Cheah Jun Yitt (WQD180107)

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

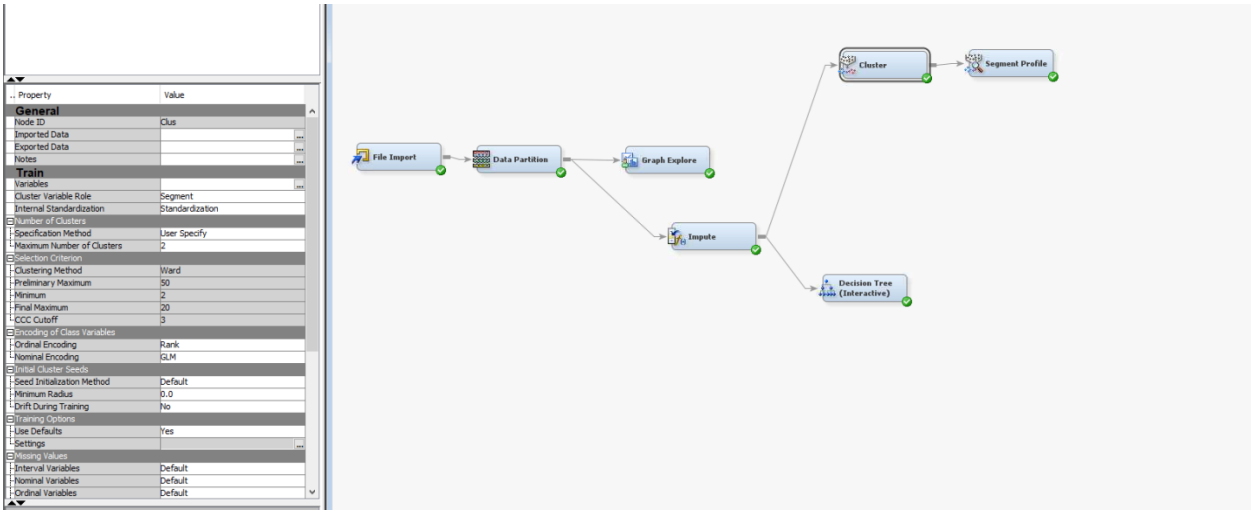
Recap

Previous milestones have established the following:

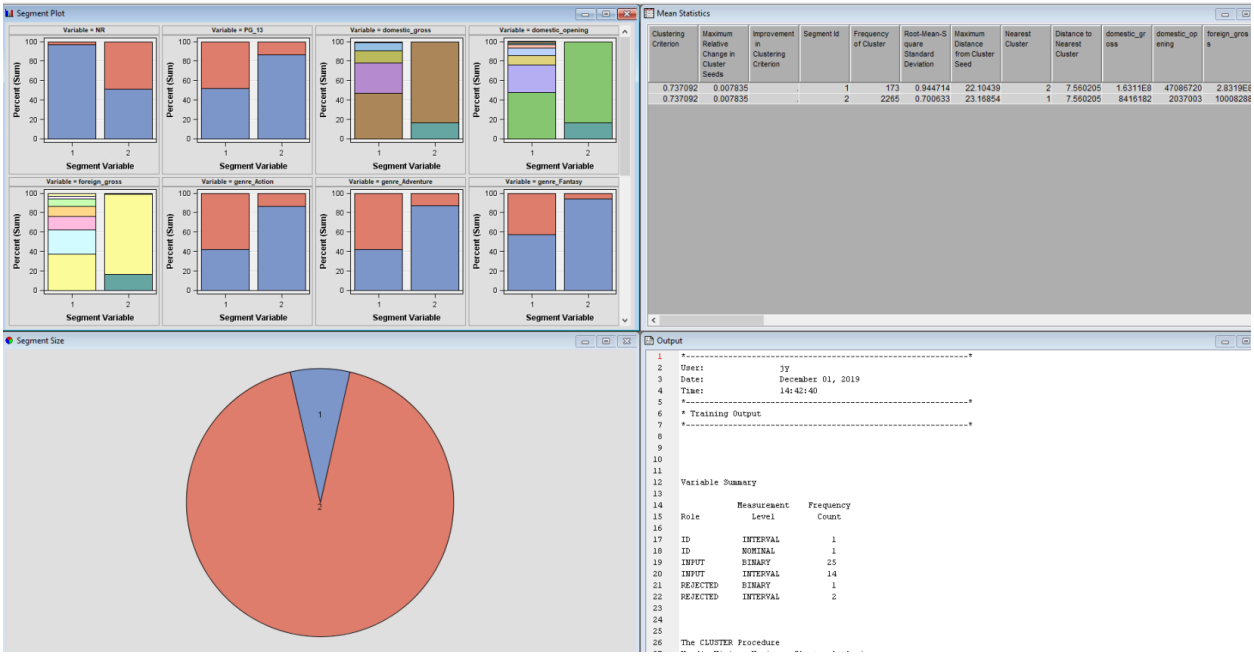
1) The need to have various sources of data, both structured and unstructured data. The data that we have collected are as shown below:
    1) Structured data: Movie information scraped from rottentomatoes.com
    2) Unstructured data (text data): Movie reviews scraped from rottentomatoes.com
    3) Structured data: Movie box office data scraped from boxofficemojo.com
2) Target variable (*audience_score_positive*), which indicates whether a movie is good or bad based on audience ratings on the movie.
3) Analysis Goal: To use sentiment score of user reviews on a movie, movie information and box office data to predict the user ratings of a movie.
4) SEMMA (Part 1: SEM):
    a. Sample – Import data from a CSV file
    b. Explore – Use histograms to identify missing values, inconsistencies or hidden patterns in the data.
    c. Modify – Impute missing values.
5) SEMMA (Part 2: MA): Interpretation of data by partitioning the data into training and validation set, and model the data using decision tree. Prediction of whether a movie is good or bad is performed using the decision tree. The nodes of the decision tree will give decision maker insights on what kind of movie is considered good or bad.

