Assignment 1: Cluster Analysis

INFS 5102 – Unsupervised Methods in Analytics, SP2, 2023

Student Name: Wangjun Shen

Student ID: 110248810

The University of South Australia

目录

What Is Cluster and How Basic K-means Algorithm Works

Cluster analysis is a technique in machine learning that groups data points with similar characteristics into distinct categories or clusters. This involves partitioning a large dataset into smaller subsets with similar characteristics.

One of the most popular clustering algorithms is the k-means algorithm. It is used to divide datasets into k number of clusters. The algorithm starts by randomly selecting k initial cluster centroids within the dataset. Each data point in the dataset is then assigned to the cluster whose centroid is closest to it. The centroid of each cluster is recalculated based on the average position of all data points in that cluster. These steps are repeated until the cluster assignments stop changing or a predefined stopping criterion is met.

To illustrate the k-means algorithm, we'll use the Iris dataset, which is built into R. This dataset contains 150 observations of 4 features of iris flowers: sepal length, sepal width, petal length, and petal width, as well as a categorical variable indicating the species of each flower. We'll use the first two features (sepal length and sepal width) to cluster the flowers into two groups.

We can perform k-means clustering on the Iris dataset in R using the kmeans() function. First, we need to extract the relevant features from the dataset and normalize the data:

```
> # Load the Iris dataset
> data(iris)
>
> # Extract the first two columns (sepal length and sepal width)
> iris_features <- iris[,1:2]
>
> # Normalize the data
> iris_norm <- scale(iris_features)
```

Next, we can apply the kmeans() function to the normalized data with k=2:

```
# Perform k-means clustering with k=2
kmeans_result <- kmeans(iris_norm, centers=2)
```

We can then plot the clusters using ggplot2:

```
# Plot the clusters
library(ggplot2)
iris_cluster <- data.frame(iris_norm, cluster=kmeans_result$cluster)
ggplot(iris_cluster, aes(x=Sepal.Length, y=Sepal.Width,
color=factor(kmeans_result$cluster))) +
  geom_point(size=3) +
  labs(color="Cluster")
```

The resulting plot shows the two clusters identified by the k-means algorithm, where each point is colored according to its assigned cluster:

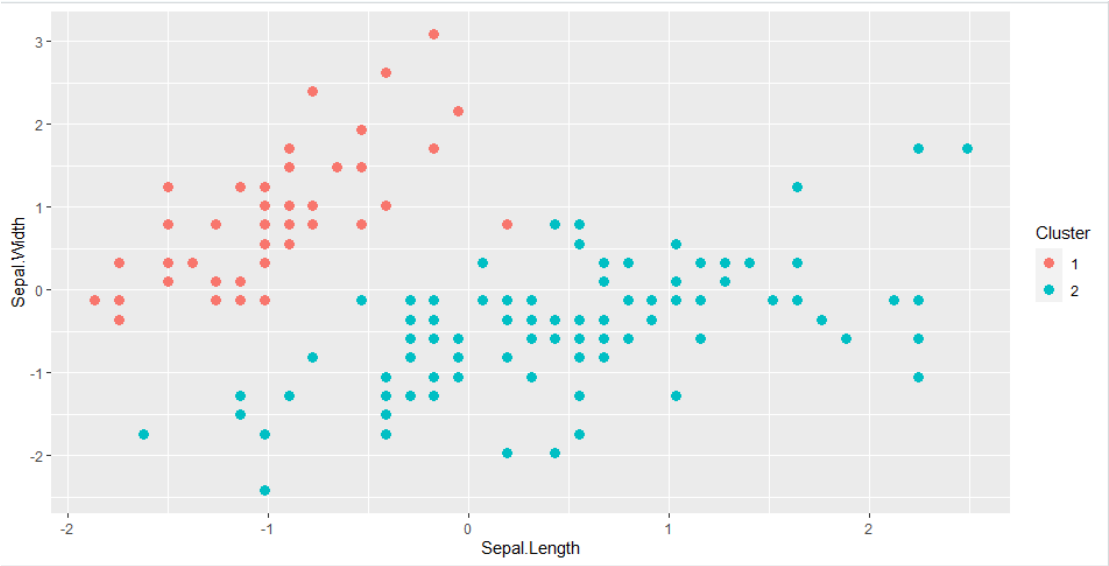

Figure 1: Result of K-mean for Iris Dataset

The resulting plot shows the two clusters identified by the k-means algorithm, where each point is colored according to its assigned cluster.

Data Exploration and Pre-processing

To properly import the dataset, it is important to understand what each attribute represents. Referring to the source website of the dataset, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original), we can find the attribute information:

Attribute	Description
Sample code number	id number
Clump Thickness	1 - 10
Uniformity of Cell Size	1 - 10
Uniformity of Cell Shape	1 - 10
Marginal Adhesion	1 - 10
Single Epithelial Cell Size	1 - 10
Bare Nuclei	1 - 10
Bland Chromatin	1 - 10
Normal Nucleoli	1 - 10
Mitoses	1 - 10
Class	2 for benign, 4 for malignant

Table 1: Attribute and Description for Original Dataset

Based on the information provided in the table, it can be determined that the original dataset consists of 10 attributes, with "Sample code number" serving as the ID number and "Class" being the target variable.

To begin, create a SAS EM Project called "assignment#01" and create a diagram. Next, create a File Import Node to import the local dataset. The process is shown below:

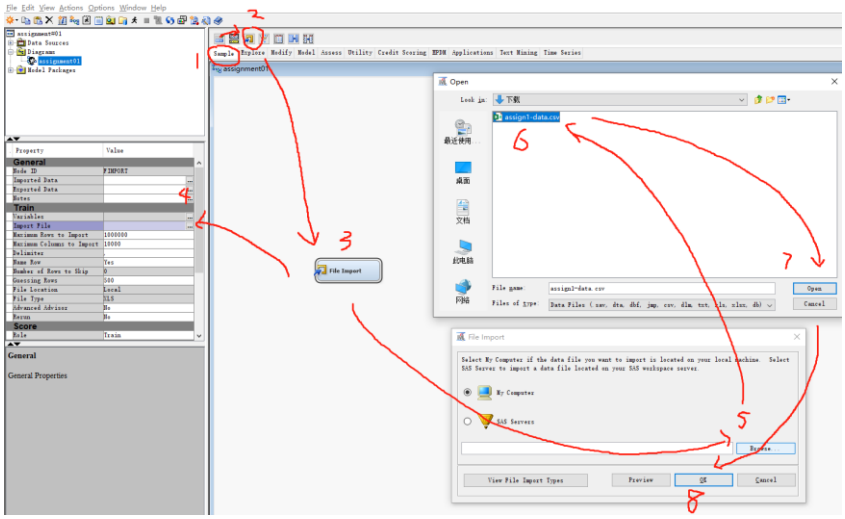

Figure 2: How to Import Local CSV Dataset into SAS EM

Right-click on the File Import Node, select Edit Variables to view and modify the Attributes of the imported dataset:

