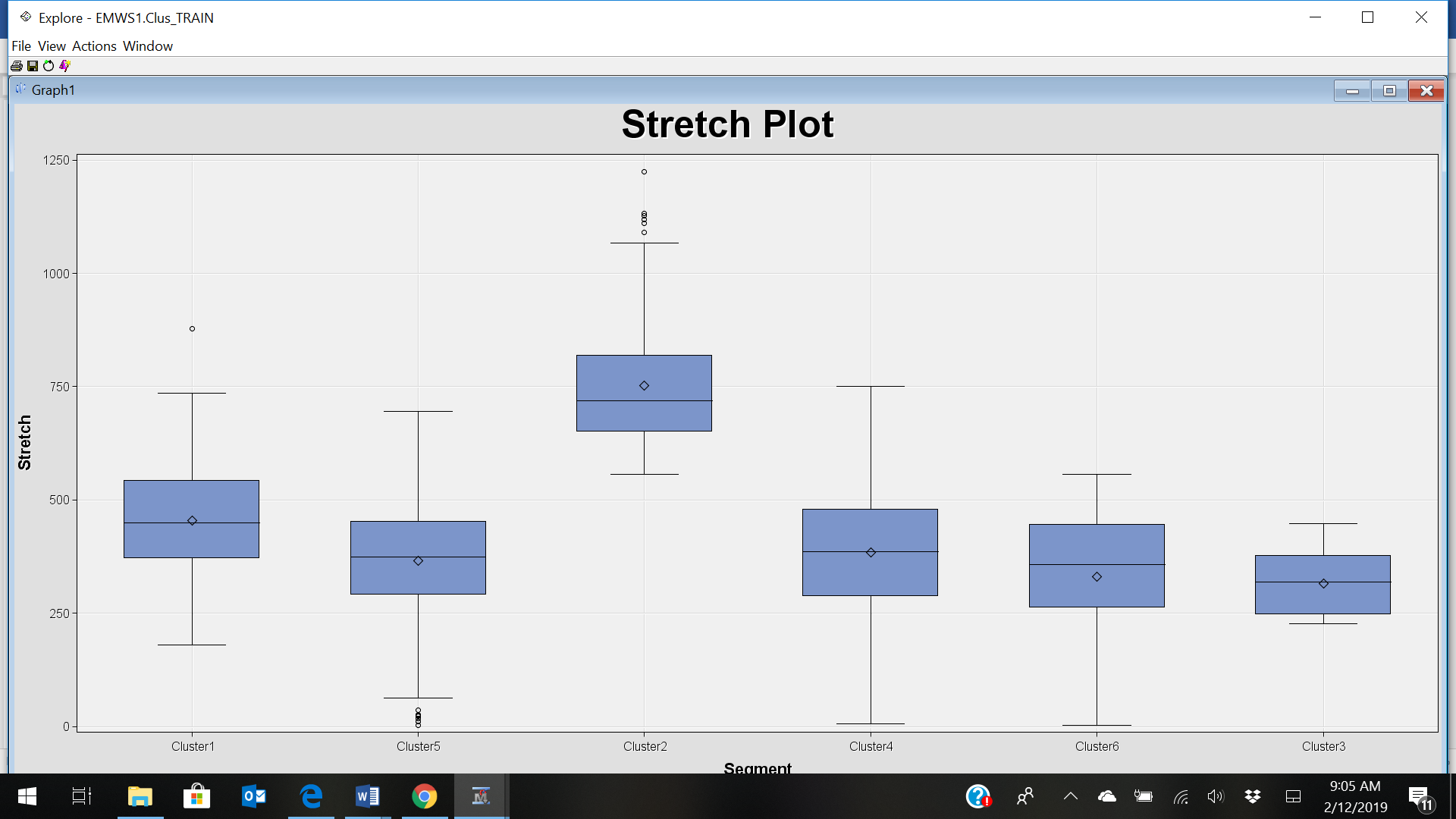
In the Chart Legends window, click **Finish**.

Box Plots









How does the number and quality of clusters compare to that previously obtained? There is a reasonable number and they depict logical clusters that will better portray the type of each product purchases.

Use the Segment Profile node to summarize the nature of the clusters. Box plots or description of each segment’s percentages

**Exercise Two**

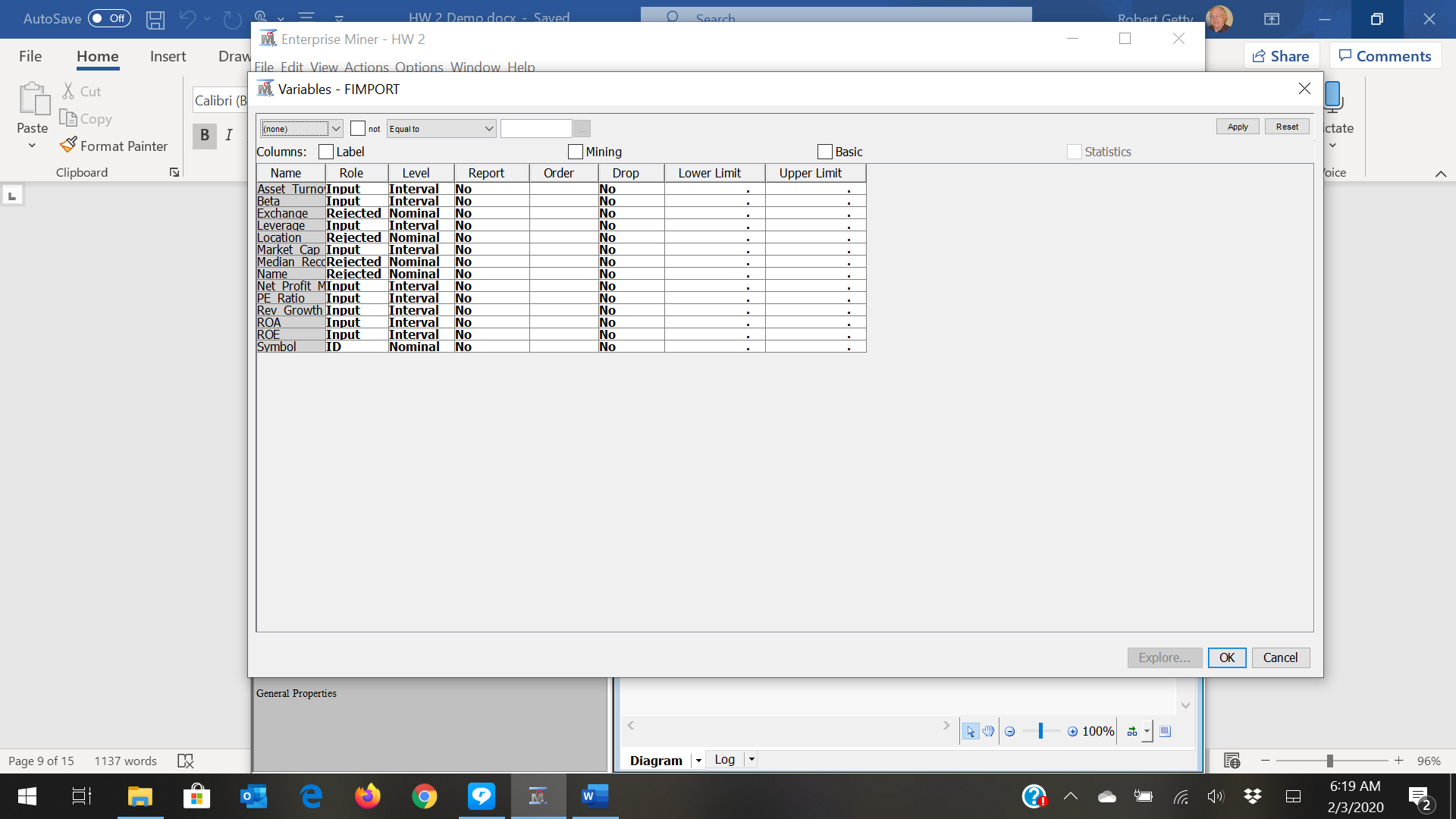
Financial data gathered on 21 firms in the pharmaceutical industry are available in the file **Pharmaceuticals.xls.**

