Tan Yin Yen (WQD180108)

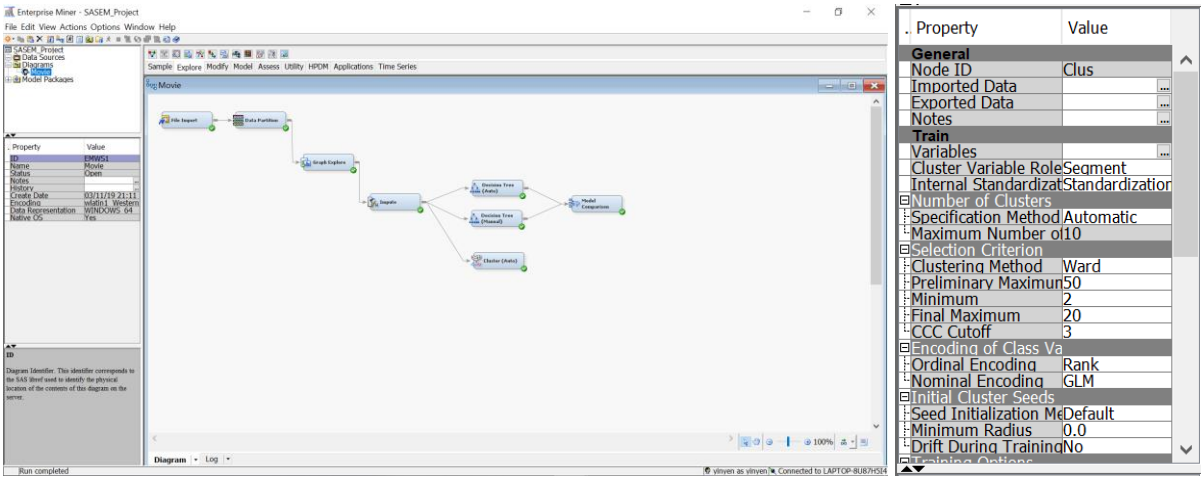
Data Mining Project: Milestone 5 (Communication of Insights of Data)

Analysis Goal

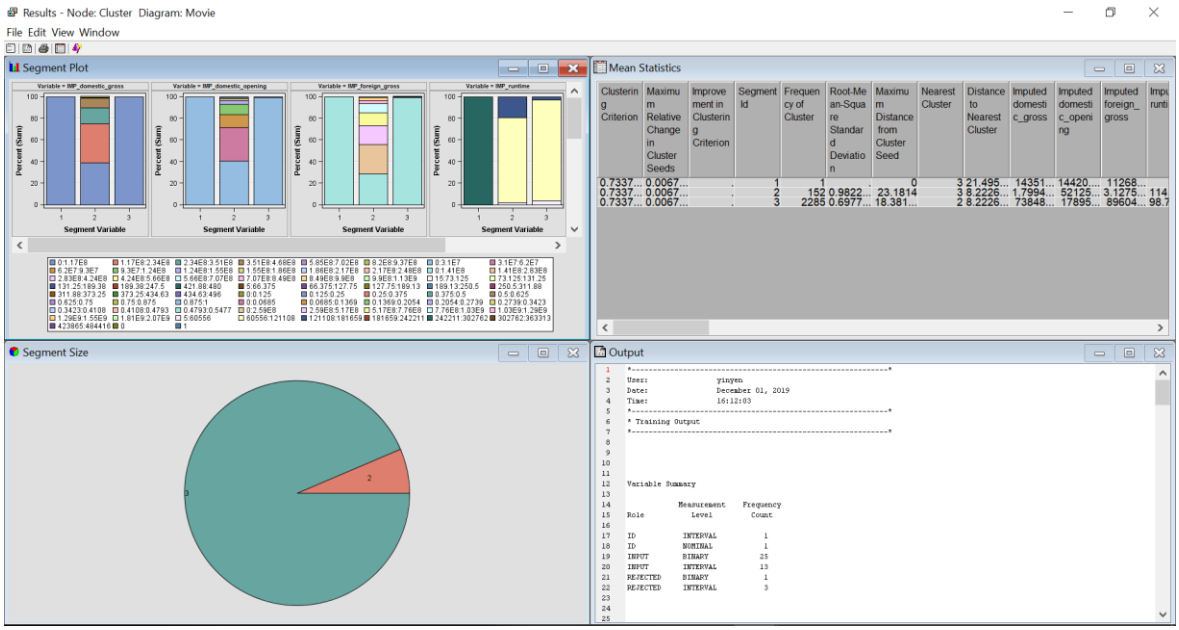
A movie streaming company (Netflix) seeks to maximize customer's retention by recommending highly rated movies with DVD or streaming options available to their users. Use sentiment score of user reviews on a movie, movie information and box office data to predict the user ratings of a movie.

