Table of Contents

Technical Handoff

Database Topic: Streaming platform content and IMDb ratings

Database Purpose: (1) Analysis of quality of streaming platform movie content (2) Strategic analysis of competitive market positioning

Operating System/Software: Microsoft Windows 10 Pro, Microsoft SQL Server v17.7

Database Sensitivity: None

Database Permission: Unrestricted/Public Access

Data Processing Software: Jupyter Notebook, Python

Data Processing: Data was obtained through a public source. It contains errors, missing values, and inconsistent formatting. The data was cleaned in Python and datasets were combined, where possible, using combination keys to relate the two primary keys. Approximately 32% of the streaming platform data has a corresponding record in the IMDb dataset. (Review comments in attached IPYNB file)

Overview of Database Tables: Database contains twelve tables. Six of the tables detail information about movies on the top four streaming platforms; Netflix, Hulu, Prime Video, and Disney+ (n=16,744). Five of the tables provide details of rating statistics for movies on the popular movie rating site, IMDb (n=85,855). These subsets of the database are connected through the 'IDs' table (n=5,5421). Review the Database Metadata section for more information.

Database Maintenance: Database should be updated yearly. Datasets should be checked for updates. Where applicable, updated data should be added by downloading updated CSV files, running through python code, and uploading into the database.

Future Considerations: The database would be more useful if there were a higher percentage of streaming platform records with a corresponding record in the IMDb dataset. Time and resources currently prevent further progress on this issue. As resources become available, efforts should be made to correlate more streaming platform records to the IMDb database. This will require checking the accuracy of the 86,000 records in the IMDb dataset, as well as additional data processing and standardization.

Executive Summary

The video streaming market is a hugely competitive and lucrative industry. According to Statista (2020), the video streaming market is expected to generate over $71 million in revenue in 2021. This market is expected to grow between 2021 and 2025, with an estimated CAGR of 11.04% (Statista, 2020). The database addressed in this document contains data pertaining to the quality of movie content on top-selling streaming platforms; Netflix, Hulu, Prime Video, and Disney+. In the ever more competitive video streaming market, this database will benefit stakeholders.

Historically, media consumption has been controlled by a strong oligopoly. Often, these companies both own content creation as well as the platforms on which it is consumed. This means they benefit substantially from selling content, popular and not, in a bundled format. So, even though consumers have long called for the option to selectively choose channels or shows, the oligopoly has steadfastly resisted this shift. When Netflix popularized monthly subscription streaming, it disrupted the media market. Now, major players are fighting for a customer base that wants access to the best content at the lowest cost and that has the option to flit between services on a monthly basis. This database relates content library information from top streaming providers with IMDb ratings and demographic information. This will enable an analysis of the quality of content across platforms and inform strategic analysis to identify growth opportunities.

This database will be beneficial to stakeholders that wish to better understand the current state of the market and growth opportunities. Stakeholders include existing competitors (i.e., Netflix, Hulu, Disney+, PrimeVideo), new platforms hoping to break into the market, and customers that want to better understand their streaming options. Existing competitors and new market entrants can utilize the information in the database to understand where different demographic markets are strongly engaged and where there is space to capture demographic markets that are not having their preferences met. Statista (2020) projects that the user base will grow from 14.3% in 2021 to 18.2% in 2025. This database is crucial in determining which platforms will capture that new audience.

This database was constructed using publicly available datasets and should remain available to the public moving forward. Datasets were collected from Kaggle, a public data repository. These datasets were transformed and combined in Python to enable a cross analysis of their content. Moving forward updates to these datasets should be downloaded, run through the python code, and added to the SQL database. Finally, the database should be uploaded to Kaggle to contribute to the publicly available data surrounding this topic.

Data Sources

The Movie Dataset (Banik, 2017) is a public data set on the popular data website, Kaggle. It was updated by the author three years ago and assembled using data from TMDb and GroupLens. This data includes 45,000 movie listings and 26 million ratings from over 270,000 users. The data was downloaded on March 1, 2021, as a series of CSV files.

Movies on Netflix, Prime Video, Hulu and Disney+ (Bhatia, 2020) is a public data set on the popular data website, Kaggle. It was updated a year ago by the author and was scraped using Beautiful Soup and the IMDbPy module. This data contains information for which platforms host the almost 17,000 movies in the dataset. The data was downloaded on March 1, 2021, as a series of CSV files.

IMDb Movies Extensive Dataset (Leone, 2020) is a public data set on the popular data website, Kaggle. IT was updated by the author seven months ago and assembled through scraping the IMDb website. This dataset contains IMDb movie and rating information for almost 86,000 movies. The data was downloaded on March 1, 2021, as a series of CSV files.