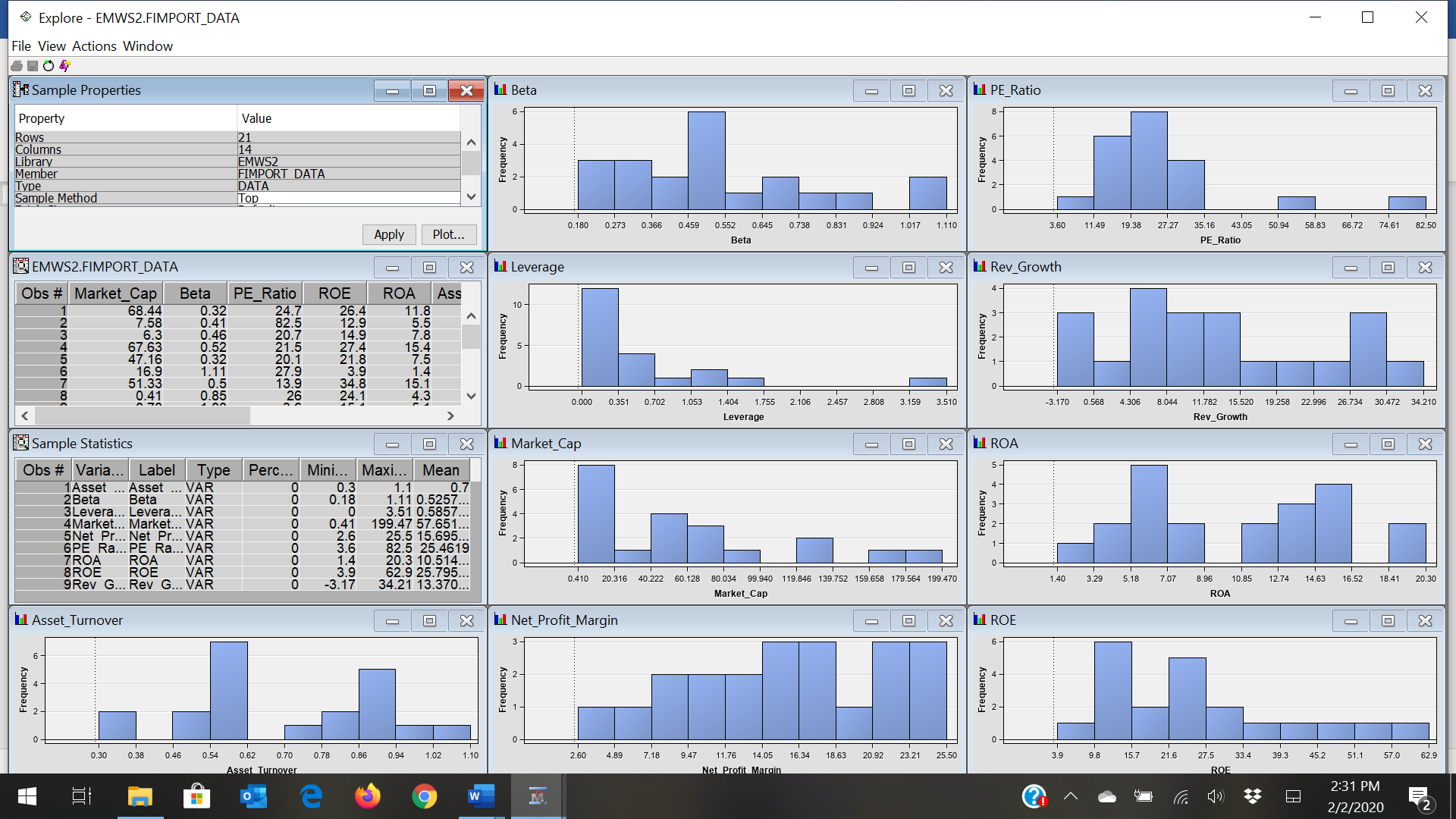
For each firm, the following variables are recorded.

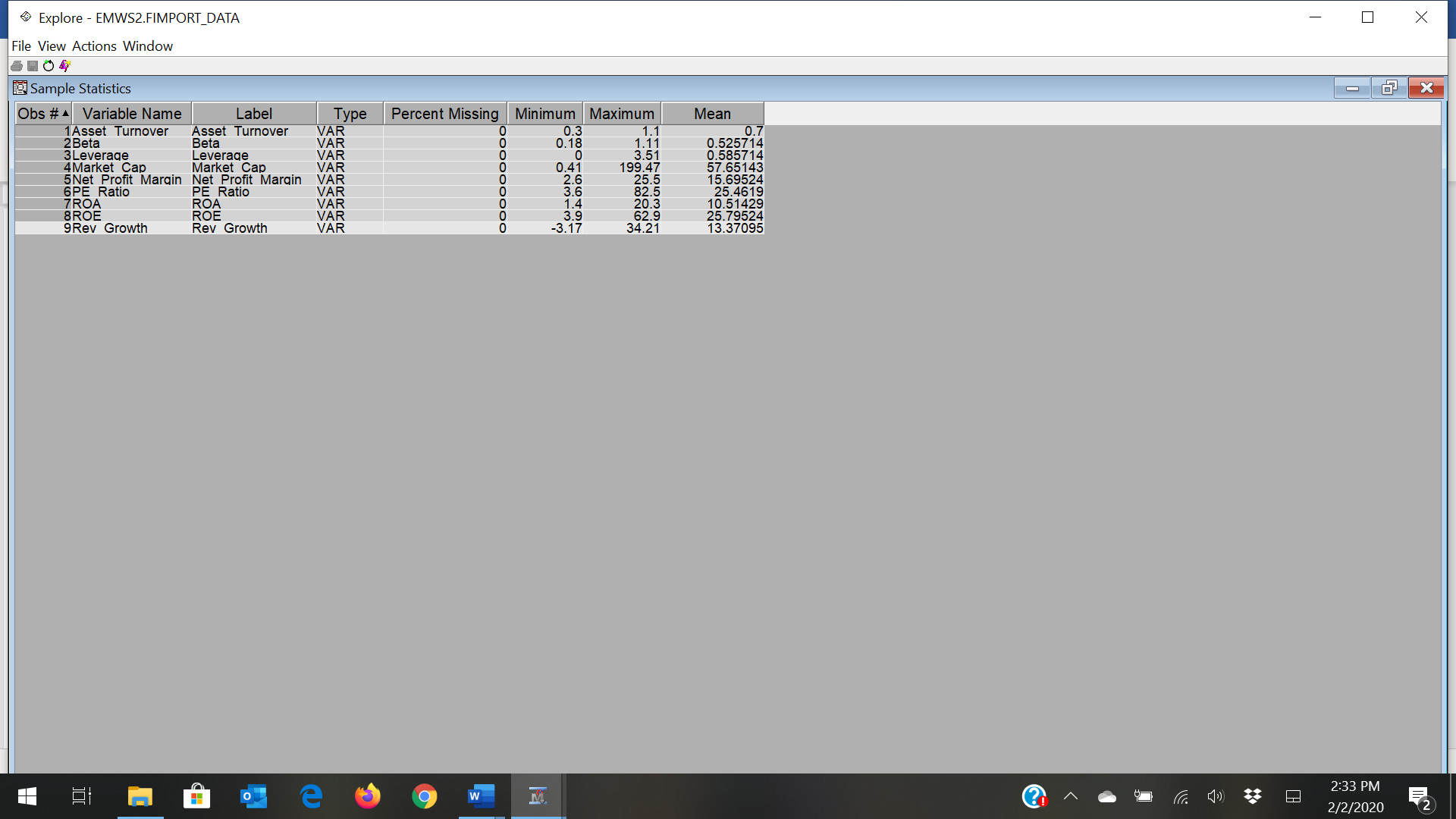
* 1. Market capitalization (in billions of dollars) (Mcap)
  2. Beta (Beta)
  3. Price/earnings ratio (PER)
  4. Return on equity (ROE)
  5. Return on assets (ROA)
  6. Asset turnover (AT)
  7. Leverage (Leverage)
  8. Estimated revenue growth (ERG)
  9. Net profit margin (NPM)
  10. Median recommendations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