By predicting the user ratings of a movie based on its reviews and box office achievement, the movie streaming company can filter out latest movies with DVD or streaming options available that are highly rated and recommend them to its users. Customers who are satisfied with the movie recommendations are more likely to subscribe to the movie streaming service in the next month.

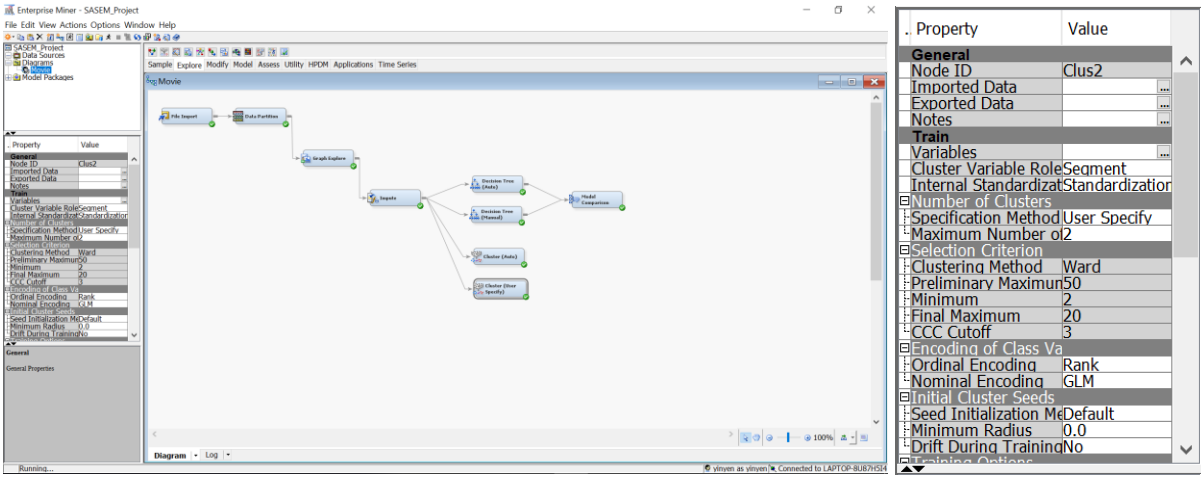
SEMMA which stands for sample, explore, modify, model and assess refers to the core process of conducting data mining. SEMMA process is used in the assignment to predict the user ratings of a movie. In this milestone, the Explore stage of SEMMA process is conducted to identify the patterns and insights of the data. The procedures of exploration are shown, and insights of the data identified are discussed as follows.