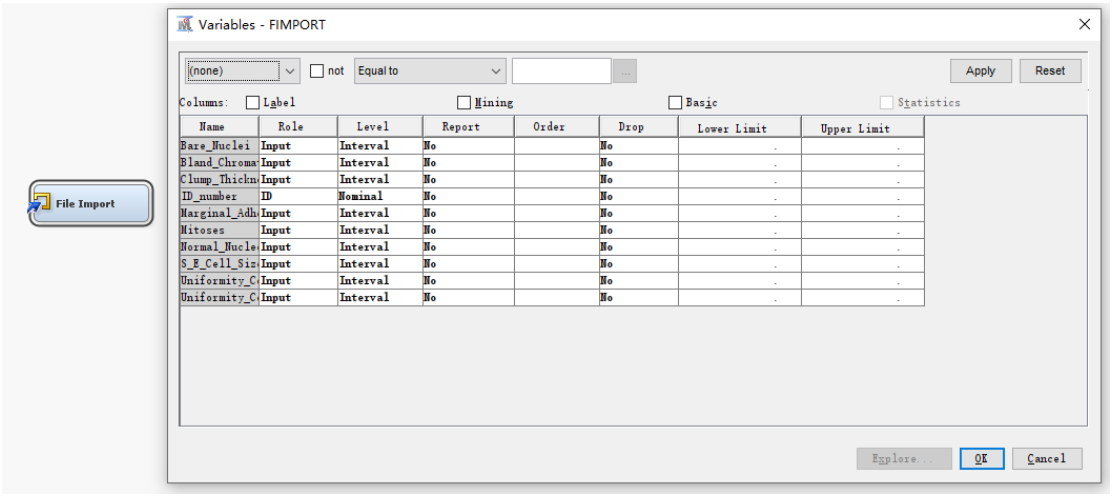

Figure 3: Check the Default Values for Role and Level for Imported Dataset

From the above figure, it can be observed that the difference between the local dataset and the original dataset is that the local dataset does not include the "Class" attribute, as this variable appears as a target and its use does not comply with the unsupervised learning method of clustering.

All attributes except ID_number have been set to level "Interval". Interval refers to numerical variables, usually used for continuous data. However, setting the level of these attributes to "Interval" does not match their actual meaning. Need to change the attributes outside of the sample code number from the default "Interval" to "Ordinal" because these variables have a meaningful order, and the numerical values have significance.

So, all Attributes except for ID_number have had their level modified from interval to Ordinal:

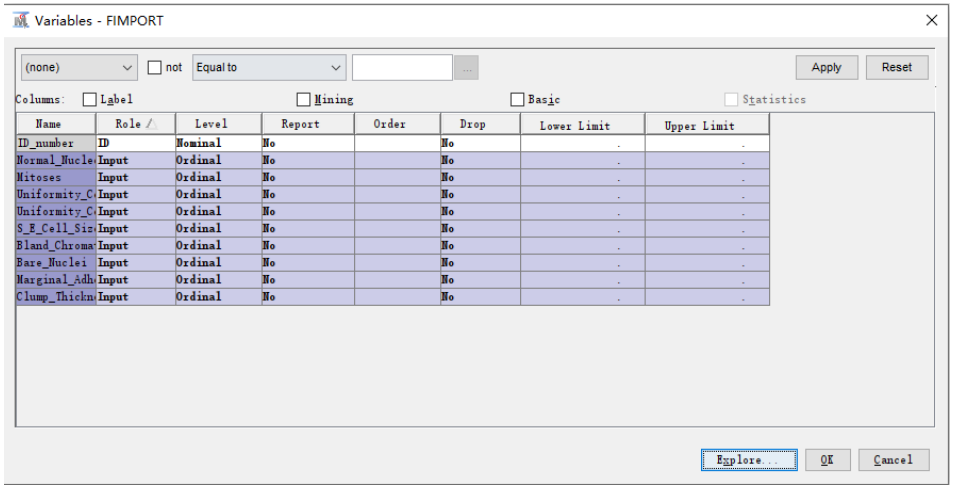

Figure 4: Change Roles and Levels for Variables

Next, click on Explore to conduct data exploration. The results obtained are:

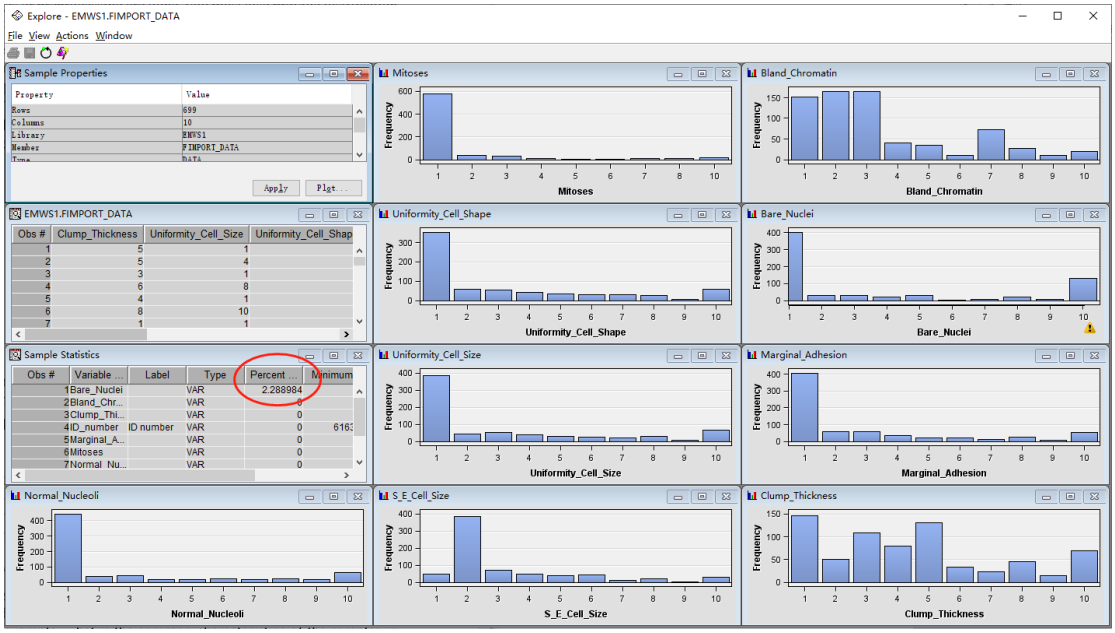

Figure 5: Check the Statistical Result for the Data

This graph shows the distribution of each of the 9 attributes, but the most important information is the part circled in red. The Bare_Nuclei attribute has about 2.29% missing values, which means that there are approximately 16 objects with missing values in this attribute.

Missing values can have an impact on the results of clustering, and how to handle missing values can also affect the results of clustering. The methods used to handle missing values and their impact will be explored in subsequent experiments.

