DS 2010
Final Data Project

What makes a Bad Movie?

Team 13

https://www.kaggle.com/datasets/octopusteam/imdb-top-1000-worst-rated-titles

Lucas Sorge
Tia Nagle
Ibrahim Aldulaimi
Aiden Streicher

What makes a Bad Movie?

Dataset Description (https://www.kaggle.com/datasets/octopusteam/imdb-top-1000-worst-rated-titles):

The IMDb Worst Rated Titles Dataset includes movies that received the lowest ratings on IMDb. It contains details such as movie title, genre, average rating, number of votes, and release year, providing a comprehensive view of factors that may correlate with poor audience reception. This dataset enables us to analyze patterns and trends across various attributes associated with low-rated films.

Objective:
To explore the IMDb Worst Rated Titles Dataset and identify key factors that contribute to low ratings.

Key Questions:

- What patterns in genres, release years, runtime, and cast are common in poorly rated movies?
- Can we predict if a movie will receive a low rating based on these attributes?

Expected Findings:

- Certain genres and release periods may be more common among low-rated titles.
- Sentiment analysis could reveal frequent critiques like plot, acting, or production issues.
- Clustering may identify a "typical" profile of low-rated movies.

Benefits:

- Insight for filmmakers on common pitfalls to avoid.
- A unique perspective for audiences on elements linked to a negative viewing experience.
- Experience in machine learning through building a predictive model for low ratings.

Research Questions and Expected Insights

1. **What patterns in genres, release years, and runtime are common in poorly rated movies?**

   We aim to identify recurring trends among the worst-rated movies, such as specific genres or release periods that frequently appear in this list.

2. **Can we predict if a movie will receive a low rating based on its attributes?**

   By analyzing features like genre, release year, and number of votes, we want to develop a predictive model to estimate a movie's likelihood of receiving a poor rating.

3. **Are certain genres more prone to receiving consistently low ratings?**

   We will explore if specific genres (e.g., comedies or documentaries) are overrepresented among low-rated titles, suggesting a trend in audience preferences or expectations

4. **Do low-rated movies tend to have fewer user votes, indicating lower popularity?**

   We want to analyze the relationship between a movie's rating and its number of votes to determine if unpopular movies are more likely to receive lower ratings or if certain low-rated films are still widely viewed despite negative reviews.

Dataset Overview

Dataset Description

We analyzed a combined dataset derived from two IMDb sources: the **IMDb Worst Rated Titles** and **IMDb Top 1000 Movies** datasets. This unified dataset contains **2,000 movies**, offering insights into the factors distinguishing top-rated and poorly-rated films.

Key Features

- **Structure**: The dataset includes the following columns:
  - **id**: Unique identifier for each movie (categorical).
  - **title**: Movie title (categorical).
  - **genres**: List of genres for each movie (categorical).
  - **averageRating**: IMDb user rating on a scale of 1 to 10 (numeric).
  - **numVotes**: Number of user votes (numeric).
  - **releaseYear**: Year of movie release (numeric).
- **Statistics**:
  - **Ratings**: Range from **1.0 to 9.3**, with low-rated movies clustering around **2.0** and high-rated movies around **8.5**.
  - **Votes**: Median vote count for poorly rated movies is approximately **20,000**, compared to **1,000,000** for highly rated ones, reflecting the greater popularity of top-rated films.

Data Collection and Purpose

- The data was sourced from IMDb, focusing on **1,000 lowest-rated** and **1,000 highest-rated** movies.
- The dataset is cleaned and standardized, ready for analysis to identify trends in **ratings**, **genre preferences**, and **temporal patterns**.

Conclusion

This comprehensive dataset provides a robust foundation for comparing the characteristics of top-rated and worst-rated movies, uncovering the factors that contribute to audience reception.