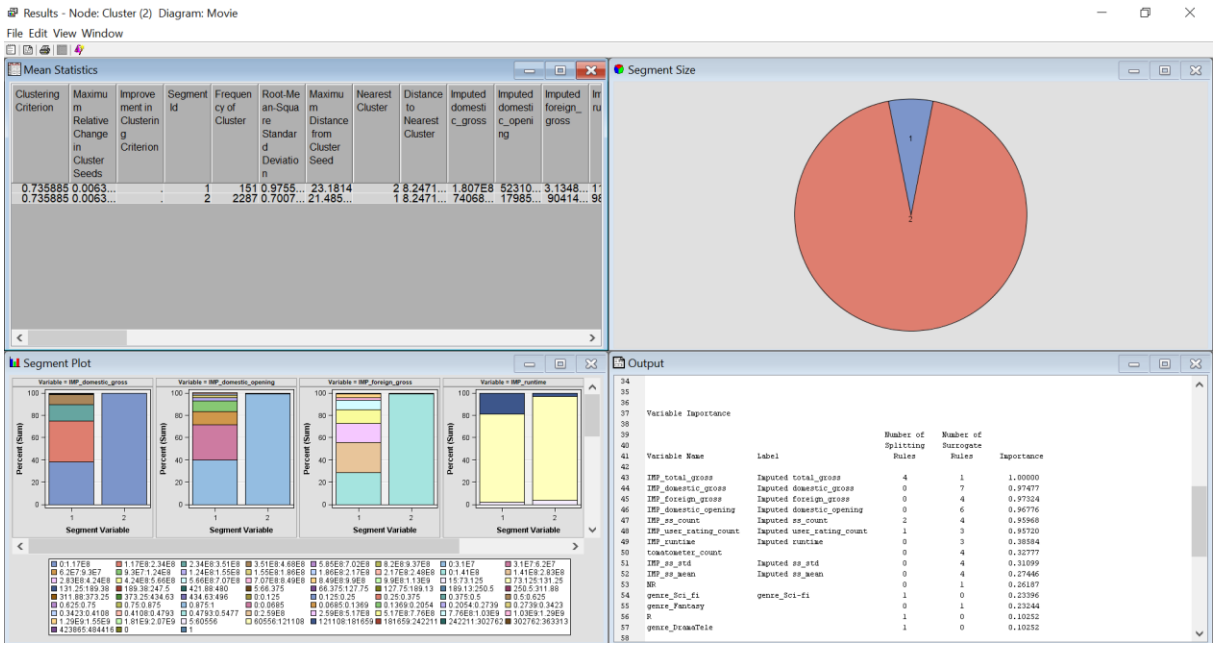
Hierarchical clustering is used to explore the data in this milestone. The cluster node is first added to the Movie diagram. The specification method is set to be automatic and maximum number of clusters is set to be 10.



From the figure shown above, the data is clustered into three segments. The third segment can be hardly seen from the segment size plot as the size of the segment is too small (1 case). Thus, the maximum number of clusters is suggested to be 2 instead.



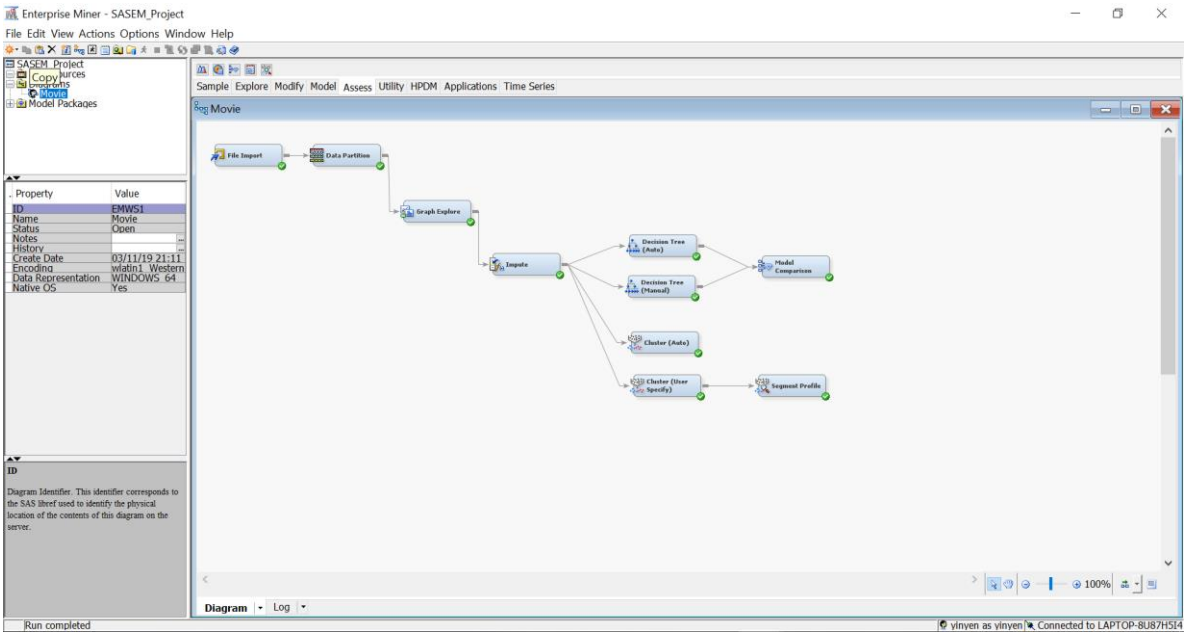
Another cluster node is added to the Movie diagram. The specification method is set to be user specify and maximum number of clusters is set to be 2.



