

**TDS3301**

**Data Mining**

Group Assignment

|  |  |
| --- | --- |
| **Name** | **Student ID** |
| Jaiprashanth A/L Ramalingam | 1131122948 |
| Farhanuddin bin Hamdan | 1132702175 |
| Kevin Lee Xin Zhe | 1132700900 |

Table of Contents

[**Part 1 : Exploratory Data Analysis** 3](#_Toc470894536)

[**About The Dataset** 3](#_Toc470894537)

[**Possible Insights** 4](#_Toc470894538)

[**C. Data Mining Approaches** 5](#_Toc470894539)

[**D. Data Quality Issues** 6](#_Toc470894540)

[**E. Preprocessing Task** 7](#_Toc470894541)

# **Part 1 : Exploratory Data Analysis**

## **About The Dataset**

In general, the people in our group are not really movie-goers. However, when we do want to watch a movie, we would prefer it to be good and interesting. We cannot even rely on famous movie rating review websites like IMDB and Rotten Tomatoes because the users may generate biased reviews. The dataset that we have obtained for this assignment is the IMDB 5000 Movie Dataset.

The IMDB 5000 Movie Dataset was created by chuansun76 from the Kaggle website. This dataset comprises of 5, 043 different movies which is scraped from the IMDB website using a Python library called “scrapy”. Using this library, he managed to obtain 28 unique variables from the 5, 043 movies which spanned across 100 years from 66 different countries. In simpler words, there are 28 columns which describes 5, 043 movies, e.g. color, director’s name, duration and genres.

There are a few insightful columns such as the columns which hold values for the amount of Facebook “Likes” obtained by either the movie, the cast, actor, or director. We hope to identify an interesting pattern revolving the social media and movie rating.

## **Possible Insights**

For each year, several thousands of movies are produced. After movies are released, many reviewers on IMDB will give their reviews, opinions and critics so that the future viewer can use that as a guide on whether the movie is worth the time to be watched or not. So, is there a better way for us to identify a movie rating without relying on critics? With this dataset, we seek to find possible patterns. Based on the chosen dataset, we can acquire various information related to past movies. Then we intend to relate the past movies’ information for predicting the rating of future releases. For example, the new movie features Uvuvwevwevwe Onyetenyevwe Ugwemuhwem Osas as the cast for the main character. If that actor has a high amount Facebook “Likes”, it might be possible to assume that the movie will turn out good and the gross revenue generated will be high as well. Of course, other factors can affect the result like the involvement between the actors and directors. For example, by having “Stan Lee” as the director and “Dwayne ‘The Rock’ Johnson” as the main cast will likely to be a good movie due to their massive experience in the filming industry. Thus, we can assume that the new movie also is going to be good.

Aside from a movie rating, we want to uncover a possible tie with the duration of the movie and its success in the box office after being released. This can help set an accurate benchmark to predict how long a movie should be to ensure that it will get a good review or rating. Some viewers might dismiss unacceptably short or long movies because the content was either was not expanded properly or too detail was given on insignificant aspects.

## **Data Mining Approaches**

In our opinion, we find *Association rule mining* as the most relevant data mining technique to be used on our selected dataset. The reason is because we would like to analyze and predict the movie watcher’s reaction with the variables given in the IMDB dataset. As a possible end-result, a model can formed and it may be able to predict movie ratings before a movie is released in theatres. For example, we can associate the IMDB rating with the sum of Facebook Likes of every lead actors involved of the movie, the budget allocated for each movie and the country of origin or etc. . . . Another use of the model may be beneficial to the movie production team as the model might be able to predict the gross income.

However, clustering can be considered as another relevant data mining technique that we can consider to uncover new patterns. For example, we can map out a possible relationship between the number of users of IMDB with the number of Facebook likes for either the movie, the cast, director or individual actors. This could contribute to identify the authenticity or legitimacy of IMDB movie rating.

## **Data Quality Issues**

Within the dataset, we can assume the data is “dirty” and issues exist in most of the columns. This results in inconsistent evaluations or calculations. These dirty data plays the part of making the dataset becoming untidy. For example, there are Not Available(NA) data exist in the column named “Gross”, this means that some of the movies gross will be unknown to us and hence missing one important factor of determining the movie’s popularity and related attributes.

There were multiple values in a single observation. For example, the column “Genres” holds multiple values in most rows. For example, while there are rows with value just like ‘Action’, there are also rows with multiple ones such as ‘Action|Adventure|Fantasy|Sci-Fi’. Arrangement like this will make the dataset to be more difficult to sort which producing inaccurate information.

Another issue that was present is inconsistency in the data existed in the Color column. There are two labels for the Color variable, “Color” and “Black and White” but there are another label present which is “color”. This causes inconsistency in data.

## **Preprocessing Task**

Please kindly refer to the link below for the Github repository of our R scripts which we made to preprocess the dataset.

<http://bit.ly/2iuJaI5>