**Group Members:**

Cheah Jun Yitt (WQD180107)

Tan Yin Yen (WQD180108)

**Data Mining Project: Milestone 3 (Processing of Data)**

**1. Analysis Goal**

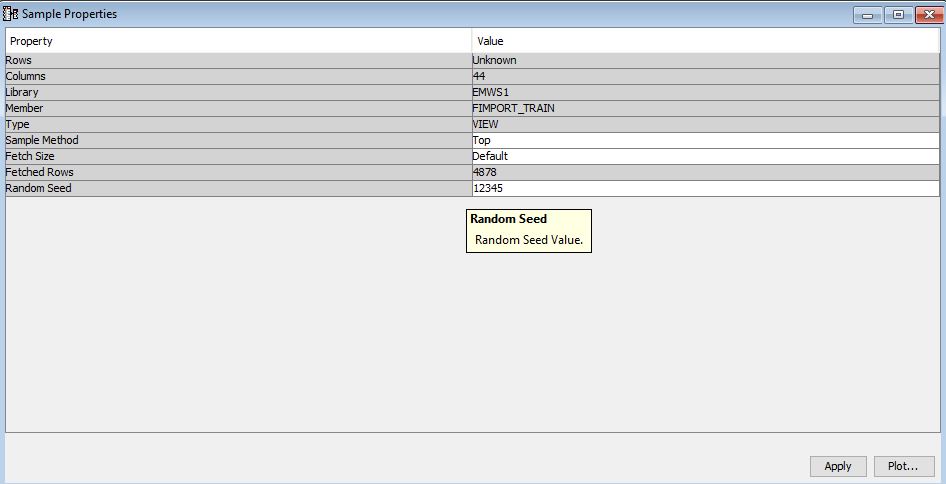
A movie streaming company (Netflix) seeks to maximize customer's retention by recommending highly rated movies with DVD or streaming options available to their users. Use sentiment score of user reviews on a movie, movie information and box office data to predict the user ratings of a movie.

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.