Benefits of the Project

Why it matters:

- The film industry makes 10’s of billions each year
- Over 500 movies were made in 2024 in the US and Canada alone and are viewed by millions
- A successful movie can gross over $1 billion while a flop can lose money

Why the data is useful to a producer or film company:

- We will show trends of which genres and combinations are likely to be poorly received
- We will show trends of which genres and combinations are likely to be positively received
- We will show what trends are more likely to be commercially successful
- We will show how much correlation there is between movie rating and movie popularity
- We will create a model to predict the rating of a movie based on its genre combination
- Overall this data can be used for film producers to make better decisions about what kinds of movies they should and shouldn’t make

Getting the Data

Data Source

The dataset combines two IMDb datasets:

- **IMDb Top 1000 Movies**: A collection of the highest-rated movies, providing insights into the factors that contribute to exceptional audience reception.
- **IMDb Worst Rated Titles**: A list of the lowest-rated movies, capturing the characteristics of poorly received films.

Both datasets include key variables such as:

- **title**, **genres**, **averageRating**, **numVotes**, and **releaseYear**, offering a comprehensive basis for analyzing trends in ratings, genres, and popularity over time.

Data Acquisition

- The data was sourced from **Kaggle** as CSV files for both the **top-rated** and **worst-rated** movies.
- By utilizing pre-existing datasets, we focused on analysis rather than web scraping, ensuring efficient use of resources and time.

Potential for Further Enrichment

- While the current dataset is robust, integrating additional variables (e.g., cast details or budget data) could provide deeper insights into the factors influencing movie success or failure.

Preparing the Data

Dataset Overview

We combined two datasets, **IMDb Top 1000 Movies** and **IMDb Worst Rated Titles**, into a single dataset containing 2,000 entries. The dataset includes essential columns such as:

- **id**, **title**, **genres**, **averageRating**, **numVotes**, and **releaseYear**, offering comprehensive information for our analysis.

Data Cleaning Steps

- **Duplicate Removal**:
  - We checked for duplicate rows to ensure accuracy in our analysis. No duplicate entries were found in the combined dataset, so no further action was required.
- **Handling Missing Values**:
  - **Genres**: Missing values in the genres column were filled with the placeholder **"Unknown"**, ensuring no gaps during genre-based analysis.
  - **numVotes**: Missing values were replaced with the **median**, minimizing the impact of outliers while preserving the dataset's integrity.
- **Standardizing Data Types**:
  - The releaseYear column was converted to **integer**, and averageRating to **float** for consistency, allowing precise statistical and numerical operations.

Preparing the Data

- **Outlier Detection**:
  - We examined:
    - **releaseYear** for any unusual values (e.g., years beyond 2025 or before 1900).
    - **averageRating** for values outside the 1-10 range.
  - No significant outliers were identified, ensuring the data's reliability.
- **Verification of Changes**:
  - After cleaning, we verified the dataset to confirm that all columns had consistent data types and no anomalies.

Final Cleaned Dataset

- The dataset is now free of duplicates, missing values, and outliers, with standardized data types. It is well-prepared for deeper exploratory data analysis, including trends in ratings, genres, and popularity over time.

Descriptive Analysis of the Dataset

We conducted a thorough analysis of both the **IMDb Worst Rated Titles** and **IMDb Top 1000 Movies** datasets to understand their structure and key statistics.

Dataset Structure:

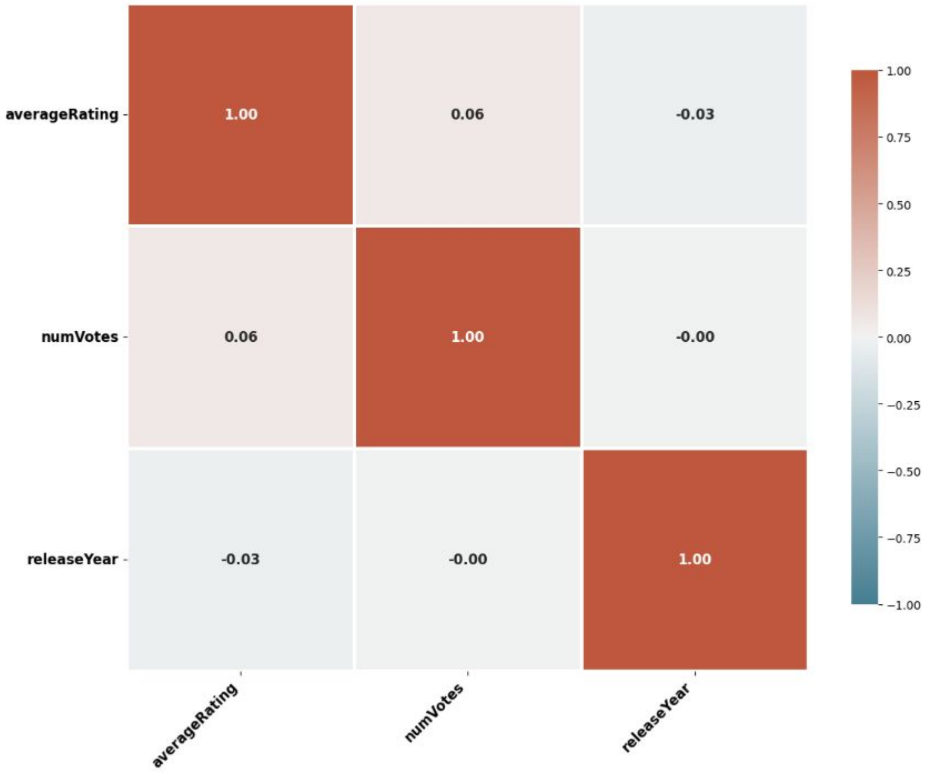
- Both datasets contain the following key columns: **id**, **title**, **genres**, **averageRating**, **numVotes**, and **releaseYear**.
- The **IMDb Worst Rated Titles** dataset includes information on 1,000 movies with the lowest ratings on IMDb, while the **IMDb Top 1000 Movies** dataset focuses on the highest-rated titles.

Basic Statistics:

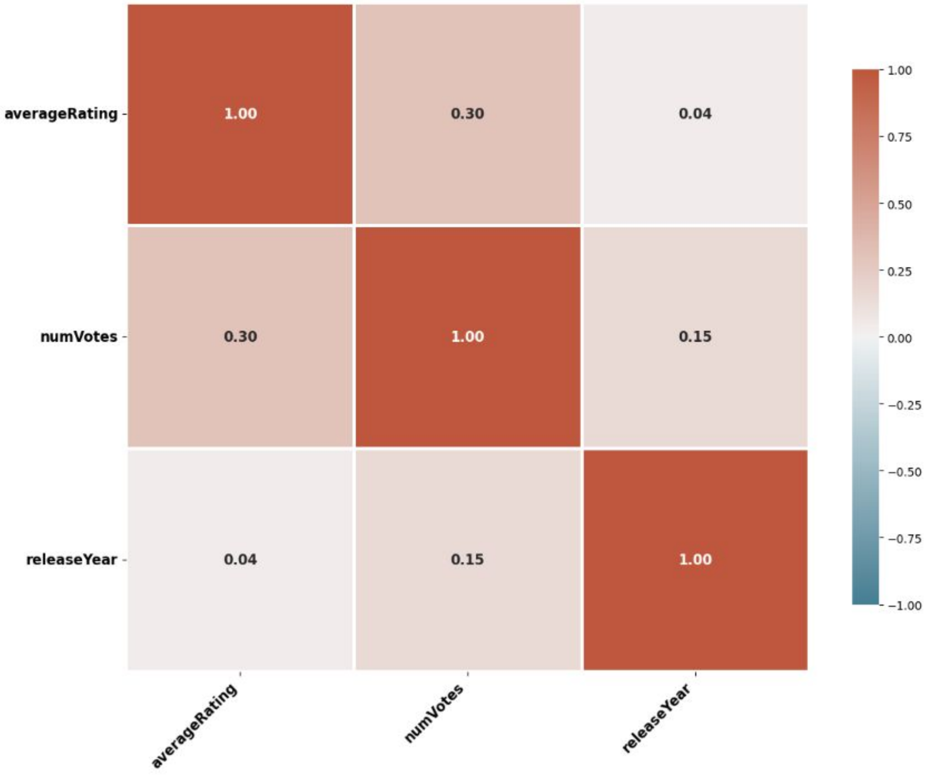
- The **averageRating** for the worst-rated titles ranges from **1.0** to **3.5**, with a median rating of around **2.0**, indicating a concentration of very low ratings.
- For the top-rated movies, the **averageRating** spans from **7.8** to **9.3**, with a median rating of **8.5**, highlighting consistently high audience scores.
- The number of votes (numVotes) varies significantly between datasets. The worst-rated movies have fewer votes on average (median: ~20,000), while top-rated movies tend to have a higher vote count (median: ~1,000,000), reflecting their popularity.

