Group Members:

Cheah Jun Yitt (WQD180107)

Tan Yin Yen (WQD180108)

Milestone 3 (Processing of Data)

1. Analysis Goal

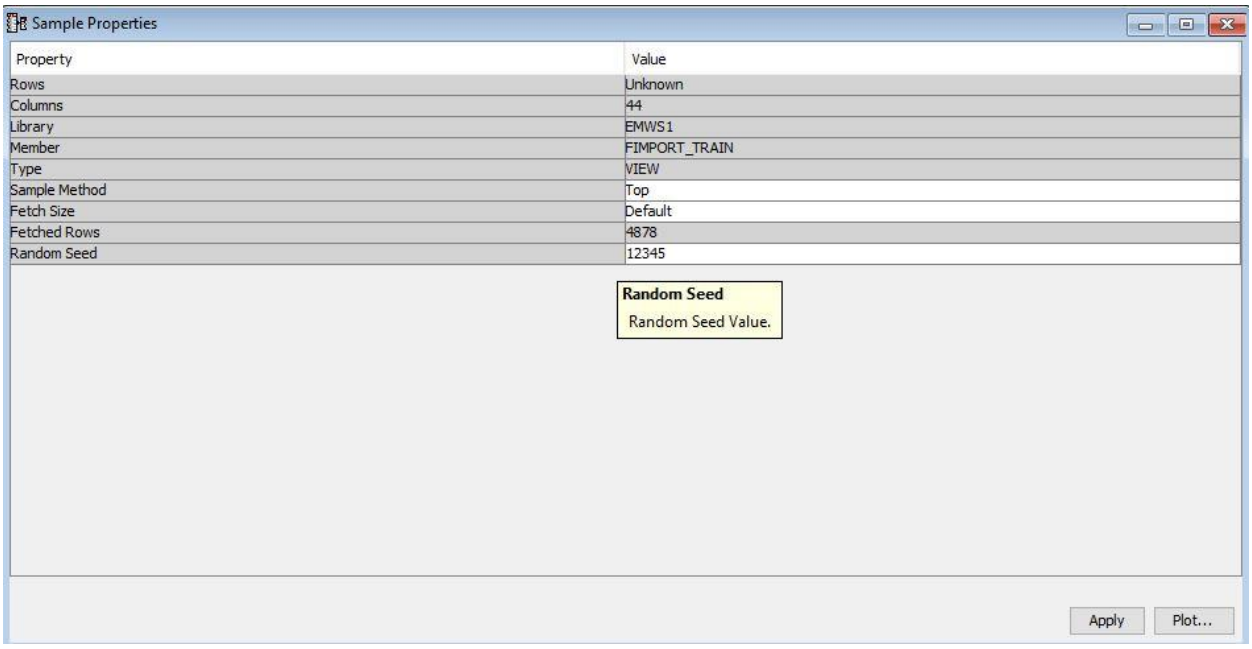
A movie streaming company (Netflix) seeks to maximize customer's retention by recommending highly rated movies with DVD or streaming options available to their users. Use sentiment score of user reviews on a movie, movie information and box office data to predict the user ratings of a movie.

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.

