**Part 1 : Exploratory Data Analysis**

Prepare a report on the chosen dataset by completing the following tasks:

A. Describe the dataset in your own words.

**In general, the people in our group are not really a movie person. But, when we do want to watch a movie, we want it to be good and interesting. How are going to do that, if the movie have not even been released in cinemas yet? We cannot even rely on famous movie rating reviewers like Rotten Tomatoes because they cannot review a movie that is not in cinemas yet. How we wish there is a way to predict whether this movie is good without relying on critics or our instinct. This is where IMDM 5000 Movie Dataset comes in.**

**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 uniques variables from the 5, 043 movies which spanned across 100 years from 66 different countries.**

B. What possible insights can be obtained from mining the chosen dataset? (**Marak** - half a page)

Give your own unique insight.

**Every year, thousands of movies are made**. **Once the movies are released, many reviewers 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. But, is there a better way for us to tell the greatness of movie without relying on critics or our own instincts (Assuming that the movie have not been released in cinemas yet)? With this dataset, we can solve the problem. By mining the chosen dataset, we can acquire various information related to past movies. Then we relate the past movies’ information with the current movie. For example, the new movie features Uvuvwevwevwe Onyetenyevwe Ugwemuhwem Osas as the cast for the main character. By mining the dataset and search for the word “Uvuvwevwevwe Onyetenyevwe Ugwemuhwem Osas”, we can find out about what movie does “Uvuvwevwevwe Onyetenyevwe Ugwemuhwem Osas” acted in. Then we find out that any movie he acted in, tends to be a billion dollars worth of profit by the end of the movie’s first month of release. Thus, we can assume that the new movie also is going to be good.**

C. What type of data mining technique (association rule mining, classification or clustering) would be relevant? Give an example, for example, if you think classification is suitable, describe what will be classified and what the possible classes are.

**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.**

D. Describe data quality issues, and be specific. Identify which attribute (column) has issues, or if the structure of the data has problems.

**Within the dataset, not available data exists 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 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.**

**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.**

E. Perform a pre-processing task on the dataset chosen. (**All of us**)

Data Cleaning

Jai - First 10 columns

Kevin - Second 9 columns

Marak - Third 9 columns

**READ THIS BEFORE CODING:**

We each focus on following the tutorial codes and apply the code on our dataset to achieve this preprocessing techniques:-

* **Data cleaning** → Tut 2 and 3
  + fill in missing values,
  + smooth noisy data,
  + identify or remove outliers,
  + resolve inconsistencies.