Conclusion: The cleaned and standardized datasets are well-prepared for deeper analysis, enabling us to explore trends in ratings, genre preferences, and temporal patterns effectively. The differences in rating distributions and vote counts between the two datasets provide a solid foundation for comparing the factors that distinguish top-rated and poorly-rated movies.

Visual Analysis - Correlation Heatmap



Correlation Heatmap for IMDb Worst Rated Titles



Correlation Heatmap for IMDb Top 1000 Movies

Conclusion - Correlation Heatmap

The heatmaps display the correlation between three main numeric features: averageRating, numVotes, and releaseYear. The values range from -1 to 1, where:

- **1.0** indicates a perfect positive correlation.
- **0.0** indicates no correlation.
- **-1.0** indicates a perfect negative correlation.

Moderate Positive Correlation in Top Movies:

- In the **IMDb Top 1000 Movies** (Figure 2), there is a **moderate positive correlation (0.30)** between averageRating and numVotes. This suggests that highly-rated movies tend to receive more votes, indicating a relationship between a film's popularity and its rating.

Weak Correlations in Worst Rated Titles:

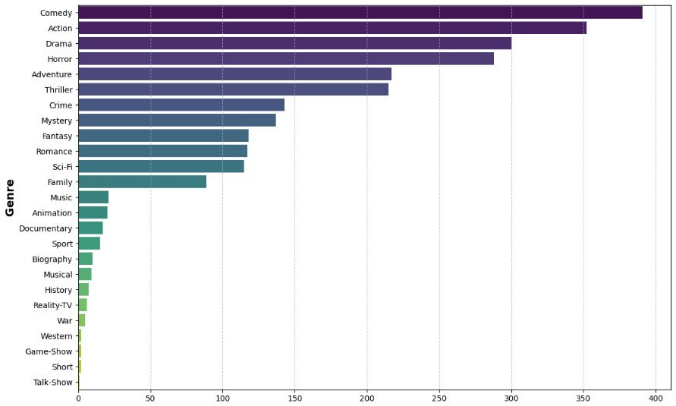
- In the **IMDb Worst Rated Titles** (Figure 1), the correlation between averageRating and numVotes is much weaker (0.06), implying that even poorly rated movies can still attract a significant number of votes, possibly due to curiosity or infamy rather than quality.

Impact of Release Year:

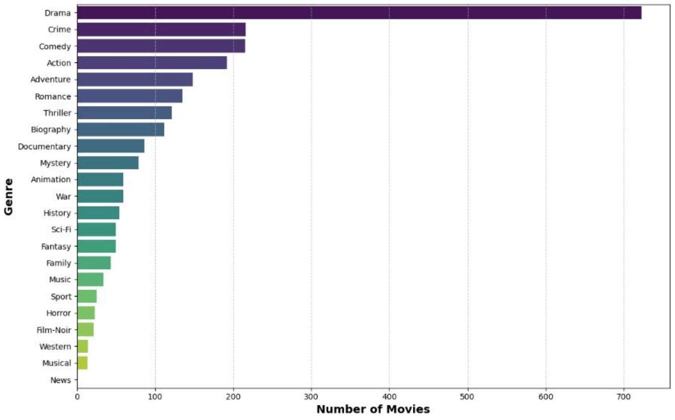
- For both datasets, the correlation between releaseYear and the other features (averageRating and numVotes) is close to **zero**, indicating that the year of release does not significantly impact the movie ratings or the number of votes received.

