Analysis of the relationship between movies and reviews

Pasquale Gravante [896983]

Antonio Mastroianni [898723]

The following report describes the procedure used to obtain a database containing all the useful information for a study regarding the relationship between movies and its reviws. The review dataset was taken from IMDb while the movies dataset was taken from Kaggle at the folling link:
`https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata`

Contents

1 Introduction and Research Question

1.1 TMDB

TMDB, short for The Movie Database, is an online platform and community-driven database that provides comprehensive information about movies, TV shows, and other forms of visual entertainment. Founded in 2008, TMDB has become one of the most popular and widely-used sources of film-related data, serving as a valuable resource for enthusiasts, professionals, and developers in the entertainment industry. The primary objective of TMDB is to offer a centralized and accessible repository of information related to movies and TV shows. The database includes details such as titles, release dates, cast and crew information, plot summaries, genres, production companies, posters, and trailers. It covers a vast range of films and series, spanning different genres, languages, and time periods. What sets TMDB apart is its collaborative and community-driven nature. Users can contribute to the database by adding and editing information, submitting reviews, rating movies, and uploading images. This collaborative approach has allowed TMDB to grow into a rich and constantly expanding collection of film-related data, with a focus on accuracy and completeness. In summary, TMDB is a prominent online platform and community-driven database that provides a wealth of information about movies and TV shows. With its collaborative nature, comprehensive data coverage, and developer-friendly API, TMDB has become a go-to resource for film enthusiasts, professionals, and developers seeking accurate and up-to-date information about the world of visual entertainment.

1.2 IMDb

IMDb, also known as the Internet Movie Database, is an online database and platform dedicated to providing information about movies, TV shows, actors, filmmakers, and other industry professionals. It serves as a go-to resource for film enthusiasts, professionals, and casual viewers seeking comprehensive and up-to-date information about the world of entertainment. IMDb's extensive database covers a wide range of content, spanning across different genres, languages, and time periods. It includes a vast library of movies and TV series from various countries and offers insights into the careers of countless actors, directors, writers, and other industry professionals. One of the key features of IMDb is its comprehensive and user-friendly interface, allowing visitors to easily search for movies, explore filmographies of actors and filmmakers, discover popular and trending titles, and access trailers and promotional material. Additionally, IMDb provides industry news, interviews, and editorial content to keep users informed about the latest developments in the entertainment world.
IMDb's popularity and credibility stem from its commitment to accuracy and its robust user community. Users can contribute by submitting information, rating movies, and writing reviews, contributing to the overall depth and richness of the database. This collaborative approach has made IMDb a reliable source of information and a platform for film enthusiasts to share their opinions and insights.

1.3 Research Question

The report mainly focuses on answering the following research question:

"How can the film data from Kaggle be effectively integrated with the review data from IMDb APIs to provide a comprehensive analysis of movies reviews? And what is the relationship between the reviews and the success/popularity of the movies?"

In other words, the aim is to find the best possible ways to effectively integrate our data and explore the connection between IMDb ratings, which represent the audience's perception of quality, and the popularity of TMDB titles. By analyzing an integrated database, we have access to more information that we can use to perform a lot of tasks, one being for example sentiment analysis.

2 Data Acquisition

2.1 Kaggle

As previously anticipated, the first dataset was download from Kaggle and it is a collection of 5000 movies from TMDB. From the initial dataset, only 1000 movies were taken into account otherwise the process would have been too much resource/time-consuming.
It contained the followed informations:

- Budget
- Genres
- Homepage
- ID
- Keywords
- Language
- Original Title
- Overview
- Popularity
- Production Company
- Production Country
- Release Date
- Revenue
- Runtime
- Spoken Languages
- Status
- Tagline
- Title
- Vote Average
- Vote Count

2.1.1 IMDb APIs

To extract the information from the IMDb website, we used the IMDb library, that is a Python package that connects to IMDb APIs and provides a convenient way to extract data from the website. It allows to retrieve information about movies, TV shows, actors, directors, and other related details that are available on IMDb.
From that, we extracted the following information:

- Title
- Rating
- Review

3 Data Quality Evaluation

After extracting the data we needed, we wanted to check one of the most important aspects, its quality. In particular, we used objective measures like Accuracy and Completeness.

