CSC 501 - Algorithms and Data Models

Assignment 1: Relational Data Model

University of Victoria
Faculty of Engineering
Department of Computer Science
Victoria, British Columbia

Instructor: Dr. Sean Chester

Aren Beagley - V00851240
Finn Archinuk - V00878751
Peter Li - V01009728
Chenuri Gunaratne - V00938807

Date Submitted: October 30, 2022

Contents

List of Figures

1 Introduction

The data contained within the six comma-separated-values (CSV) files available from MovieLens which covers movies, a random sampling of users, user ratings/tags, movies genres, machine learning tag relevance scores, and links to additional databases such as the Internet Movie Database (IMDB) and The Movie Database (TMDB). Given the range of information available, it was decided to investigate the perception and ratings of movie genres that have changed over time. This focus was chosen as what a genre is can be quite nebulous and interacts with the dominant narratives of a society leading to considerable variety with many possible causes.

2 Modelling

While it would be possible to investigate the rating/perception of movie genres in a naive way by analyzing movie ratings for each genre by year this would provide little insight as it is a purely descriptive analysis that lumps all possible factors affecting the perception of genre together and could possibly fail to represent the actual change in perception over time as ratings may be applied to movies many years after they are released. To provide greater insight into why the perception and ratings of each genre may change over time the following factors were identified and selected to be included in the modeling for analysis.

Movie Director: Popular directors tend to have large followings that will watch any movie they direct regardless of the genre as the appeal comes from the signature style of the director. As such, directors who make movies in a variety of genres may contribute to positively rated outliers within a genre.

External Economic Factors: Access to visual entertainment has varied throughout history as has the frequency of production and movie releases. People's lives and health are also greatly impacted by economic booms and recessions, meaning that the perception of each movie genre may be affected by the economy they live in. As such, inflation data was determined as a measure of economic trends. Inflation data was readily available for the United States of America, which led to the decision to limit our analysis to movies released in North America and in English. It is worth noting that the sociopolitical state people live in also dramatically impacts them and would be worth modeling but we were unable to find a reasonable measure of this.

Classic and Cult Status: Some movies achieve status as a "Classic" and can become considered required viewing and usually enjoy great acclaim. Others are released to relatively little acclaim but slowly amass a passionately dedicated fanbase that supports the franchise, thereby becoming a "Cult Classic". In either of these cases if a genre contains a number of classic or cult films the ratings of these movies could skew the perception of the entire genre when averaging together ratings.

Overton Window: Over time the discourse within a society shifts and what is considered sensible and acceptable versus radial and outrageous changes. This means that the perception of movies and various genres may also shift, leading to rises and falls in popularity as the themes within a genre or movie shift from "ahead of the times" to "uniquely relevant" or even "a relic of the past". The possibility of a genre or movie's perception changing over time as society itself changes should not be ignored as it means averaging many years worth of ratings together may not adequately capture the actual trends.

Accounting for these factors and tracking user ratings assigned to movies of various genres over time gave rise to the following requirements:

- Movies must be uniquely identified such that the directors, genres, region and language of each release, year of release, and user ratings can be determined from the identifier of the movie.
- Inflation data should be available for each year that the movie was released.
- The time that user ratings were applied in should be tracked.

2.1 Conceptual Model

Prior to creating the logical schema for the relational model that would be used to perform the investigation, the conceptual model shown in fig. 1 as an entity relationship diagram (ERD) was developed to create a schema that satisfied the analysis requirements and allowed enumeration of the attributes required to perform the analysis, so that any need for additional data sources other than MovieLens could be identified. While not officially present in the MovieLens data, the release year of a movie was included as part of the movie title and could be acquired through string parsing. In contrast, it was determined that data regarding release region and language, yearly inflation, and movie directors would need to be sourced from additional sources. Data regarding movie directors and releases were sourced from the Internet Movie Database (IMBD) as the MovieLens data contained a table providing a mapping from MovieLens movie ids to IMDB movie ids. Inflation data was provided by the US Department of Labour and aggregated into yearly averages (source: https://www.in2013dollars.com/). The data available from each of these sources informed the choice of attributes for each entity. For example, although merging ML_ID and IMDB_ID would result in a smaller set of attributes it would negatively impact the traceability of where data came from and therefore reduce the scalability of any implementation informed by the conceptual model. The choice to keep releases as a weak entity set was similarly informed as that is how IMDB modeled the data and preserving that structure facilitates the inclusion of future data. Based on an inspection of the MovieLens data an additional constraint regarding the 'Dates' relation was determined; while a user may rate multiple movies and may rate a movie multiple times, only the newest rating for any given movie by a user is kept. Consequently, while the conceptual model suggests that the movie identifier, user identifier, and timestamp identifier would be required to determine the rating, for the actual data available only the movie identifier and user identifier are required and together they determine both the rating and timestamp.