Exploration and Comparison of Experimental Results With Different Maximum Number of Clusters

To begin, set up a Cluster node and connect it to the File Import node. Next, create a Segment Profile node and connect it to the Cluster node, following the diagram below:

Figure 6: Create and Link with Cluster Node and Segment Profile Node

Then modify the properties of the Cluster Node as shown in the following image:

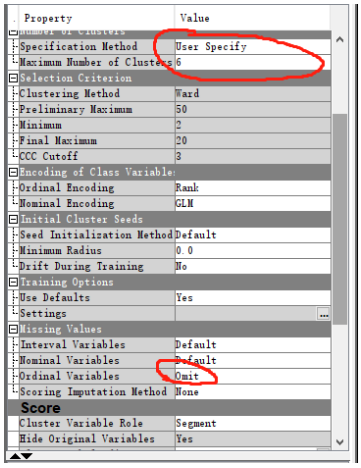

Figure 7: Set the Maximum Number of Clusters as 6 and pre-processing

Since the missing value makes up only 2.29% of the total, its impact on overall performance is minimal. Therefore, we will adjust the setting for Missing Values' Ordinal Variables to 'Omit', which will enable the Cluster Node to omit these missing values during execution.

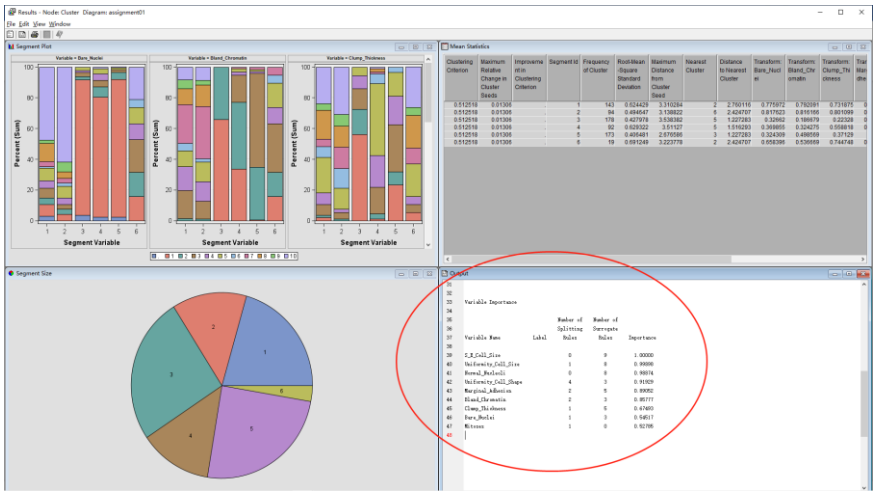

Figure 8 : Result of Cluster Node for Maximum 6 Clusters

The data has been divided into six categories. Segment 3 has the highest proportion, but

the difference with segment 5 is not significant, while segment 6 has the smallest proportion. S_E_Cell_Size is the most important variable from the Variable Importance perspective, with an importance score of 1.0. It is used to split a node in each constructed decision tree. Uniformity_Cell_Size is the second most important variable, with an importance score of 0.99898. It is used to split a node in all nodes except one decision tree. Normal_Nucleoli has an importance score of 0.98874 and is chosen to split a node in all nodes except two decision trees. On the other hand, Mitoses has the lowest importance score of 0.52785. It is only used to split a node in one decision tree and has never been used as an alternative variable.

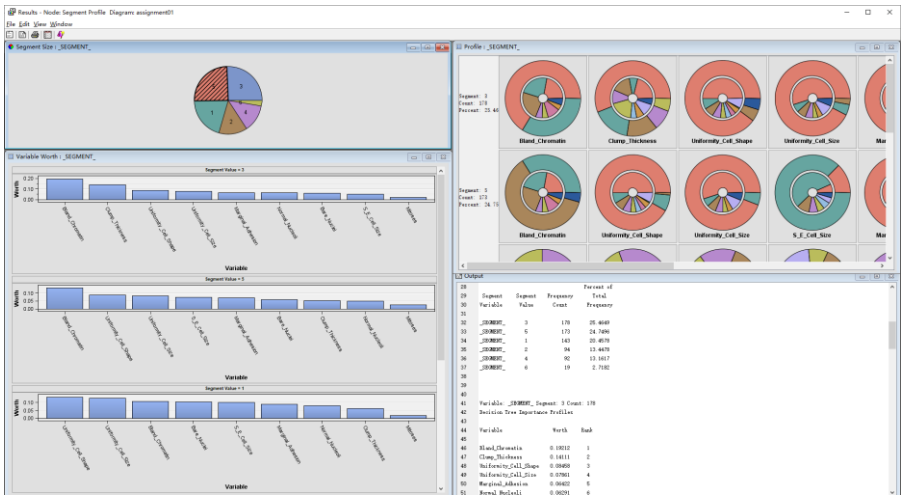

Figure 9 : Result of Segment Profile for Maximum 6 Clusters

Several variables, including Bland_Chromatin, Clump_Thickness, Uniformity_Cell_Shape, and Uniformity_Cell_Size, are important in multiple segments, while Mitoses and Marginal_Adhesion are more significant in certain segments. This output provides insights into the relative importance of variables in decision tree models for different subsets of data, informing decision-making.

Continuing our discussion on the Maximum Cluster Number, we will gradually decrease it from 6 to 3 and observe the results at each step.

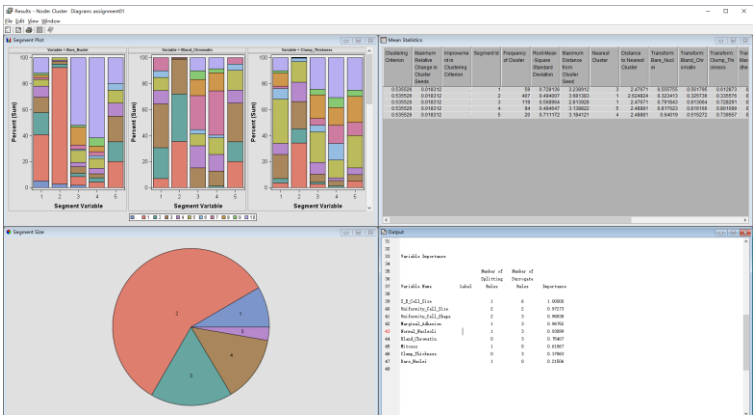

Figure 10 : Result of Cluster Node for Maximum 5 Clusters

Segment 2 is the largest, exceeding half of the total proportion. Next comes Segment 3 and

Segment 4, while Segment 5 has the smallest proportion, almost the same as the segment with the smallest proportion when the Maximum Cluster Number is set to 6. S_E_Cell_Size still holds the highest importance of 1.0, followed by Uniformity_Cell_Size. Normal_Nucleoli's importance ranking has started to decline and is now third, while Uniformity_Cell_Shape is the third most important. The importance ranking of Mitoses has begun to rise and Bare_nuclei is the least important.

