**HOMEWORK 02**

**Group 09:** Anji Lanke, Chris Benton, Raheyma Khan, Srinivas Rao Kolla, Cody Holmes

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Exercise 1**

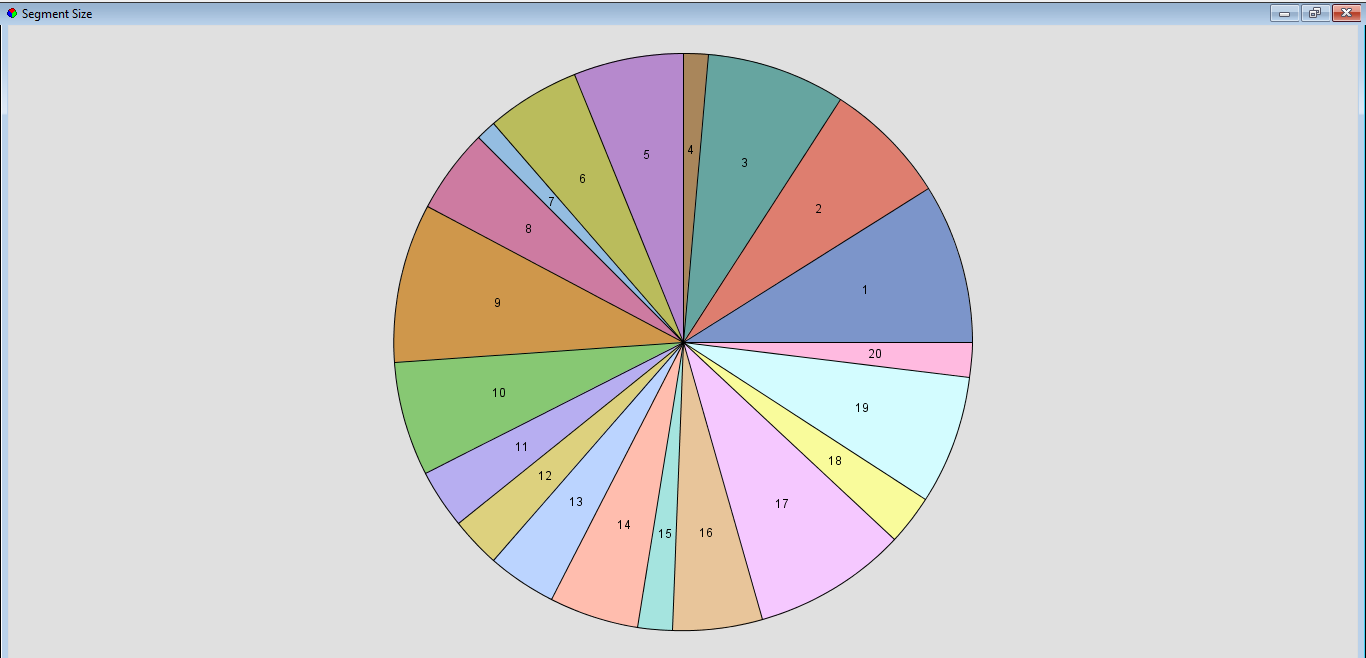
A screenshot of a computer

Description automatically generatedFor the first part of the assignment, we will use a DUNGAREE dataset. It has six variables: STOREID, FASHION, LEISURE, STRETCH, ORIGINAL and SALESTOT, which describe four types of dungaree stores. The diagrams below show histograms and descriptive statistics for the six variables.