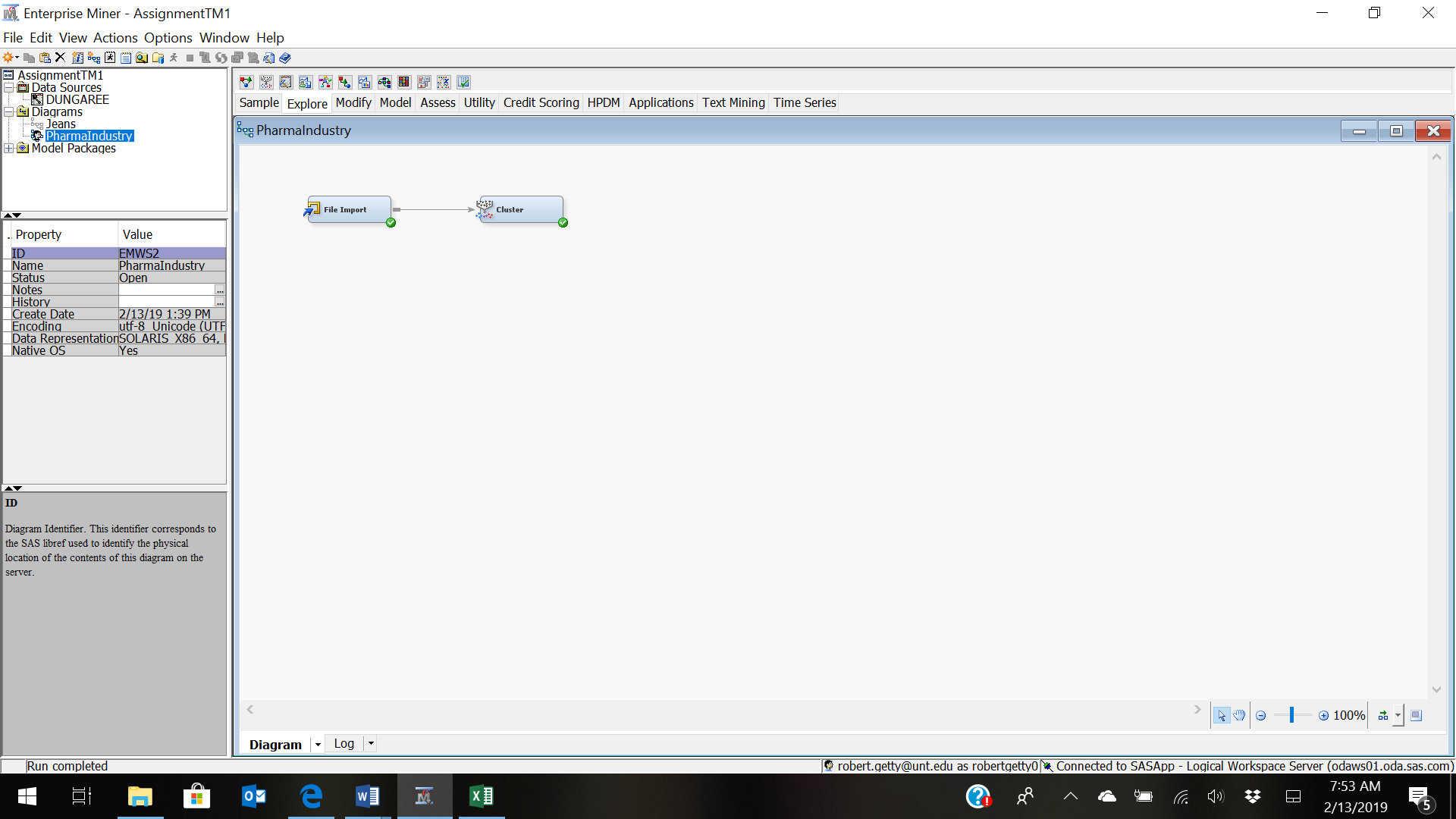
Use cluster analysis to explore and analyze the given dataset as follows (you need to use the File Import node as explained in Homework 1)

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).