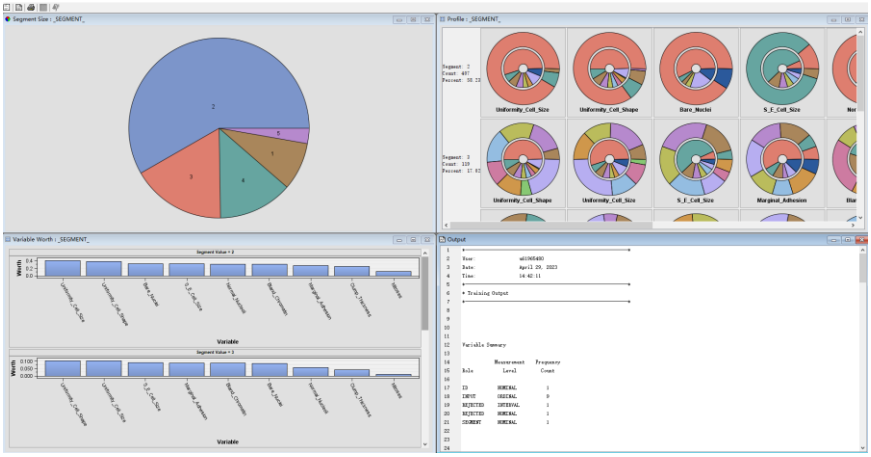

Figure 11 : Result of Segment Profile for Maximum 5 Clusters

Uniformity_Cell_Size and Uniformity_Cell_Shape are highly ranked variables across different segments, while Bare_Nuclei is highly ranked in high proportion segment but not in low proportion segment. In comparison to 6 clusters, Uniformity_Cell_Size and Uniformity_Cell_Shape remain the most important variables, but Bland_Chromatin and Clump_Thickness are now considered less important. Although Mitoses is generally a low priority variable, it emerges as a top important variable in the low proportion segment.

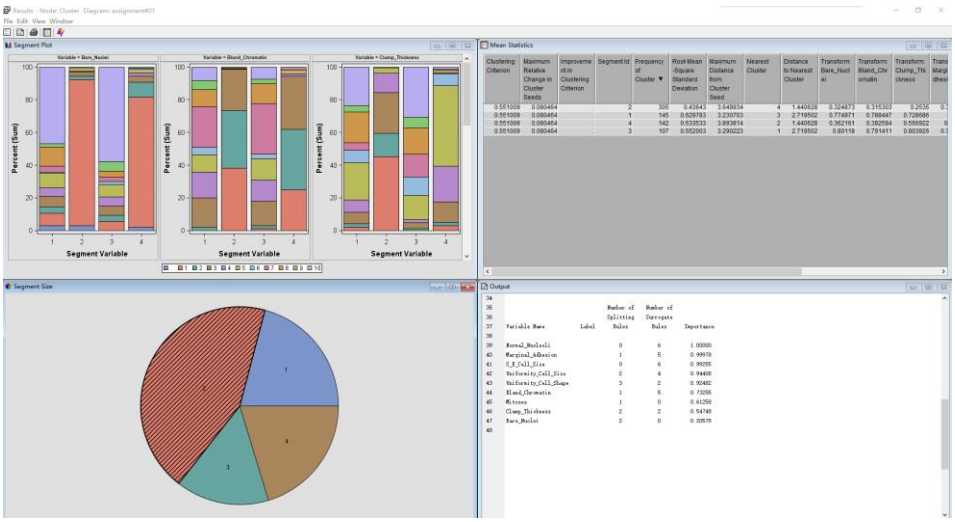

Figure 12 : Result of Cluster Node for Maximum 4 Clusters

Segment 1 has the largest share, but it's less than half of the total. The next two largest segments are segments 1 and 4, but the difference between them is very small, only 4. The segment with the smallest share is segment 3, which has a significantly higher share than the smallest segment of clusters 6 and 5.

Normal_Nucleoli is the most significant variable, with a value of 1, replacing the position of S_E_Cell_Size. Marginal_Adhesion and S_E_Cell_Size are closely following. Uniformity_Cell_Size and Uniformity_Cell_Shape are also highly ranked, while Bare_Nuclei has the lowest importance score. Mitoses and Clump_Thickness are considered relatively less important compared to other variables. The importance of Mitoses is still rising.

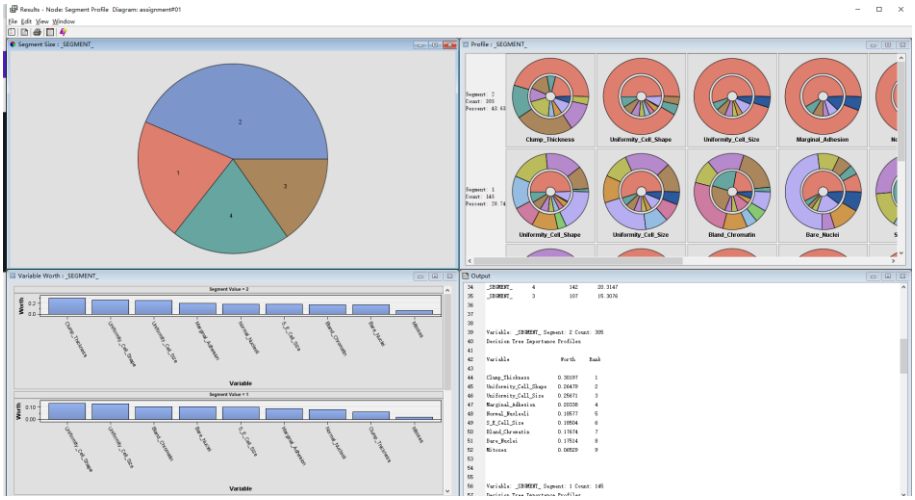

Figure 12: Result of Segment Profile for Maximum 4 Clusters

Both segment 2 and segment 1 place high value on Uniformity_Cell_Shape and Uniformity_Cell_Size, while Clump_thickness is highly valued only in segment 2 and segment 4. Meanwhile, Mitoses is most highly valued in segment 3, and has the lowest value in other segments.

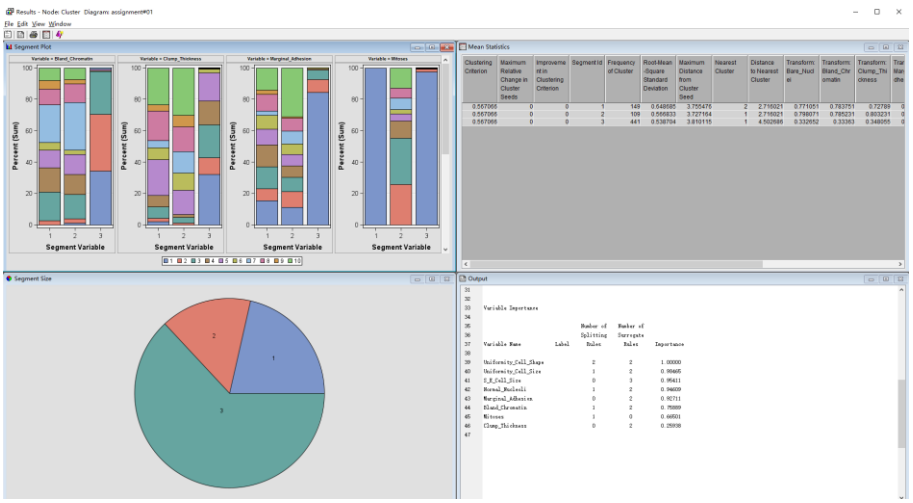

Figure 13 : Result of Cluster Node for Maximum 3 Clusters

The significance of Uniformity_Cell_Shape has been raised to 1.0, and Uniformity_Cell_Size has been elevated to 0.98. Normal_Nucleoli's importance has decreased from 1.0 to 0.95. Mitoses remains unchanged, and Bare_Nuclei, which may be considered an insignificant variable, has been removed.