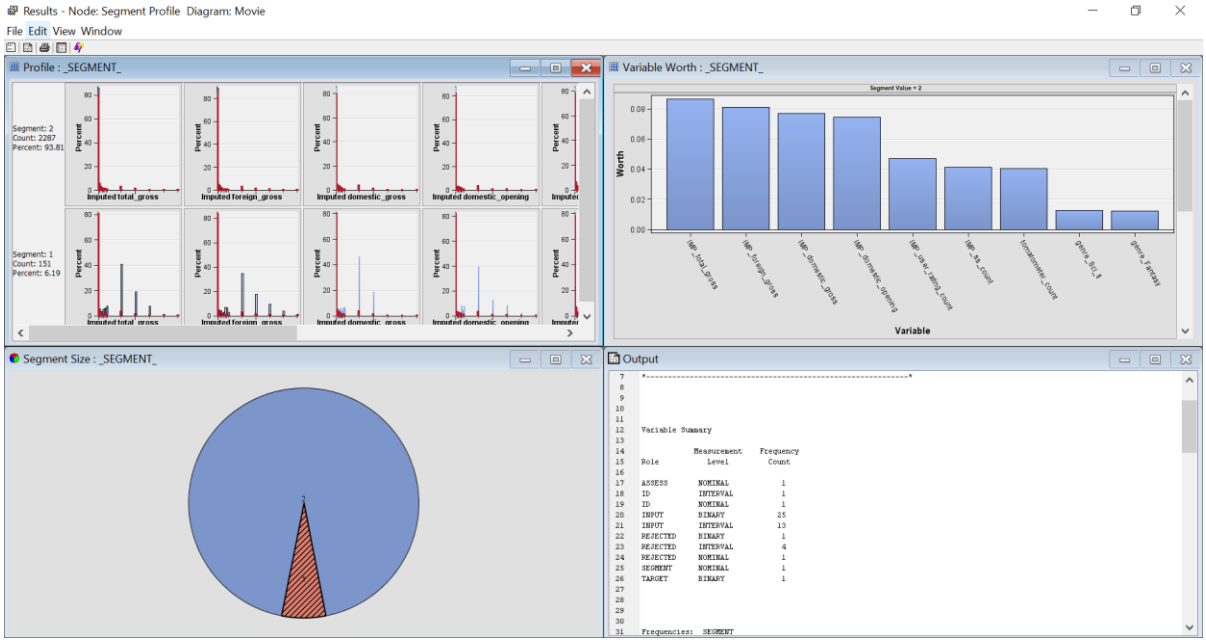
The results of hierarchical clustering is shown as above. The Cluster window contains four embedded windows.

- The Mean Statistics window lists various descriptive statistics by cluster.
- The Segment Plot window shows the distribution of input variables by cluster.
- The Segment Size shows a pie chart describing the size of each cluster formed.
- The Output window shows the output of various SAS procedures run by the Cluster node.