A screenshot of a cell phone

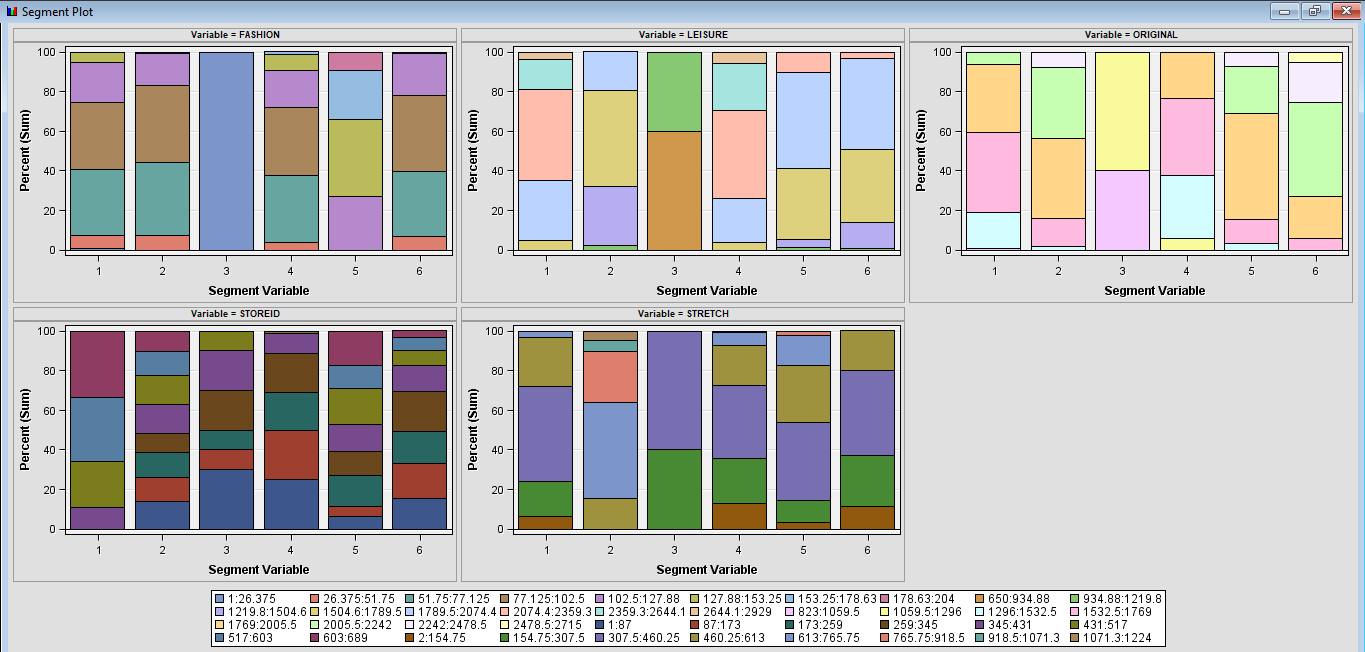
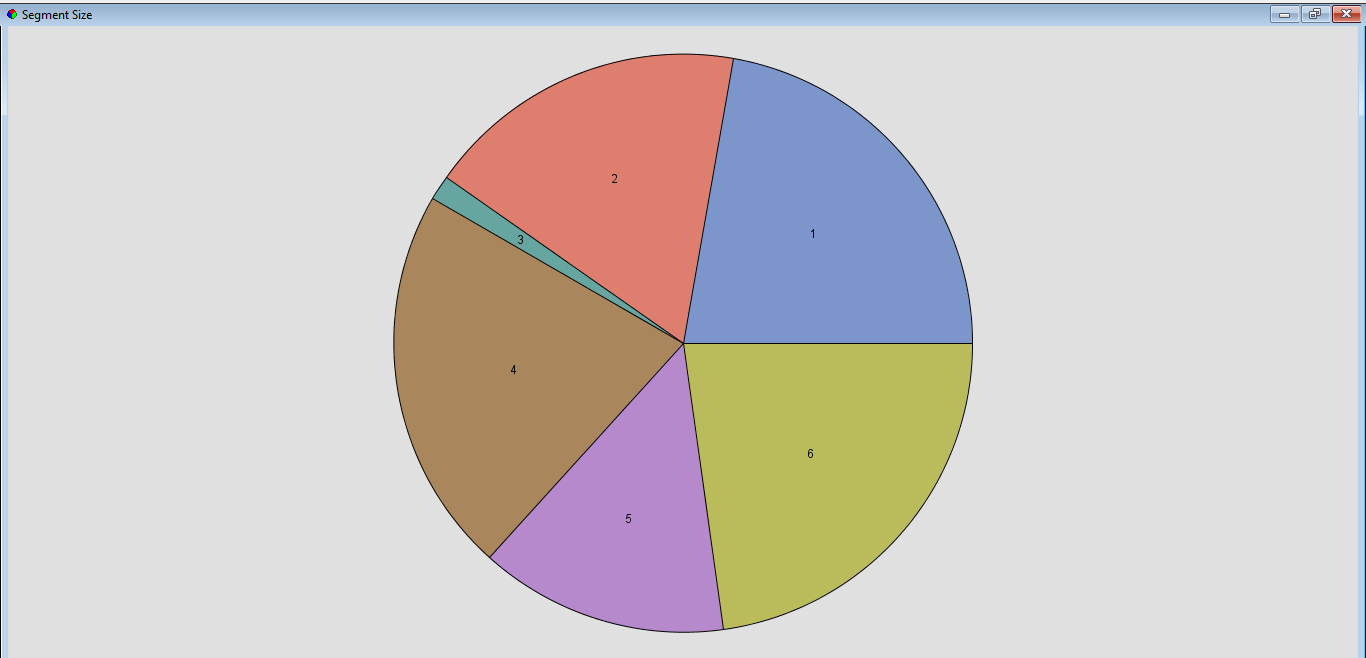
Description automatically generated

Looking at the diagrams, there do not seem to be any unusual or missing values in the dataset. So, we will move on and reject SALESTOT because it is the sum of the other input variables and will not add value to our model. We will also change STOREID’s role to ID. Then we will conduct a cluster analysis for our data and standardize the inputs. Standardization is important because it ensures that variables with larger ranges, LEISURE and ORIGINAL in our case, do not have larger effects than they actually should. Running this cluster analysis gives the following results.