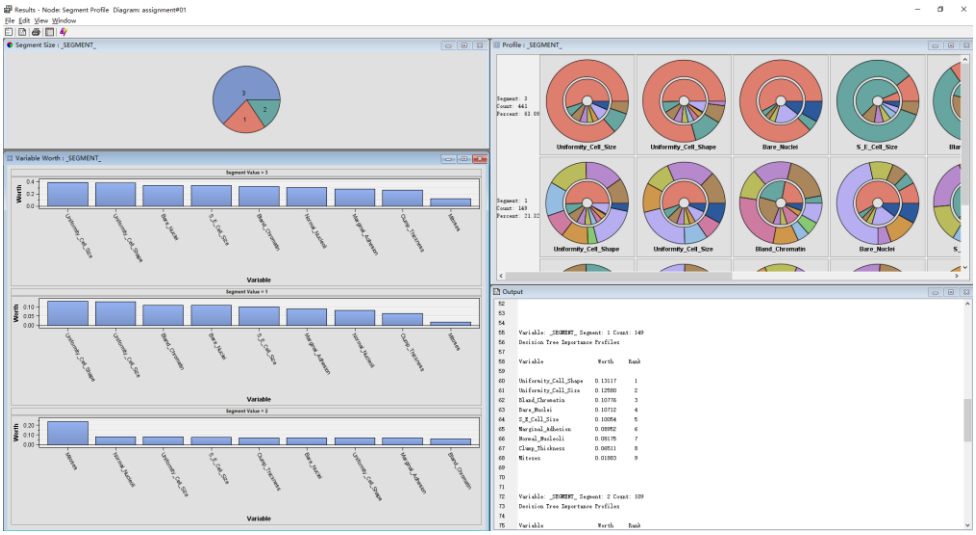

Figure 14: Result of Segment Profile for Maximum 3 Clusters

Although mitoses remain the least valuable in the high percentage segment, they are still more valuable than any other feature in the low percentage segment.

To sum up, the importance ranking of variables varies depending on the maximum value of different clusters. Certain variables, such as Uniformity_Cell_Size and Uniformity_Cell_Shape, consistently maintain high importance scores in most submarkets. Meanwhile, the importance ranking of other variables, such as Normal_Nucleoli and Mitoses, fluctuates between different cluster maximums and segments. Therefore, knowing the ranking of variable importance provides valuable information for decision-making for various subsets of data.

Exploration and Comparison of Experimental Results With Different Settings for Automatic Cluster Using SAS EM

First, change the Specification Method to Automatic, set Missing Values for Ordinal Variables to default, keep the Selection Criterion setting unchanged, run the Cluster Node and get the result:

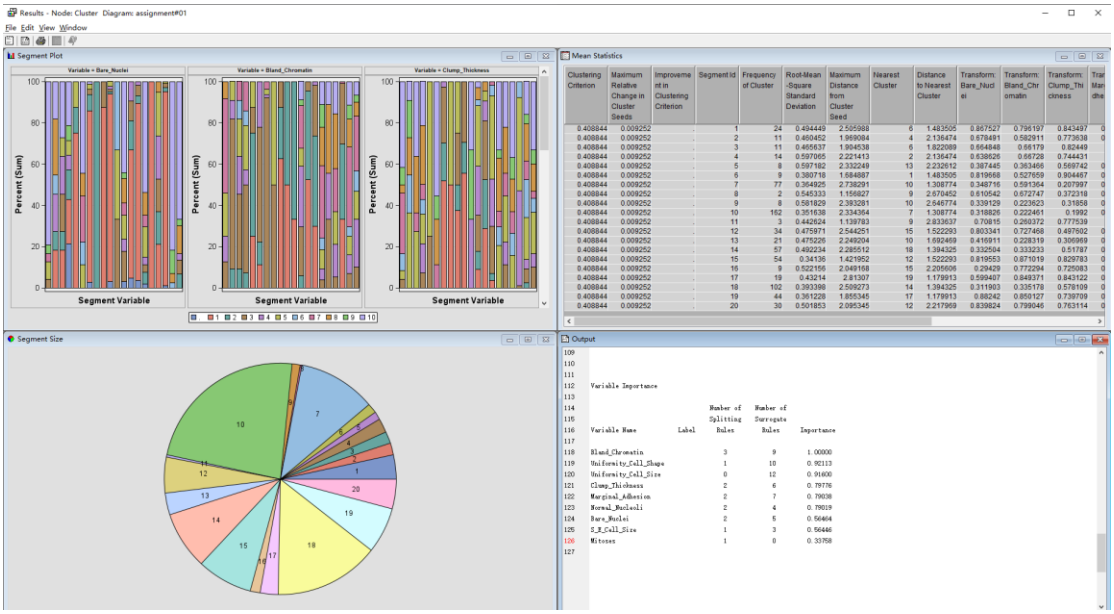

Figure 15 : Result of Cluster Node for Automatic Specification Method and default setting for others

Based on the results, it is apparent that the dataset has been separated into a total of 21 segments, with one segment designated as "others" to encompass all remaining datasets. The number of segments is significantly higher than that achieved through manual setting of the maximum cluster number, and many of the resulting segments are small in size. As such, these results cannot be regarded as optimal.

Next, check the result of the Segment Profile Node:

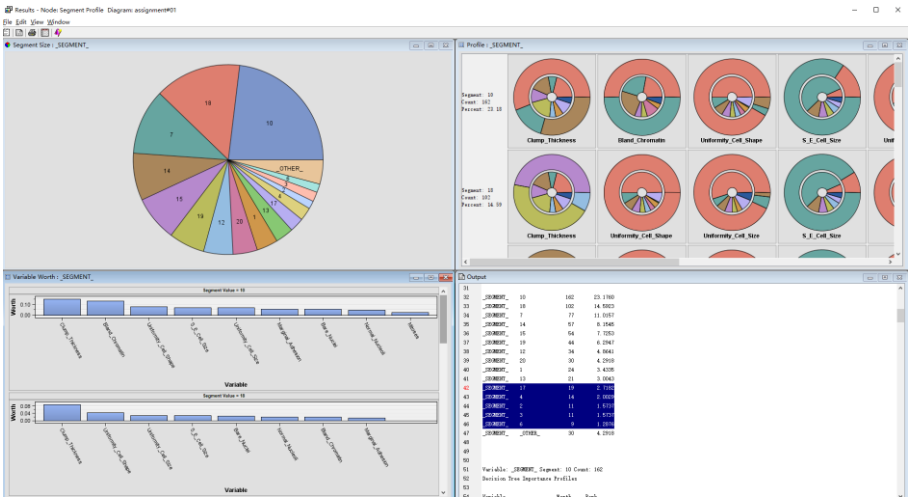

Figure 16 : Result of Segment Profile Node for Automatic Specification Method and default setting for others

Bare_Nuclei, Bland_Chromatin, Uniformity_Cell_Shape, and Uniformity_Cell_Size are all attributes that have high worth in segments with high percentages. Mitoses, on the other hand, ranks lower in worth in these segments, but ranks higher in segments with low percentages. This is consistent with trends previously explored manually.