Conclusion: The heatmap analysis highlights a key difference between top-rated and poorly rated movies: popularity and high ratings are more closely linked in successful films, while poorly rated movies do not exhibit this trend. This suggests that while good movies often gain popularity, the reasons for attention on bad movies might be different, such as controversy or notoriety.

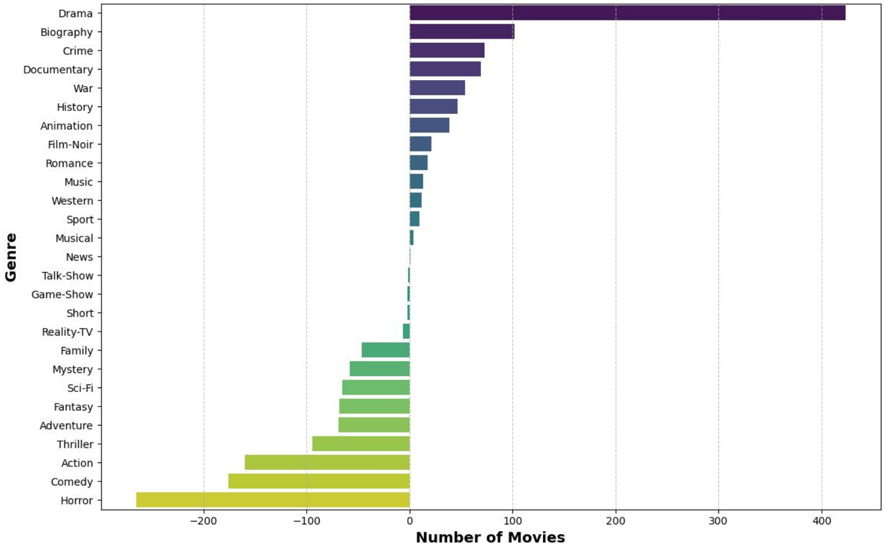
Visual Analysis - Genre Distribution Analysis



Genre Distribution for IMDb Worst Rated Titles



Genre Distribution for IMDb Top 1000 Movies



Genre difference between high and low rated movies

Conclusion - Genre Distribution Analysis

Genre Distribution Bar Chart

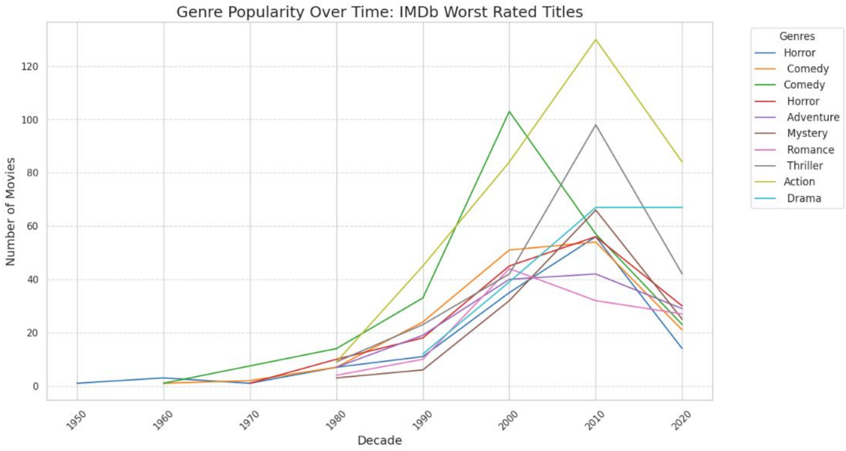
The Genre Distribution Bar Chart visualizes the frequency of movie genres across the IMDb Top 1000 Movies and Worst Rated Titles. Each bar represents a genre, with its length indicating the number of movies belonging to that genre.