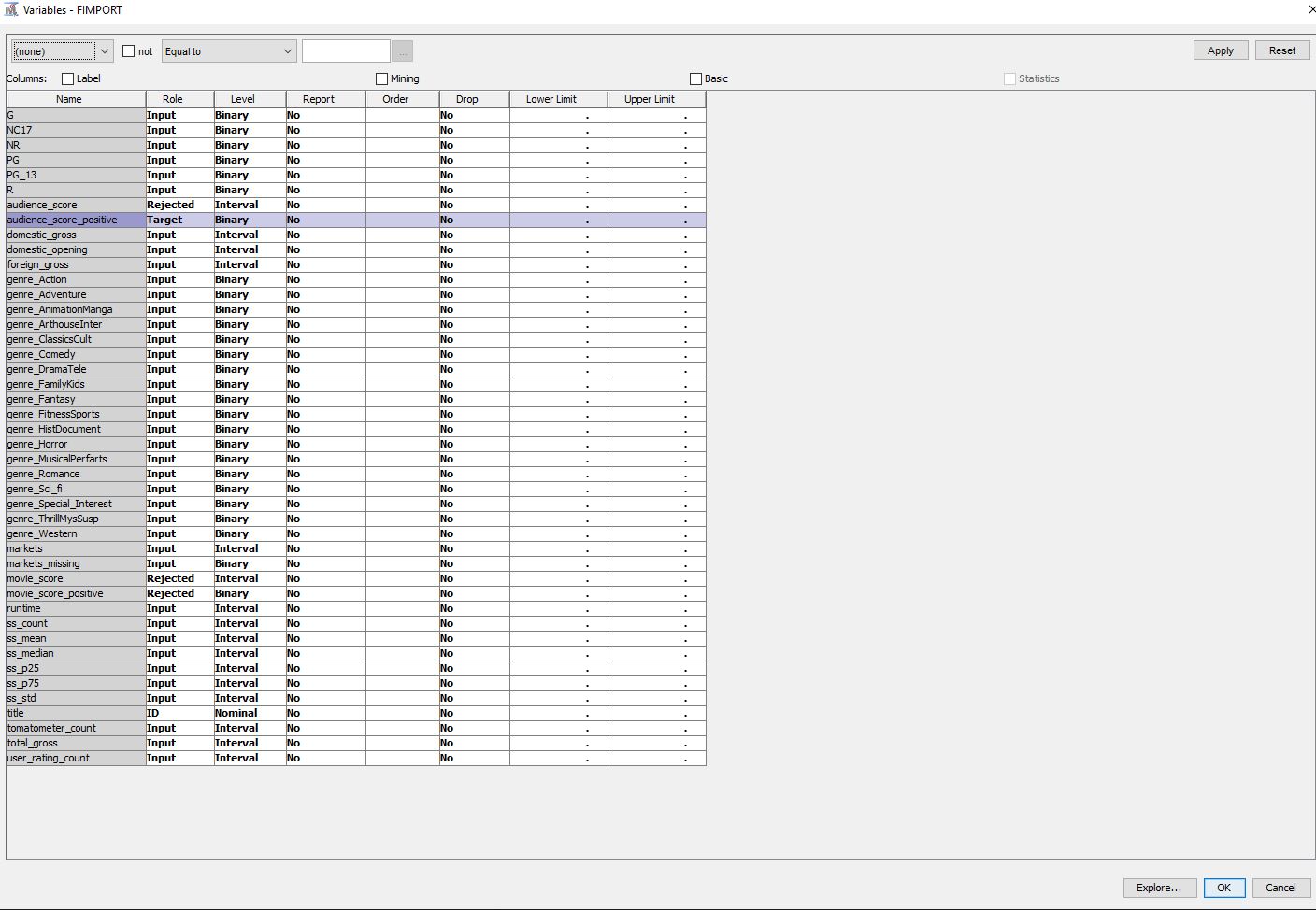
**2. Analysis Data**

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\_positive*) 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**

****

For attributes that represent the MPAA (Motion Picture Association of America) film rating system, such as G, NC17, NR, PG, PG\_13, and 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:

1. genre\_Action: Action (movies that exhibit action theme)
2. genre\_Adventure: Adventure (movies that exhibit adventure theme)
3. genre\_AnimationManga: Animation, Manga (movies that are animated or have japanese manga reference)
4. genre\_ArthouseInter: Art House, International (international movies)
5. genre\_ClassicsCult: Classics, Cult Movies (movies that exhibit classical or are cult classics)
6. genre\_Comedy: Comedy (comedy movie)
7. genre\_DramaTele: Drama, Television (movies that are drama or TV series based)
8. genre\_FamilyKids: Family, Kids (movies for family and kids)
9. genre\_Fantasy: Fantasy (movies that exhibit a fantasy theme)
10. genre\_FitnessSports: Fitness, Sports (movies that exhibit fitness or sports theme)
11. genre\_HistDocument: History, Documentary (documentary films or movies that are based on history)
12. genre\_Horror: Horror (horror movie)
13. genre\_MusicalPerfarts: Musical, Performing Arts (movies that exhibit musical or performings arts theme)
14. genre\_Romance: Romance (movies that exhibit a romance theme)
15. genre\_Sci\_fi: Sci-fi (Science fiction movies)
16. genre\_Special\_Interest (miscellaneous movies)
17. genre\_ThrillMysSusp: Thriller, Mystery, Suspense (movies that exhibit thriller, mystery or suspense theme)
18. 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.

