Midterm Project Presentation

Movies Entertainment

By: Dylan Lam & Eric Nguyen

Course: C S 329E

Time: Friday 2:00-5:00PM

Table of Contents

Overview

Datasets

Content

Scope

Challenges

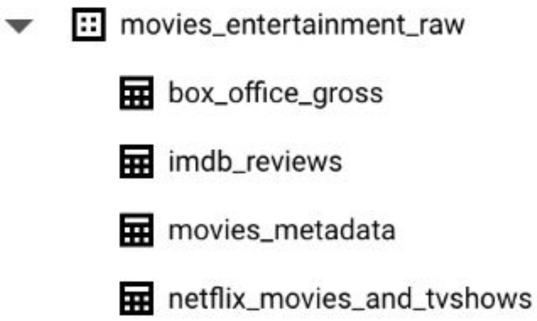
Approach

Overview

Our project aims to build a data warehouse and analytical data marts to analyze key trends in the movie industry, including company success, genre popularity, and more. By integrating data from box office earnings, Netflix metadata, and company production details, we create optimized marts that provide insights into which companies generate the most revenue, how budgets affect financial success, and which genres dominate different audience ratings (PG, PG-13, R, etc.). Through data transformation and pre-aggregation, we enable efficient business intelligence reporting, allowing stakeholders to identify high-performing genres, investment-worthy movie types, and audience preferences. These insights help movie studios, investors, and analysts make data-driven decisions on budgeting, production, and marketing strategies. Our approach ensures scalable, accurate reporting for long-term industry analysis.

Datasets

1. **movies_metadata (Kaggle)**: contains detailed metadata on movies, such as titles, genres, release dates, production companies, and ratings.
2. **box_office_gross (Kaggle)**: data on movie earnings, including domestic and international box office revenues.
3. **netflix_movies_and_tvshows (Kaggle)**: provides information about content available on Netflix, including titles, genres, and release years.
4. **imdb_reviews (Stanford AI)**: a set of 50,000 There is additional unlabeled data for use as well. Raw text and already processed bag of words formats are provided.

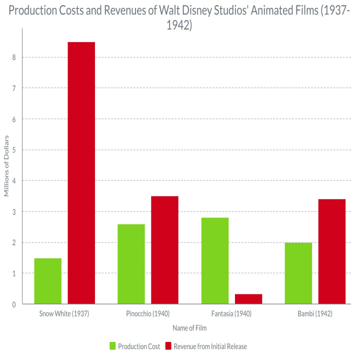


- movies_entertainment_raw
  - box_office_gross
  - imdb_reviews
  - movies_metadata
  - netflix_movies_and_tvshows

Content

Our project focuses on building a data warehouse with multiple data marts to analyze trends in the movie industry. The datasets we are working with include:

1. **Movies Metadata** – General information on movies, including **title, budget, language, and genre**
2. **Box Office Gross Data** – Revenue figures for theatrical releases
3. **Netflix Movies & TV Shows Data** – Streaming content metadata, including **audience ratings (PG, PG-13, R, etc.)**
4. **Company Information** – Data on **production companies, total releases, and financial performance**



THEATRE TICKETS HALF PRICE & DISCOUNT
LEICESTER SQUARE BOX OFFICE
OFFICIAL
HALF PRICE THEATRE TICKET BOOTH
OFFICIAL HALF PRICE & DISCOUNT BOOTH
AND FULL PRICE ON SELECTED SHOWS SUBJECT TO AVAILABILITY
No.1 LEICESTER SQUARE
NEVER BEFORE HAS A SHOW GIVEN MORE!
Oliver!
TICKETS AVAILABLE TODAY!
THE LION KING
BUREAU de CHANGE NO COMM
PRISCILLA QUEEN OF THE DESERT the musical
PALACE THEATRE
MIND THE STEP
PRICE LIST
DAVID ESSEX
FUN FAIR

Scope

In-scope (What We Analyze):

- **Company-Level Performance** → Which companies generate the most revenue and release the most movies.
- **Budget vs. Revenue Trends** → How movie budgets correlate with box office success.
- **Genre & Audience Rating Impact** → Understanding **which genres** are most produced and how **audience ratings (PG, PG-13, R)** affect budget and revenue.
- **Historical Movie Trends** → How movie releases and financial performance have evolved over time.

Out-of-Scope (What We Do Not Analyze):

- **Streaming Platform Performance** → We do not differentiate between Netflix, Hulu, Disney+, etc. due to a lack of platform-specific data.
- **Viewer Engagement & Ratings** → We do not analyze **user reviews or star ratings**, only **audience classification ratings (PG, PG-13, R, etc.)**.
- **Marketing & Advertising Impact** → We do not track promotional budgets or campaign effectiveness.

Challenges

1. Data Integration and Data Hygiene Issues
2. Time Constraints & Scope Management
3. Gaps in the data (budget, revenue, country, director, cast, etc.)
4. Relating the data across the entire schema

```
('require_name_email') && !$user
strlen($comment_author_email) ||
__('<strong>ERROR</strong>: please
is_email($comment_author_email))
__('<strong>ERROR</strong>: please
content )
```

Approach

1. Data Collection & Integration
2. Data Cleaning and Transformation
3. Data Mart Design & Optimization
4. Business Question Refinement
5. Business Intelligence and Reporting

DEMO

Thank you.