Group 8 Project 4

Netflix Recommender

Project Team:
Andrew Montemayor, Mason DeJesus,
Judith Landa, Mathias Carnero

Introduction: (what is your dataset and the purpose of the project)

Our project explores a dataset of over 16,000 movies spanning from 1910 to 2024, sourced from Metacritic. This dataset contains a rich variety of data points, including movie titles, directors, ratings, and user votes, providing a comprehensive view of how movie reception has evolved over time. The purpose of this project is to analyze trends in movie ratings, popularity, and directorial impact while exploring how these factors differ across genres, time periods, and cultural backgrounds. Our analysis uses machine learning models and interactive visualizations to uncover insights into the influence of directors and genres on audience reception.

Data Cleaning/Database Creation:

To ensure the integrity of our analysis, we performed extensive data cleaning on the original dataset. This process involved removing duplicates, handling missing values, and normalizing data fields to ensure consistency. We also standardized director names and genres to eliminate variations in spelling and format. After cleaning the data, we created a relational database to allow for efficient querying and analysis. This setup enabled us to filter data by time period, director, and genre, making it easier to extract meaningful insights for both machine-learning experiments and visualizations.

Color Design:

Our design choices were inspired by Netflix's sleek and modern aesthetic. We opted for dark, contrasting color schemes to create a cinematic feel, drawing attention to key data points without overwhelming the viewer. Palettes were chosen from popular color generators like Coolors, featuring shades of blue, black, and gray to maintain visual consistency. Each visualization type—bubble charts, heatmaps, and line charts—was designed with clarity and

engagement in mind, ensuring the data is both informative and visually appealing. The aim was to create dashboards that are easy to navigate while still maintaining a professional, polished look.

ML Experiment:

To explore patterns in the dataset, we applied machine learning models, including regression analysis and clustering techniques. These models helped us predict future trends in movie ratings and genre popularity. For example, we used linear regression to identify factors that most influence a movie's rating, such as director influence, release year, and genre. Clustering allowed us to group similar movies based on user ratings and votes, offering deeper insights into audience preferences. The results of our ML experiment provided a foundation for answering key research questions about directorial success and genre popularity.

Dashboard design concepts:

Our dashboard design was guided by the need to present complex data in an accessible way. Inspired by popular Tableau Public examples, we created dashboards that showcase long-term trends in director ratings, audience votes, and genre popularity. Key visualizations include:

Bubble Charts: Showing the relationship between director ratings and audience votes.

Heatmaps: Comparing directors based on average ratings and total votes.

Line, Bar, and Bubble Charts: Tracking genre popularity over time and showing distribution across different periods.

These visualizations allow users to explore how different directors and genres have performed over the years, highlighting trends that would be difficult to spot in raw data alone.

Call to action: (How does your dashboard and ML answer any research questions?)

Our dashboards and machine learning models offer concrete insights into our guiding research questions by transforming raw data into visually intuitive stories. Through these tools, users can quickly see patterns and correlations that would otherwise remain hidden in the dataset.

- *Research Question 1:* "Which directors have the highest average ratings and most audience votes?"

  This project section delves into the analysis of prominent film directors based on key metrics such as average ratings, total votes, and overall popularity. Using various visualizations in Tableau, the insights reveal patterns in audience perception and engagement with these directors' work.

  Firstly, the Heated Bar Chart (Director vs. Average Rating) provides a clear picture of the directors who have maintained consistently high ratings across their films. Directors clustered toward the top of this chart (when sorted descending) have demonstrated an ability to produce movies that resonate positively with audiences, reflecting the quality and craftsmanship of their work. These high ratings signify a level of consistency in their filmmaking, which often correlates with a director's reputation for delivering films that meet or exceed viewer expectations. Such directors are seen as reliable in terms of critical acclaim and audience satisfaction.

  Moving to a broader audience metric, the Density Map (Director vs. Total Votes) shifts focus from quality to quantity, highlighting which directors have attracted the largest number of votes across their films. This chart is particularly useful in identifying directors with widespread appeal and mass audience engagement. Directors like Christopher Nolan, Steven Spielberg, and Peter Jackson feature prominently in this

analysis, with dense clusters of votes surrounding their names. This suggests that their films not only reach large audiences but also encourage higher participation in terms of audience ratings, likely due to their blockbuster status and widespread recognition in the industry.

To assess overall popularity, the Bubble Chart (Director Rating vs. Number of Votes) offers a combined view by considering both the volume of votes and the average ratings. Larger bubbles in this chart indicate directors who are highly popular, as they are able to garner a high number of votes (the larger the bubble, the more votes received). The color vibrance of the bubbles further illustrates the correlation between a director's popularity and the quality of their films (the more vibrant the color, the higher average rating). Directors like Christopher Nolan, Steven Spielberg, and Peter Jackson once again emerge as dominant figures, with large, vibrant bubbles signifying their strong presence in both areas. This indicates that their films not only draw large audiences but also receive positive feedback, underscoring their sustained popularity and success in the film industry.