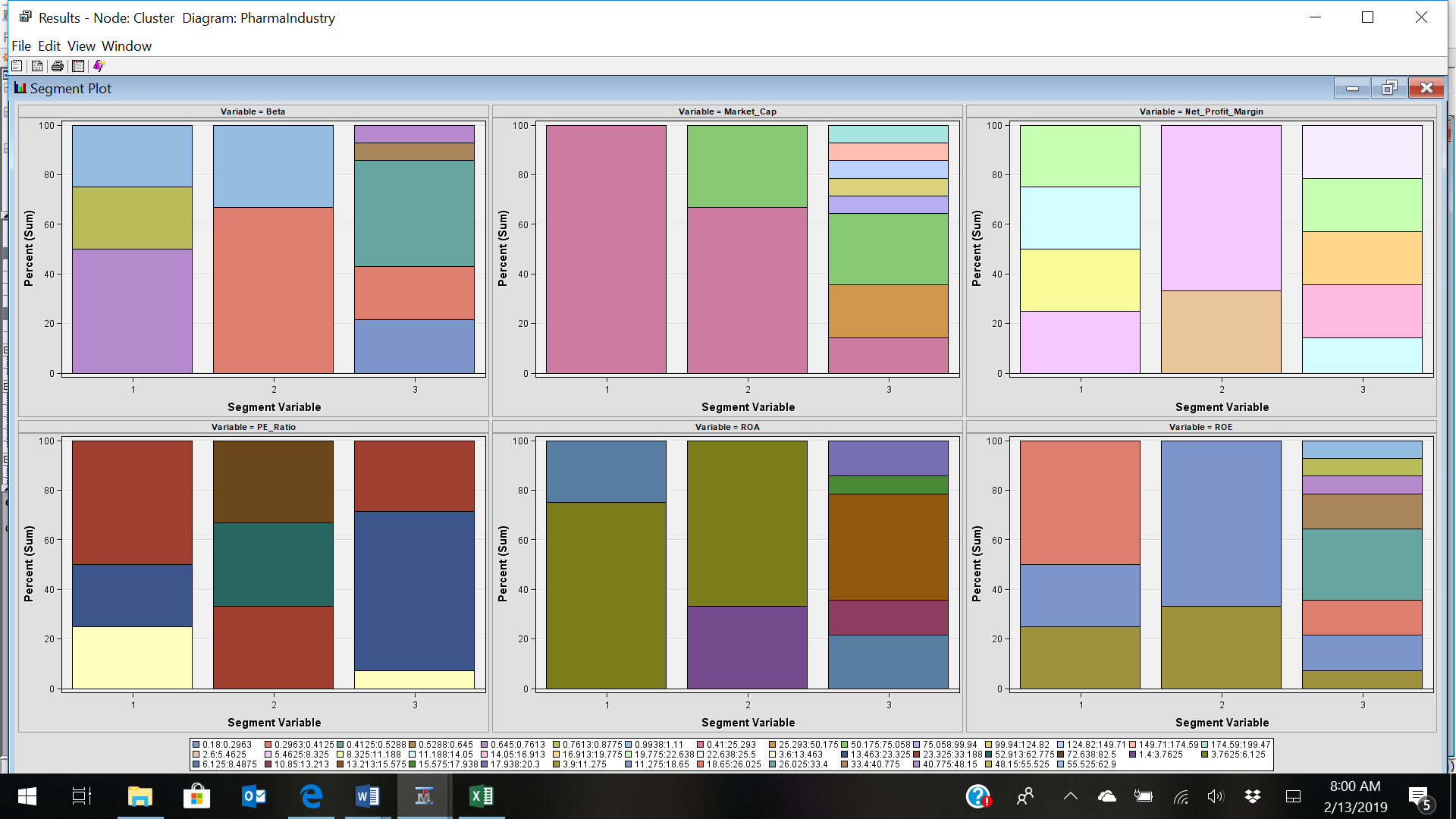


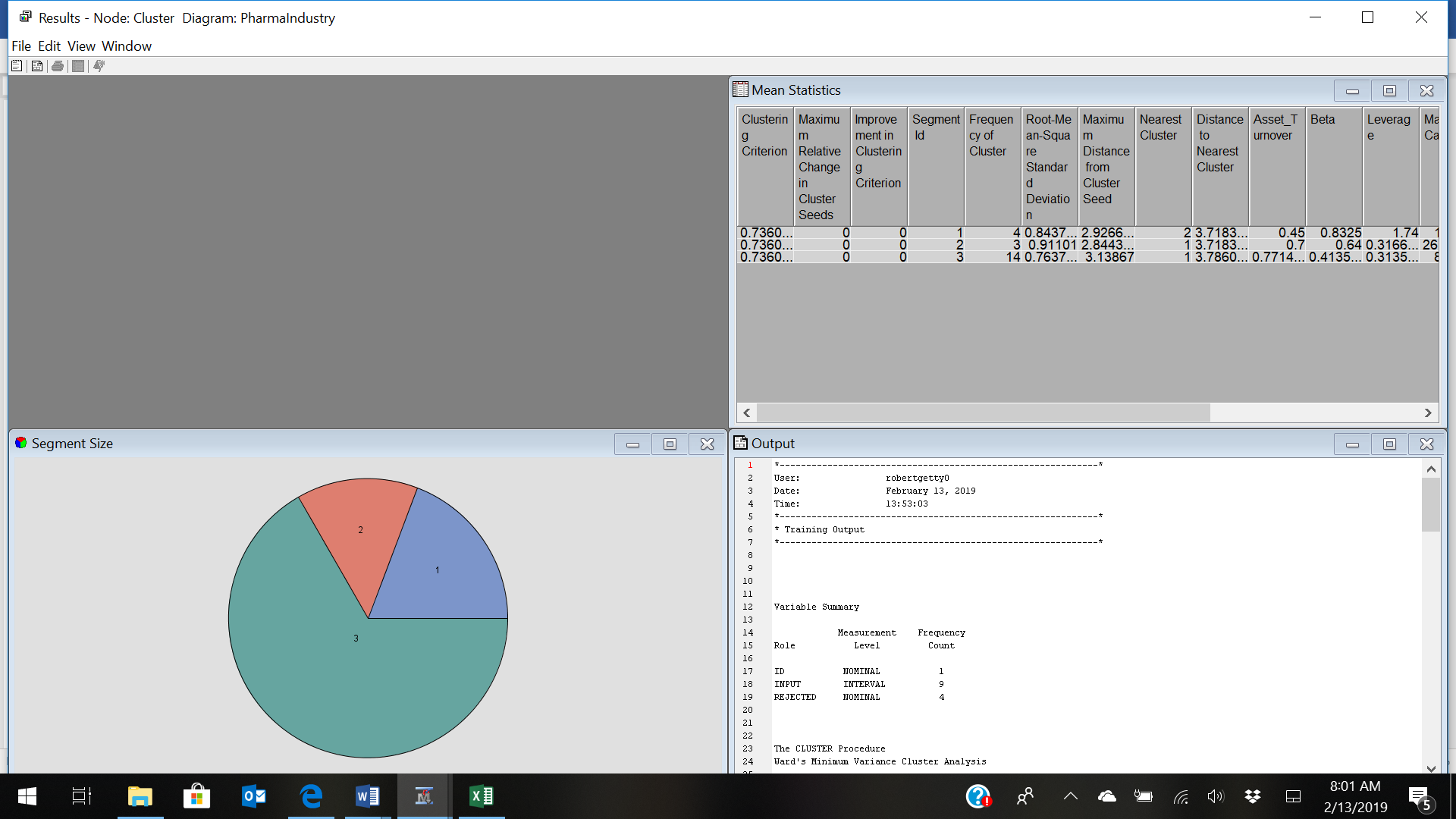


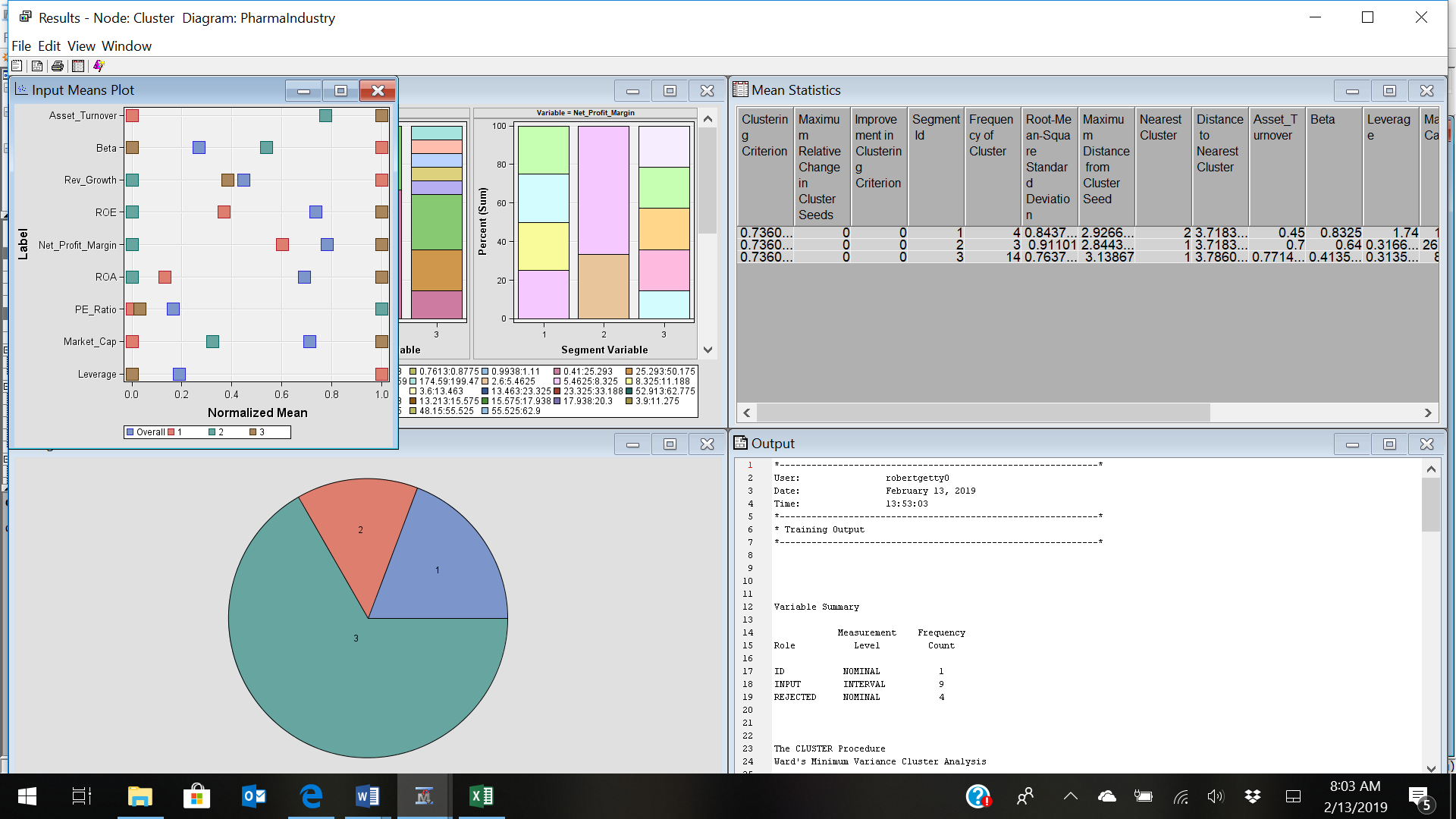


