University of Toronto
CSC343, Fall 2020

Project - Phase 1:

Changyan Xu (1004802181), Jiaming Yang(1006458575)

Domain

Entertainment, Ratings and Reviews on Movies

Dataset

- The link to the dataset we have identified:
  - https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+ratings.csv
- What information in this dataset is relevant to the project:
  - All four `.csv` files are relevant to the proposed project, but only selected columns will be deployed.
  - For `IMDb movies.csv`, we will use the columns as follows:
    - ∗ imdb_title_id, title, date_published, genre, duration, country, language, director, writer, production_company, actor, description, avg_vote, budget, usa_gross_income, worldwide_gross_income
  - For `IMDb names.csv`, we will use the columns as follows:
    - ∗ imdb_name_id, name, birth_name, height, date_of_birth, place_of_birth
  - For `IMDb ratings.csv`, we will use the columns as follows:
    - ∗ imdb_title_id, genders_0age_avg_vote, genders_18age_avg_vote, genders_30age_avg_vote, genders_45age_avg_vote
  - For `IMDb title_principals.csv`, we will use the columns as follows:
    - ∗ imdb_title_id, imdb_name_id, category, job, characters
- Any learning we will have to do in order to interpret the data:
  - The tables that that we chose are quite large with many attributes. This means that we need to learn about what kind of information can each attributes provide us and how are the attributes related to each other.
- Any cleaning up we think we will have to do in order to use the data:
  - learn how to import `.csv` files with `SQL`
  - break one table into two or more tables, and store the tables
  - We found some columns with mostly empty values and thus we should split those columns into separate table(s).

Questions

Three investigative questions we plan to answer using this dataset:

1. Who are the top 10 actors/actresses who casts the most among the movies?
2. What is the relation between the movie genre and movie gross revenue?

3. What are the top 30 highest rating movies?
4. How does people's age affect the type of movie they like? (i.e. What is the most preferable type of movie for each of the age stages?)

Schema

- $IMDb\_movies(\underline{imdb\_title\_id}, title, date\_published, genre, duration, country, language, director, writer, production\_company, actor, description, avg\_vote, budget, usa\_gross\_income, worldwide\_gross\_income)$
- $IMDb\_names(\underline{imdb\_name\_id}, name, birth\_name, height, date\_of\_birth, place\_of\_birth)$
- $IMDb\_ratings(\underline{imdb\_title\_id}, genders\_0age\_avg\_vote, genders\_18age\_avg\_vote, genders\_30age\_avg\_vote, genders\_45age\_avg\_vote)$
- $IMDb\_title\_principals(\underline{imdb\_title\_id}, \underline{imdb\_name\_id}, category, job, characters)$

can be turned to:

- $Movies\_info(\underline{imdb\_title\_id}, title, date\_published, genre, duration, country, language, description)$
- $Movies\_workers(\underline{imdb\_title\_id}, director, writer, production\_company, actor)$
- $Movies\_vote(\underline{imdb\_title\_id}, avg\_vote)$
- $Income(\underline{imdb\_title\_id}, usa\_gross\_income, worldwide\_gross\_income)$
- $Budget(\underline{imdb\_title\_id}, budget)$ – Splitting for the reason of missing values appearance
- $Actor\_info(\underline{imdb\_name\_id}, name, birth\_name, height, date\_of\_birth, place\_of\_birth)$
- $IMDb\_age\_ratings(\underline{imdb\_title\_id}, 0age\_avg\_vote, 18age\_avg\_vote, 30age\_avg\_vote, 45age\_avg\_vote)$
- $IMDb\_title\_principals(\underline{imdb\_title\_id}, \underline{imdb\_name\_id}, category, job, characters)$