Next, change the setting of Missing Values for Ordinal Variables from default to Ignore, Median, Mode, and Omit.

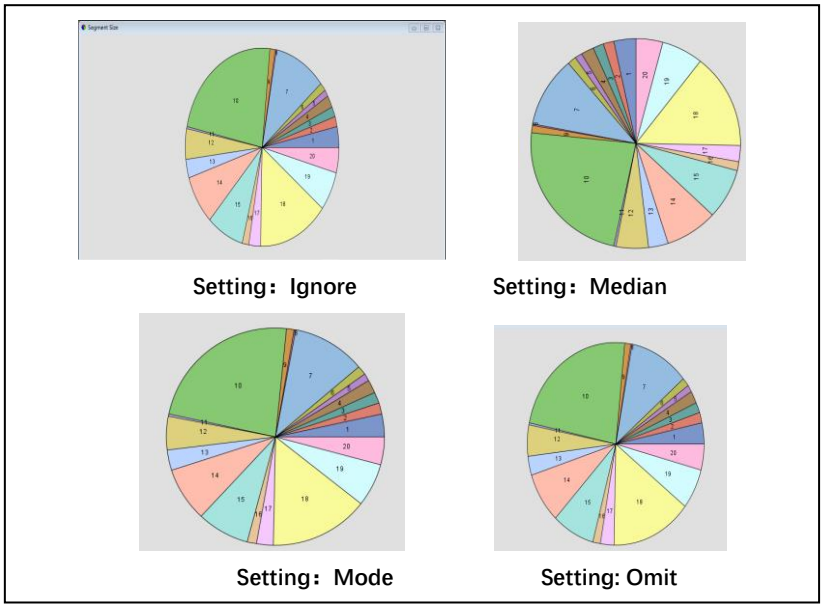

Figure 17: Results of Different Setting for Handle Missing Values

The figure above shows the clustering results of four different methods for handling missing values of Ordinal Values. It can be found that there is no significant difference in general. Based on these findings, it appears that the method of handling missing values has little effect on the clustering results. This aligns with previous results that suggest the low incidence of missing values has a minimal impact overall.

Continuing with the Clustering Method, modify the settings to "Average" and "Centroid" to observe their impact on the Cluster results.

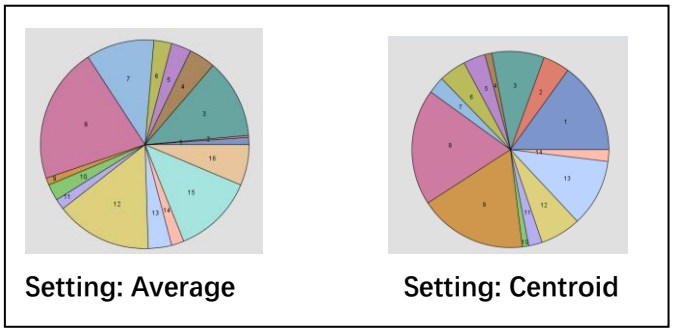

Figure 18: Result of Different Setting for Clustering Method

It can be found that the number of cluster segments has not been significantly reduced, and there are still many small segments.

Finally, try changing Internal Standardisation from Standardisation to Range and check the

result of the Cluster Node:

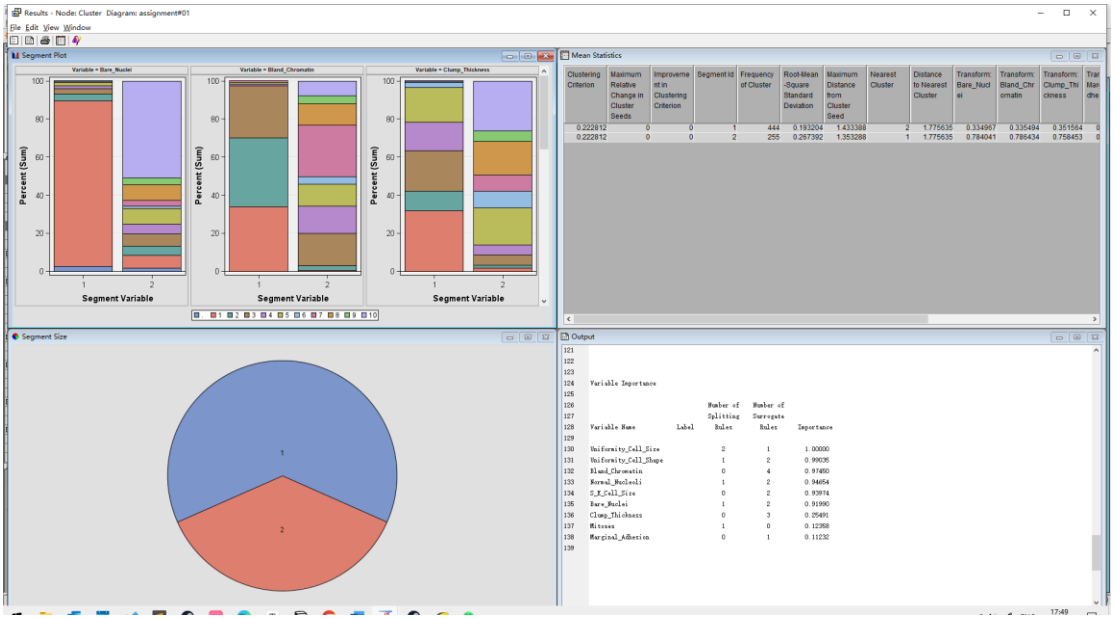

Figure 19 : Result of Cluster Node for Automatic Specification Method and Range as Setting for Internal Standardisation

It can be observed that changing the internal standardisation from "Standardisation" to "Range" resulted in significant changes in the cluster results, from nearly 20 segments to only 2 segments. Comparing this with the cluster result with a maximum number of clusters set to 3 (Figure 13), it can be considered that the two are very similar, and the importance of each attribute is also approximately the same.