3.1 Quality Measures

- **Completeness** To evaluate if there are missing values in some of the fields, with the formula:

$$1 - \left(\frac{\text{Missing Values}}{\text{Total Number of Values}} \right) \% \tag{1}$$

- **Semantic Accuracy** To evaluate the difference between two strings (Movie Titles) by means of the normalized edit distance:

$$ED_{norm} = 1 - \frac{ED(v1, v2)}{n} \tag{2}$$

  In particular, the ED_{norm} was used to evaluate how difference the movie titles in the Kaggle dataset were from the original ones taken from the IMDb official website.

3.2 Results evaluation

By applying our evaluation criteria we obtained the following results:

Variable	**Completeness**
Budget	100%
Genres	100%
Homepage	52.4%
ID	100%
Keywords	100%
Language	100%
Original Title	100%
Overview	100%
Popularity	100%
Production Company	100%
Production Country	100%
Release Date	100%
Revenue	100%
Runtime	100%
Spoken Languages	100%
Status	100%
Tagline	96.6%
Title	100%
Vote Average	100%
Vote Count	100%

In addition, the accuracy for the Title variable is reported:

Variable	Accuracy
Title	99%

3.3 Overall Quality Evaluation

By applying our evaluation criteria we obtained the overall following results:

- Overall Accuracy: 99.00%
- Overall Completeness: 97.40%

4 Data Storage

We decided to use MongoDB for storing our data because it is a NoSQL database that offers flexibility, particularly when we were uncertain about the data schema beforehand. Since Movies can have multiple genres of varying lengths, a NoSQL database allowed us to create documents with ease. Among the different types of NoSQL databases, we specifically chose a document-based one because its JSON-like format seemed ideal from the beginning. During the API acquisition process, we found that using Python's lists and dictionaries to temporarily store the data was very convenient. As a result, the scraped data was in the form of a list of dictionaries, with each dictionary representing a game. Storing the data permanently in a JSON format seemed like the most obvious choice to us.
Furthermore, considering our research focused heavily on movies, we knew that our queries would involve both the characteristics of the movies and their review scores. Instead of solely querying the review table, we would be querying the movies themselves. This is why we believed that the "embedding" approach offered by a document-based database would be more practical and efficient than the separate table structure of a relational database, which would require frequent join operations for our queries.

5 Data Integration

5.1 Schema Transformation

To achieve a good data integration, the first step is to schema transformation which consists in homogenize the schema of the two different sources of data. In order to do so, we have manipulated data to guarantee consistency and compatibility.

In order to achieve this goal, we had to solve conflicts among data. In particular, semantic conflicts and naming conflicts. These problems are faced in the following section.

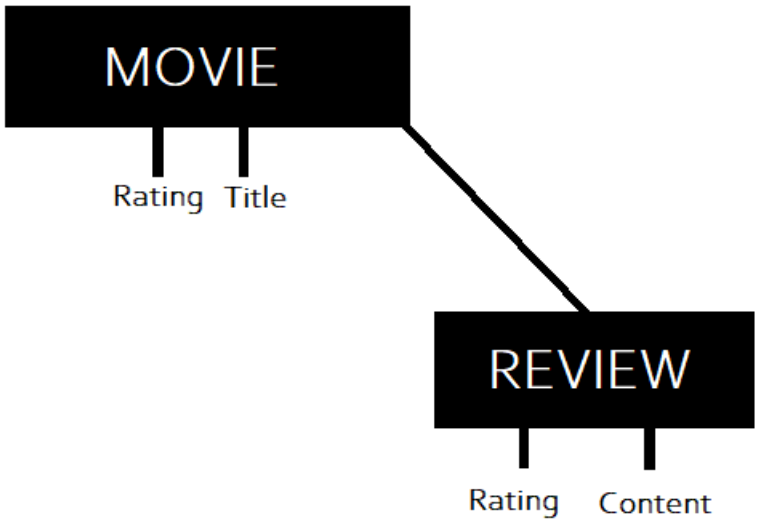


Figure 1: Json schema of IMdb data

5.2 Data Cleaning and Conflicts Resolution

Since the movie industry doesn't have a universal Id to identify their products, the decision was to use the movie titles to merge the two datasets. This implied the necessity of the normalization of the titles in order to prevent conflicts during the merge of the two datasets.
Application of the transformations included:

- Removal of special characters.
- Transformation of the titles in lowercase.
- Removal of whitespaces.