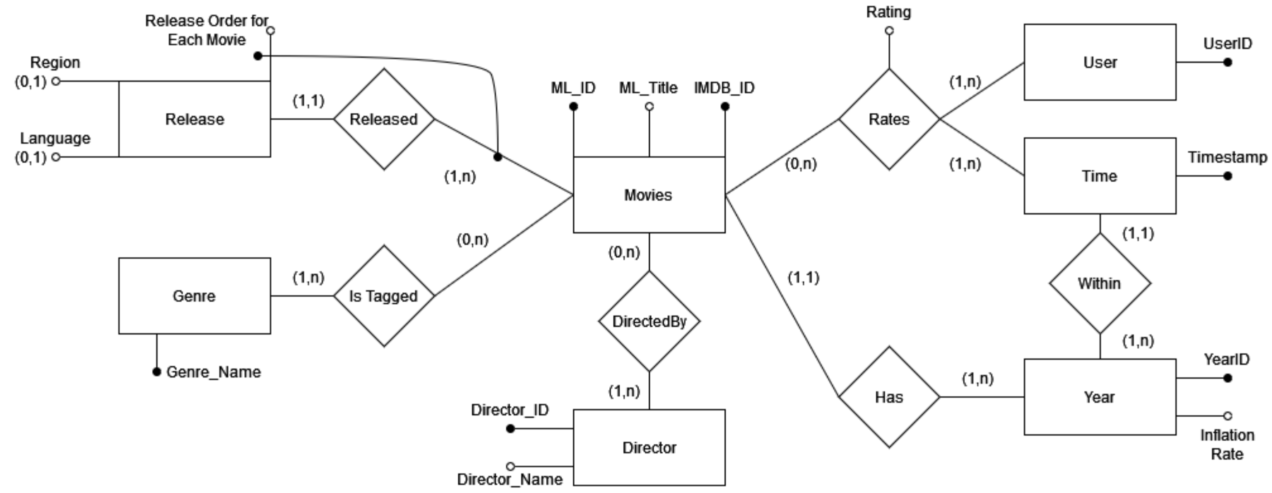

Figure 1: Conceptual schema used to inform implementation and analysis of movie genre perception over time using a relational model, presented as an entity relationship diagram. All attributes without an explicit representation of cardinality are assumed to have a cardinality of (1,1).

2.2 Logical Schema

From the conceptual schema, a logical schema was then developed to ensure that implementation of the relational model used for analysis would be suitable, as the conceptual model already satisfied our analysis requirements, and was theoretically sound. To begin, all of the attributes from the ERD in fig. 1 were enumerated and grouped into a relation. The following functional dependencies were determined from the conceptual schema:

- ML_ID → ML_ID, IMDB_ID, ML_Title
- IMDB_ID → ML_ID, IMDB_ID, ML_Title
- Director_ID → Director_ID, Director_Name
- Timestamp → Timestamp, YearID
- YearID → YearID, Inflation_Rate
- UserID, ML_ID → Rating, Timestamp

The following set of constraints was also determined from the cardinality of the relationships in fig. 1 along with inspection of the data. The constraints can be expressed in natural language:

1. Each user may only have one rating per movie.
2. Every movie must have both a unique MovieLens and IMDB identifier.
3. Each genre must have at least one movie within it, but movies may have no genres.
4. Genre names must be from the set {'Action', 'Adventure', 'Animation', 'Children', 'Comedy', 'Crime', 'Documentary', 'Drama', 'Fantasy', 'Film-Noir', 'Horror', 'IMAX', 'Musical', 'Mystery', 'Romance', 'Sci-Fi', 'Thriller', 'War', 'Western'}.
5. Ratings must be from the set {0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0}.
6. All directors must direct at least one movie.
7. All movies must have at least one release.

Given the set of attributes along with their associated functional dependencies and constraints, the Boyce–Codd Normal Form (BCNF) decomposition algorithm was applied to create a set of relations that would be theoretically sound, resulting in the logical schema shown in fig. 2 where the only functional dependencies for each relation are that the primary key determines every other attribute within the relation. Constraints 1 through 4 are enforced by the relations themselves, the ENUM type for genre names, and the use of foreign keys, while the other constraints must be enforced using additional checks or assertions. Constraint 5 is enforceable as a CHECK constraint in MySQL or using relational algebra. Constraints 6 and 7 require assertions, which are not supported in MySQL but can be expressed in relational algebra by eqs. (1) to (3) and enforced at a higher level of implementation than MySQL, such as during data cleaning in Python.