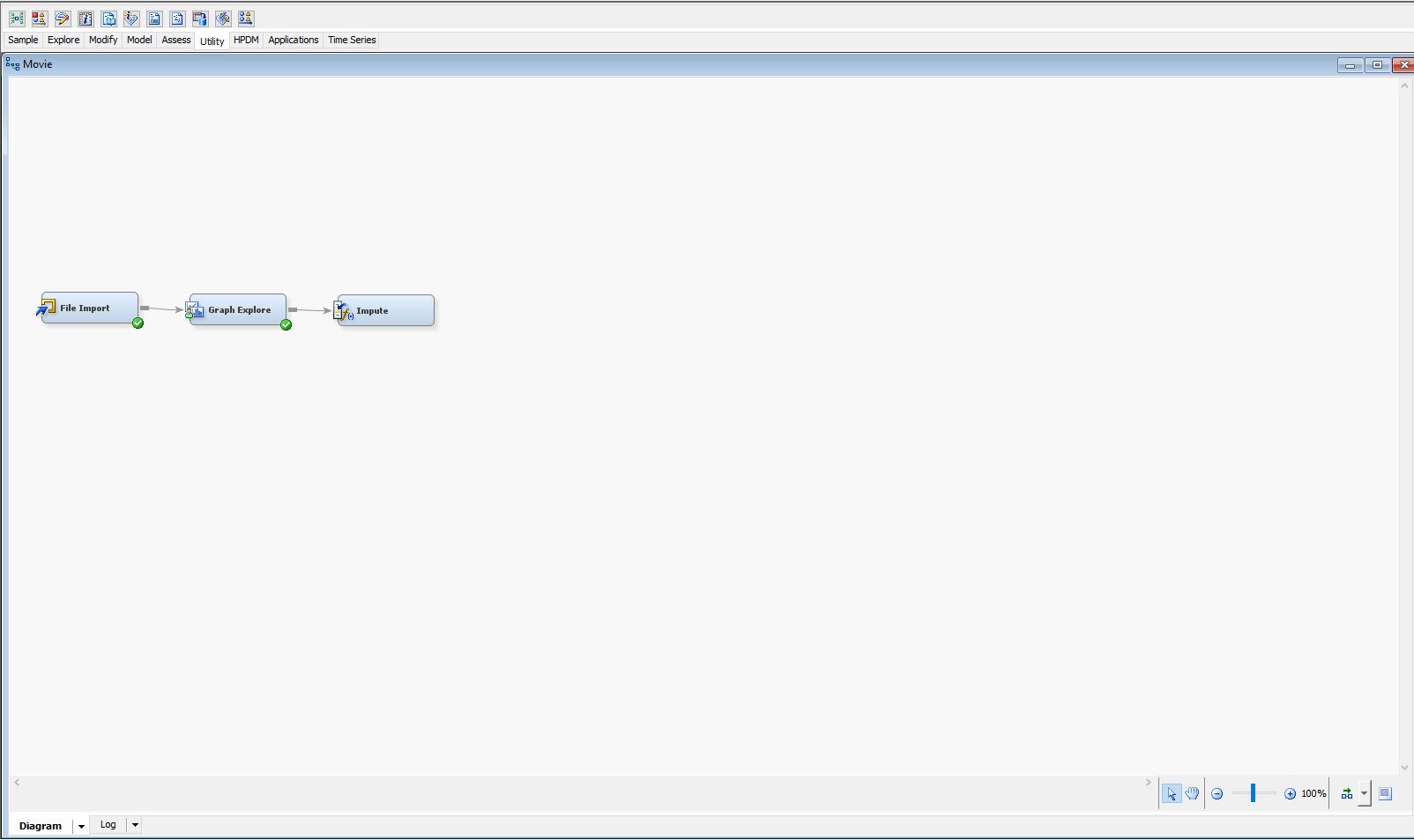
The sentiment score attributes, box office values, and number of ratings are all interval values.

*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.