A picture containing stationary, writing implement, pencil

Description automatically generated

The first cluster analysis results show an extremely high number of clusters. Therefore, we will again conduct the analysis but with a specification of only six clusters. The results are shown below.



As the diagrams show, six looks like a reasonable number of segments. Each segment, except one, looks well populated and the distance between the clusters also seems suitable. To take a closer look at these segments, we will use the Segment Profile function in SAS. We obtain the following histograms.

A screenshot of a cell phone

Description automatically generatedA picture containing indoor, shelf

Description automatically generatedA picture containing indoor, building, shelf

Description automatically generated

Segment 01 includes stores that sell a higher-than-average number of original jeans. Segment 02 has stores with higher-than-average sales of stretch jeans. Segment 03 has stores that sell much lower-than-average number of fashion, leisure and original jeans, and no stretch jeans. Segment 04 includes stores that sell higher-than-average number of leisure jeans. Segment 05 has stores with higher-than-average fashion jeans sales. Lastly, Segment 06 has slightly higher-than-average sales of original jeans while lower-than-average sales of stretch and fashion jeans. We then plot Variable Worth graphs shown below.

A picture containing screenshot

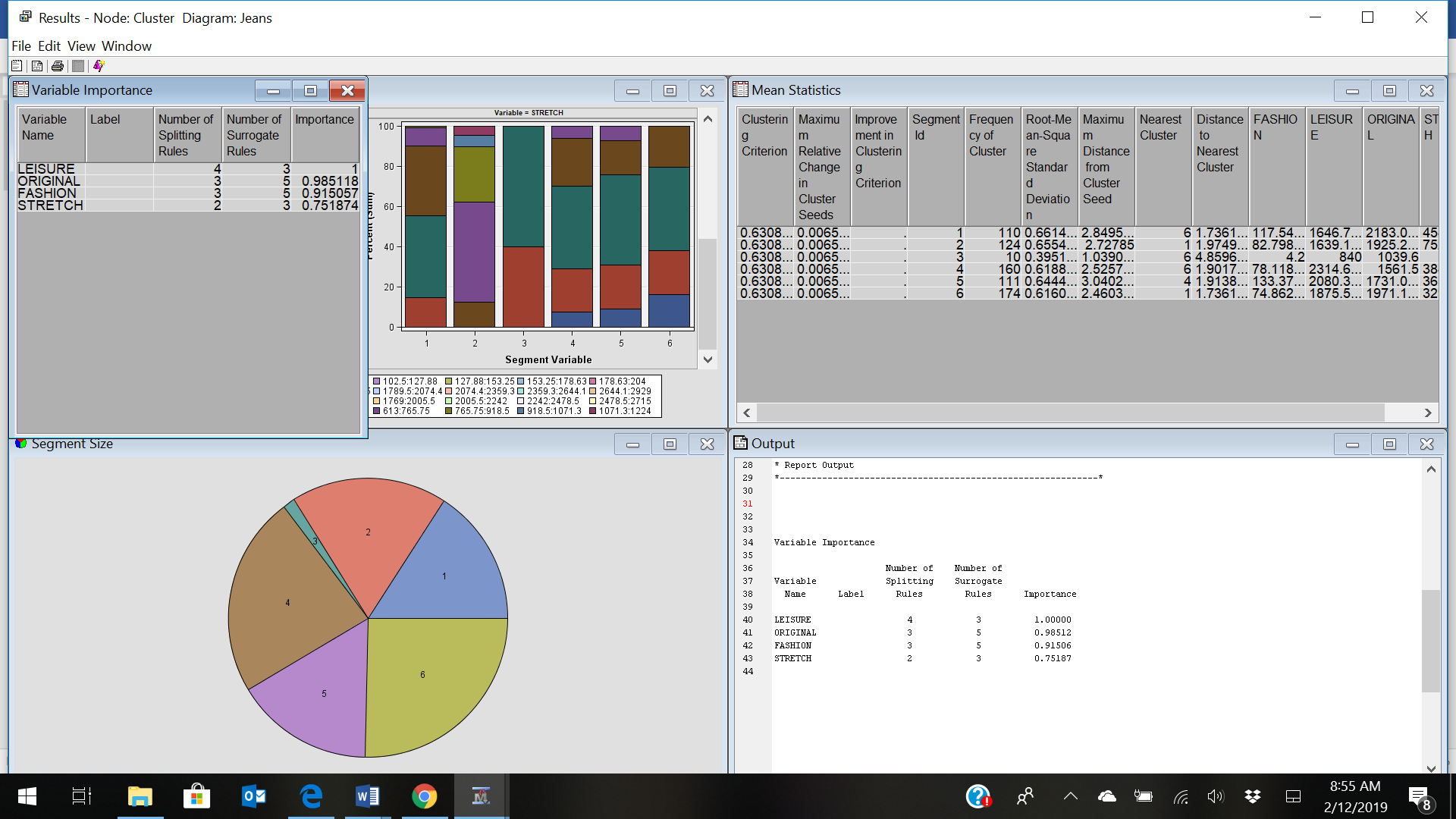
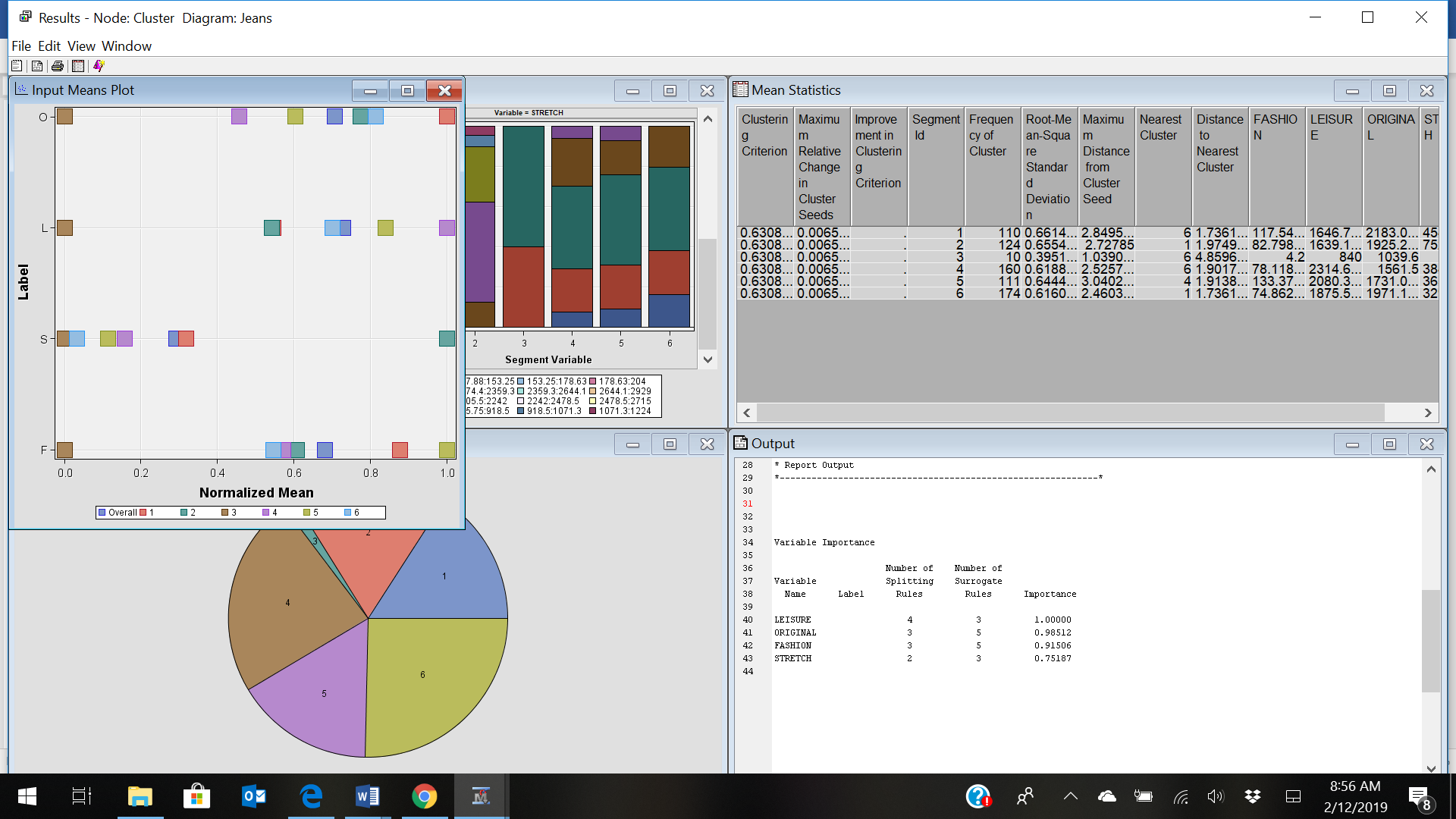
Description automatically generatedA picture containing screenshot