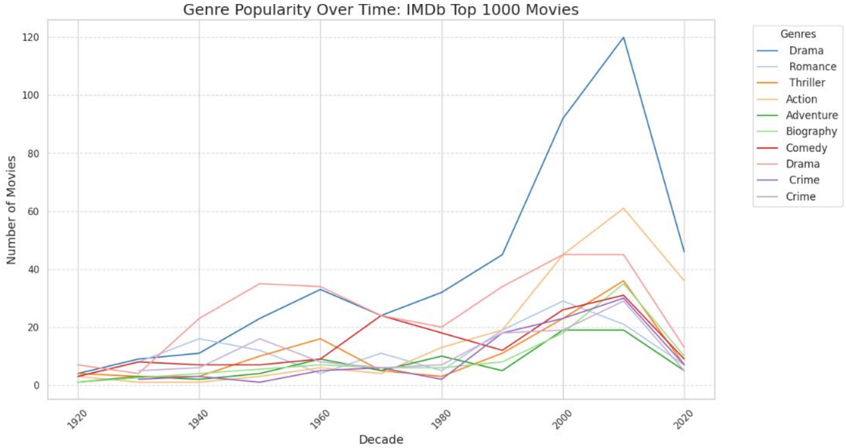
Key Insights

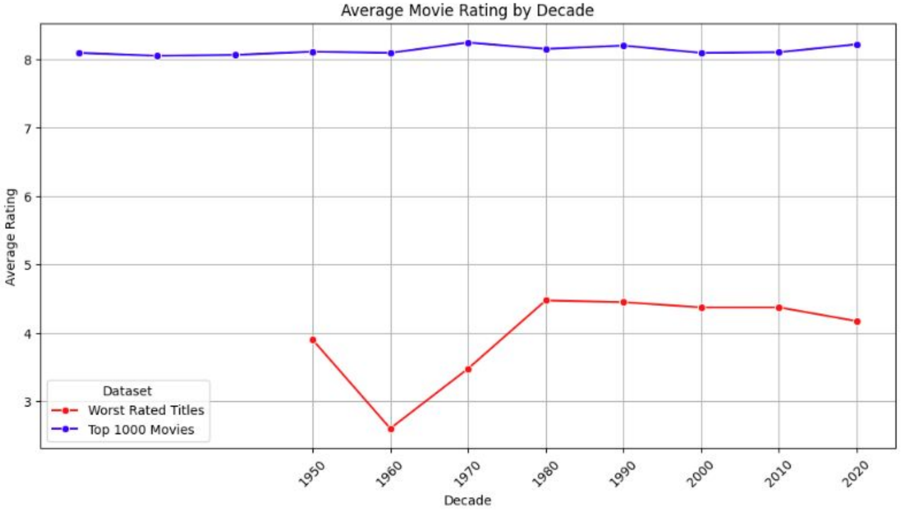
- **Worst Rated Movies**:
  - Dominated by **Comedy**, **Action**, and **Drama**.
  - **Horror** frequently appears, reflecting its challenge in achieving high audience reception.
- **Top Rated Movies**:
  - Led by **Drama**, **Crime**, and **Comedy**.
  - **Drama** is particularly successful, appearing significantly more in high-rated lists.
- **Comparing Trends**:
  - **Comedy** is polarizing, common in both lists but with high variability.
  - Genres like **Horror** are skewed toward poor ratings, while **Biography** and **Documentary** favor high ratings.

Conclusion: Genre plays a critical role in movie success. While **Drama** shows strong positive trends, genres like **Comedy** and **Horror** pose higher risks of audience dissatisfaction.

