2.2. Analyses and results

2.2.1. Manual Cluster Analysis

Initially, the number of clusters was set manually to a specific value, such as 6 clusters. Subsequently, the number of clusters was systematically reduced, iteratively exploring different cluster counts from 5 clusters down to 2 clusters. The Cluster tool was used to perform cluster analysis with each specified number of clusters. Concurrently, the Segment Profile tool was employed to investigate the characteristics and profiles of the data segments created by each cluster configuration. This manual approach allowed for a range of cluster configurations to be considered, offering insights into how different cluster counts might reveal distinct patterns in the data.

- For each clustering in this study, we use interval standardization as None, because the inputs were all on the same measurement (Scale 1-10).

Analysis with 6 Clusters

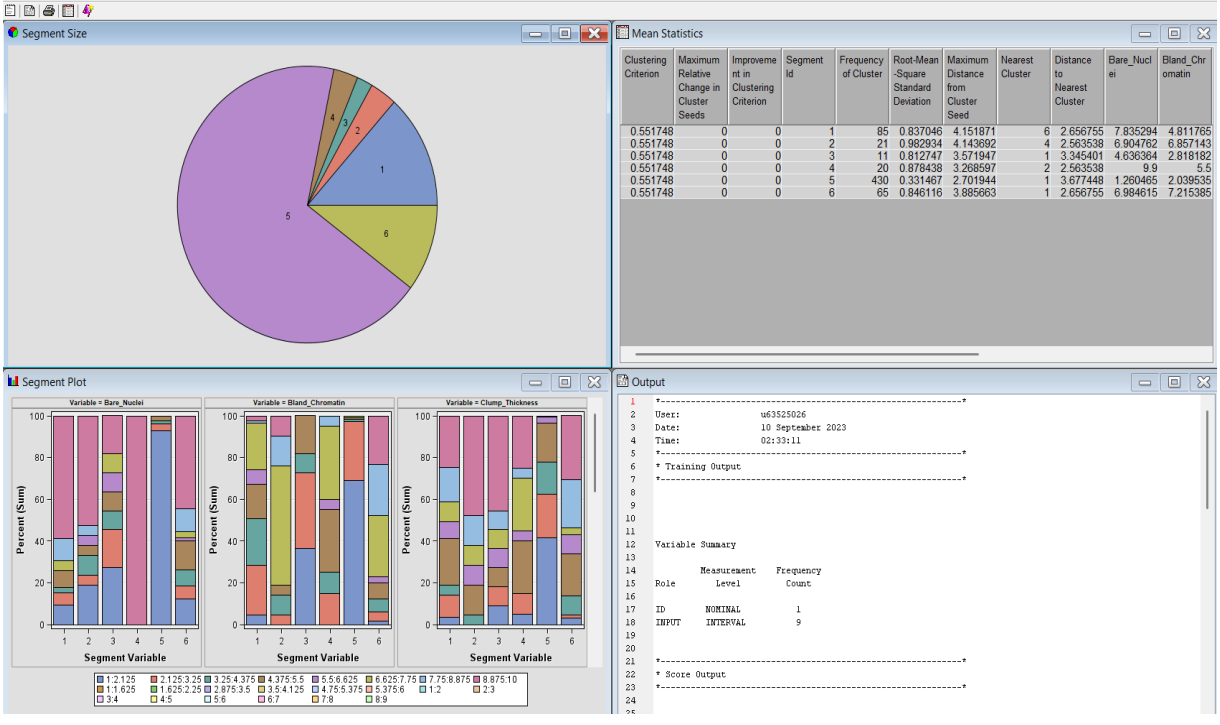

Figure 5: Results of running the Cluster node with Maximum Cluster size 6.