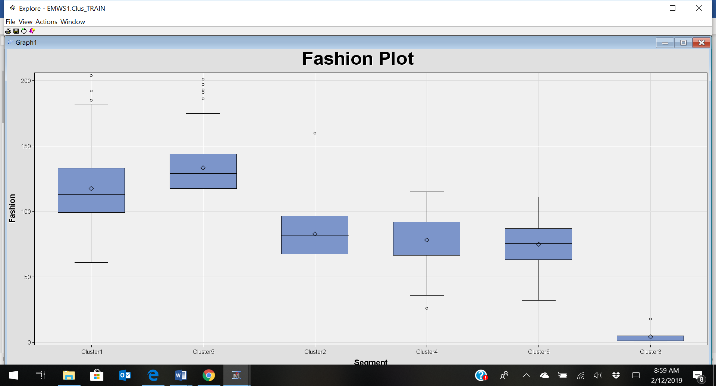
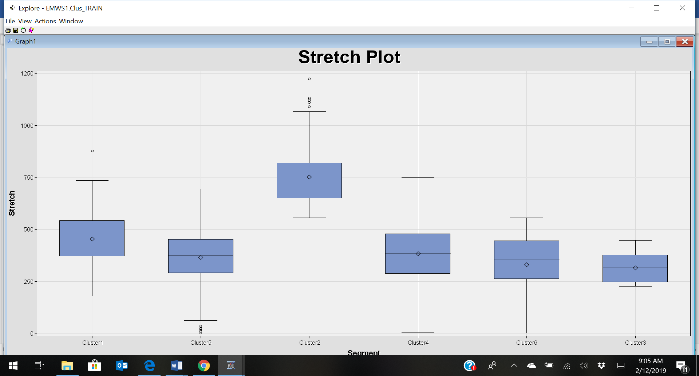
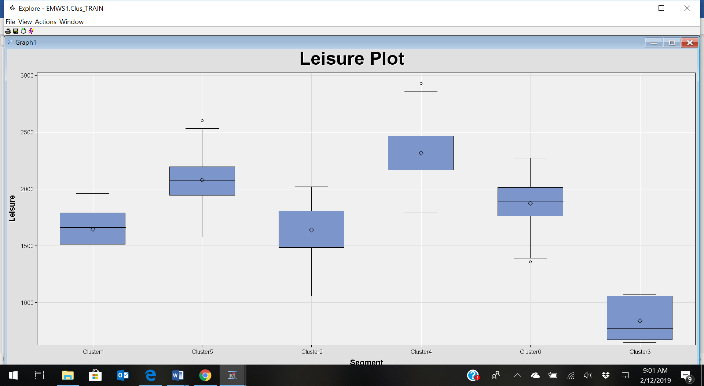
Description automatically generated

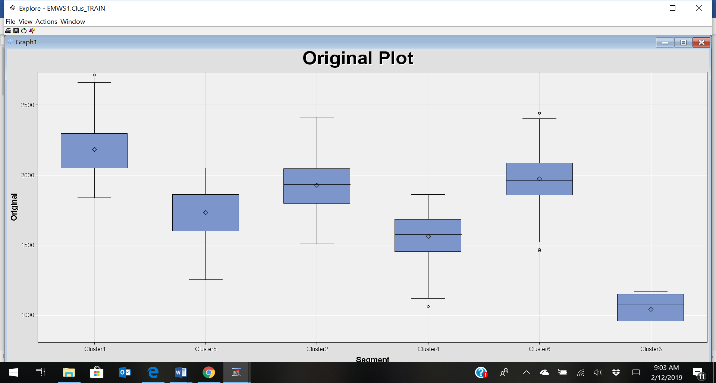
A picture containing screenshot

Description automatically generated

As we can see, Segment 01 has highest original jeans sales; Segment 02 has highest stretch jeans sales; Segment 03 has similarly low sales for fashion, leisure and original jeans, and no sales for stretch jeans; Segment 04 has highest leisure jeans sales: Segment 05 has highest fashion jeans sales; and Segment 06 has similar jeans sales for every variety but leisure is the highest.