Visual Analysis - Genre Popularity Over Time


Genre Popularity Over Time: IMDb Worst Rated Titles


Genre Popularity Over Time: IMDb Top 1000 Movies


Average Movie Rating by Decade

Conclusion - Genre Popularity Over Time

Genre Trends by Decade

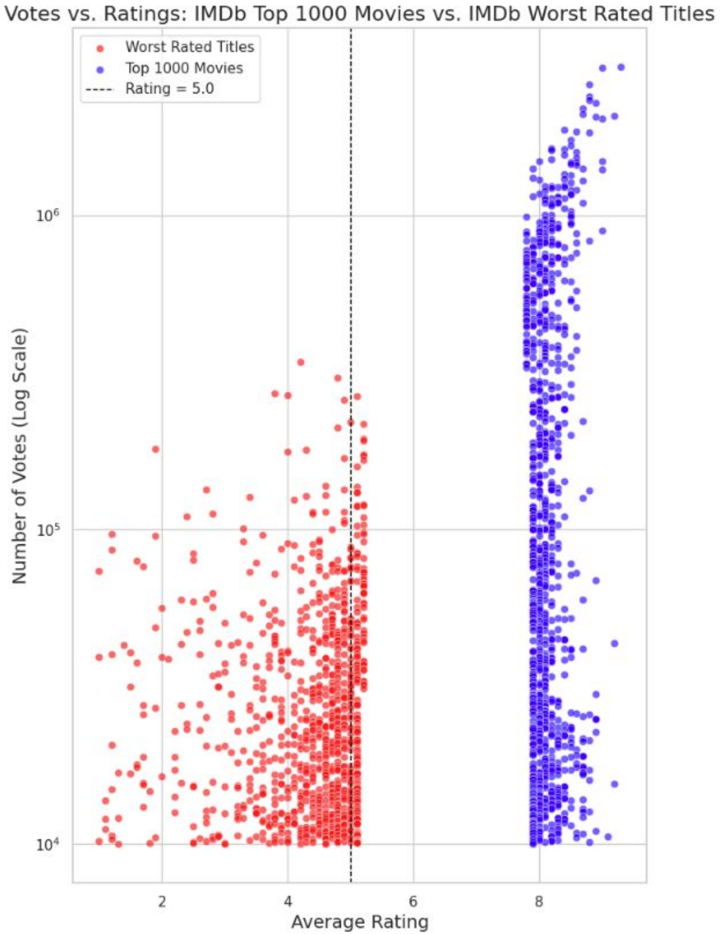
- **IMDb Worst Rated Titles**:
  - **Comedy**, **Action**, and **Drama** dominate poorly rated movies, with a significant rise from the 1990s onward, reflecting high production volumes and inconsistent quality.
  - **Horror** shows volatility, with spikes in the 2000s, while genres like **Romance** and **Mystery** appear less frequently, maintaining steadier quality.
- **IMDb Top 1000 Movies**:
  - **Drama** is the most dominant genre, consistently popular after the 1980s, due to its storytelling depth and emotional appeal.
  - **Crime** and **Romance** also appear frequently, highlighting their strong critical reception.

Average Ratings Trends Across Decades

- **Top Rated Movies**:
  - Ratings remain stable around 8.0 across decades.
- **Worst Rated Titles**:
  - Fluctuating ratings with sharp dips, notably below 3.0 in the 1960s, reflecting an era of experimental or poorly executed films.
  - A recent decline in ratings (2010s-2020s) suggests an increase in low-budget or poorly received direct-to-streaming films.

Conclusion: Drama consistently excels across decades, while Comedy and Action are more polarizing, often appearing in low-rated lists when poorly executed. High-rated movies maintain stable quality, whereas low-rated titles show greater volatility, reflecting shifts in industry trends and audience expectations.

Visual Analysis - Scatter Plot



Conclusion - Scatter Plot

The **Votes vs. Ratings Scatter Plot** visualizes the relationship between the number of votes (a measure of popularity) and the average rating (a measure of quality) for both datasets:

- **IMDb Top 1000 Movies** (blue) tend to have higher ratings and a wider range of vote counts, with many movies receiving millions of votes. The higher density of points around **8.0 to 9.0** ratings shows consistent audience approval, even as vote counts increase.