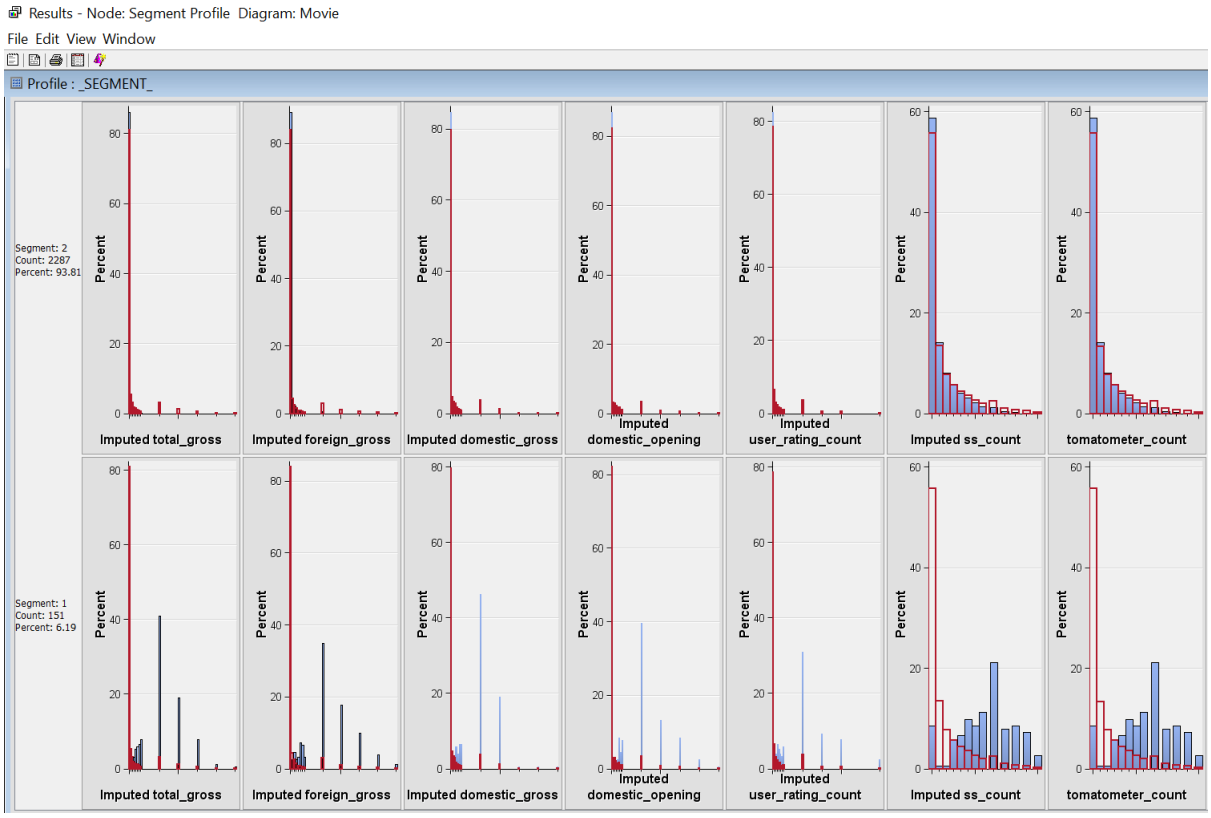
As seen in the Segment Size window, two clusters are formed, representing good and bad movie respectively. From the Mean Statistics windows, the segment frequency counts are 151 and 2287 cases respectively.



A segment profile node is then added to the Movie diagram and connected to the Cluster node.



The results of the Segment Profile are shown as above.



As seen in the Profile window, box office and number of reviews are more significant for clustering compared to the genres of the movie. The red graph represents the original distribution and the blue graph represents the distribution of the clustered data. Segment 1 contains only 151 cases whereas segment 2 contains 2287 cases. For segment 1, the overall box office distribution and number of reviews are higher than average. For segment 2, the overall box office distribution and number of reviews are

lower than average. Therefore, segment 1 can be considered as good and recommended movie as it has a higher-than-average box office and higher-than-average number of reviews, whereas segment 2 slants towards bad movies and requires further investigation.