The diagrams above show Variable Importance and Input Means Plot. Variable importance gives the overall importance of a variable in the model. We can see that LEISURE has the most importance or effect, but ORIGINAL and FASHION are not that far behind. STRETCH has the least importance in the analysis. The Input means plot displays the normalized mean value for each variable, both inside each cluster and for the complete data set. Cluster 3 has normalized mean that is always lower than the overall mean, which shows it has below average sales for each variable. Moreover, each variable has the highest normalized mean in the cluster where they have higher-than-average sales. For instance, Segment 01 has higher-than-average sales of original jeans, which leads to original jeans having the highest normalized mean in cluster 1.

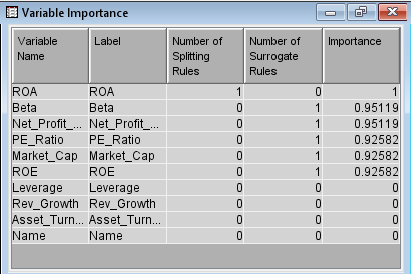
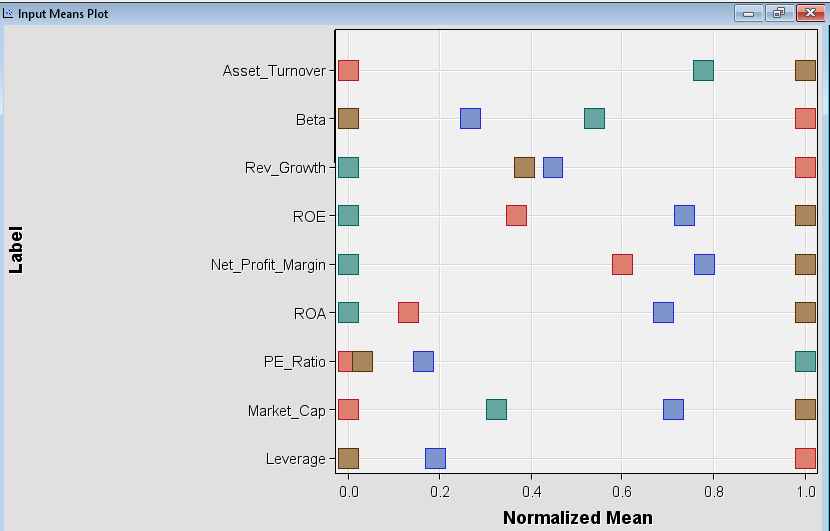
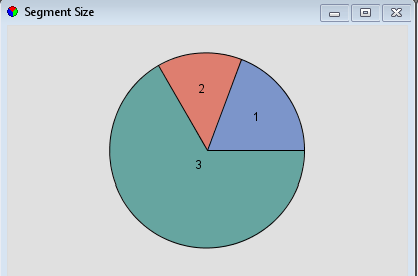
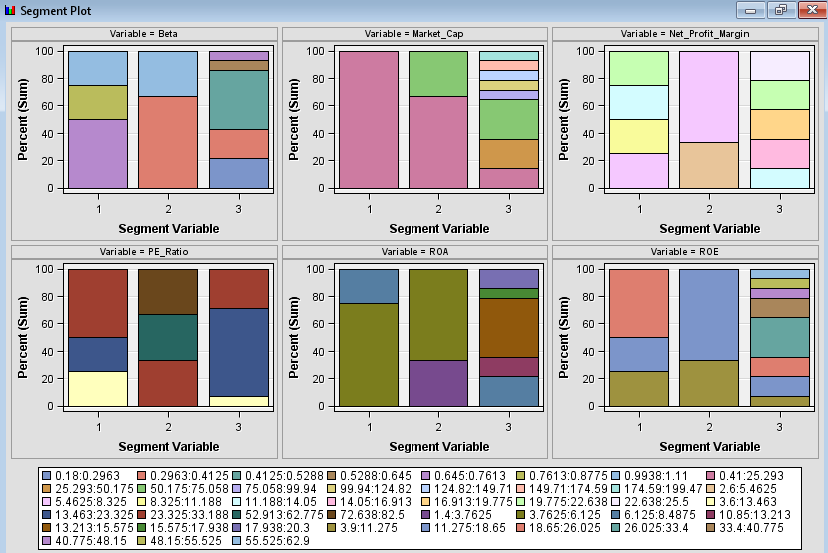