- **IMDb Worst Rated Titles** (red) cluster heavily below a **rating of 5.0**, regardless of vote count. This suggests that even with high popularity (high vote count), these movies fail to achieve positive ratings, highlighting consistently negative reception.

- The **log scale** for the vote count helps visualize the wide range of popularity in both datasets, from niche films with few votes to highly popular titles with millions of votes.

Conclusion - Scatter Plot

- **Audience Engagement**: Highly rated movies attract more viewers due to positive reviews, awards, and strong recommendations. This increases their visibility and leads to a larger pool of ratings.

- **Selection Bias**: Viewers are more likely to watch movies they expect to enjoy. Poorly rated movies tend to be avoided based on their negative reputation, resulting in fewer ratings overall.

Conclusion: Popularity (measured by vote count) does not guarantee high ratings. Even movies with a significant number of votes in the Worst Rated Titles dataset struggle with poor ratings, while movies in the Top 1000 dataset consistently maintain high ratings despite varying popularity levels. This analysis suggests a need to further investigate factors influencing audience perception beyond just the vote count.

Hypothesis Development and Analysis

Hypothesis 1: Positive Audience Perception Correlates with Movie Popularity

- **Hypothesis**: High ratings correlate with higher vote counts in top-rated movies, suggesting that positive audience perception drives popularity.
- **Testing Plan**:
  - Use **correlation analysis** to measure the strength of the relationship between averageRating and numVotes for top-rated movies.
  - Build a **regression model** to predict vote counts based on ratings and test its predictive power.
  - Evaluate how well positive ratings explain audience engagement trends through the model's performance.

Hypothesis 2: Genre Selection Significantly Impacts Audience Reception

- **Hypothesis**: Certain genres (e.g., Comedy and Horror) show polarized audience reception, with high variance in ratings depending on execution quality.
- **Testing Plan**:
  - Perform a **genre-based statistical analysis** to compare the rating distributions for each genre.
  - Use a **classification model** to predict whether a movie is top- or worst-rated based on its genre and evaluate the importance of genre features in the model.
  - Explore how genre contributes to audience polarization by analyzing variance in ratings within specific genres.

Issues Found and Improvement

Potential Bias in Data Collection:

- **Problem**: The dataset was compiled based on user ratings from IMDb, which might be influenced by selection bias. Users who choose to rate movies may not represent the general audience, leading to skewed ratings (e.g., fans rating their favorite genres or movies disproportionately high).

- **Recommendation**: Future datasets should aim to mitigate selection bias by integrating data from multiple sources, such as Rotten Tomatoes or Metacritic, for a more balanced perspective.

Stage 4: Model the Data

(Describe your plan)

Setting up the Model

Deciding which model to use:

- We decided to use a regression model. This is because our target variable ‘averageRating’ while discrete still works well with regression models.

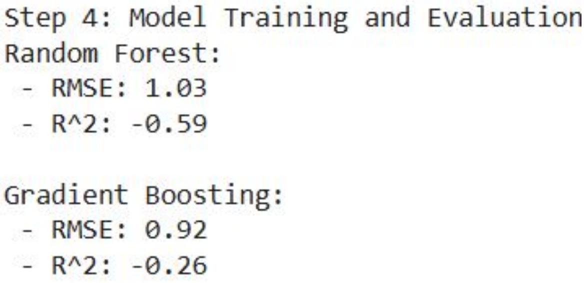
Setting up the model:

- We had a sample size of 1000, so we used 800 movies for training and 200 for testing
- Hot one encoding was used to ensure consistency between training and testing sets.
- There were two regression models we decided to use. A random forest regression model and a gradient boosting regression model.

Evaluating the models:

- The two things we decided to evaluate our model with was
- RMSE and R^2

- As you can see neither model performed great in their R^2 score
- but the gradient boosting model performed notably better



```
Step 4: Model Training and Evaluation
Random Forest:
 - RMSE: 1.03
 - R^2: -0.59

Gradient Boosting:
 - RMSE: 0.92
 - R^2: -0.26
```

Stage 5: Communicate the Data

(Describe your plan)