$$\textbf{Constraint 5: } \pi_{Rating}(Ratings) - \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0\} = \emptyset \tag{1}$$

$$\textbf{Constraint 6: } \pi_{IMDB\_ID}(Movies) - \pi_{IMDB\_ID}(Releases) = \emptyset \tag{2}$$

$$\textbf{Constraint 7: } \pi_{Director\_ID}(Directors) - \pi_{Director\_ID}(DirectedBy) = \emptyset \tag{3}$$

2.3 MySQL Implementation

After acquiring all of the required data from MovieLens, IMDB, and the US Department of Labour in the form of CSV or tab-separated value (TSV) files the data was cleaned using Python and reformatted into a single CSV file for each relation in the logical schema from fig. 2. During the cleaning process, string parsing was used to determine the year of each movie along with the conversion from timestamps to dates to allow the mapping of timestamps to years. Constraints 6 and 7 were enforced during data cleaning as they could be implemented through MySQL. The rest of the constraints were also enforced during data cleaning to increase the ease of importing data into the MySQL database. Relational tables were created through Python scripting with the MySQL-connector package and data was entered row by row from each

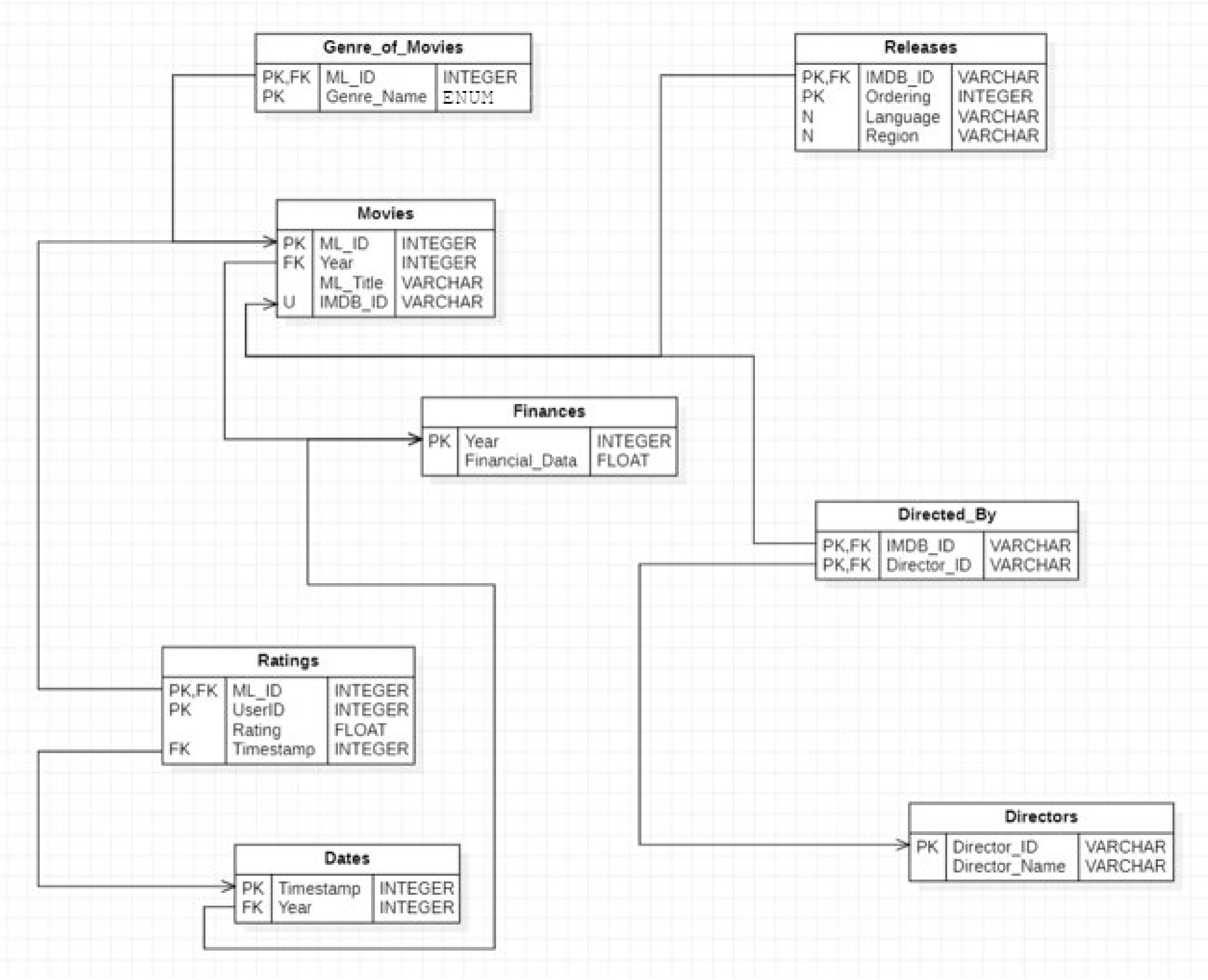

Figure 2: Logical schema derived from the conceptual model of movies through Boyce–Codd Normal Form decomposition. On the left-hand side column of each relation PK denotes the attributes that compose the primary key for a relation, U denotes that all values for that attribute must be unique (ie a secondary key), N denotes that the default value is None and the attribute is optional, and FK denotes a foreign key and the arrows trace back to the source of the foreign key. On the right-hand side, the variable type for each attribute is listed.