2. Analysis Data

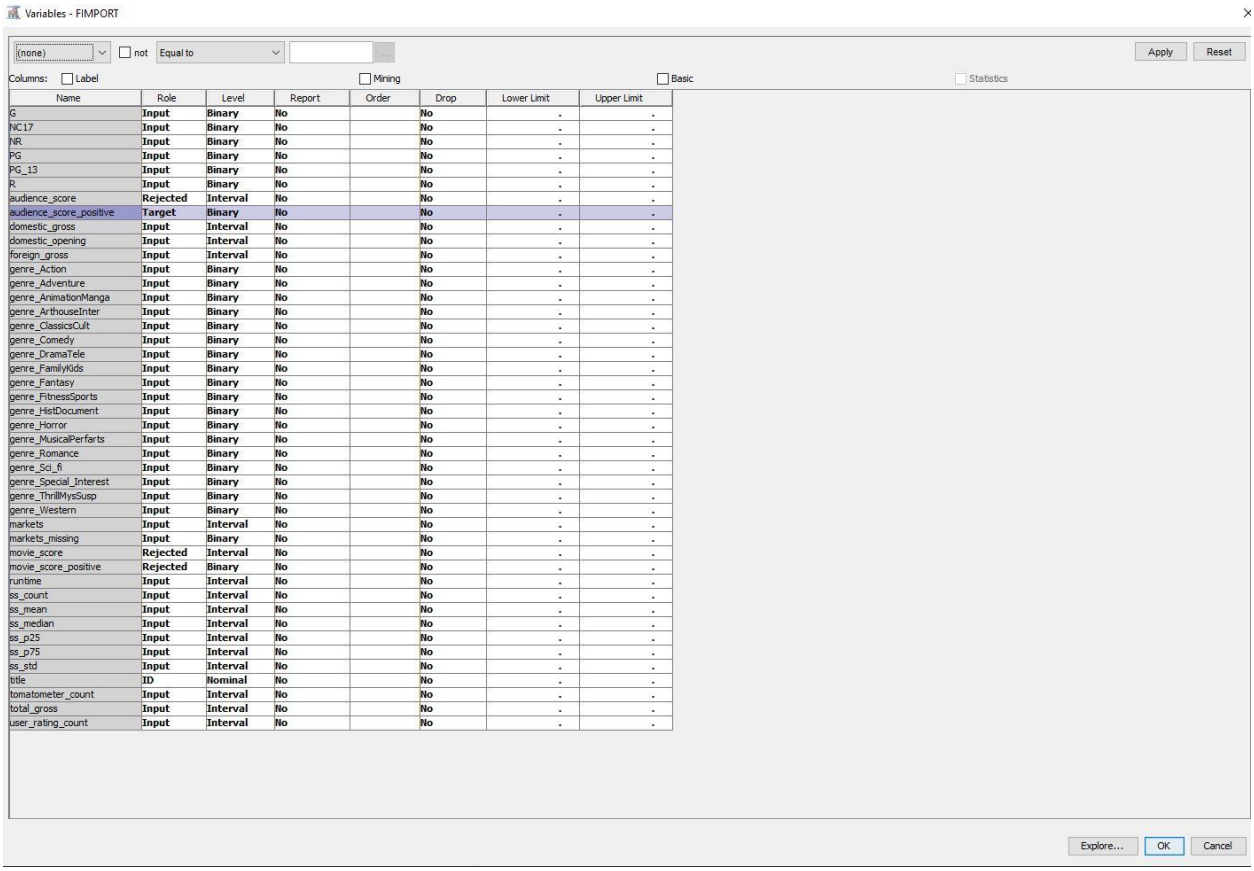
Movie information and movie reviews data were scraped from rottentomatoes.com. Movie box office data were scraped from boxofficemojo.com. The binary target variable (audience score indicator) is balance, i.e. 50% good and 50% bad.

3. Table properties



The input data has a total of 4878 rows (observations) and 44 columns (variables/attributes).