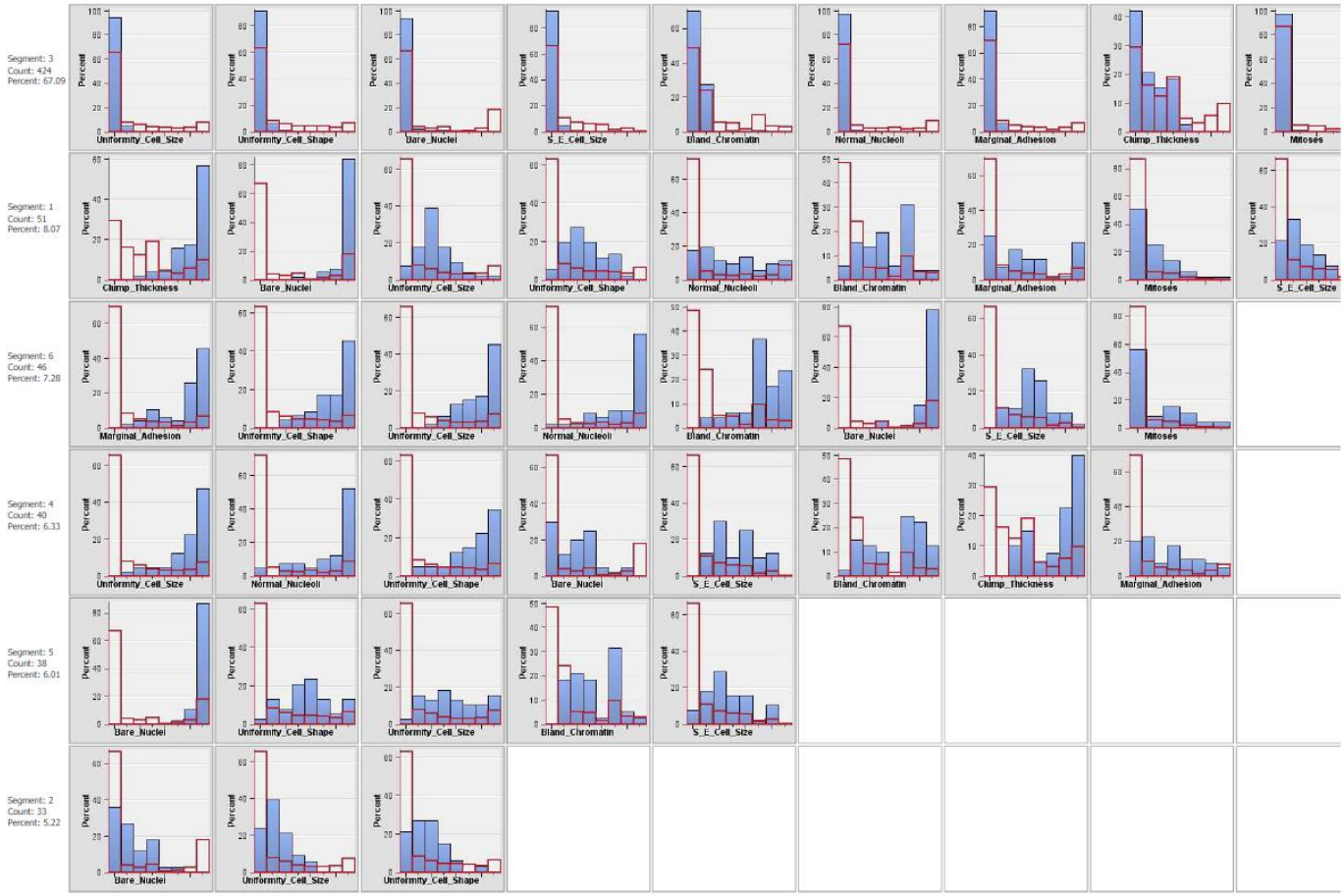

Figure 6: Segment Profile Results with Maximum Cluster size 6

When clustering with a maximum of six clusters, segments 1 & 3 only had data from all variables, and Segment 2 only contained three variables. Segment 3 had a significantly lower average in each variable. Segments 4 & 6 both had higher than average uniformity cell size and cell shape while these two variables show similar distribution within each segment.