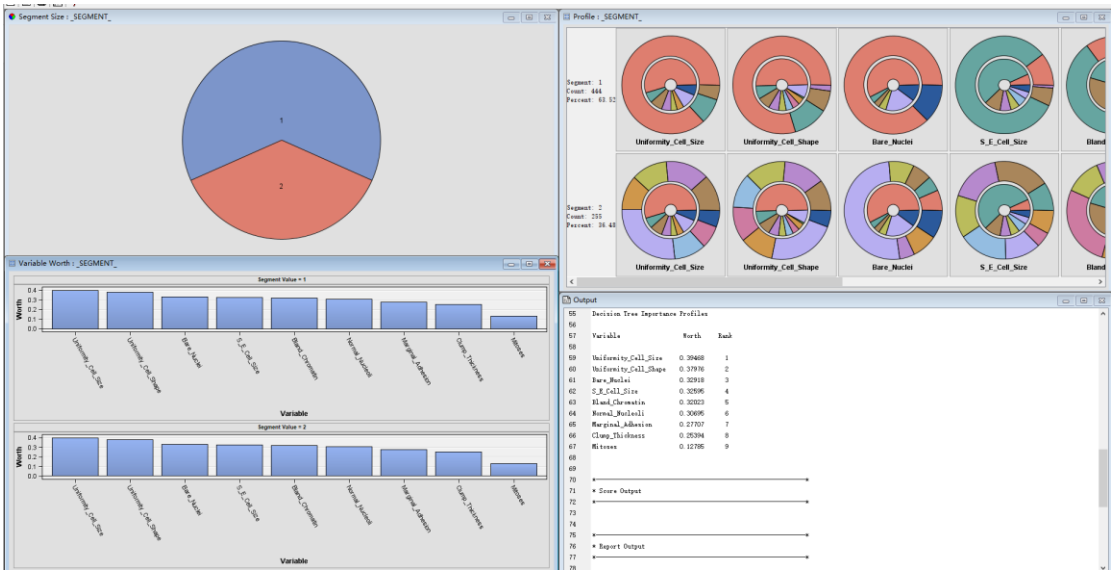

Figure 20 : Result of Segment Profile Node for Automatic Specification Method and Range as Setting for Internal Standardisation

The segment with the highest percentage in the cluster is characterized by high values for Uniformity_Cell_Size, Uniformity_Cell_Shape, and Bare_Nuclei, and low values for Mitoses. These variables are ranked highest in worth in this segment, indicating that they are the

most important predictors for this group. Other attributes, such as S_E_Cell_Size, Bland_Chromatin, and Normal_Nucleoli, also have high worth in this segment, but to a lesser extent. Marginal_Adhesion and Clump_Thickness have lower worth in this segment.

Analysis by Best Cluster Result with Real-World Problems

Best Clustering: Number of Clusters = 2. Clearly, this is an improvement. The original dataset had a 'class' attribute that only indicated two classification results for the 'target', which is in accordance with the number of segments in the cluster. As shown in the figure, the distribution of attributes in the dataset is not normal. Therefore, using 'range' is a better approach because it is like normalization and can be beneficial when variables have similar scales. This ensures that each variable has equal weight in the segmentation process.

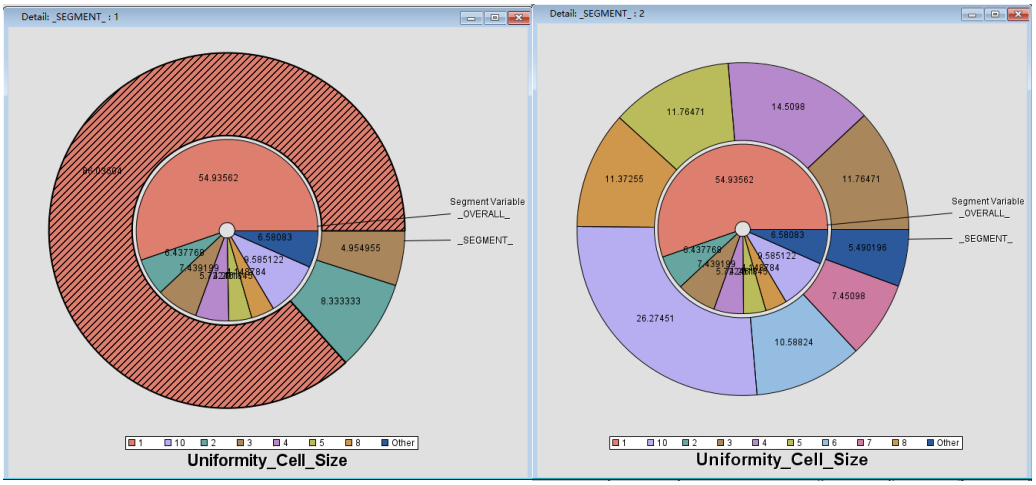

Figure 21: Segment Profile for Uniformanity_cell_size

Based on the figure above, it is apparent that all Uniformity_cell_size values of 1, 2, and 3 are assigned to Segment 1, while values greater than 3 (including some 3s) are assigned to Segment 2. Hence, we can deduce that if the Uniformity_cell_size is below 3, it will always be identified as Segment 1. If it exceeds 3, it will always be identified as Segment 2, and if it is exactly 3, it may belong to either segment.

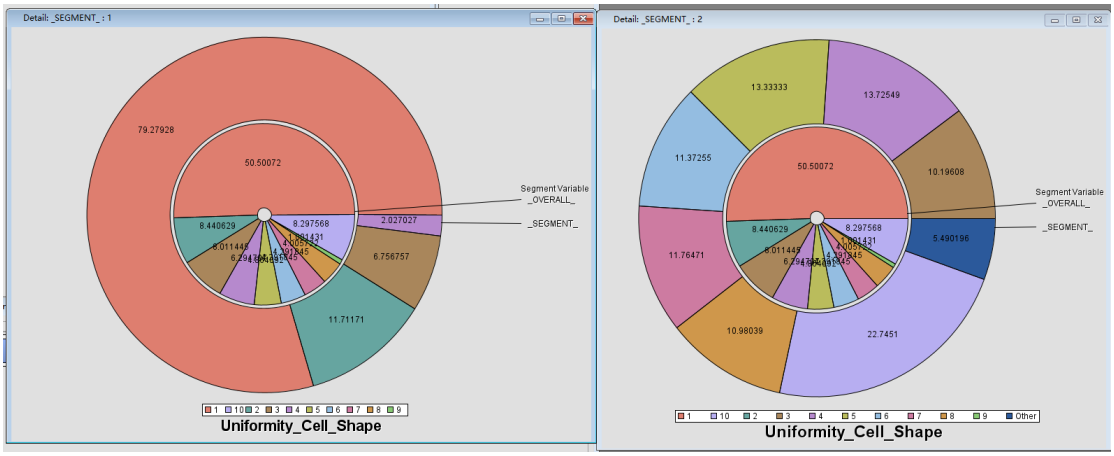

Figure 22: Segment Profile for Uniformity_Cell_Shape

Regarding Uniformity_Cell_Shape, it seems to be consistent with Uniformity_cell_size, potentially indicating a strong correlation between these two variables.