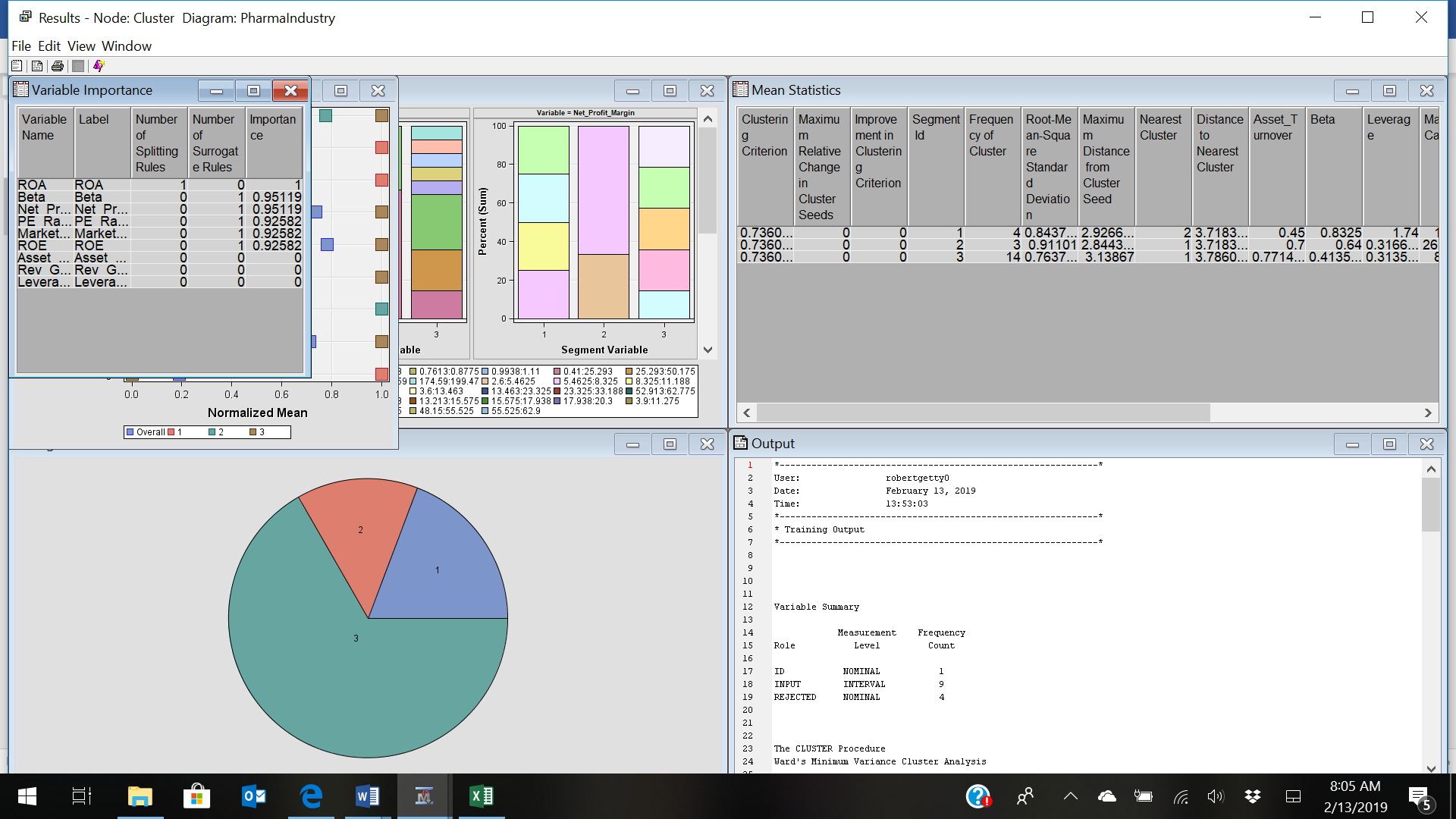


Automatic Cluster Results



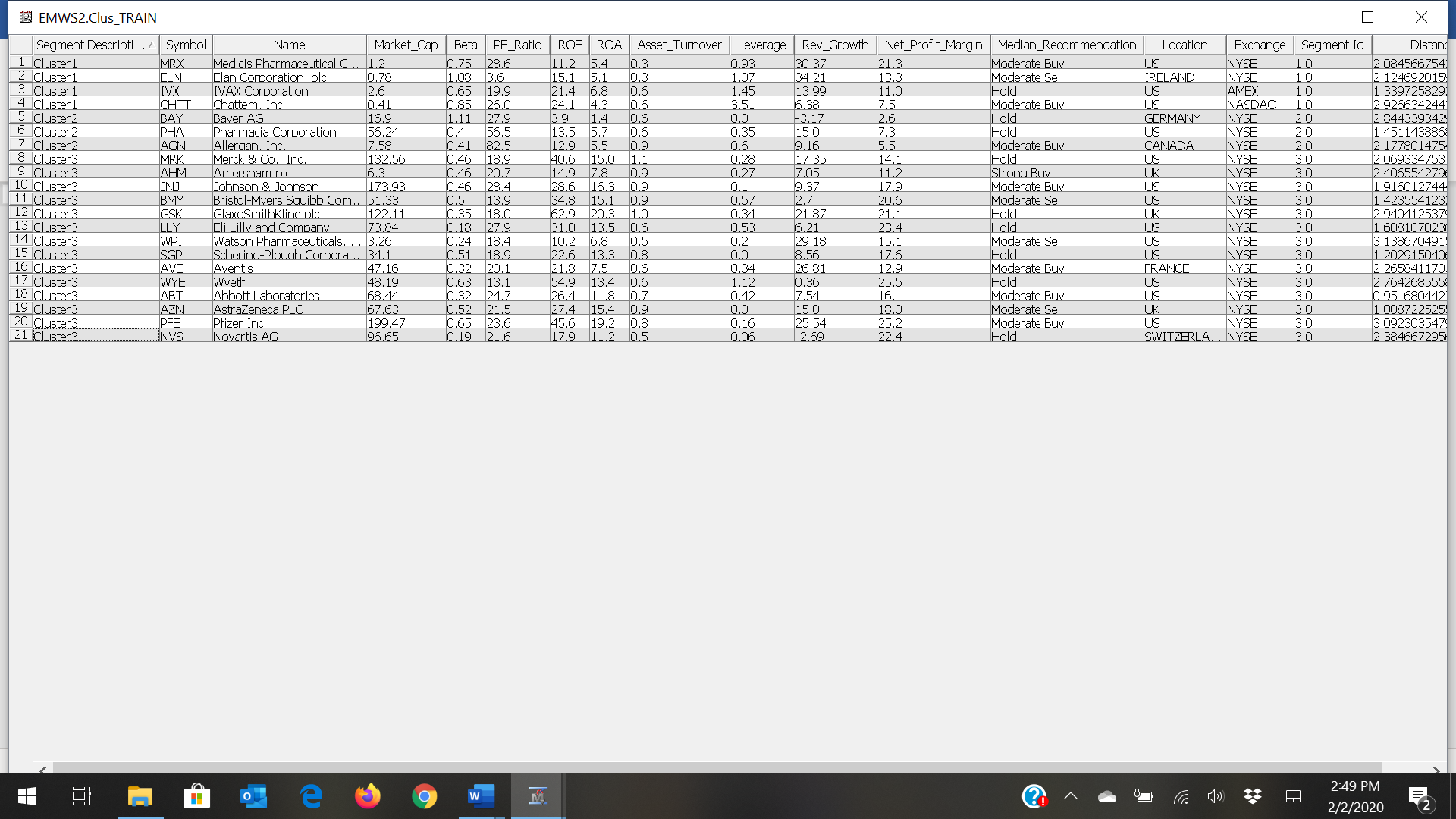




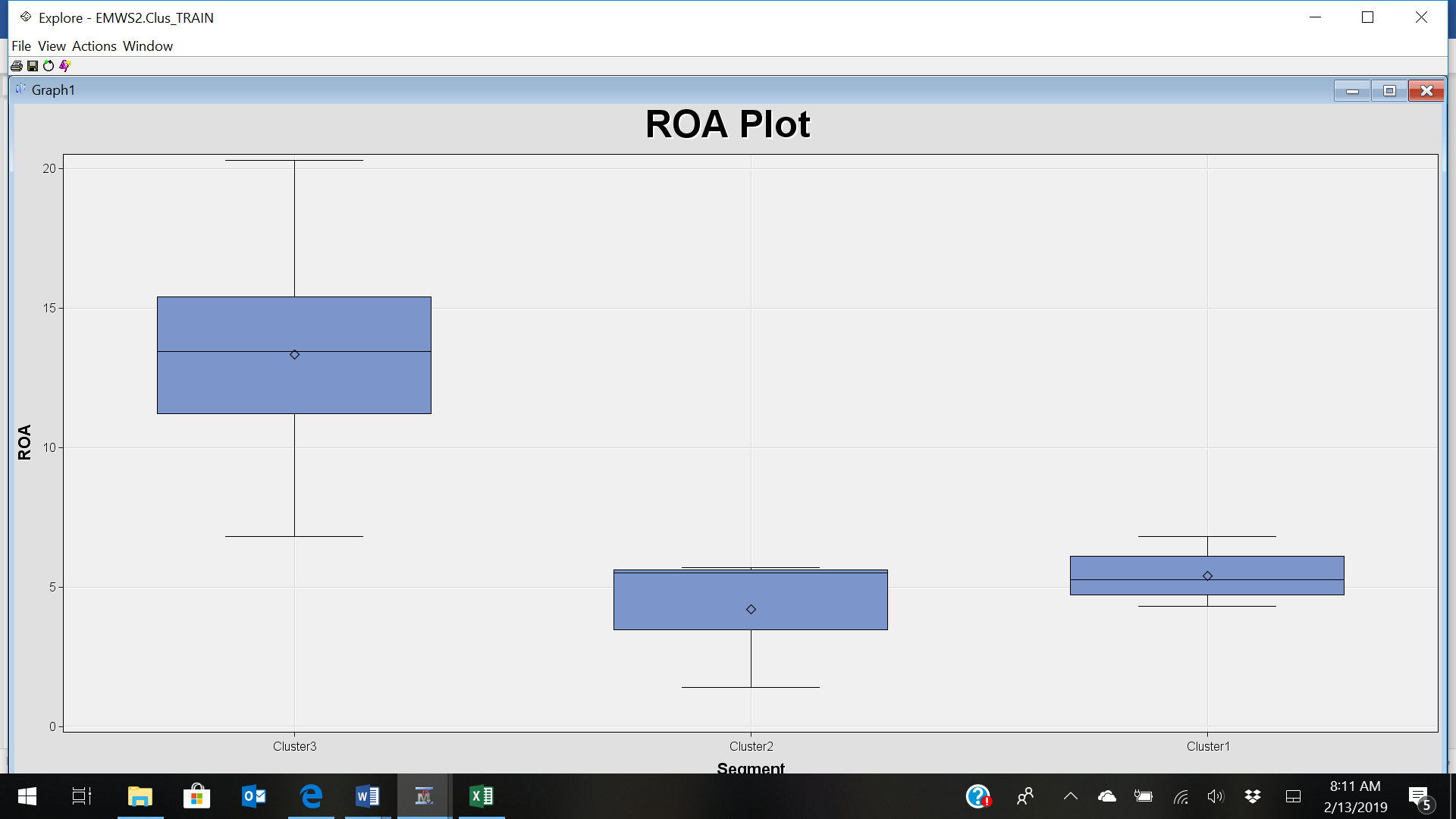