4. Column Metadata



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
G	Input	Binary	No		No	.	.
NC17	Input	Binary	No		No	.	.
NR	Input	Binary	No		No	.	.
PG	Input	Binary	No		No	.	.
PG_13	Input	Binary	No		No	.	.
R	Input	Binary	No		No	.	.
audience_score	Rejected	Interval	No		No	.	.
audience_score_positive	Target	Binary	No		No	.	.
domestic_gross	Input	Interval	No		No	.	.
domestic_opening	Input	Interval	No		No	.	.
foreign_gross	Input	Interval	No		No	.	.
genre_Action	Input	Binary	No		No	.	.
genre_Adventure	Input	Binary	No		No	.	.
genre_AnimationManga	Input	Binary	No		No	.	.
genre_ArthouseInter	Input	Binary	No		No	.	.
genre_ClassicsCult	Input	Binary	No		No	.	.
genre_Comedy	Input	Binary	No		No	.	.
genre_DramaTele	Input	Binary	No		No	.	.
genre_FamilyKids	Input	Binary	No		No	.	.
genre_Fantasy	Input	Binary	No		No	.	.
genre_FitnessSports	Input	Binary	No		No	.	.
genre_HistDocument	Input	Binary	No		No	.	.
genre_Horror	Input	Binary	No		No	.	.
genre_MusicalPerfarts	Input	Binary	No		No	.	.
genre_Romance	Input	Binary	No		No	.	.
genre_Sci_fi	Input	Binary	No		No	.	.
genre_Special_Interest	Input	Binary	No		No	.	.
genre_ThrillMysSusp	Input	Binary	No		No	.	.
genre_Western	Input	Binary	No		No	.	.
markets	Input	Interval	No		No	.	.
markets_missing	Input	Binary	No		No	.	.
movie_score	Rejected	Interval	No		No	.	.
movie_score_positive	Rejected	Binary	No		No	.	.
runtime	Input	Interval	No		No	.	.
ss_count	Input	Interval	No		No	.	.
ss_mean	Input	Interval	No		No	.	.
ss_median	Input	Interval	No		No	.	.
ss_p25	Input	Interval	No		No	.	.
ss_p75	Input	Interval	No		No	.	.
ss_std	Input	Interval	No		No	.	.
title	ID	Nominal	No		No	.	.
tomatometer_count	Input	Interval	No		No	.	.
total_gross	Input	Interval	No		No	.	.
user_rating_count	Input	Interval	No		No	.	.

For attributes that represent the MPAA (Motion Picture Association of America) film rating system, such as G, NC17, …, R, they are binary attributes and act as predictors to the target variable (audience_score_positive). For example, if G is true, then the movie's rating is classified as General Audience; while if NC17 is true, then the movie should not be viewed by children under the age of 17. The details are:

- **G**: General audiences – All ages admitted
- **PG**: Parental guidance suggested – Some material may not be suitable for children.
- **PG-13**: Parents strongly cautioned – Some material may be inappropriate for children under 13.
- **R**: Restricted – Under 17 requires accompanying parent or adult guardian.
- **NC-17**: No one 17 and under admitted.
- **NR:** Not Rated

Similarly, the 11 genre clusters are binary attributes. The genre clusters are 'genre_Action', 'genre_Adventure', 'genre_Comedy', 'genre_Fantasy', 'genre_Horror', 'genre_Romance', 'genre_Sci-fi', 'genre_Special Interest', 'genre_Western', 'genre_FamilyKids', 'genre_AnimationManga', 'genre_FitnessSports', 'genre_DramaTele', 'genre_MusicalPerfarts', 'genre_ClassicsCult', 'genre_ArthouseInter', 'genre_ThrillMysSusp', 'genre_HistDocument'.

These genre clusters were identified using domain knowledge, where similar genres were group into a genre cluster, as follows:

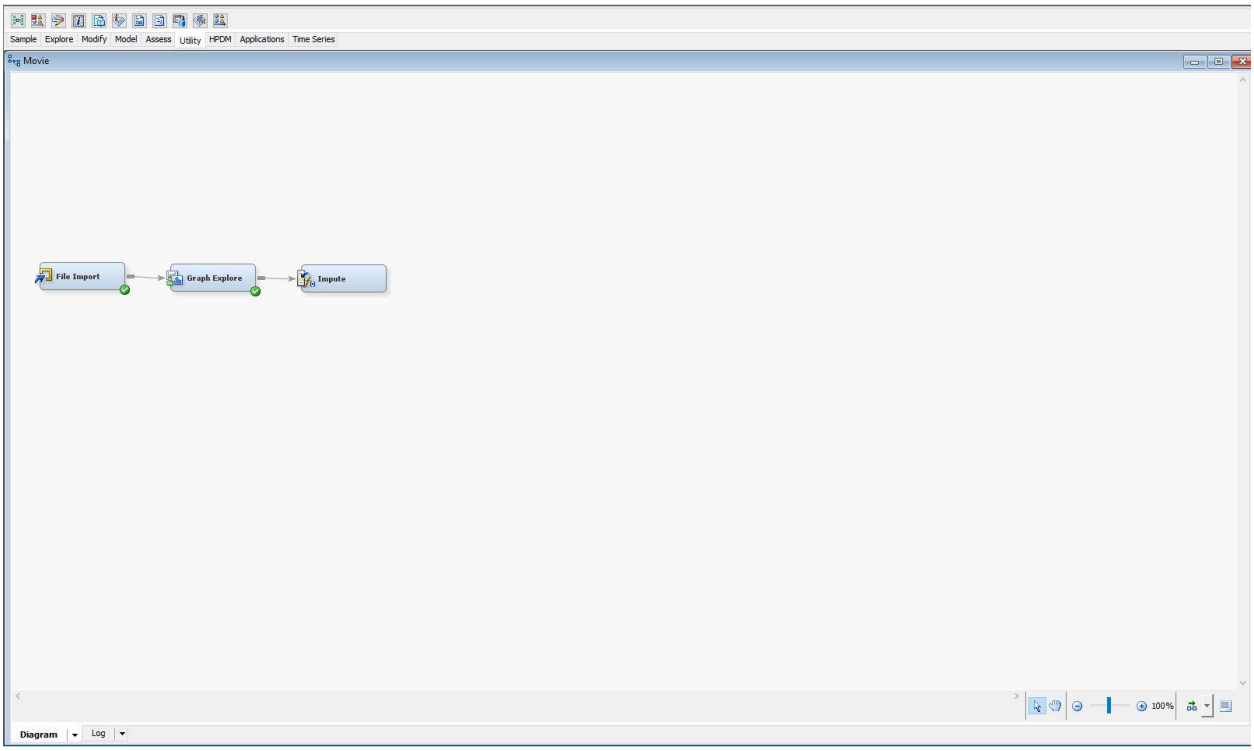
i. genre_Action: Action (movies that exhibit action theme)
ii. genre_Adventure: Adventure (movies that exhibit adventure theme)
iii. genre_AnimationManga: Animation, Manga (movies that are animated or have japanese manga reference)
iv. genre_ArthouseInter: Art House, International (international movies)
v. genre_ClassicsCult: Classics, Cult Movies (movies that exhibit classical or are cult classics)
vi. genre_Comedy: Comedy (comedy movie)
vii. genre_DramaTele: Drama, Television (movies that are drama or TV series based)
viii. genre_FamilyKids: Family, Kids (movies for family and kids)
ix. genre_Fantasy: Fantasy (movies that exhibit a fantasy theme)
x. genre_FitnessSports: Fitness, Sports (movies that exhibit fitness or sports theme)
xi. genre_HistDocument: History, Documentary (documentary films or movies that are based on history)
xii. genre_Horror: Horror (horror movie)
xiii. genre_MusicalPerfarts: Musical, Performing Arts (movies that exhibit musical or performings arts theme)
xiv. genre_Romance: Romance (movies that exhibit a romance theme)
xv. genre_Sci_fi: Sci-fi (Science fiction movies)
xvi. genre_Special_Interest (miscellaneous movies)
xvii. genre_ThrillMysSusp: Thriller, Mystery, Suspense (movies that exhibit thriller, mystery or suspense theme)
xviii. genre_Western: Western (movies that exhibit a western theme)

The title (movie title) of nominal data type is set as the ID, used to identify an observation, hence should not be used in the analysis.