CSV file of the relational tables. Details of the implementation can be found at the GitHub repo link: https://github.com/cgunaratne/CSC-hockey-data.

3 Analysis Methodology

As the data was modeled relationally using MySQL, it is possible to efficiently query the database without loading everything into memory. Accessing parts of the database are done with 'Queries' that allow a user to define which section of the database to select. The Python package pymysql (https://pymysql.readthedocs.io/en/latest/) was used to perform queries and return the data as variables in Python for ease of visualization. All analysis was performed through a combination of queries and basic statistical operations on attributes found in our data model.

3.1 Relational Queries

Presented below is a selection of relational queries in algebraic format we used to access our database. Demonstrating the relational algebra allows others to reproduce our methods in their desired framework. A major consideration is that we are limiting our movies to the United States. From our logical model (fig. 2), the table that holds regional information is 'Releases' therefore, movies that are only included in the US region:

$$\pi_{(IMDB\_ID)}(\sigma_{Region='US'}Releases)) \bowtie Movies \tag{4}$$

where σ is the selection operator, π is the projector operator which isolates the unique tuples, and $\bowtie$ is the natural join operator which merges each tuple that has the same value for the common attribute and in this example, it is the IMDB_ID. Our analysis regularly needs to access the many-to-many table 'Genre_Of_Movies'. For a given movie (identified by its MovieLens identifier 'ML_ID0')

$$\pi_{Genre\_Name}(\sigma_{(ML\_ID=ML\_ID0)}Genre\_Of\_Movies) \tag{5}$$

4 Results and Discussion

In the previous sections, we have outlined how we designed our logical model and developed a MySQL database to facilitate our analysis. In this section, we outline how we use this database to answer the questions related to our analysis, specifically, how external factors may impact the ratings of movies.

4.1 Movies

First, we explore the number of movies and when they were released, presented in fig. 3. We see a bump in movies in 1930, and the number of movies only increases until the 2000s. The early development of this technology was only around 1900, though some experimental and early work existed as early as 1874 with 'Passage de Venus', a 6-second clip of the transfer of Venus. While the COVID-19 pandemic had a major impact on the number of movies released [1] the drop in movies in the late 2010s is instead caused by the last update to this dataset. The most recent movies in the MovieLens dataset were released in 2018, and the apparent drop is likely an artifact of how the most recent movies were collected.

4.2 Financial

Inflation data can act as an imperfect measure of the economy. Our analysis focuses on the United States due in part to the availability of financial information. Figure 4 shows average annual inflation for the period between 1874 (the first film in our analysis) and 2022. Inflation is the change in purchasing power between two years and tends to be slightly positive. Periods of economic uncertainty are reflected in fig. 4 with the World Wars, early Cold War, and 1970s oil crisis being notable deviations. More recently we see the 2009 recession and the post-pandemic inflationary period. Since movies have high budgets and are luxury goods, we expect the number of movies released to be lower in times of economic uncertainty.

4.3 Genre

Our conceptual and logical frameworks were developed to be able to investigate genre. Many movies have been assigned to multiple genres, which complicates our analysis. Consider 'Jumanji (1995)' which has been assigned 3 genres; 'Adventure', 'Children', and 'Fantasy'. We have decided to consider this movie

[1] A comprehensive overview of digital entertainment in 2020: https://www.motionpictures.org/wp-content/uploads/2021/03/MPA-2020-THEME-Report.pdf