Moreover, we deleted:

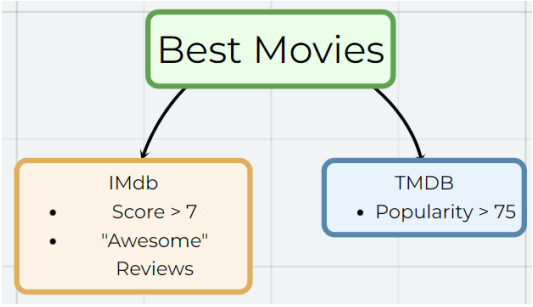
- Movies for which the status was not equal to "released".
- Columns with the initial titles since we created new normalized ones.

To solve the naming conflict between the variables (rating in the IMDB and vote in the TDBM dataset), it was chosen to use the arithmetic mean between the two values.

5.3 Data Merging

In this study, the data has undergone a merging process based on the "Normalized Title" column which serves as a key or identifier for matching and combining related data from different sources or datasets. Using a merging algorithm or technique, we successfully merged the data based on the values in the "Normalized Title" column. Out of the 1000 records considered, we were able to match and merge 942 records, resulting in a match rate of 94.2%. After completing the merging process, the merged data was uploaded to MongoDB. A new collection named "Merged Collection" was created to store the merged data. By merging the data based on the "Normalized

Title" column and storing it in the "Merged Collection" in MongoDB, we have consolidated the related information from multiple sources into a single, unified dataset. This merged dataset can now be used for further analysis, exploration, and deriving valuable insights.



In this section, we describe the process of querying and merging the databases to obtain a dataset of the most appreciated movies. Our objective was to consolidate information from multiple sources and create a unified dataset that represents the most highly regarded movies.
In particular, the queries performed were:

1. Get the movies with a score greater than 7 for the IMdb database and whose review contains the word "awesome".

2. Get the movies with a score greater than 75 for the TDBM database.

The queries generated two individual dataframes, each representing a subset of the most appreciated movies from its respective database. In order to consolidate the data and create a comprehensive dataset of the most liked movies, we merged the individual dataframes obtained from the queries. In particular,

6 EDA

Exploratory analysis and Visualizations is carried on in this section. First of all, the bar plot showing the genres of the most popular movies is showed in Fig. 2

From the bar plot, it is visible that, among the most popular movies, the most frequent genres are Action and Adventure while the least frequents are Mystery and Western movies.
In addition, the distribution of the rating score for each genre of the movies is investigated. The results are provided by a grouped boxplots chart which is showed in Fig. 3

The movie with the highest median is Romance, probabily indicating that the even though this genre is not one of the public most favorite, they are highly appreciated by the small niche of aficionados. The lowest variation genre corresponds to Western. This is attributed to the presence of only one movie in the database of the most appreciated movies. The highest variation genre is instead is Action with a median of

