CS432/532: Final Project Report

Project Title: TMDB Movie Data Analysis

Abhang Tejas Dnyaneshwar, Birudala Nikitha Reddy, Nerkar Kaushal Vinay

***Abstract*—The TMDB Movie Data Analysis project endeavors to delve into the vast repository of movie metadata encapsulated in the TMDB 5000 Movie Dataset. This report provides an exhaustive examination of the project's objectives, methodologies, analysis techniques, and potential challenges, aiming to unearth valuable insights for stakeholders in the movie industry.**

I. Introduction

The TMDB (The Movie Database) analysis project involves examining and deriving insights from a vast dataset of movies and TV shows. This project typically entails tasks such as data collection, data cleaning, exploratory data analysis (EDA), and potentially building predictive models or creating visualizations to better understand trends and patterns within the entertainment industry. By analyzing various attributes such as genre, release year, ratings, and popularity scores, researchers or enthusiasts can uncover valuable insights about audience preferences, industry trends, and the factors that contribute to the success of a movie or TV show.

Here merging of two collections from the dataset, namely "movies" and "credits," enables comprehensive analysis by consolidating information about both the movies themselves and their associated credits (cast, crew, etc.). This integrated dataset provides a rich resource for querying and deriving insights through MongoDB aggregation pipelines.

II. NoSQL Database and Dataset

A. NoSQL Database: MongoDB

MongoDB offers a flexible and scalable platform for analyzing the TMDB movie dataset due to its document-oriented nature and powerful aggregation framework. The dataset, sourced from TMDB, provides a rich array of attributes encompassing movie details, cast, crew, ratings, and more. MongoDB's ability to handle complex nested data structures aligns well with the diverse information present in the TMDB dataset. Its aggregation pipeline allows for sophisticated querying, facilitating tasks such as grouping, filtering, and statistical analysis. Additionally, MongoDB's horizontal scalability ensures efficient handling of large volumes of data, making it suitable for processing the extensive TMDB dataset. Overall, the combination of MongoDB and the TMDB movie dataset offers a potent platform for conducting comprehensive analyses, extracting valuable insights into movie trends, audience preferences, and the factors influencing movie success.

B. Dataset: TMDB 5000 Movie Database

The TMDB 5000 Movie Dataset stands as a comprehensive repository of movie metadata, encompassing a vast array of attributes crucial for detailed analysis and exploration of the film industry's dynamics. This dataset presents a treasure trove of information spanning various facets of movie production, distribution, and reception. Comprising around 5000 rows and 20 columns, it offers a rich and diverse resource for researchers, analysts, and enthusiasts alike.

At its core, the dataset provides essential details about individual movies, including their titles, release dates, and genres. These fundamental attributes serve as the foundation for deeper exploration into the cinematic landscape. Additionally, the dataset captures intricate details about the cast and crew involved in each film, shedding light on the talent behind the scenes. By delving into information about directors, actors, writers, and other crew members, analysts can gain valuable insights into the collaborative efforts driving movie production.

One of the dataset's significant strengths lies in its inclusion of financial data, such as budget and revenue figures. These metrics offer vital insights into the economic aspects of filmmaking, allowing researchers to explore correlations between production costs, box office performance, and profitability. Furthermore, the dataset includes information on movie popularity and user ratings, providing a glimpse into audience preferences and reception.

III. NOSQL QUERIES

Query 1 : Identifying trends in actor rating based on movie popularity, revenue, budget, user votes.

The query aims to uncover trends in actor performance based on various aspects of a film's success, including budget, revenue, popularity, and user votes. By generating actor ratings in relation to these factors, we can gain valuable insights into the connection between actor performance and the overall success of a movie.

To execute this query, we leverage attributes such as budget, revenue, vote count, vote average, and popularity from the structured data, along with the unstructured data of the movie's cast. We calculate an actor rating by aggregating relevant metrics across all movies they've appeared in. This actor rating serves as a proxy for their performance and influence on a movie's success.

Using a scatter plot, we visualize the relationship between actor ratings (representing popularity and vote average) and movie success scores (derived from budget and revenue). This visualization allows us to identify correlations or patterns between actor performance and movie success metrics. For instance, we may observe that movies featuring highly-rated actors tend to have higher budgets or revenues, indicating the potential influence of star power on a film's financial performance.

Overall, this query and visualization provide a comprehensive analysis of the interplay between actor performance and various aspects of a movie's success, offering valuable insights for filmmakers, studios, and industry professionals seeking to optimize casting decisions and enhance the commercial prospects of their films.

Query 2 : Explore how frequently certain cast and crew members collaborate across multiple movies.

This query delves into the collaborative patterns between cast members, directors, and production companies across multiple movies, aiming to identify the most frequent pairs of collaborators. By leveraging attributes such as cast, crew, and production companies from the dataset, we analyze the relationships between individuals and entities involved in movie production.

To execute this query, we first extract the unstructured data of cast, crew, and production companies from the dataset. We then analyze the frequency of collaboration between pairs of cast members, directors, or cast-director pairs across different movies. This involves counting the number of times two individuals or entities have collaborated within the dataset.

