**Project Proposal**

**Problem description**

Since the beginning of the 20th century, the motion pictures industry invests increasingly large amounts of money into its cinematic productions. Movie studios have tried different approaches to maximize ROI on theatrical releases but pleasing the public with motion pictures continues to be a risky endeavor. One such approach is the use of private screenings and reviews. In these events, movies are exhibited before their final editing or release and the general feeling of a hand-picked audience towards their potential is "captured". Nonetheless, this seems to have little effect on box office results. Mega-productions, known as blockbusters, often fail to meet box office expectations, while low-budget movies sometimes perform beyond the wildest hopes and dreams. Addressing how the general sentiments of the public towards specific theatrical releases relate to box office results could bring substantial benefits to the movie industry.

Thus, this project aims to measure how reviews and ratings of movies released in theaters relate to their sales at the box office.

**Data Sources**

**Primary dataset: Rotten Tomatoes movies and critics review Dataset**

This dataset is taken from the website <https://www.rottentomatoes.com> as of 2020-10-31. It contains the ratings given by the regular users and by the critics who are certified members of film critic associations.

This dataset contains two parts:

1. **Movies Dataset:** This set contain details of movies with columns such as the URL used for the scraping, movie tile, description, genres, duration, director, actors, content rating, user’s ratings, and critic's ratings.
2. **Critics Dataset:** This set contains details of the critic reviews with columns such as the URL used for scraping, critic name, review publication, date, score, and content.

Link to the Rotten tomatoes’ dataset: [**https://www.kaggle.com/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset**](https://www.kaggle.com/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset)

**Contextual dataset: IMDB Dataset:**

This dataset has 5 files and are as follows:

1. **Title information:** This set contains information related to the title of the movies and has the columns - titleId, ordering, title, region, language, types, attributes, and isOrginalTitle
2. **Title basics:** This set contains data related to movie title and contains following columns – tconst, titleType, primaryTitle, originalTitle, isAdult, startYear, endYear, and runtimesMinutes
3. **Title principals:** This file contains data related to principal cast/crew for titles. Its attributes are tconst, ordering, nconst, category, job, and characters
4. **Title Ratings:** This set contains data on ratings and votes for the title and its columns are tconst, averageRating, and numVotes
5. **Names:** This file contains information for names and has the attributes- nconst, primaryName, birthyear, deathYear, primaryProfession, and knownForTitles

Linkto IMDB dataset**:** [**https://www.kaggle.com/ashirwadsangwan/imdb-dataset**](https://www.kaggle.com/ashirwadsangwan/imdb-dataset)

**Project Outline:**

1. **Problem definition and planning:**

As the problem statement has been identified, we have formed a team of five to work on the analysis.

1. **Data preparation:**

There were various tables present in the collected datasets, we will be integrating the tables, categorizing the attributes and transform them into a form which will be appropriate for analysis. We will be checking for any inconsistencies or errors in the datasets and make appropriate changes without losing the important features. Overall, our goal here is to make dataset amenable for proper analysis.

1. **Analysis:**

In this step, we will be visualizing the dataset to see the relationships between the attributes to present conclusions and summarizations of the data. Using data mining techniques, we will be providing the facts/rules from the data and provide exploratory analysis.

1. **Deployment:**

We are going to submit a detailed final report on the analysis performed on sales of movie in box office related to movie ratings/reviews.

This project’s success can be measured based on the quality of analysis, reaching the milestones on time without any delay and the satisfaction rate on the analysis report provided.

**Team Members:**

Anusha Dasari

Bhavi A. Shah

Mauro Jasse

Sai Ruchith Reddy Ginnavaram

Vinjal Doshi