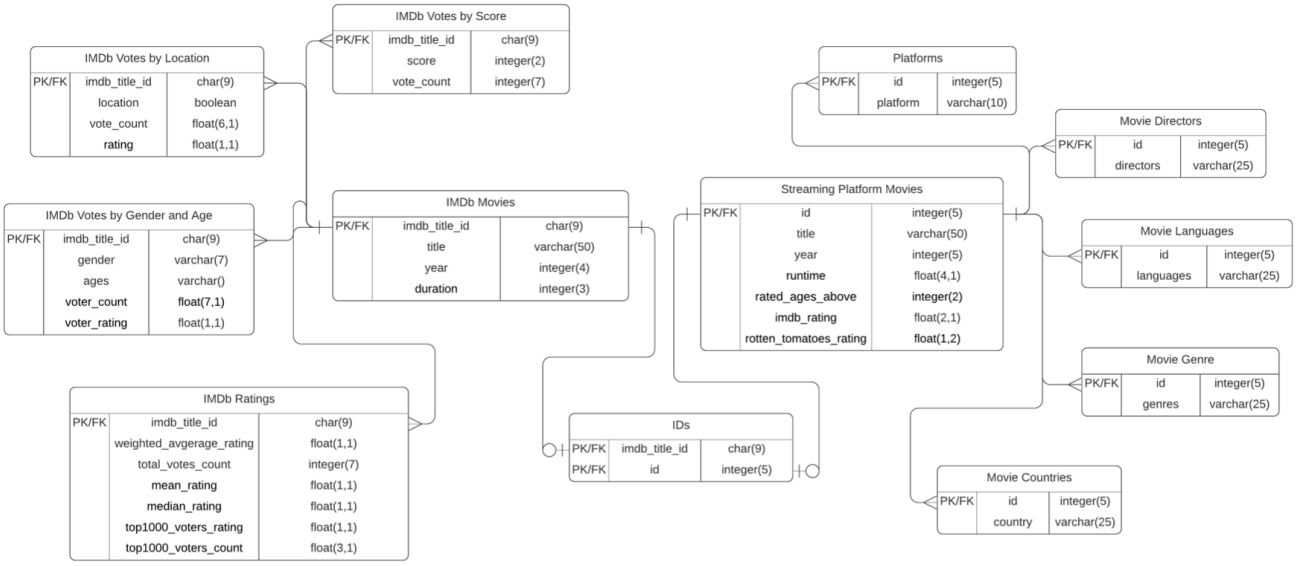
ERD Introduction

The ERD for the streaming platform database contains twelve tables derived from the three different datasets. Six of the tables detail information about movies on the top four streaming platforms; Netflix, Hulu, Prime Video, and Disney+ (n=16,744). Five of the tables provide details of rating statistics for movies on the popular movie rating site, IMDb (n=85,855). These subsets of the database are connected through the 'IDs' table (n=5,421). This database allows users to explore movies on popular streaming platforms and also movies on the IMDb website. It also enables analysis of movies that exist in both datasets.

Database Tables

1. IMDb Movies
2. IMDb Ratings
3. IMDb Votes by Gender and Age
4. IMDb Votes by Location
5. IMDb Votes by Score
6. IDs
7. Streaming Platform Movies
8. Platforms
9. Movie Genres
10. Movie Directors
11. Movie Languages
12. Movie Countries

ERD Diagram



ERD Technical Details

Each table is connected to one or more other tables through lines with varying endpoints. These line endpoints denote the type of relationships between the tables. For more information about ERD table connections, look at the LucidChart website page for ERD Symbols and Notation (Lucid Software Inc, n.d.). Additionally, the far right hand column in every table details information about the data type, format, and length for that column. For more information about SQL Server data types, check out the blog article 'SQL Server Data Types you Must Know' (Wordpress, 2018). Finally, the far left column denotes the primary and foreign keys in the tables.

Database Metadata

Table Name: IMDb Movies
Number of Records: 85,855
Memory: 2.6 MB

Column	Non-null count	Dtype	Key Type	Notes
imdb_title_id	85,855	varchar(9)	PK / FK	
title	85,855	varchar(30)		title of movie
year	85,855	integer		year of movie release
duration	85,855	integer		length of movie

Table Name: IMDb Ratings
Number of Records: 85,855
Memory: 4.6 MB

Column	Non-null count	Dtype	Key Type	Notes
imdb_title_id	85,855	varchar(9)	PK / FK	
weighted_avg	85,855	real		
total_votes	85,855	integer		
mean_votes	85,855	real		
median_votes	85,855	real		
top1000_voters_rating	85,176	real		average rating from "top 1000" IMDb raters. "Top 1000" is a title given to IMDb users who have voted on the highest number of titles.
top1000_voters_rating	85,176	real		count of votes from "top 1000" IMDb raters. "Top 1000" is a title given to IMDb users who have voted on the highest number of titles.