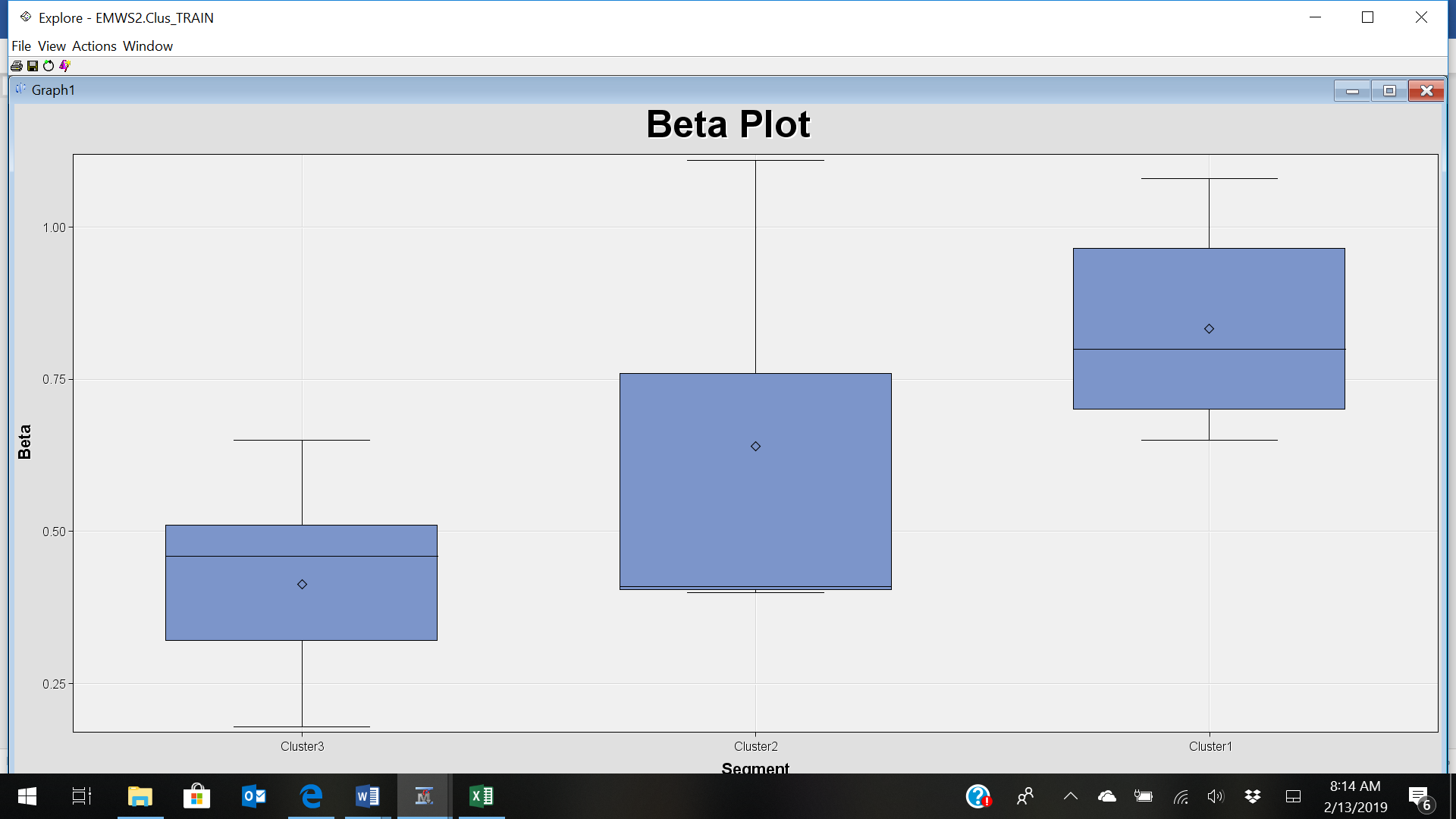
Interpret the clusters with respect to the quantitative variables that were used in forming the clusters.