Analysis with 5 Clusters

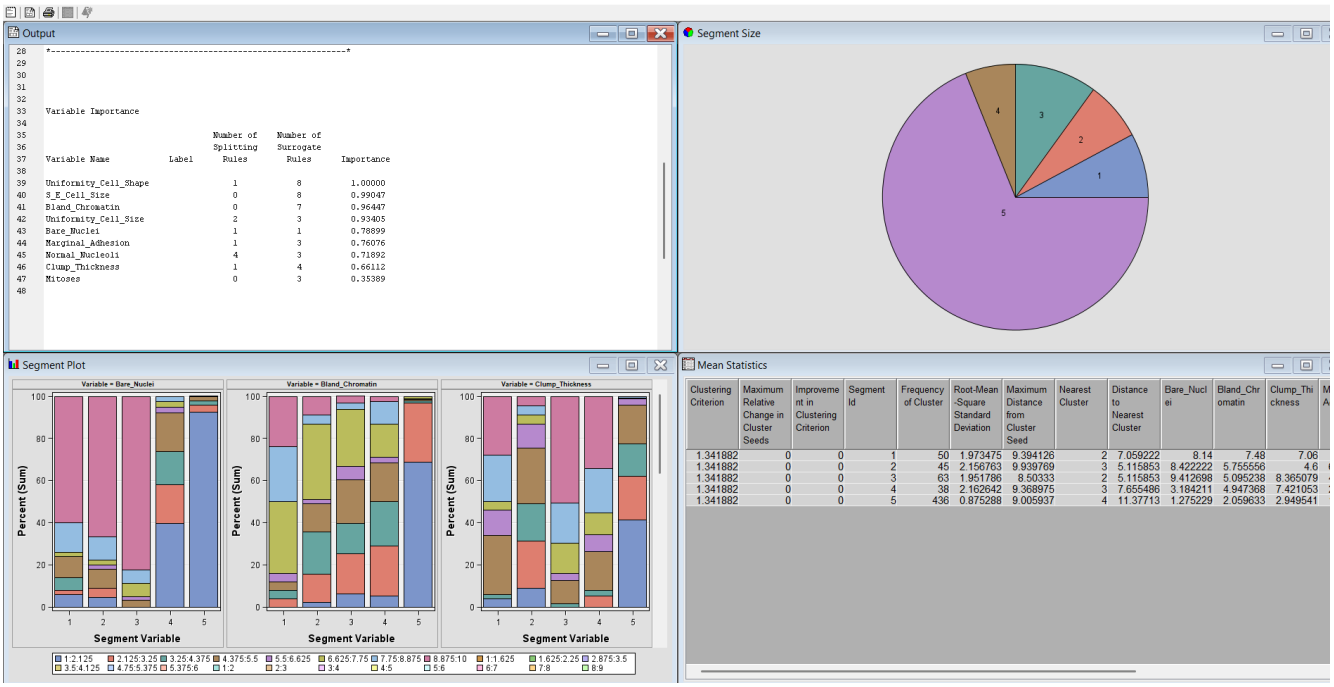

Figure 7: Results of running the Cluster node with Maximum Cluster size 5

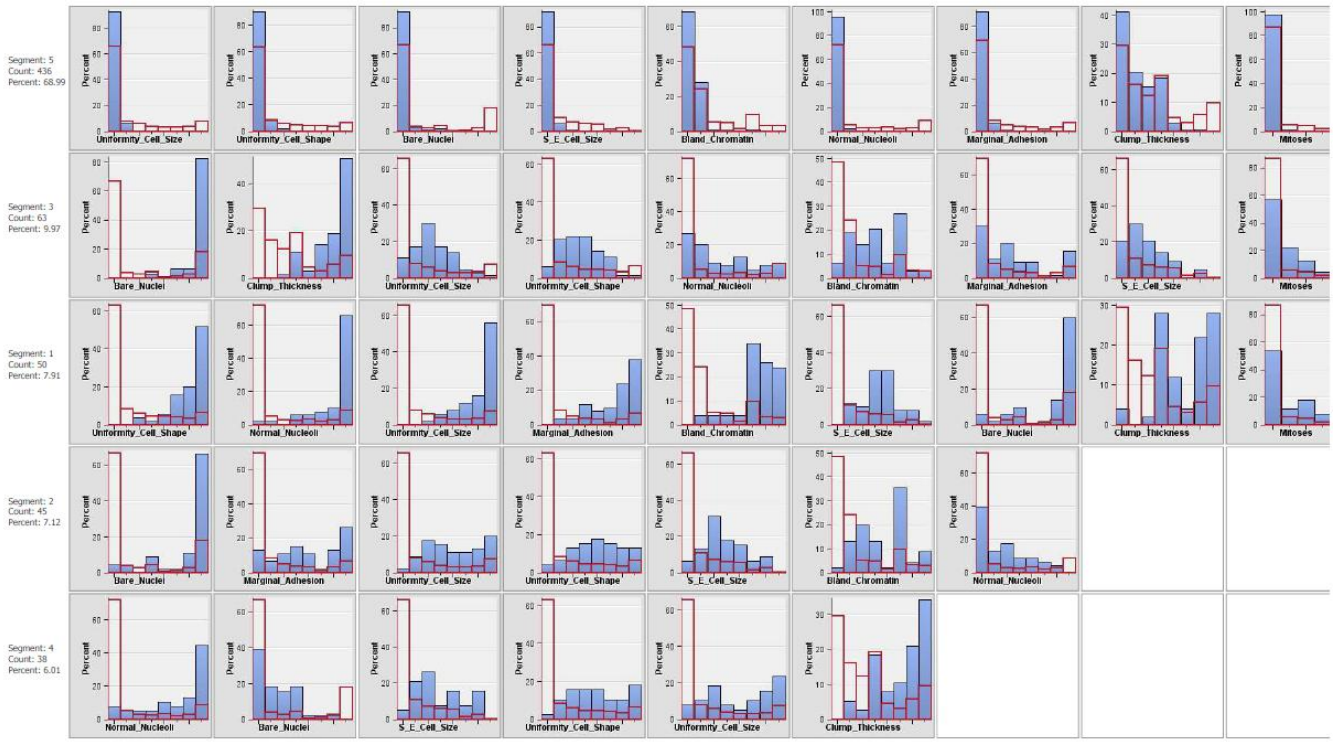

Figure 8: Segment Profile Results with Maximum Cluster size 5

Grouping with 5 clusters segmentation segment 5 had a lower percentage in all variables. In all other segments each variable had an initial point lower average than overall and then distributed in several ways.

