**Group 7 Project 2 - Movies Database**

Members: Lisa Santonato, Minna Wu, Dian Wang, Suchitra Sharma

**Abstract**

The objective of this project is to compare gross movie earnings data with ratings and popularity rankings from various online resources. Our process involves extracting the Movie earnings and rating data from our sources, transform it into a format that is comparative and then load it to the MySQL workbench or MangoDB. The project aims to answer several possible questions as outlined below and will include a summative and comparative analysis.

Possible Research Questions:

* What are the top grossing films per week?
* What are the top ranked films?
* Do the film rankings differ between websites?
* Are the film ratings for each website normally distributed? (Or skewed/bimodal)
* What does the comparison between critic scores and user scores look like?

Additional Questions May Include:

* What is the distribution of popularity between shows in differing markets?
* What are the most popular TV shows by genre?
* What time slots are most popular, how do time slots affect popularity?
* Look at popularity of shows comparing male or female characters.

**Breakdown of Roles & Tasks:**

Extract:

Lisa: Scrape the-numbers.com for top 100 grossing movies as the base data

WangDian: Scrape Rotten Tomatoes for user vs. critic ratings

Sue: Download IMDB CSV with ratings and use movie IDs to find titles in IMDB API

Minna: Scrape Metacritic for user vs. critic ratings

Transform:

All members will be using Pandas to clean the data

TBA: Merge the data into one table using

Load:

TBA: Load into MySQL

1. **Introduction**

There are many websites online that offer movie reviews. However, how each website is unique in its review content is not known. This project is a small effort to compare the review ranking from different websites and conclude which review is linearly correlated with the gross earnings. The project is divided into three main components: Extract, Transform and load. For this, data is extracted from four different websites, transformed into the required format and then load it to the MongoDB. The analysis is done on Pandas IDE and plotting is done using Matplotlib library.

The questions that this project proposes to answer are:

1. What are the top grossing films per week?
2. What are the top ranked films?
3. Do the film rankings differ between websites?
4. Are the film ratings for each website normally distributed? (Or skewed/bimodal)
5. What does the comparison between critic scores and user scores look like?
6. **Resources**

The websites that are used in this project are:

1. <https://www.the-numbers.com/>
2. <https://www.imdb.com/>
3. <https://www.rottentomatoes.com/>
4. <https://www.metacritic.com/>

Technologies Used are:

1. Pandas,Jupyter
2. Matplotlib
3. MongoDB
4. Beautiful Soup
5. **Methods and Discussion**
   1. The data from Numbers website is extracted with web scraping using BeautifulSoup. The box office chart table is directly extracted from the website and used for analysis. For data cleaning, column Headers are added manually replacing the column numbers. Subsequently ‘$’ sign is removed from the Gross and Total Gross columns to enable analysis. The data is then grouped by “Distributer column and gross earning per distributer data is extracted.