According to Bouchet-Marquis et al. (2008), quantitative analysis of breast histopathology can be used to detect ductal carcinoma in situ (DCIS) of the breast. The study found that uniformity of cell size and shape were important features in distinguishing DCIS from normal breast tissue.

Based on the data, it is reasonable to speculate that Segment 1 indicates benign breast cancer. In the medical field, doctors may consider a lower value for Uniformity cell shape and Uniformity cell size as an indication of benign breast cancer. Conversely, higher values may suggest malignant breast cancer.

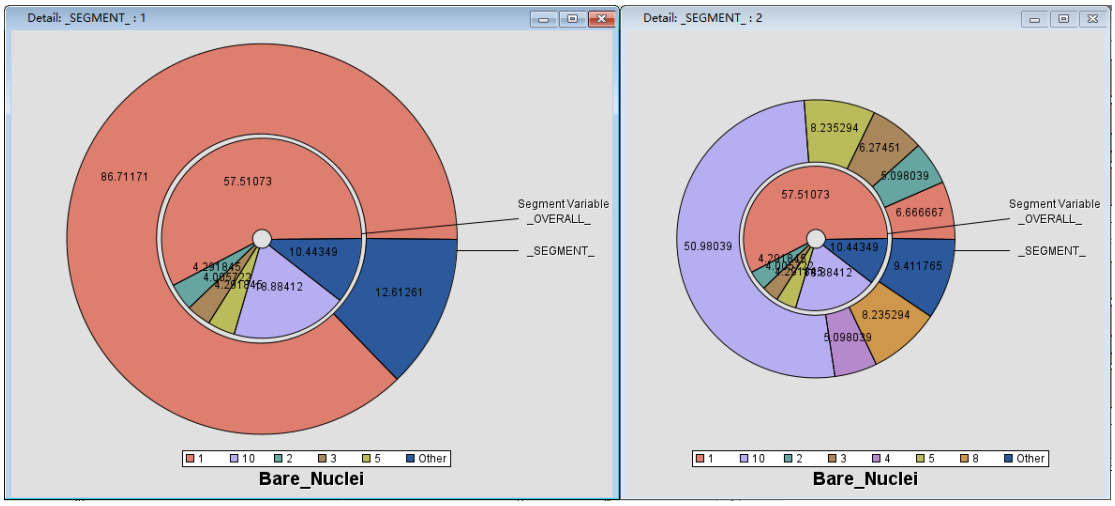

Figure 23: Segment Profile for Bare_Nuclei

Barenuclei also exhibits a similar pattern, but not all cases with Bare_Nuclei values of 1 and 2 are assigned to Segment 1. This is only a relatively higher probability scenario.

Based on the literature related to Uniformanity Cell, we can conclude that the lower the value of Bare Nuclei, the lower the possibility of having malignant breast cancer, and vice versa.

In order to further confirm this viewpoint, relevant literature will be explored. Jebarani, Umadevi, Dang, and Pomplun (2021) proposed a novel hybrid machine learning model using k-means and Gaussian mixture models for breast cancer detection. The study confirmed that Bare Nuclei is one of the important features for predicting benign and malignant breast cancer.

Next, let's take a look at the following four important attributes:

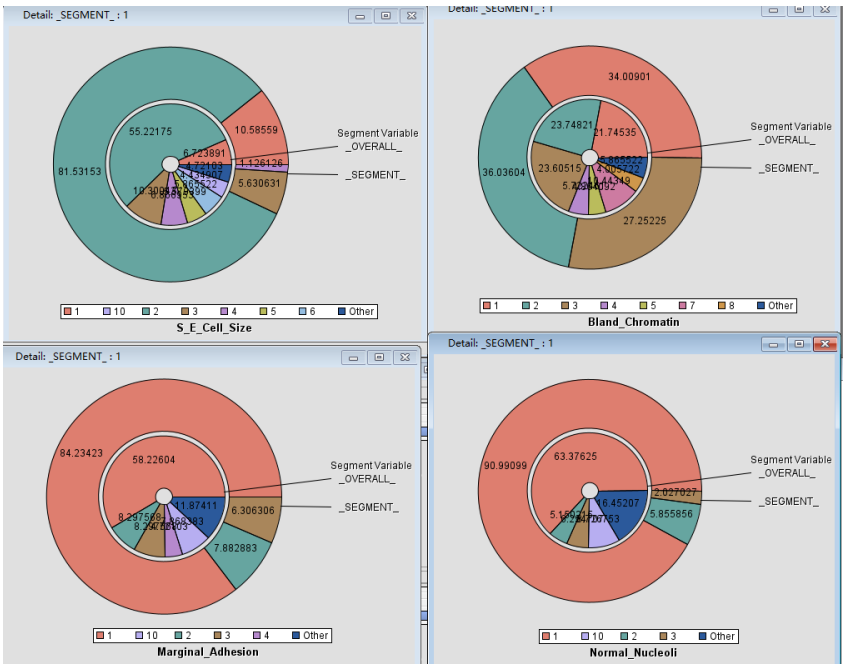

Figure 24: Segment Profile for S_E_Cell_Size, Bland_Chromatin, Marginal_Adhesion and Normal_Nucleoli

It can be concluded that lower values for these four attributes increase the chances of being assigned to segment 1 and decrease the likelihood of having malignant breast cancer. Conversely, higher values increase the chances of being assigned to segment 2 and increase the likelihood of having malignant breast cancer.

Reflection

When performing K-means calculation on data, SAS EM provides a visual interface and omits the specific code, which is extremely convenient for data exploration. However, SAS EM does not automatically explore all possibilities to set parameter values and obtain optimal results, which means that users need to have relevant knowledge and continuously try various settings to obtain the best results. In this experiment,

By setting the maximum cluster number to automatic, SAS EM will automatically select the optimal result based on the settings, eliminating the need to specify a predetermined value for k. However, determining which result is optimal and suits actual needs requires users to have sufficient understanding of the data and background knowledge. For example, when internal standardization is set to standardization, nearly 20 segments will be obtained as the clustering result regardless of other parameter settings. However, by reading the description of the original dataset, we can determine that the actual target has only two possibilities. Based on this information, we can determine that the optimal clustering solution is k=2 when internal standardization is set to range.

Furthermore, to determine which segment represents benign and which represents malignant, it is necessary to read related research literature. This information cannot be directly obtained through k-means.

[Word Count: 2433, excluding references, figures/diagrams, tables, figure/table captions, coversheet and table of contents.]

References

Jebarani, P. Esther., Umadevi, N., Dang, H., & Pomplun, M. (2021). A Novel Hybrid K-Means and GMM Machine Learning Model for Breast Cancer Detection. *IEEE Access*, 1–1. https://doi.org/10.1109/access.2021.3123425

Liu, S., Zeng, J., Gong, H., Yang, H., Zhai, J., Cao, Y., Liu, J., Luo, Y., Li, Y., Maguire, L., & Ding, X. (2018). Quantitative analysis of breast cancer diagnosis using a probabilistic modelling approach. *Computers in Biology and Medicine*, *92*, 168–175. https://doi.org/10.1016/j.compbiomed.2017.11.014