Communication of Insights

Hierarchical Clustering

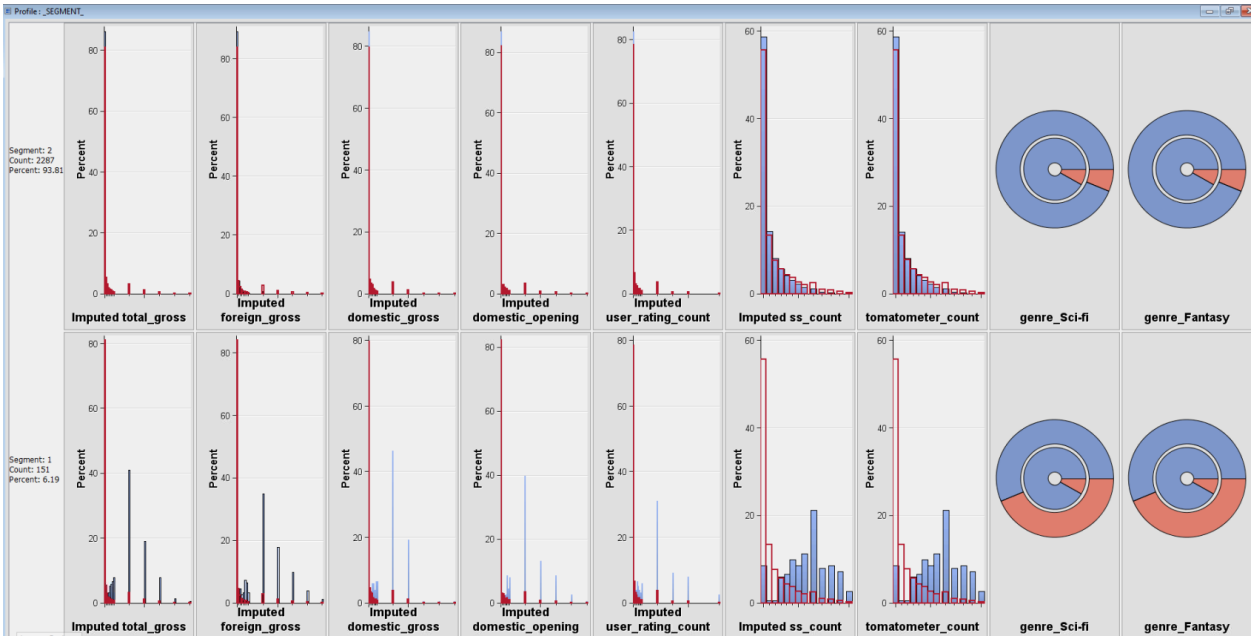


In this milestone, hierarchical clustering is performed on the data to produce segments useful for decision makers in movie streaming company to suggest movies to user. The *Cluster* and *Segment Profile* is added to the diagram as shown above. The clustering method is selected as "Ward", which represents hierarchical clustering. The number of clusters is specified to 2 to represent good and bad movie.



The number of segments obtained from the hierarchical clustering is 2 as shown above. Segment profiling is performed to determine whether the two segments represent good and bad movie respectively.

Segment Profile



The two segments can be described with a few attributes with the highest worth, as follow:

Segment 1: This segment contains movies with higher than average total gross amount, foreign gross amount, domestic gross amount, domestic opening amount, number of user rating, number of reviews and number of official reviews. All these suggest that the segment contain movie that are widely recognized with dominant box office performance, that are worth recommending to the users.

Segment 2: This segment contains movies with slightly lower than average and with a similar distribution with the overall sample for attributes like total gross amount, foreign gross amount, domestic gross amount, domestic opening amount, number of user rating, number of reviews and number of official reviews. All these suggest that the segment contains movies with slightly lower box office performance and should be further analyzed before suggesting them to a user.