The sentiment score attributes, box office values, and number of ratings are all interval values.

audience_score, audience_score_positive, movie_score, movie_score_positive are the target variables. Currently, we are only interested in the audience_score_positive binary target, hence the other 3 target variables were rejected.

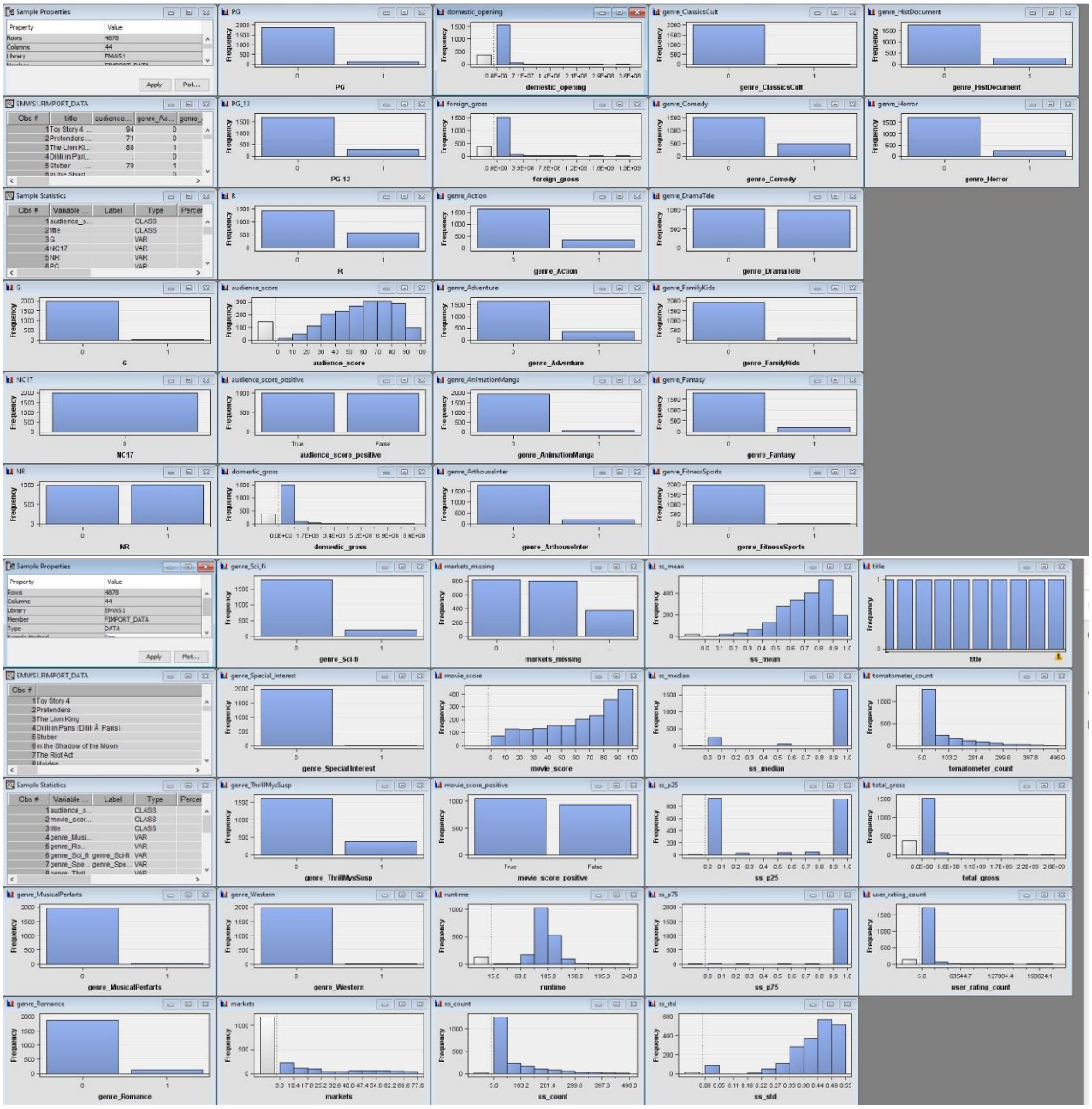
5. Diagram (SEM)



SEM (Sample, Explore, Modify) is performed on the input data.