Using a Sankey plot, we visualize the relationships between actors and directors based on their collaborative history. This visualization highlights the flow of collaborations between various cast members and directors, illustrating the strength and frequency of their partnerships. For instance, we may observe prominent actors consistently working with specific directors across multiple movies, indicating strong professional relationships or creative synergy.

Overall, this query and visualization provide valuable insights into the collaborative dynamics within the film industry, shedding light on the patterns of cooperation between cast members, directors, and production companies. This information can inform casting decisions, production strategies, and industry partnerships, fostering greater efficiency and creativity in movie production processes.

Query 3 : Identify the most dominant genres in each decade based on the number of movies and analyze the evolution of genre popularity

This query aims to identify the most dominant genres in each decade based on the number of movies and analyze the evolution of genre popularity over time with respect to production companies. By leveraging attributes such as release date, genres, and popularity from the dataset, we can uncover trends in genre preferences and their association with production companies across different decades.

To execute this query, we first extract the structured data of release dates and popularity, along with the unstructured data of movie genres. We then group the movies by decade and analyze the frequency of each genre within each decade. This allows us to determine the most dominant genres in each decade based on the number of movies associated with them.

Using this information, we further analyze the popularity of genres over time with respect to production companies. We examine how the popularity of genres fluctuates across decades and whether certain production companies specialize in particular genres or exhibit trends in genre preferences over time.

A bar plot is employed to visualize the top three genres over decades based on popularity. This visualization illustrates the evolution of genre popularity over time and highlights any shifts or trends in audience preferences. Additionally, it provides insights into the strategies of production companies in catering to changing audience tastes and preferences.

Overall, this query and visualization offer valuable insights into the dynamics of genre popularity across different decades and their relationship with production companies, facilitating a deeper understanding of the evolution of the film industry over time.

Query 4: Analyze the relationship between movie revenue, genres, and user ratings to identify trends in audience preferences.

This query aims to analyze the relationship between movie revenue, genres, and user ratings to identify trends in audience preferences, with a specific focus on a chosen genre. By leveraging attributes such as revenue, genres, user ratings, and movie overviews from the dataset, we can uncover patterns and insights into audience preferences within the selected genre.

To execute this query, we first filter the dataset to include only movies belonging to the specified genre. We then examine the distribution of movie revenue within this genre, calculating metrics such as average revenue. Additionally, we analyze user ratings for movies within the genre, computing metrics like average rating.

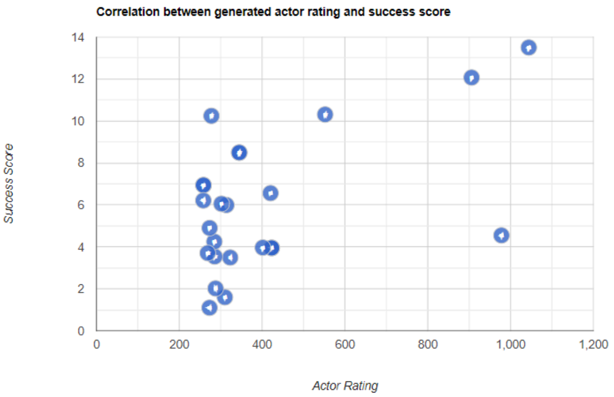
Incorporating the movie overviews, we extract revenue and ratings-related patterns by analyzing textual information for potential correlations with revenue and user ratings. This step involves natural language processing techniques to extract relevant insights from the unstructured data.

Using a bar plot, we visualize the top 10 genres alongside their respective average revenue and average rating. This visualization enables us to compare audience preferences across different genres based on revenue and user ratings, identifying trends and potential correlations.

Overall, this query and visualization offer valuable insights into audience preferences within a specific genre, shedding light on the relationship between movie revenue, user ratings, and textual information from movie overviews. By understanding these trends, filmmakers and studios can make informed decisions regarding genre selection, content creation, and audience engagement strategies.

IV. Project Outcome

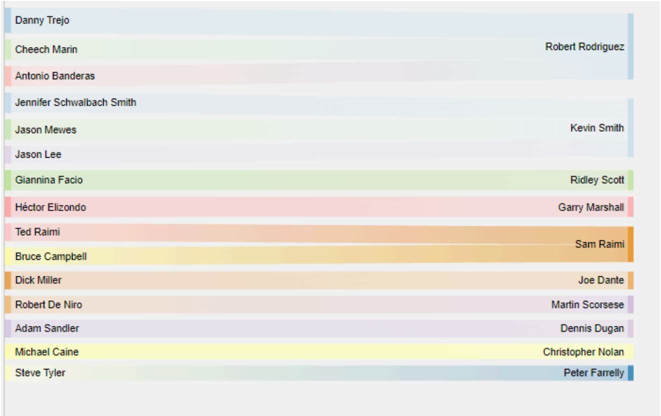
Query 1 : Identifying trends in actor rating based on movie popularity, revenue, budget, user votes.