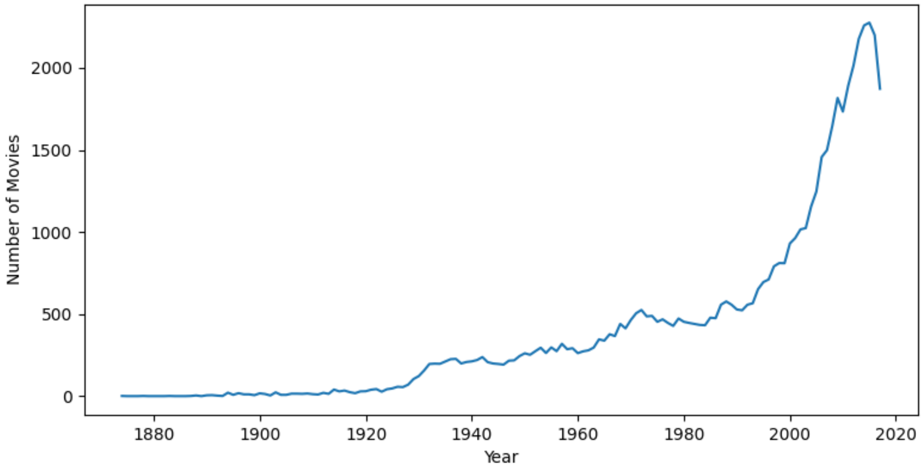

Figure 3: Number of movies released each year

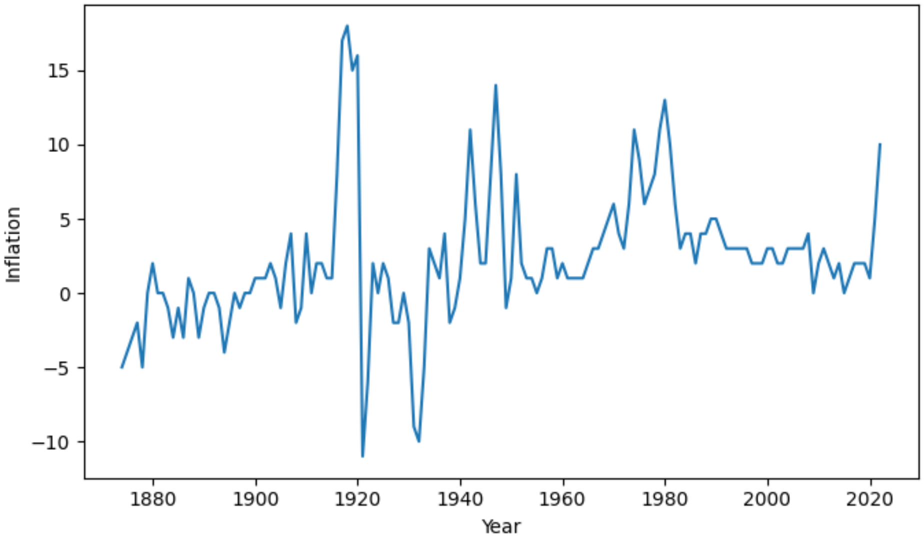

Figure 4: US inflation rate since 1874

as being counted as a full movie in each genre. Therefore, if we were to divide the fraction of movies by genre in a year, we would end up with fractions that add up to greater than 1 (see fig. 7 for an example of this). We believe this is appropriate when making comparisons between genres because it is maximally permissive; that is, an ‘Adventure’ isn’t penalized for also being a ‘Fantasy’. An alternative approach would be to weigh the movie relative to the number of genres. In that case, ‘Jumanji (1995)’ would be counted 1/3 for each of ‘Adventure’, ‘Children’, and ‘Fantasy’.

We investigated how ratings are distributed among genres. Figure 5 shows that most ratings follow a similar distribution. There appears to be a bias to whole number scores (this may be an artifact of how the data was collected, especially if partial scores were not available in one of the sources). The panel in the lower right shows the average rating regardless of the genre across this dataset.

We have selected the genres of all movies produced and presented this in fig. 6. Here we see that ‘Drama’ and ‘Comedy’ are the most numerous, with ‘Thriller’, ‘Romance’, and ‘Action’ completing the top-5 list.