**Metadata Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **NAME** | **ROLE** | **LEVEL** | **DESCRIPTION** |
| G | INPUT | BINARY | General audiences – All ages admitted |
| NC17 | INPUT | BINARY | No one 17 and under admitted |
| NR | INPUT | BINARY | Not Rated |
| PG | INPUT | BINARY | Parental guidance suggested – Some material may not be suitable for children. |
| PG\_13 | INPUT | BINARY | Parents strongly cautioned – Some material may be inappropriate for children under 13. |
| R | INPUT | BINARY | Restricted – Under 17 requires accompanying parent or adult guardian. |
| audience\_score | REJECTED | INTERVAL | The audience rating in rottentomatoes.com |
| audience\_score\_positive | TARGET | BINARY | A binary indicator that indicates whether the movie is good or not (in the perspective of the audience) |
| domestic\_gross | INPUT | INTERVAL | Total domestic gross amount ($) |
| domestic\_opening | INPUT | INTERVAL | Total domestic opening gross amount ($) |
| foreign\_gross | INPUT | INTERVAL | Total foreign gross amount ($) |
| genre\_Action | INPUT | BINARY | Action genre |
| genre\_Adventure | INPUT | BINARY | Adventure genre |
| genre\_AnimationManga | INPUT | BINARY | Animation or Manga genre |
| genre\_ArthouseInter | INPUT | BINARY | Art House or International genre |
| genre\_ClassicsCult | INPUT | BINARY | Classics or Cult Movies genre |
| genre\_Comedy | INPUT | BINARY | Comedy genre |
| genre\_DramaTele | INPUT | BINARY | Drama or Television genre |
| genre\_FamilyKids | INPUT | BINARY | Family or Kids genre |
| genre\_Fantasy | INPUT | BINARY | Fantasy genre |
| genre\_FitnessSports | INPUT | BINARY | Fitness or Sports genre |
| genre\_HistDocument | INPUT | BINARY | History or Documentary genre |
| genre\_Horror | INPUT | BINARY | Horror genre |
| genre\_MusicalPerfarts | INPUT | BINARY | Musical or Performing Arts genre |
| genre\_Romance | INPUT | BINARY | Romance genre |
| genre\_Sci\_fi | INPUT | BINARY | Science Fiction genre |
| genre\_Special\_Interest | INPUT | BINARY | Special Interest genre |
| genre\_ThrillMysSusp | INPUT | BINARY | Thriller, Mystery or Suspense genre |
| genre\_Western | INPUT | BINARY | Western genre |
| markets | INPUT | INTERVAL | Number of markets exposure |
| markets\_missing | INPUT | BINARY | Missingness indicator of 'markets' column |
| movie\_score | REJECTED | INTERVAL | The tomatometer rating in rottentomatoes.com |
| movie\_score\_positive | REJECTED | BINARY | A binary indicator that indicates whether the movie is good or not (in the perspective of the movie critics) |
| runtime | INPUT | INTERVAL | Movie length in minutes |
| ss\_count | INPUT | INTERVAL | Number of text reviews |
| ss\_mean | INPUT | INTERVAL | Mean sentiment scores |
| ss\_median | INPUT | INTERVAL | Median sentiment scores |
| ss\_p25 | INPUT | INTERVAL | 25th percentile of sentiment scores |
| ss\_p75 | INPUT | INTERVAL | 75th percentile of sentiment scores |
| ss\_std | INPUT | INTERVAL | Aggregate standard deviation of sentiment scores |
| title | ID | NOMINAL | Movie title (ID) |
| tomatometer\_count | INPUT | INTERVAL | Number of ratings given by movie critics rottentomatoes.com |
| total\_gross | INPUT | INTERVAL | Total gross amount ($) |
| user\_rating\_count | INPUT | INTERVAL | Number of ratings given by verified users in rottentomatoes.com |