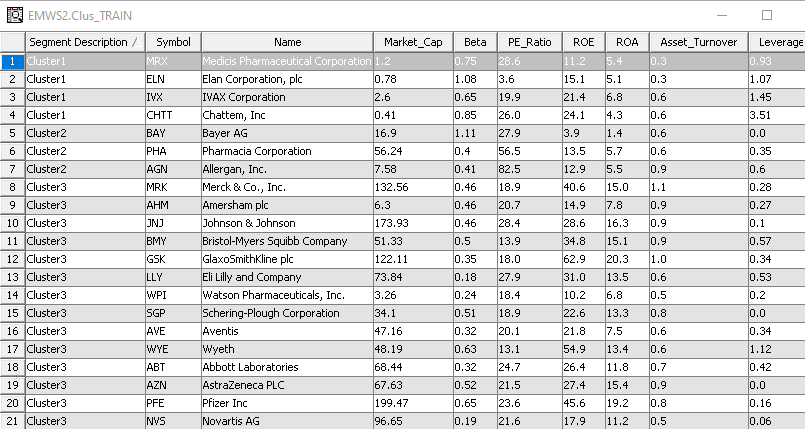
Lastly, we will use box plots to compare the independent variables in each cluster. The box plots are shown in the diagrams above. They agree with our variable worth results. Fashion jeans sales are highest in cluster 05, leisure jeans sales are highest in cluster 04, original jeans sales are highest in cluster 01, and stretch jeans sales are highest in cluster 02. Thus, we end our cluster analysis for the Dungaree dataset.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Exercise 2**

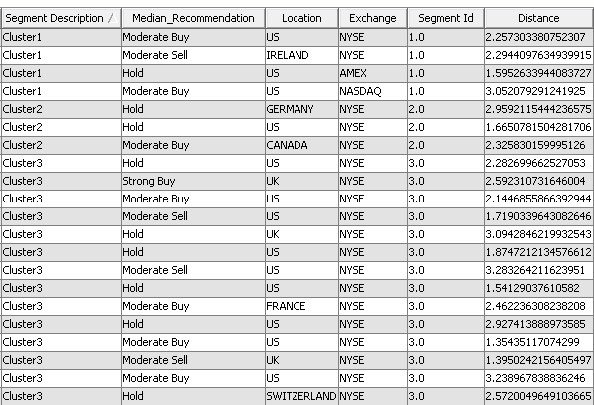
**Use only the quantitative variables (1-9) to cluster the 21 firms. Use the default settings in SAS Enterprise Miner.**





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

Segment 1 includes companies with high Leverage, Beta, and Revenue Growth. Segment 2 includes companies with high PE Ratios but low Net Profit Margin, ROE, ROA, Leverage, and Rev Growth. Lastly, Segment 3 includes companies that mostly have high Net Profit Margins, Market Cap, and Asset Turnover with Beta values more on the low end.

****

**Is there a pattern in the clusters with respect to the qualitative variables (10-12) (those not used in forming the clusters)?**

In Segment 2 & 3, all companies trade on NYSE. Three of the four locations in Segment 1 are US; however, the US location makes up almost 62% of the locations. As for the Median recommendations, Section 2 has only Hold or Moderate Buy. There are no other prominent patterns for Median recommendations.

**Provide an appropriate name for each cluster using any or all of the variables in the dataset. Don’t describe the cluster, name it.**

Section 1 – Companies with Potential for Growth

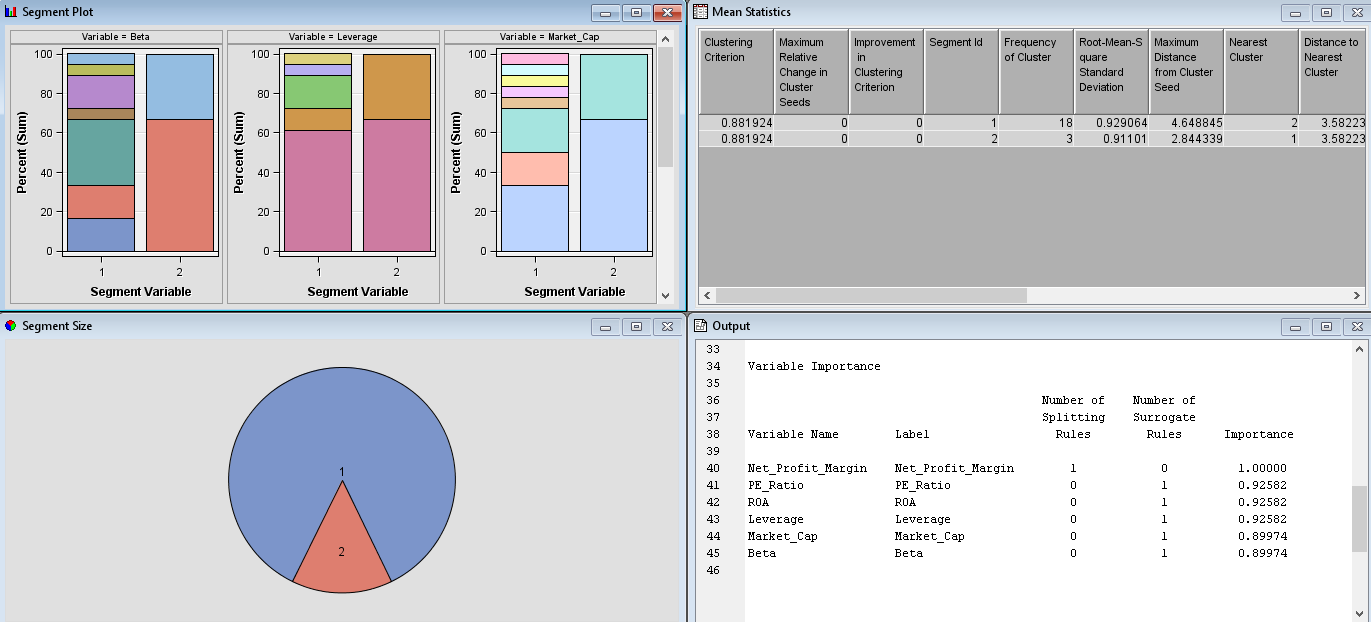
Section 2 – Companies with Top Stock Market Value (or Price/Earnings Ratio)