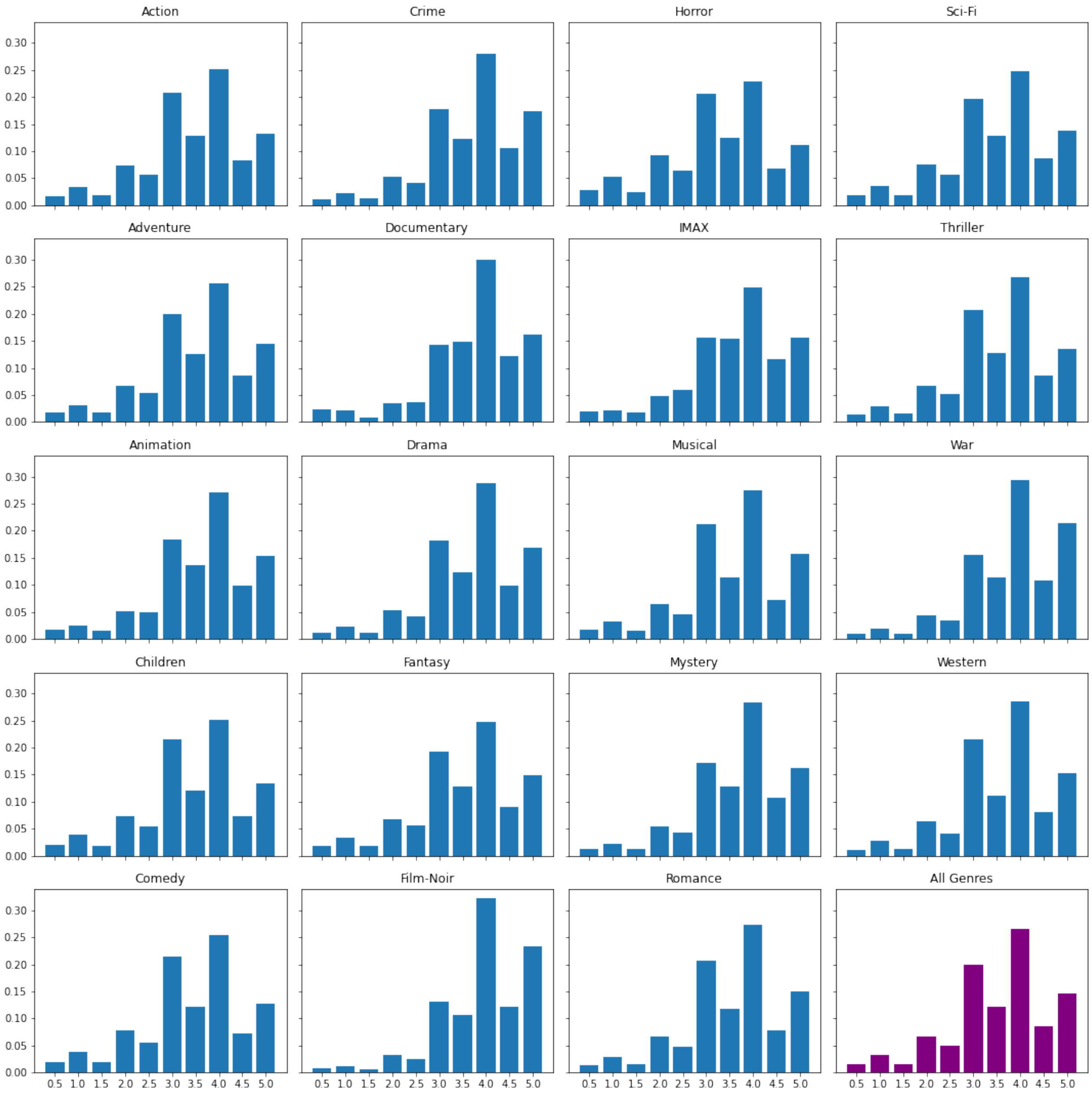

Figure 5: Distribution of ratings for each genre

From this top-5 list, we plotted the fraction of movies of that genre over time as seen in fig. 7. The goal is to determine whether movies of a certain genre have become more numerous over time and whether those changes correspond with the economic situation. We see that 'Romance' movies were most numerous around the World Wars. The number of 'Comedy' movies dropped in the early Cold War era. 'Thriller' movies have had a slow and steady incline since they first appeared. Conclusions about the earliest movies (before 1920) are difficult to determine due to how few movies there were, causing high variations.

With the constraint that we are inspecting only the releases from the United States, fig. 8 shows how ratings change over time for the top-5 most numerous genres. These ratings use the 'Dates' table to map the time the rating was submitted with the year. These five genres show correlated trends such as a dip in 2005 and a bump in 2014 with 'Drama' maintaining a consistent +0.1 buffer. The variation in 1995 is likely an artifact of how and when the data was collected.