Analysis with 4 Clusters

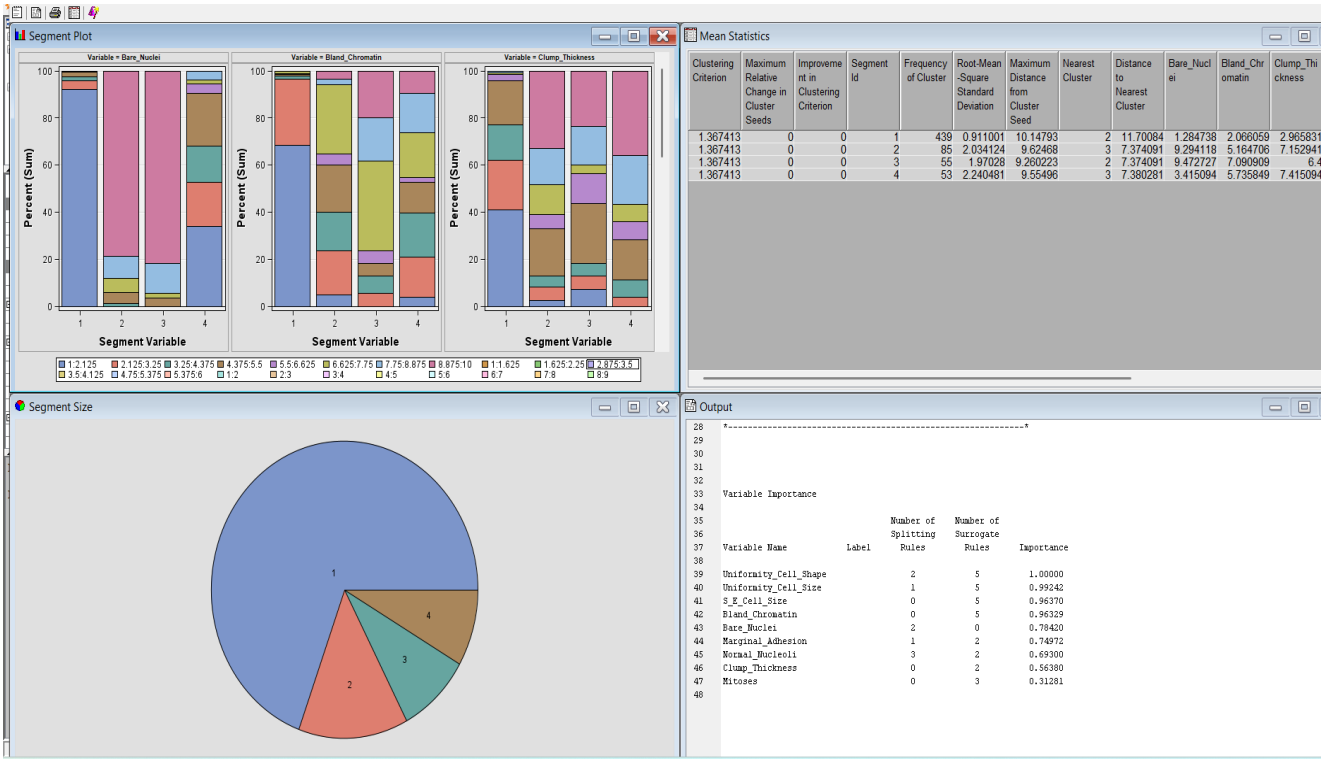

Figure 9: Results of running the Cluster node with Maximum Cluster size 4

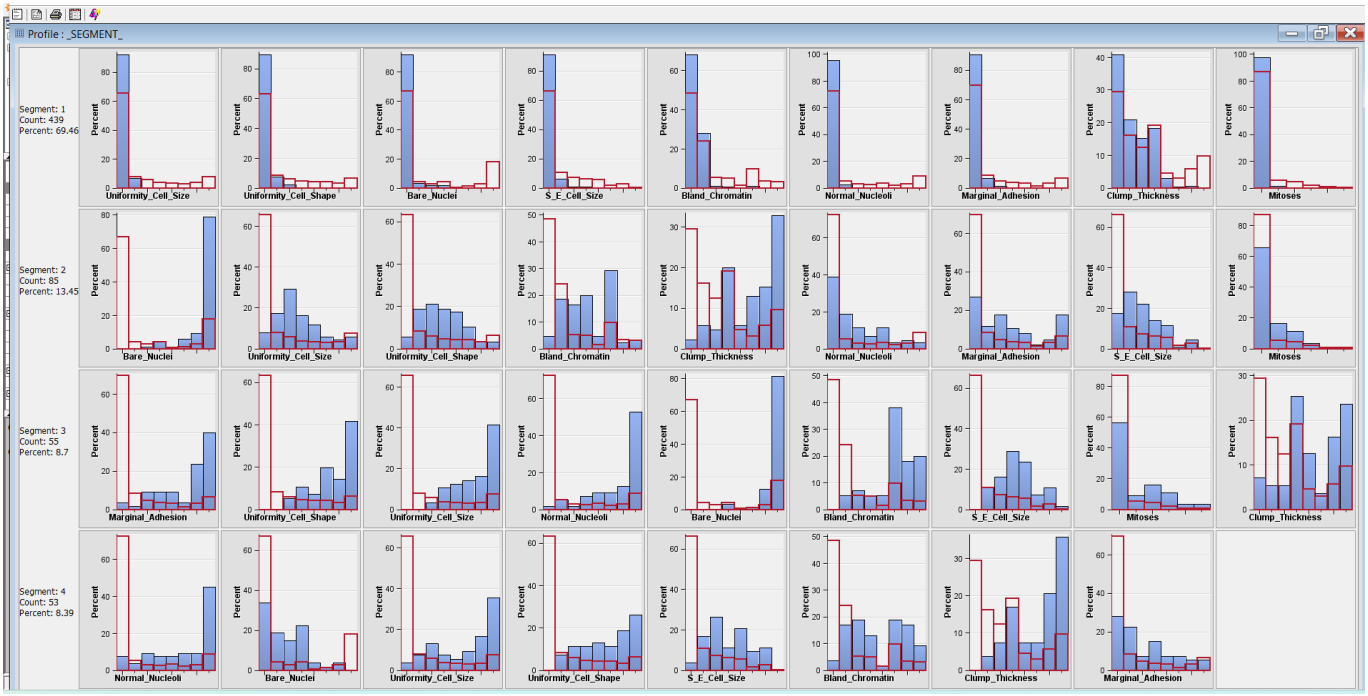

Figure 10:Segment Profile Results with Maximum Cluster size 4

Four clustering method results in the segment profile show that Segment 1 had a lower average in all variables and Segment 3 showed the highest percentage in variables other than, Metosity and Clump thickness. Segment 2 had likely central distribution for most of the variables while segment 4 had higher normal nucleoli, uniformly cell size, shape, clump thickness, and central distribution than overall in other variables.