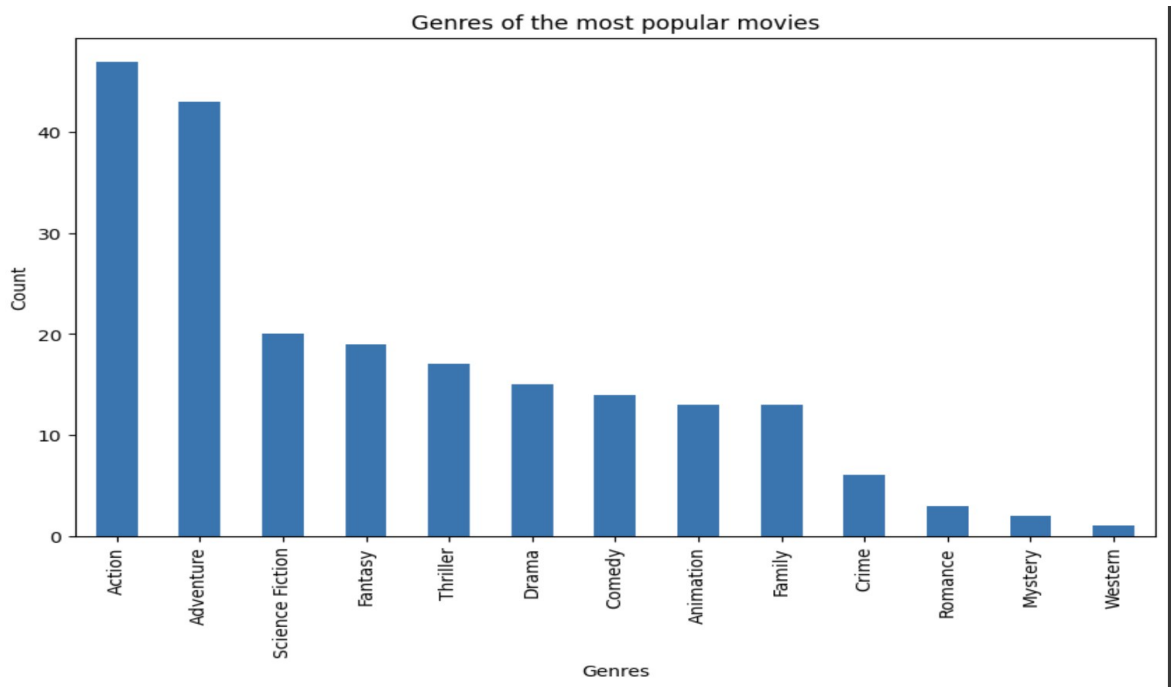


Figure 2: Most frequent genres between most popular movies

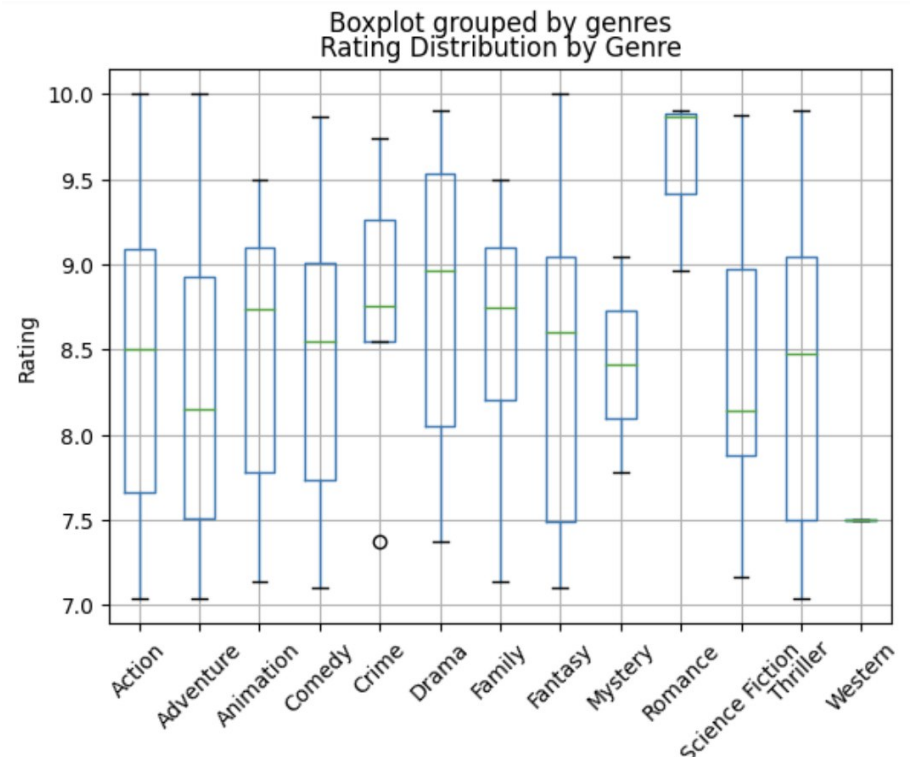


Figure 3: Distribution of Rating according the to genre

8.5. The high variation is probably cause by the fact that this is the most frequent genre in the created dataabase.
Moreover, the distribution of the Rating according to the budget spent for the movie is presented in Fig. 4