Key Findings:

1. The data suggests a positive correlation between the two variables. This means that movies with higher actor ratings tend to also have higher success scores.
2. Several factors can influence a movie's success beyond the popularity of the actors. These factors include script quality, director reputation, marketing budget, and genre. A movie with a strong script, a well-regarded director, and a successful marketing campaign may still be a hit even if the actors have low ratings.
3. The generated actor rating system might be biased towards certain genres or actor types. For instance, comedies or actors known for comedic roles might receive high ratings, while dramas or dramatic actors might receive lower ratings, even if both genres and actor types can be commercially successful.

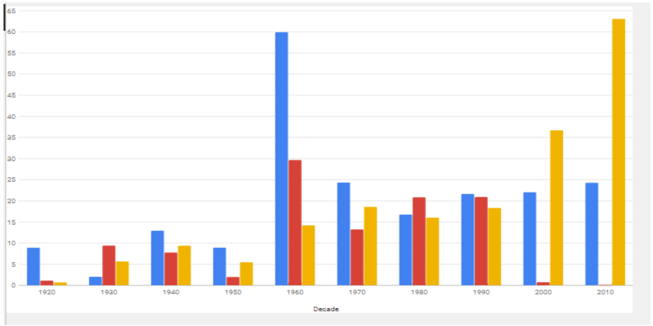
Query 2 : Explore how frequently certain cast and crew members collaborate across multiple movies.



Key Findings:

1. Actors and directors develop recurring working partnerships. The thickness of the lines connecting directors and actors indicates the number of films they have collaborated on.
2. The plot highlights prominent collaborations. For example, the data suggests a strong collaborative history between Sam Raimi and Bruce Campbell, likely exceeding the number of films other directors have made with any single actor on the list.
3. The visualization allows for identification of multiple collaborations. Directors like Robert Rodriguez and Kevin Smith have collaborated on multiple films with several actors, as evidenced by the thicker lines connecting them to actors like Danny Trejo, Antonio Banderas, Jason Mewes, and Jason Lee.
4. The plot can reveal long-standing partnerships. The Sankey plot might showcase collaborations between actors and directors who have worked together consistently throughout their careers, such as the potential connection between Martin Scorsese and Robert De Niro.

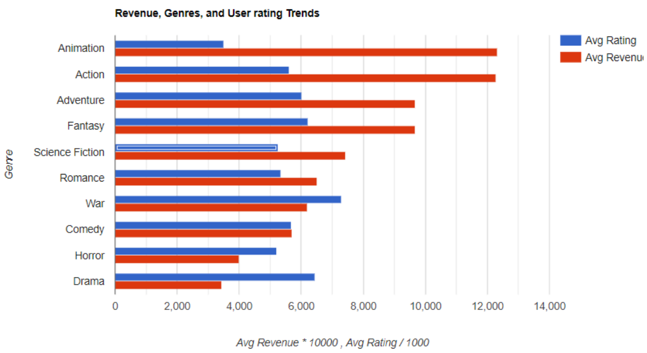
Query 3 : Identify the most dominant genres in each decade based on the number of movies and analyze the evolution of genre popularity



Key Findings:

1. As filmmaking technology advances, new creative possibilities emerge, potentially influencing the types of genres that become popular. For instance, the introduction of sound in the 1920s opened doors for more dialogue-driven genres like Dramas.
2. Movie genres can reflect the social and cultural climate of a particular era. Comedies might gain popularity during times of economic prosperity or peace, whereas Dramas might resonate more during periods of social unrest or economic hardship.
3. The success of a major franchise within a particular genre can significantly impact the overall popularity of that genre. For example, the Star Wars and Indiana Jones franchises likely contributed to the rise of Action/Adventure films in the 1980s, while the Marvel Cinematic Universe's dominance propelled Superhero movies to the top spot in the 2010s.

Query 4: Analyze the relationship between movie revenue, genres, and user ratings to identify trends in audience preferences.



Key Findings:

1. Animation, Fantasy, and Adventure appear at the top in terms of both average revenue and ratings. This suggests that audiences are drawn to these genres and are willing to pay to see them. These genres often feature fantastical elements, special effects, and exciting narratives that might be particularly appealing to a broad audience.
2. Action and War movies tend to have high average revenue but slightly lower average ratings compared to Animation, Fantasy, and Adventure. This could indicate that these genres are commercially successful but may not always be critically acclaimed. Action movies often prioritize exciting stunts and set pieces over complex narratives, while War movies might depict violence or deal with sensitive subject matters that may not resonate with all viewers.
3. Drama and Comedy appear to have lower average revenue compared to some other genres but still maintain high average ratings. This might suggest that these genres, while critically well-received, may not always translate into high box office numbers. Dramas tend to focus on character development and social issues, and comedies can be subjective depending on individual taste in humor.

References

[1] Murray, P., Collins-Thompson, T. D., & Bennett, P. N. (2015). Mining IMDb: Exploring Trends, Genre Evolution, and the Impact of Star Power. ACM Transactions on Interactive Intelligent Systems.

[2] Mednis, A., Brezhnev, R., & Romans, D. (2017). Predicting the Success of Movies Based on Genre, Cast, and Crew. Procedia Computer Science.