In conclusion, the Tableau visualizations paint a clear picture of the most prominent directors in terms of ratings, votes, and overall audience engagement. Christopher Nolan, Steven Spielberg, and Peter Jackson stand out as the most popular directors, consistently delivering films that receive high ratings and attract a large number of votes. Their repeated presence across all visualizations suggests that they have mastered the art of appealing to both critics and the general public, making them some of the most influential and well-regarded directors in modern cinema. This analysis

underscores the importance of not just producing quality films but also maintaining widespread appeal, as reflected in the enduring popularity of these renowned directors.

- *Research Question 2*: "What are the most popular movie genres over time, and how do their ratings compare?"

This analysis delves into the evolving popularity of movie genres over time and explores how their ratings and audience engagement compare. Using three distinct visualizations, we gain insights into the cultural and critical shifts in the film industry.

The first analysis focuses on the average ratings of popular genres from 1997 to 2024, depicted in a line graph. Action and Drama emerge as genres with consistently high ratings, demonstrating sustained appeal. Action, in particular, reached its peak in the late 2000s. On the other hand, genres like Comedy and Horror show greater variability. Comedy enjoys steady audience engagement, though with fluctuating ratings, while Horror struggles with lower ratings despite having a dedicated fanbase. The rise of newer genres like Animation and Fantasy, especially in the past decade, points to changing viewer preferences and the growing popularity of these once niche categories.

In a bar chart, the number of films produced across various genres provides further insight into the industry's output. Action and Drama dominate, with Action having nearly 3,000 films and Drama following closely behind. Comedy and Adventure are also well represented, while Horror and Documentary films make up a smaller portion of the overall movie count. This disparity suggests that certain genres hold greater cultural significance or commercial appeal, leading to a higher production rate. In contrast, the limited number of Horror and Documentary films indicates their niche status, even though they have passionate audiences.

A final visualization, a bubble chart, maps out the total votes each genre has received, with the size of each bubble corresponding to audience popularity. Action clearly stands out as the most voted genre, reflecting its widespread appeal. Drama and Comedy also receive significant engagement, affirming their broad fanbases. In contrast, genres like Horror, Biography, Adventure, Animation, and Crime feature smaller bubbles, signaling fewer votes and a more limited audience reach. This chart highlights how audience preferences influence the popularity and reception of different genres, with Action and Drama solidifying their status as audience favorites.

In conclusion, these findings paint a comprehensive picture of how movie genres perform in terms of ratings, production output, and audience engagement. While Action and Drama consistently lead the pack in all areas, genres like Comedy and Adventure maintain strong positions, with emerging genres like Animation gaining traction. Meanwhile, Horror, despite its loyal following, often lags behind in both ratings and production count. These insights reveal how genre popularity evolves, providing valuable context for understanding shifts in the film industry over time.

Bias/Limitations:

Despite the insights our project provides, it is crucial to address the inherent biases and limitations present in the dataset and analysis, as these factors may shape our conclusions.

- *Time-Period Bias*: Movies released in more recent years are more likely to gather a higher number of votes and ratings, largely due to the widespread availability of online rating platforms like IMDb and Rotten Tomatoes. As a result, newer films may unfairly dominate in comparison, disadvantaging older films with fewer ratings.

- *Cultural Bias:* The dataset overrepresents films and directors from certain regions, particularly from prominent film industries like Hollywood. This imbalance means that movies from lesser-known industries or regions may be underrepresented, skewing conclusions about global film trends and the true diversity of directorial success.
- *Rating Aggregation Bias*: Aggregating ratings across a director's filmography can mask the nuances of their body of work. For example, a director with one or two blockbuster films may appear to be more successful than one with a consistent track record of moderately rated films. This oversimplification could distort our understanding of a director's overall influence.

These biases highlight the need to interpret our results with caution, acknowledging the limitations they impose on our analysis.

Conclusions:

Our project successfully explored the Metacritic movie dataset, offering valuable insights into how directors and genres shape movie reception. We found that directors with higher average ratings tend to have more engaged audiences, and certain genres remain consistently popular over time. However, biases in the dataset—such as the overrepresentation of recent films and specific regions—highlight the need for caution when interpreting results. Future work could address these biases by using weighting techniques or expanding the dataset to include more global film data. Machine learning models could also be enhanced to predict future trends with greater accuracy.

Work Cited

Angela Drucioc, "Time Series Analysis Dashboard." Tableau Public

Jules Claeys, "Vancouver Canucks Season Recap Sports Viz Sunday." Tableau Public

Kashif Sahil. "16000+ Movies 1910-2024 (Metacritic)." Kaggle

Ryan Soares, "Trends in TV Show Genres." Tableau Public