Analysis with 3 Clusters

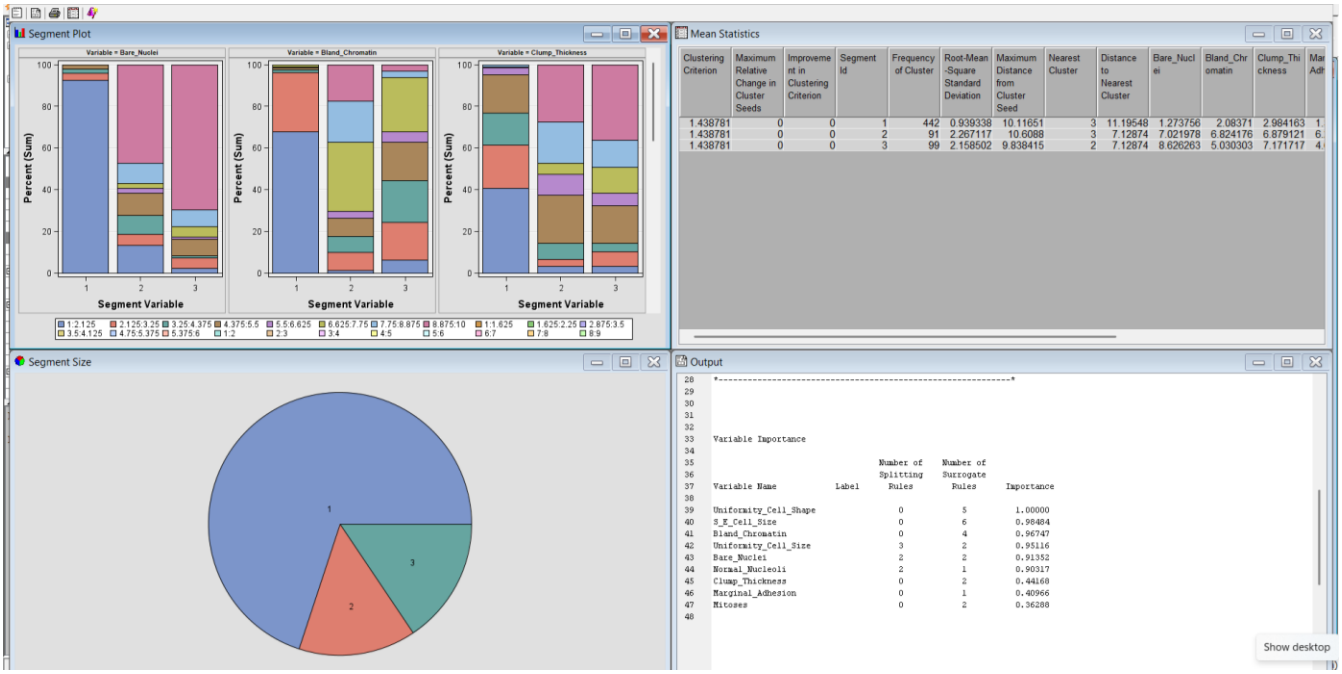

Figure 11: Results of running the Cluster node with Maximum Cluster size 3

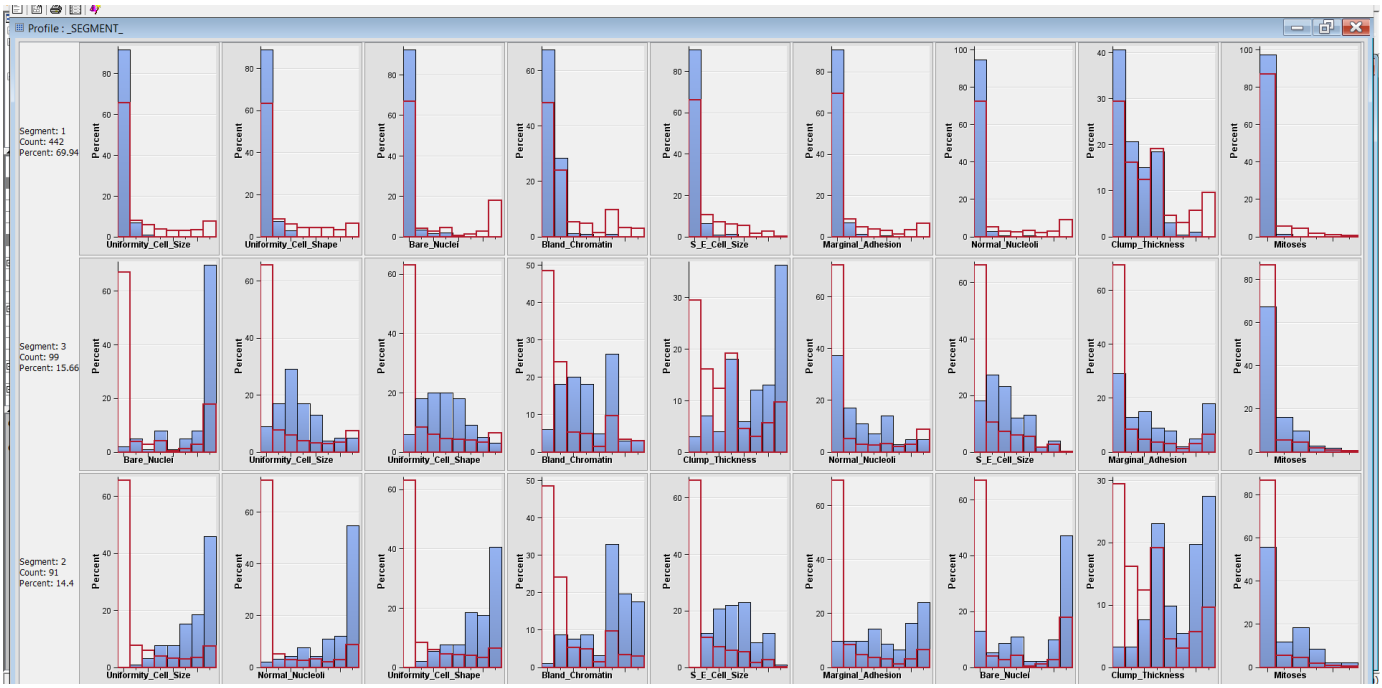

Figure 12:Segment Profile Results with Maximum Cluster size 3

Here all three segments consist of all variables. Segment 1 had lower percentages than overall in each variable. Segment 3 had lower normal nucleoli, marginal adhesion, and mitoses. Segment 2 had a lower percentage in mitoses, central single_epithelial_cell_size, and slightly higher marginal adhesion, and a higher percentage than overall in other variables.