Table Name: IMDb Votes by Age and Gender
Number of Records: 1,201,970
Memory: 55.0 MB

Column	Non-null count	Dtype	Key Type	Notes
imdb_title_id	1,201,970	varchar(9)	PK / FK	
gender	1,014,573	varchar(10)		
ages	1,201,970	varchar(8)		
voter_count	1,201,970	real		
rating	1,014,573	real		

Table Name: IMDb Votes by Location
Number of Records: 171,710
Memory: 6.6 MB

Column	Non-null count	Dtype	Key Type	Notes
imdb_title_id	171,710	varchar(9)	PK / FK	
location	171,710	varchar(2)		this is a binary field. The value can be either 'us' or 'non us', which denotes the location of voters as us residents or non us residents
vote_count	171,500	real		
rating	171,500	real		

Table Name: IMDb Votes by Score
Number of Records: 858,550
Memory: 19.7 MB

Column	Non-null count	Dtype	Key Type	Notes
imdb_title_id	858,550	varchar(9)	PK / FK	
vote_count	858,550	integer		
score	858,550	integer		rating value. IMDb has a rating system of 1-10.

Table Name: IDs
Number of Records: 5,421
Memory: 127.1 KB

Column	Non-null count	Dtype	Key Type	Notes
imdb_title_id	5,421	varchar(9)	PK / FK	primary key for the IMBd dataset tables
id	5,421	integer	PK / FK	primary key for the streaming platform dataset

Table Name: Movies
Number of Records: 16,744
Memory: 915.8 KB

Column	Non-null count	Dtype	Key Type	Notes
id	16,744	integer	PK / FK	
title	16,744	varchar(30)		title of movie
year	16,744	integer		year of movie release
runtime	16,152	real		length of movie
rated_ages_above	7,354	varchar(15)		integer representing the recommended view age or older
imdb_rating	16,173	real		
rotten tomatoes_rating	5,158	real		

Table Name: Platforms
Number of Records: 17,381
Memory: 407.4 KB

Column	Non-null count	Dtype	Key Type	Notes
id	5,421	integer	PK / FK	
platform	5,421	integer		streaming platform. Denotes the streaming platforms that host the corresponding movie id.

Table Name: Genres
Number of Records: 39,373
Memory: 407.4 KB

Column	Non-null count	Dtype	Key Type	Notes
id	39,373	integer	PK / FK	
genres	39,098	varchar(11)		genres the movie is ascribed to

Table Name: Directors
Number of Records: 18,924
Memory: 443.5 KB

Column	Non-null count	Dtype	Key Type	Notes
id	18,924	integer	PK / FK	
directors	18,198	varchar(32)		genres the movie is ascribed to

Table Name: Languages
Number of Records: 20,956
Memory: 491.2 KB

Column	Non-null count	Dtype	Key Type	Notes
id	20,956	integer	PK / FK	
language	20,357	varchar(100)		languages the movie can be watched with in the streaming platforms

Table Name: Countries
Number of Records: 21,134
Memory: 495.3 KB

Column	Non-null count	Dtype	Key Type	Notes
id	21,134	integer	PK / FK	
country	20,699	varchar(32)		countries in which the streaming platform offers the movie

References

Banik, R. (2017, November 10). *The Movies Dataset*. Kaggle. https://www.kaggle.com/rounakbanik/the-movies-dataset.

Bhatia, R. (2020, May 22). *Movies on Netflix, Prime Video, Hulu and Disney+*. Kaggle. https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney.

Leone, S. (2020, September 14). *IMDb movies extensive dataset*. Kaggle. https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset.

Lucid Software Inc. (n.d.). *Entity-Relationship Diagram Symbols and Notation*. Lucidchart. https://www.lucidchart.com/pages/ER-diagram-symbols-and-meaning#section_1.

Video Streaming (SVoD) - Worldwide: Statista Market Forecast. Statista. (2020). https://www.statista.com/outlook/dmo/digital-media/video-on-demand/video-streaming-svod/worldwide.

WordPress. (2018, May 28). *SQL Server Data Types you Must Know*. Business Intelligence A - Z. https://data-savvy.com/tag/erd/.