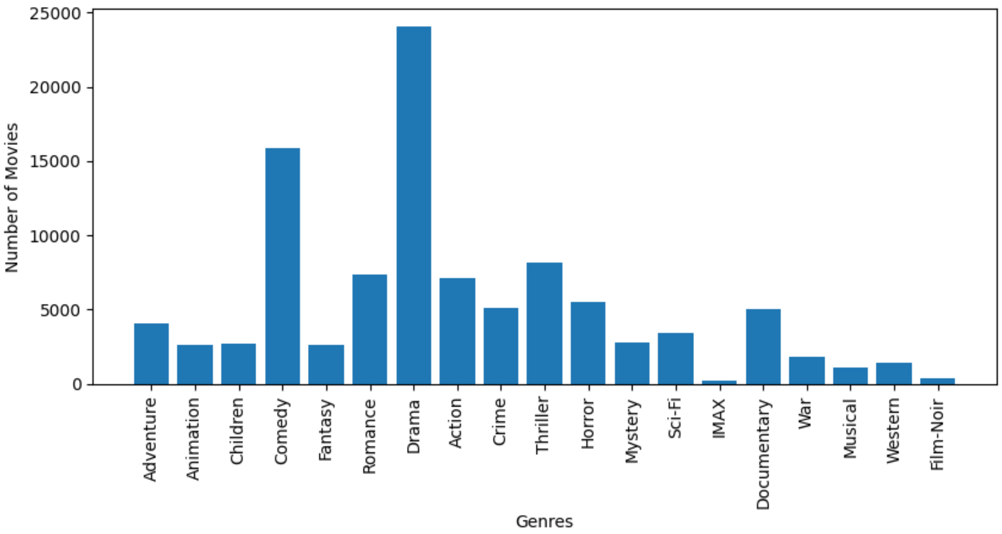

Figure 6: Total number of movies released up to 2018 of given genres

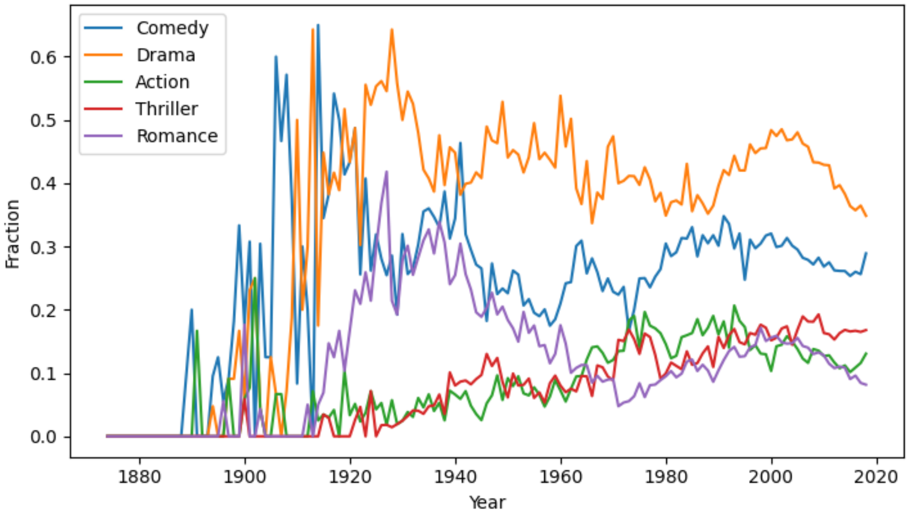

Figure 7: Fraction of movies by genre for top-5 genres

4.4 Ratings by Users

The MovieLens database contains 27 million ratings from over 28000 users. Users can only have one active rating per movie, which appears to be enforced in the data collection process. We have further enforced this through the use of the composite primary key (ML_ID, UserID) which ensures that for a given UserID and ML_ID there can be only one rating.

4.5 Directors

The number of movies each director directs must be considered. In fig. 9, we show that of the 13904 directors in our database, most directors direct only 1 movie, and there is a long tail with one director, Michael Curtiz, having 78 movies released.