Analysis with 2 Clusters

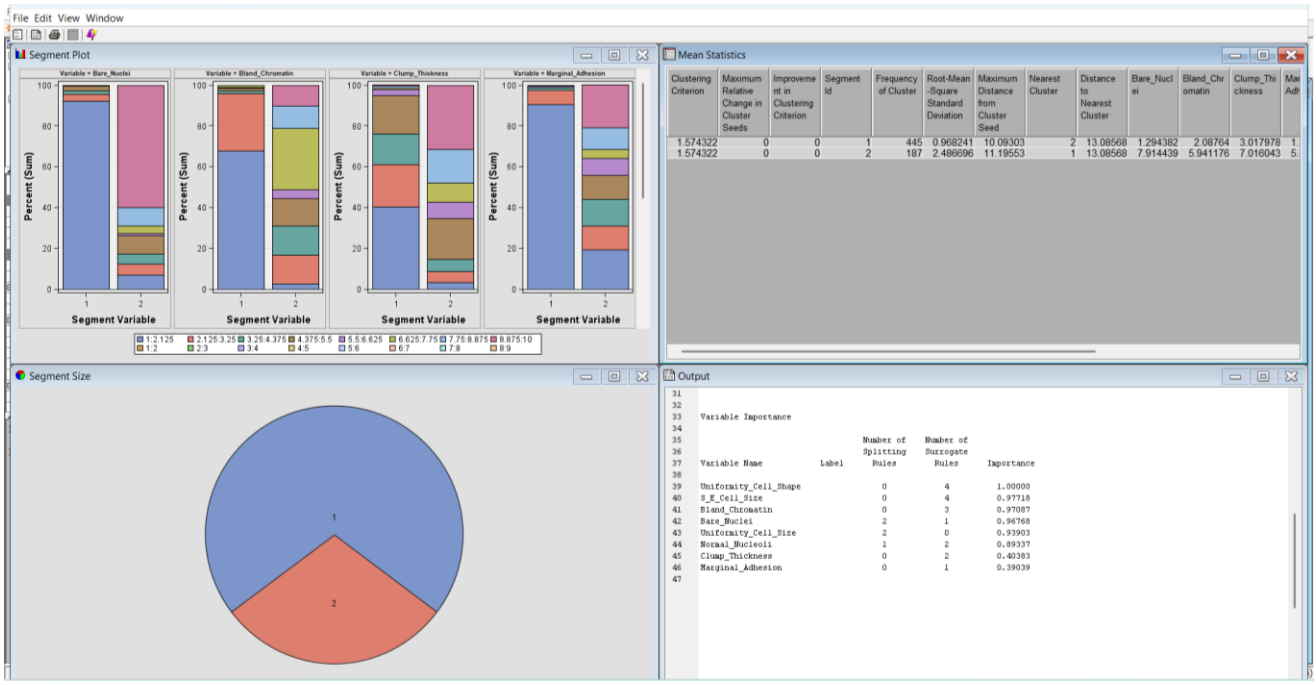

Figure 13: Results of running the Cluster node with Maximum Cluster size 2.

In the below figure: the below segmentation with 2 clusters, Segment 1 has lower than average in all variables compared to overall, and Segment 2 has only mitoses variable with a lower percentage. Segment 2 has a higher percentage bare nuclei variable and clump thickness variable and is slightly higher in uniformity cell size and shape.