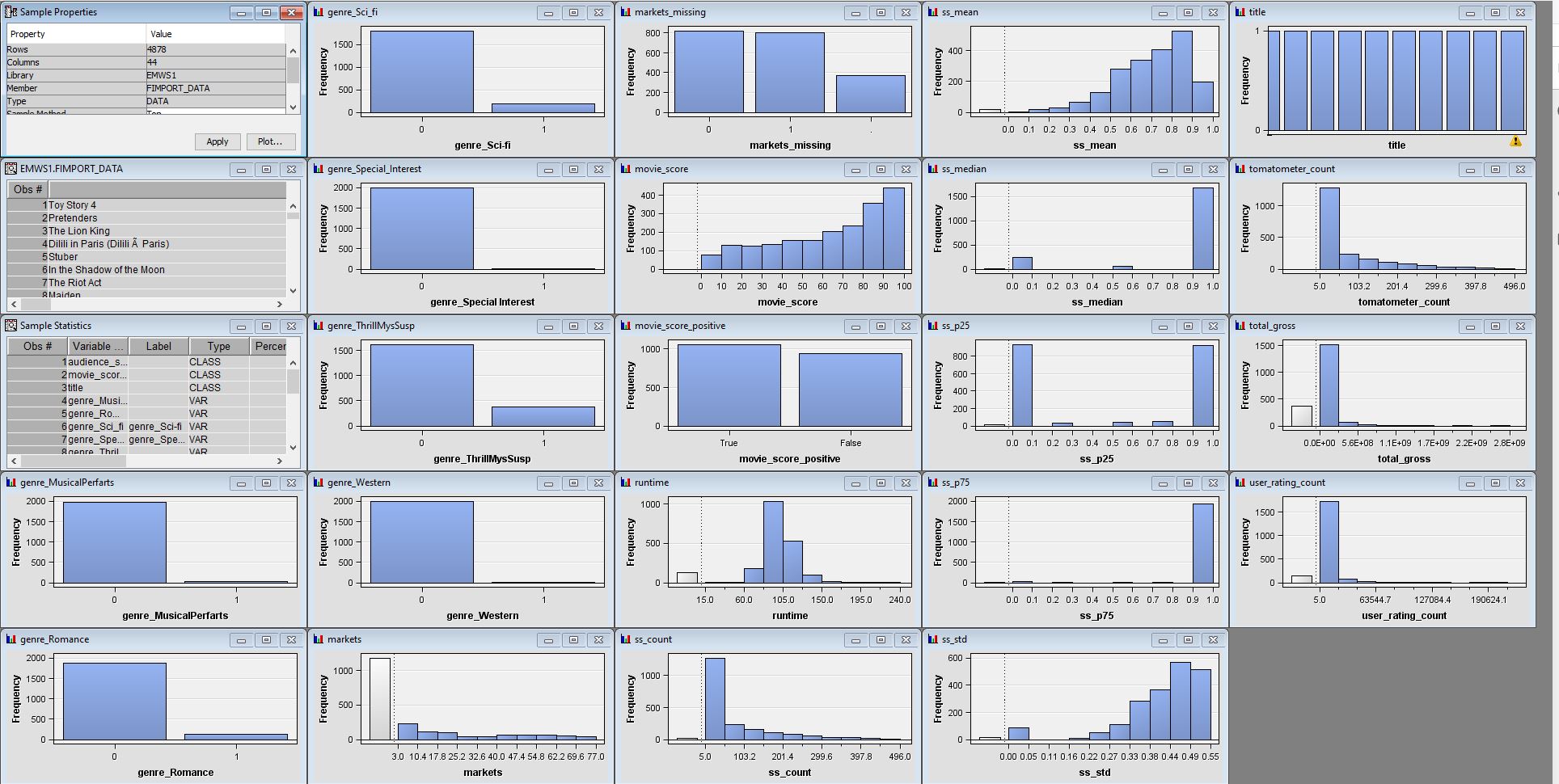
5. Diagram (SEM)