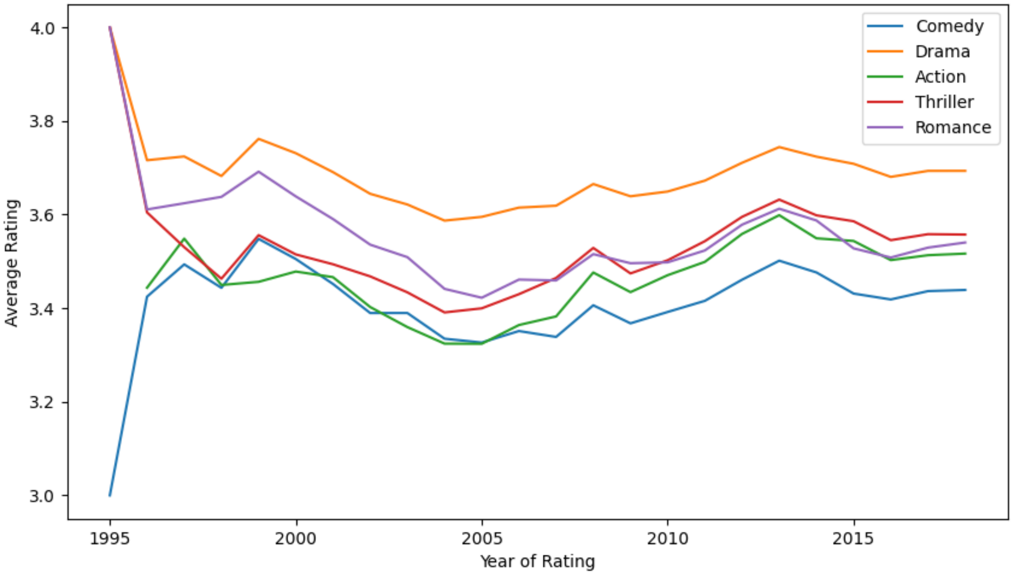

Figure 8: Average rating by genre using real-time updates

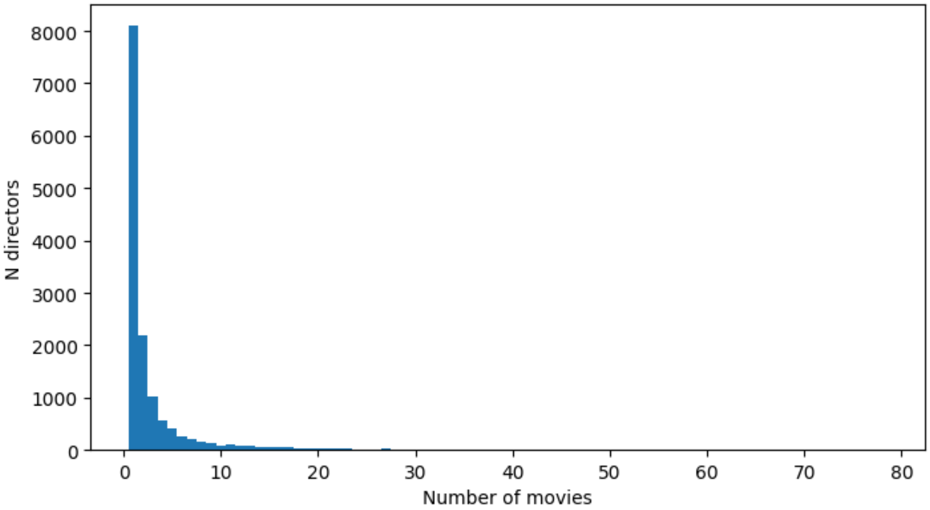

Figure 9: Number of movies per director

4.6 New Release Peak

We have access to the times when ratings were placed for a movie. As a byproduct of how these ratings were collected, no rating occurs before 1995. For movies released since then, we can inspect when movie ratings are submitted. Figure 10 looks at the 'Crash (2004)' compared with 'Crash (1996)'. Despite the name, these movies are unrelated. We can see that both movies have a peak just after their respective releases. 'Crash (1996)' has a long tail of reviews, likely from it having a cult following. 'Crash (2004)' has a second peak in 2015, likely coming from a media retrospective [2] of Oscar controversies that brought new viewers to the movie, or caused previous viewers to update their rating. We would like to note that both movies don't have a rating until the year after they were released. We expect this is an artifact of how the

[2] Media retrospective: https://www.theatlantic.com/magazine/archive/2014/03/what-was-the-biggest-oscar-mistake-ever-made/357581/

movies were added to these databases, as the lag time between release and first ratings could cross into the following year. Both films were released in the fall of their respective years.

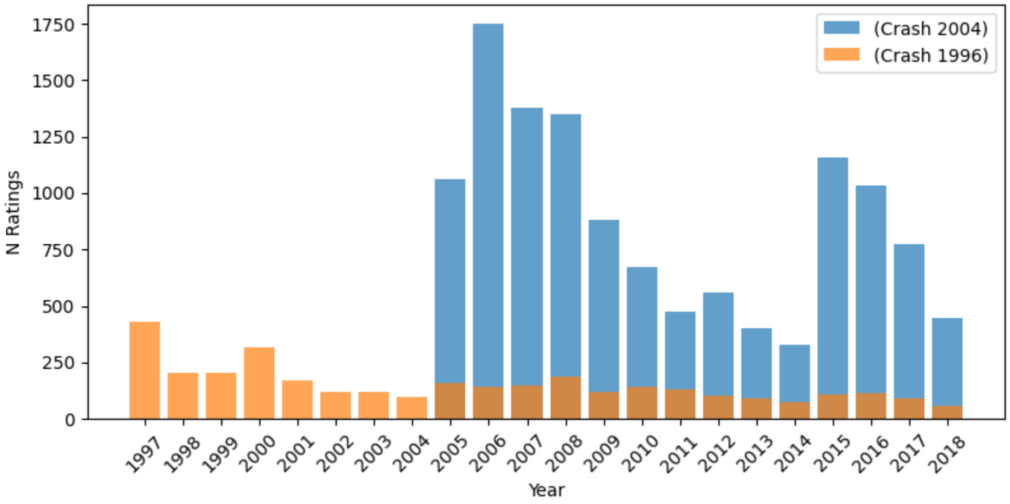

Figure 10: Timing of ratings after movie release.

5 Conclusions

In this report, we have demonstrated conceptual and logical schemata to structure MovieLens data and associated datasets to analyze external factors that impact ratings and genres. Our implementation was using MySQL and Python, and by outlining the requirements in relational algebra we have made the model generalizable so that a different framework could be built to recover the same data model.

We demonstrated our data model was appropriate for our analysis by implementing it and performing queries to extract the data and present these results through visualizations. In doing so, we have identified potential artifacts in the dataset which could be repaired through additional cleaning, or adding supplementary data to this data model. With the constraints of our model clearly outlined, the extension of this model to more recent movies would be facile.