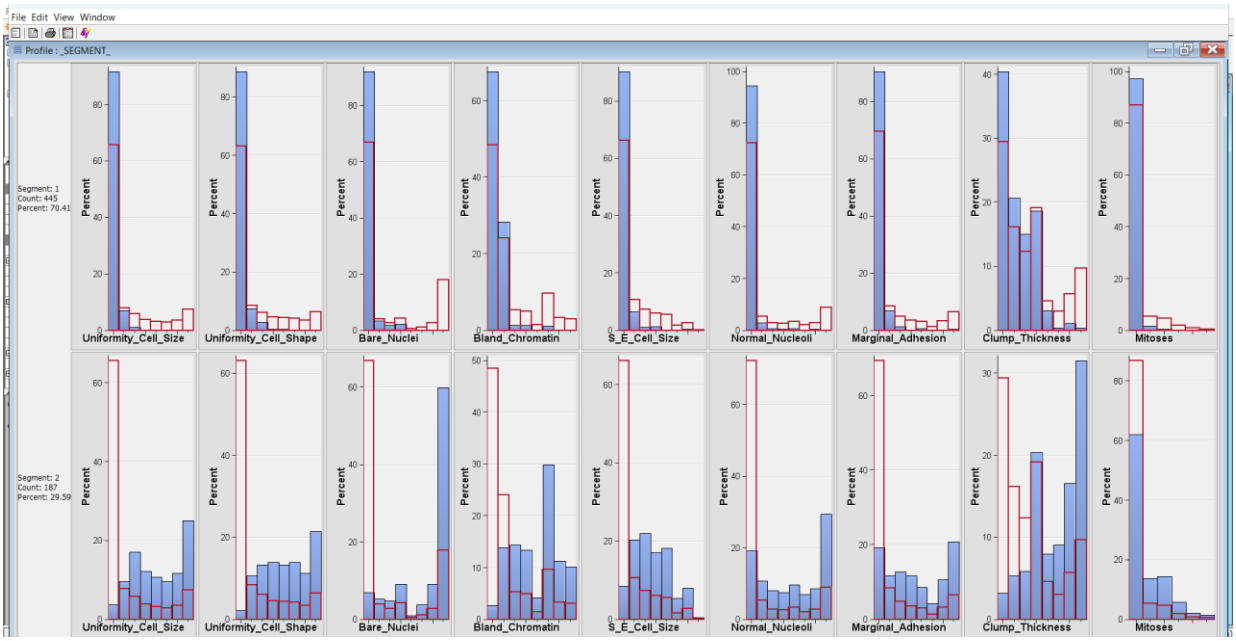

Figure 14: Segment Profile Results with Maximum Cluster size 2

2.2.2. Analysis with Automatic as the Specification Method

In this phase, the Specification Method was set to "Automatic" within the Cluster tool, enabling SAS Enterprise Miner to determine the optimal number of clusters automatically.

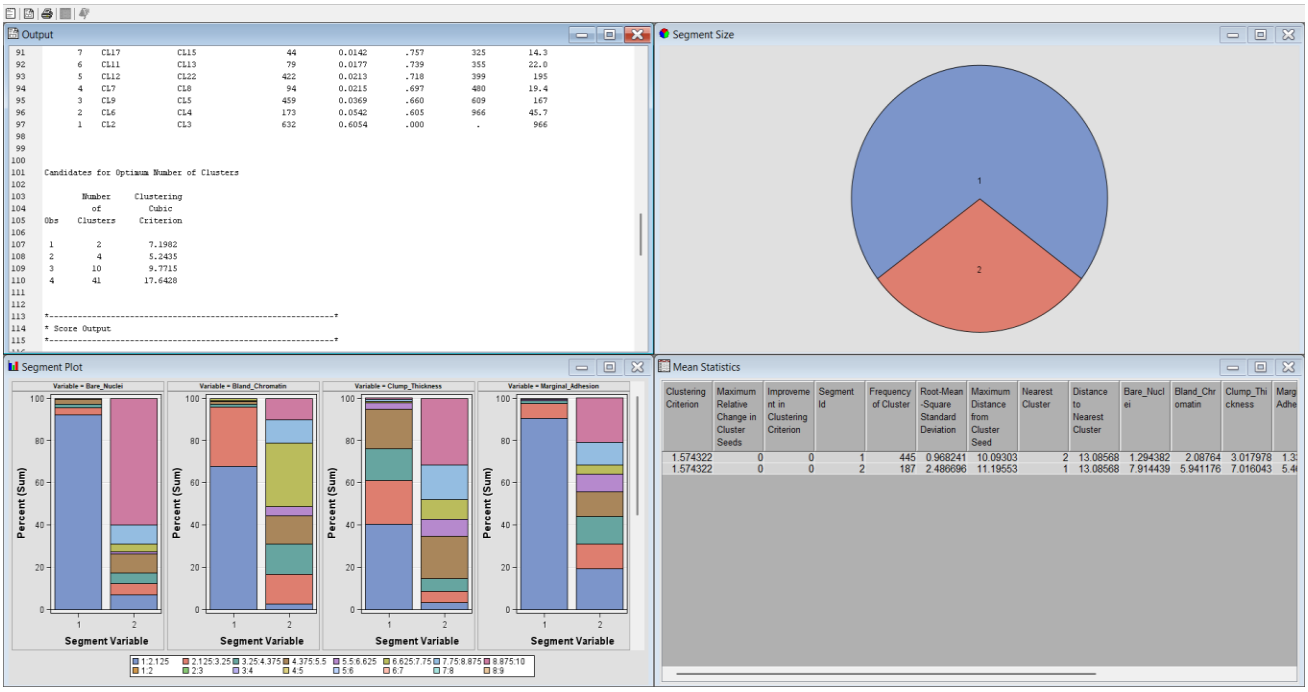

Figure 15: Results of running the cluster node without specifying number of clusters manually.