Section 3 – Companies with Top Earnings

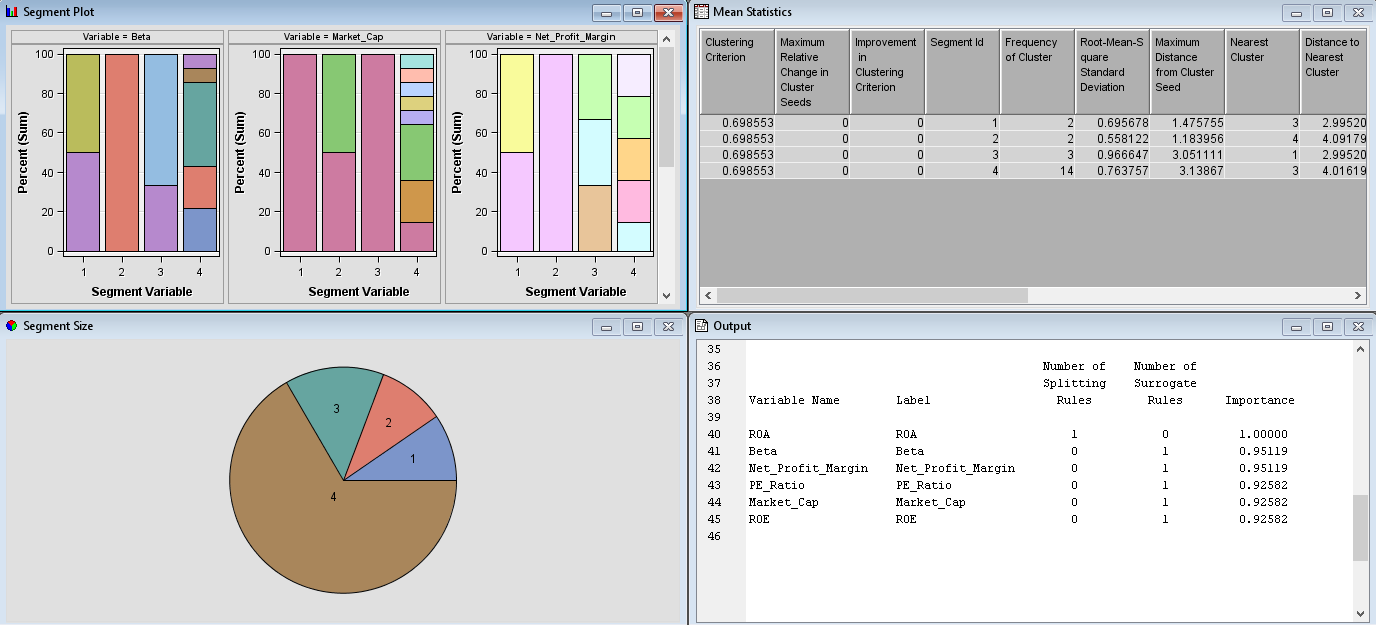
**Do the clusters formed seem reasonable? Try different numbers of clusters and examine the results. Feel free to experiment with other criteria as needed. Explain the reasons for your selections and identify the best clustering in your opinion (justify).**

Yes, three clusters are optimal even though there is some inter-class similarity present in Segments 1 and 2. While we want to avoid inter-class similarity, we are at the optimal solution. This can be seen by increasing and decreasing the number of clusters. Changing the number of clusters increases inter-class similarity. Additionally, the Percent of Total Frequency starts to become in-significant for the latter clusters as the number of clusters are changed.

**2 Clusters**

****

**4 Clusters**

****