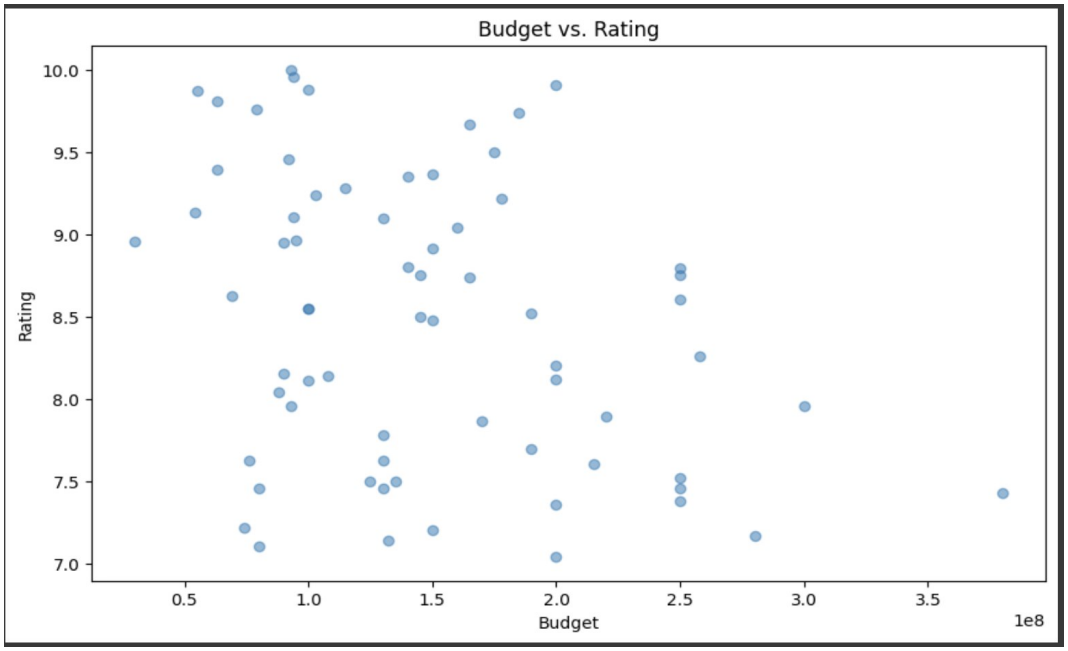


Figure 4: Scatter Plot of the Rating according to the Budget

Differently from what expected, the highest-budget movies are not the ones having the highest-score. It is shown that the movies with the highest ratings have budget values contained in the first 50% of the budget range.

7 Conclusions

We successfully merged the IMDb and TMDB databases, consolidating their respective collections of movie data. The merging process allowed us to create a unified dataset that combined information from both sources, providing a more comprehensive and enriched resource for analysis and insights.
By merging the IMDb and TMDB databases, we achieved several benefits. First, we obtained a larger dataset, incorporating a wider range of movies and associated attributes. This expanded dataset enhances the representativeness and coverage of our analysis, enabling us to draw more robust conclusions. Additionally, merging the output of the queries allowed us to obtain a database of the most appreciated movies, which might result in a treasure for all the movie lovers.
In future research, one approach to address these challenges could be adopting a "trust your friend" strategy which could help in the task of obtaining a 100% match leading to a more reliable merged database. Moreover, exploring opportunities for integration with other relevant platforms or databases can further enhance the value of the merged dataset. For example, other industry-specific databases can provide users with a more integrated and immersive experience.
The merging of the IMDb and TMDB databases has provided us with a valuable consolidated dataset that expands the scope and richness of movie information available for analysis. While the merging process brings its own challenges, the potential adoption of a "trust your friend" approach in future research holds promise

for further enhancing the quality and accuracy of merged databases. This will facilitate more comprehensive analyses and deeper insights into the world of movies.