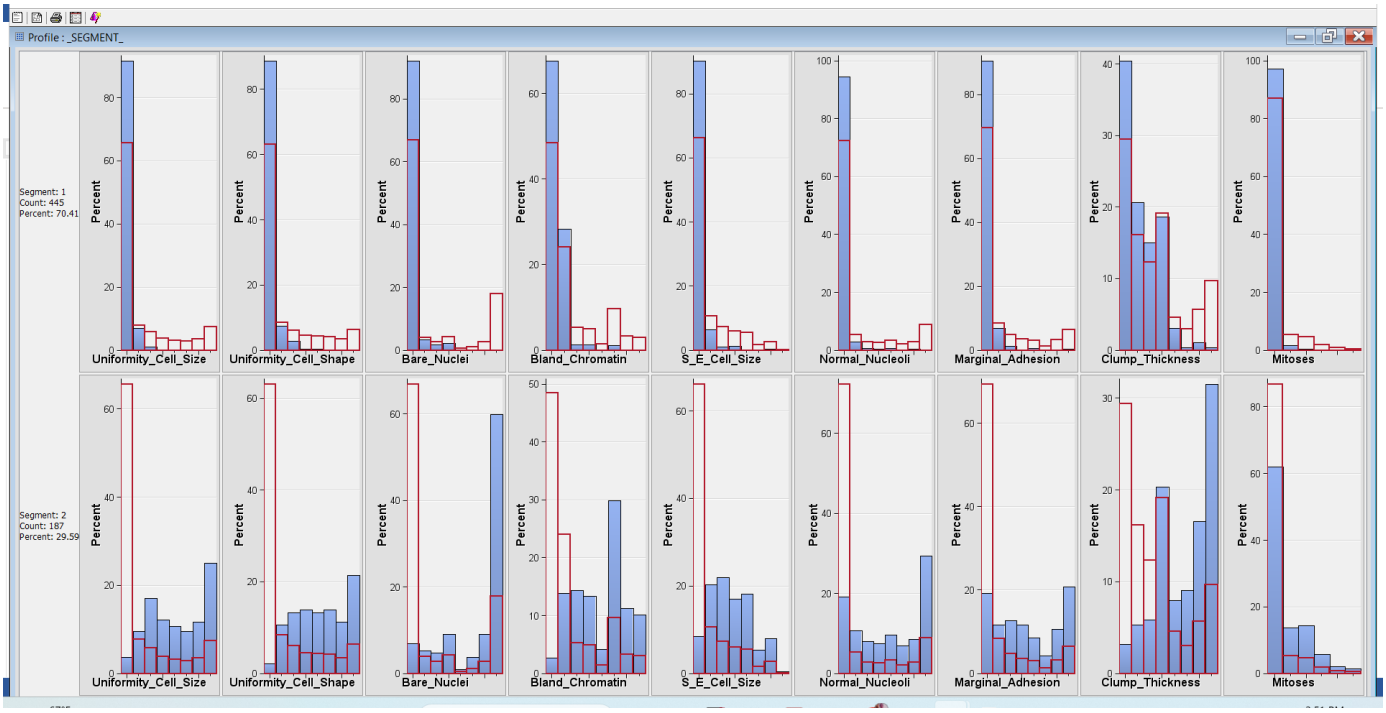

Figure 16: Segment Profile Results Without specifying maximum clusters manually.

Here we can see the SAS Enterprise Miner suggest clustering is same as the clustering with 2 clusters.

3. Result discussion

With six clusters, Segment 3 showed significantly lower averages in all variables, suggesting it represents a distinct subgroup within the data with lower values across the board. In the 5-cluster configuration, Segment 5 had lower percentages in all variables, indicating a subgroup with consistently lower values. In the 4-cluster configuration, Segment 1 had lower percentages in all variables, Segment 3 had lower values in some variables, and Segment 4 had higher values in certain variables, such as normal nucleoli and uniformity of cell size and shape.

The 3-cluster configuration revealed that Segment 1 had lower percentages across variables, Segment 3 had distinctive lower values in specific variables, and Segment 2 showed differences in multiple variables. In the 2-cluster configuration, Segment 1 had lower than average values in all variables, while Segment 2 showed differences primarily in the mitoses, bare nuclei, clump thickness, and uniformity of cell size and shape variables. The results showed that the automatic clustering result was the same as the manual clustering result with 2 clusters.

The manual cluster analysis gave important information about the potential subgroups of the dataset at various cluster counts. It exposed clear patterns and variances in the data, revealing subgroups with lower or greater values for variables, for example. According to SAS Enterprise Miner's criteria, the dataset may be best represented by two natural clusters, as indicated by the automatic clustering result aligning with the 2-cluster setup. This is consistent with the results of the manual analysis, which showed that Segment 2 had changes largely in mitoses, bare nuclei, clump thickness, and homogeneity of cell size and shape, while Segment 1 had lower values for all variables.