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

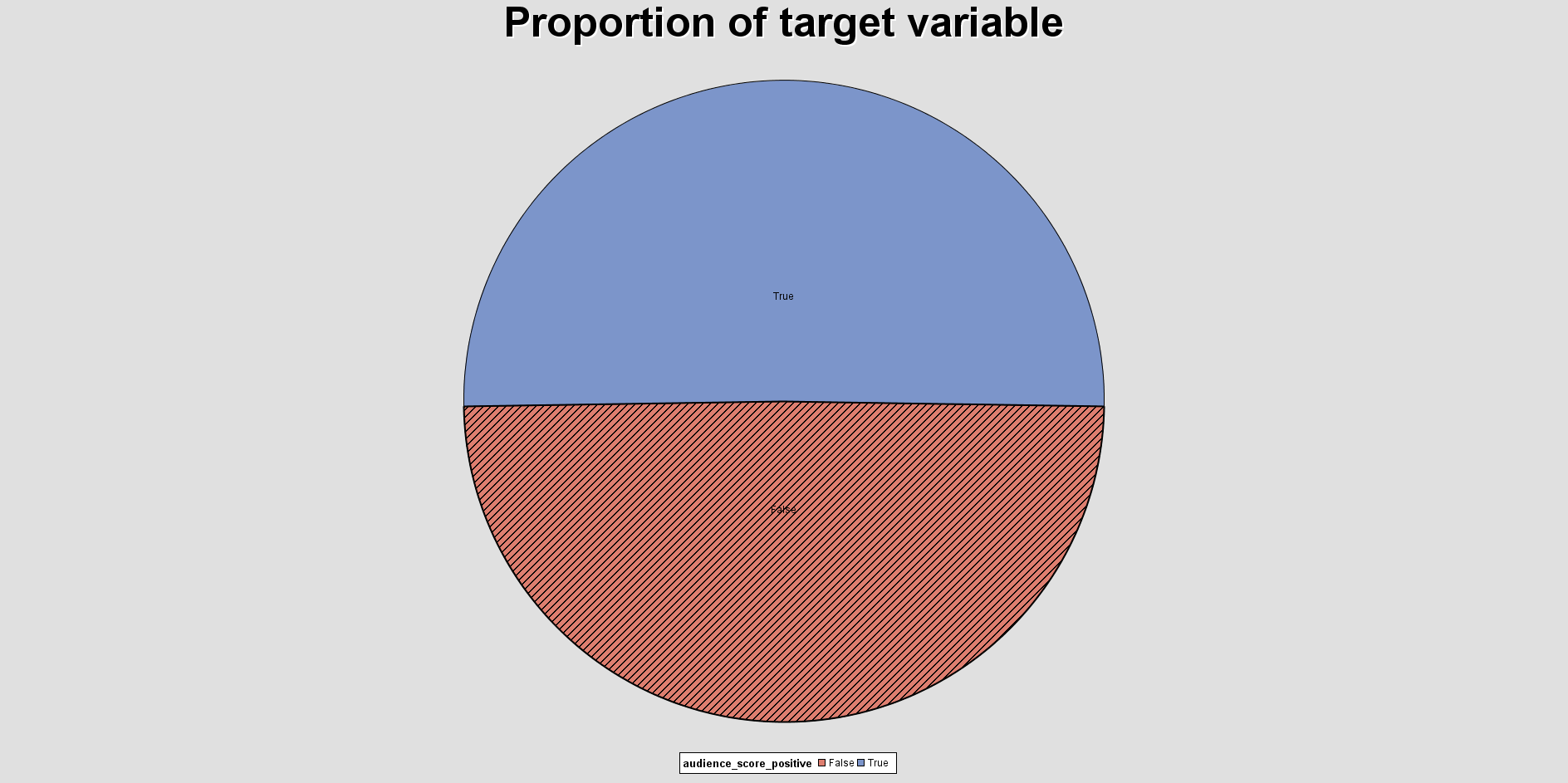


Missing data can be identified from the above histograms. For example, *domestic\_opening, foreign\_gross, audience\_score, domestic\_gross, total\_gross, runtime, user\_rating\_count, markets, ss\_mean, ss-median, ss\_p25, ss\_p75, ss\_count, ss\_std* have missing values.

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 feed into a model for training.