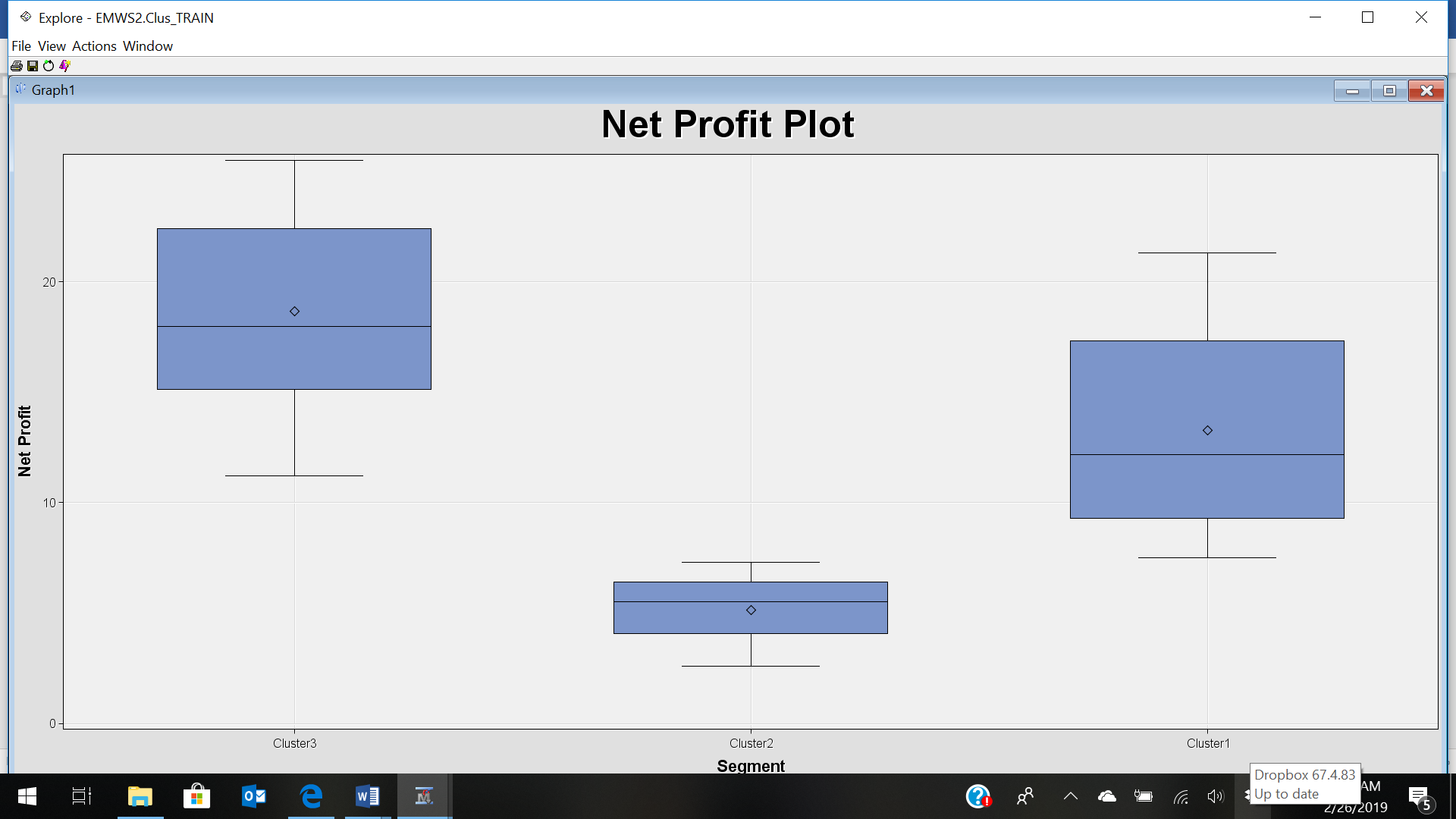
Select the **Cluster** node in your process flow diagram. Click the ellipsis button next to the **Exported Data** property. The Exported Data — Cluster window appears. Click **TRAIN** and click **Browse**.