1. Sample – Data is imported from a CSV file.
2. Explore – The attributes are explored using histograms to identify missing values, any inconsistencies in the data, or any hidden patterns.
3. Modify – Missing values were imputed using some pre-defined methods.

6. Exploring the Data Source

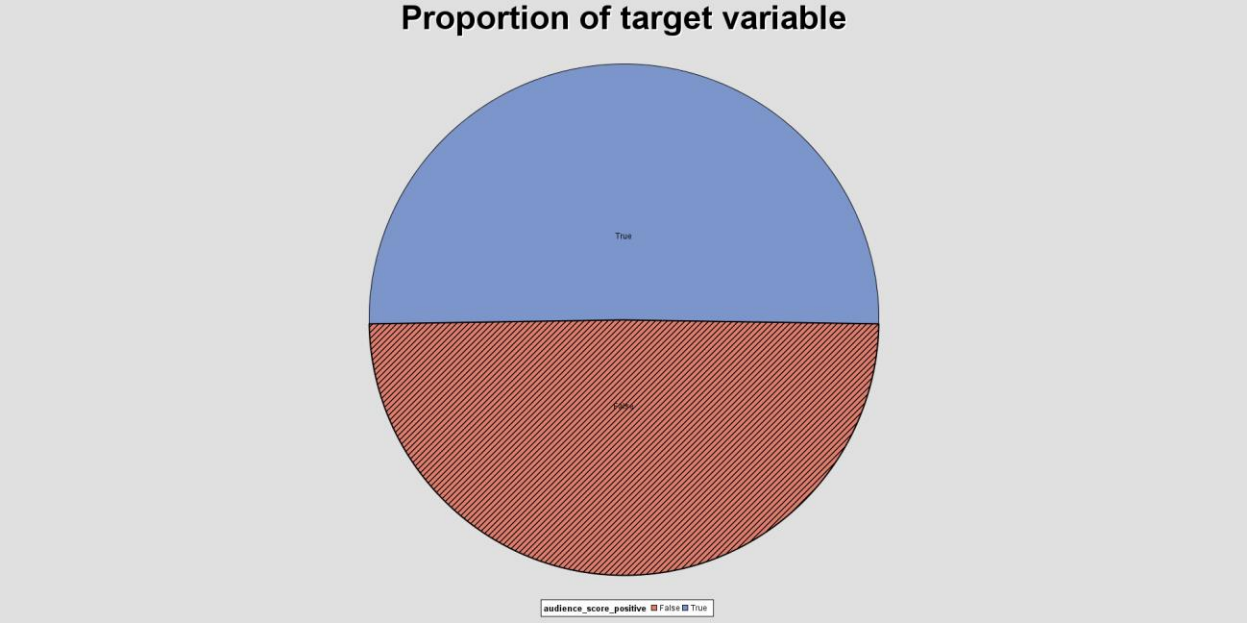


Missing data can be identified from the above histograms. For example, *domestic_opening, foreign_gross, audience_score, domestic_gross, total_gross,runtime, user_rating_count, markets, ss_mean, ss-median, ss_p25, ss_p75, ss_count, ss_std* have missing values.

7. Cleansing of Data

The missing values were imputed using the Tree Surrogate method in SAS Enterprise Miner.

8. Pie Chart of Target Variable



The pie chart above shows that the target variable (*audience_score_positive*) has a proportion of approximately 50% True (good movie) and 50% False (bad movie). This shows that the target class is balance, and can be model