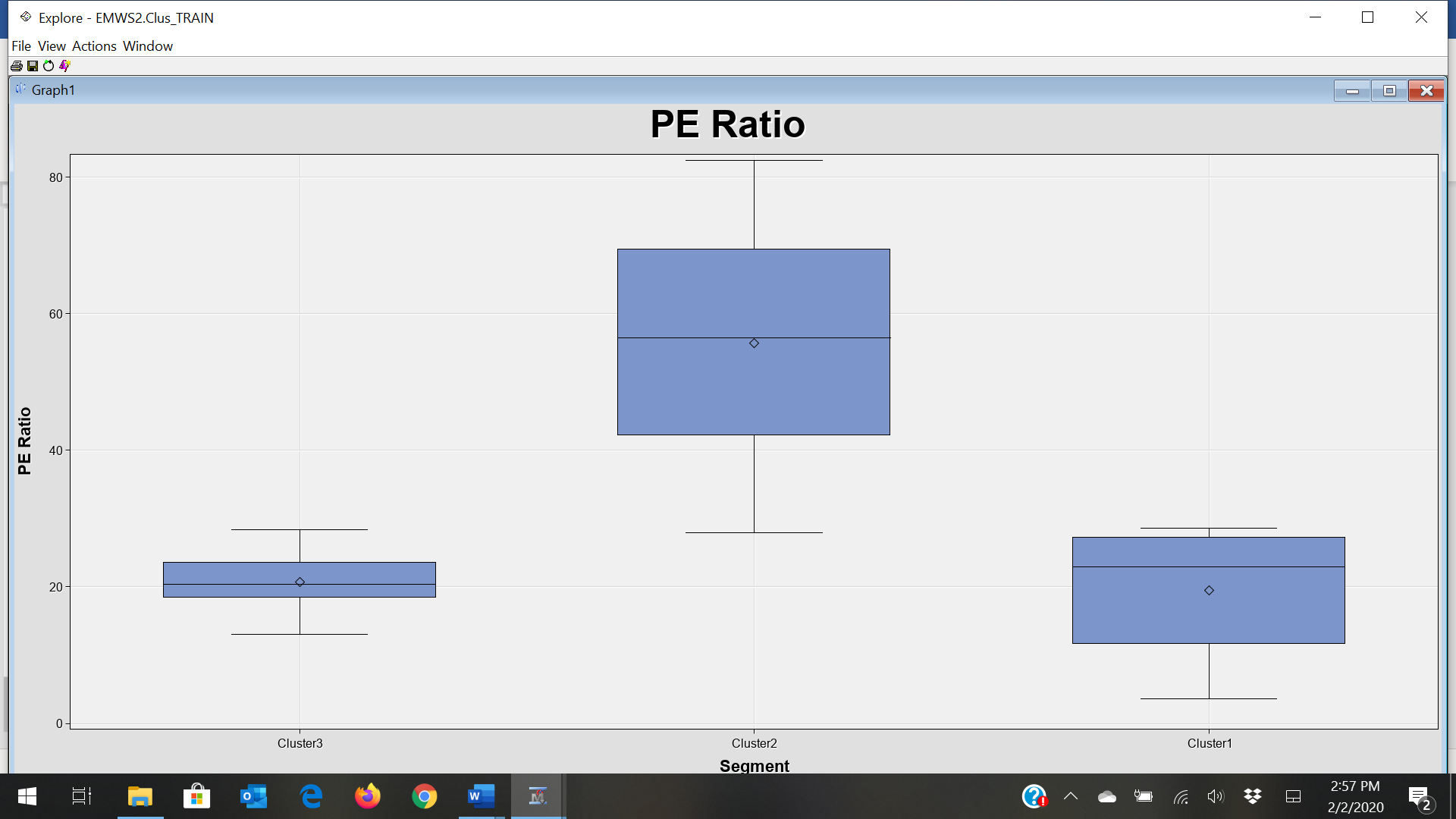


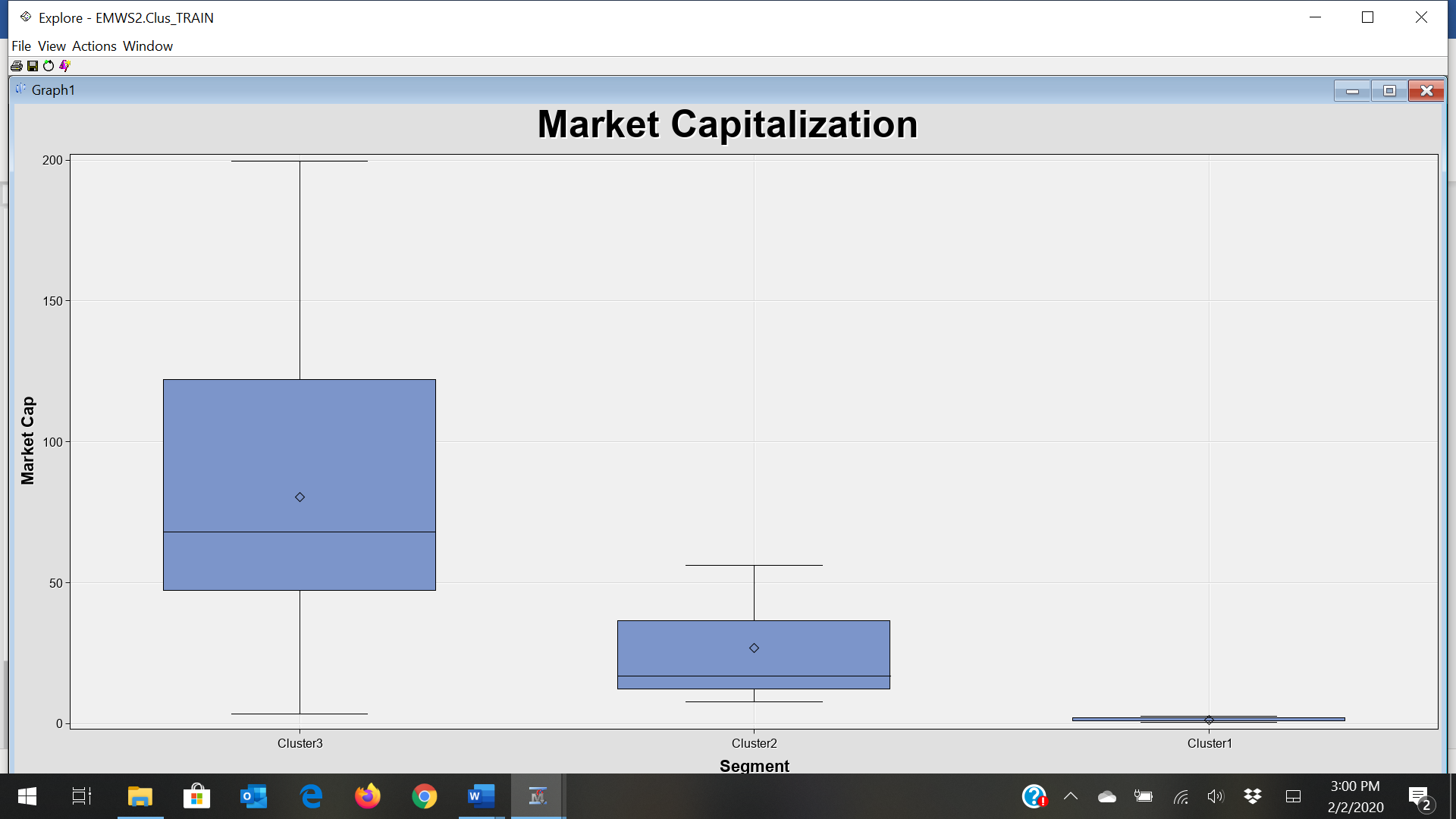
See the following boxplots:

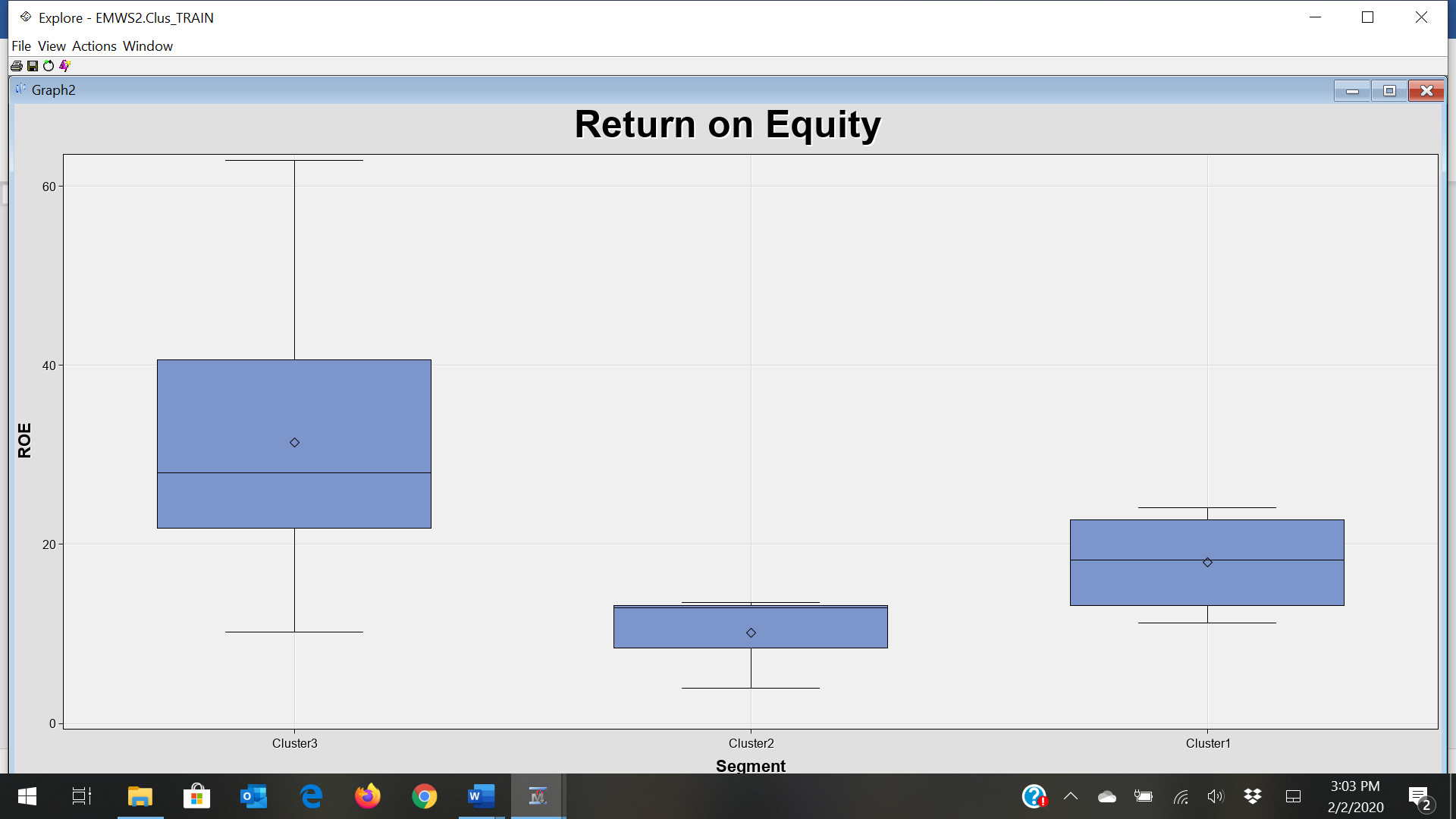












Do the clusters formed seem reasonable? Try different numbers of clusters and examine the results. Feel free to experiment with other criteria as needed. Explain the reasons for your selections, and identify the best clustering in your opinion (justify)