Data Analysis Using Rotten Tomatoes Website

Shihao Zhang, Zeqi Chen, Yifan He, Songchen Wu, Chongrui Sun

Dec 5, 2024

Introduction

Rotten Tomatoes is a website that aggregates movie reviews from professional critics and general audiences. "Tomatometer" score represents the percentage of professional critics who gave a movie a positive review."Popcornmeter" score represents the general audience's feedback based on user reviews. Other than that, the website also provides comprehensive information on Genre (e.g., Action, Comedy, Horror), Rating(e.g., G, PG, PG-13, R) , Release Date,etc. Based on these data, we will conduct our project revolving around the following 5 subject matters.

Box Office Analysis for Weekends from June-Dec 2024

In this analysis, we explored weekly box office data from June to December in 2024 to uncover trends, understand the impact of genres, and evaluate the volatility of movie performances over time. This comprehensive approach involved web scraping, data processing, and visualization techniques to generate actionable insights.

Firstly, we began with extracting weekly box office data from a collection of URLs. Using web scraping techniques, we collected information about the movie title, weekly gross and total gross of box office. Then, we structured the data into a clean dataset for further analysis. (Graph 1)

To highlight standout performances, we identified the highest-grossing movie for each week. By cleaning the gross values, we could reliably compare weekly earnings and pinpoint the top movie for every week in the dataset.

On top of this, we then retrieved genre data for the weekly top-grossing movies from the mainpages of them, to better understand what types of movies resonate most with audiences.(Table 3)

To delve deeper into the role of genres, we analyzed their frequencies among the weekly highest-grossing movies. By splitting the genres into individual categories and counting their occurrences, we observed that action, adventure, and comedy were among the most popular genres. Action movies often feature high-budget spectacles, stunts, and special effects, making them attractive to many audiences. Adventure movies often feature expansive world-building, captivating visuals, and heroic narratives that appeal to broad demographics, including families and younger audiences. Comedy provides escapism and lighthearted entertainment, making it a reliable choice for casual moviegoers.

For each movie, we calculated the percentage change in weekly gross earnings, enabling us to assess trends and patterns in performance. By focusing on movies with multiple weeks of data, we uncovered significant fluctuations in box office of some movies, emphasizing the challenges movies face in

maintaining their momentum after a strong opening week. For example, "The Strangers: Chapter 1" stands out with an extreme peak followed by a sharp decline. The spike likely represents a strong opening week, followed by a rapid drop in audience interest or limited theatrical showings in subsequent weeks. However, during the holiday season (November and December), films like "Tim Burton's The Nightmare Before Christmas" and "The Wild Robot" show a relative stability, which is probably because holiday boosts family-friendly movies and blockbusters.

One of the most intriguing aspects of our analysis was the evaluation of volatility in weekly gross performance. By calculating the standard deviation of percentage changes for each movie, we identified those with the most unpredictable earnings. The movie "Smile 2" stood out for relatively high volatility. The reason behind this is probably that as a horror movie, it might have had a strong opening due to loyal genre audiences but struggled to maintain momentum in the following weeks. Also, as a sequel, it may have relied on the goodwill of its predecessor. If the quality of the sequel was inconsistent, word-of-mouth might have caused significant volatility.

Ratings Analysis of Best Movies in Theaters (2024)

This report analyzes the relationship between audience ratings (Popcornmeter) and critics' ratings (Tomatometer) using a scatter plot, hexbin plot, and correlation matrix, each providing unique insights.

The scatter plot (Graph 3) shows a moderately positive correlation between audience and critics' ratings, especially for scores above 80. However, discrepancies arise: commercially-driven films often score higher with audiences but lower with critics, while arthouse films receive critical acclaim but less audience resonance. These differences highlight varied evaluation criteria.

The hexbin plot (Graph 4) reveals a concentration of ratings between 60 and 80, reflecting a balance of appeal to both groups. Low-scoring films (0–40) show consistent dissatisfaction. Outliers, where ratings diverge, offer insights into factors driving polarization.

The correlation matrix (Graph 5) quantifies this relationship, with a correlation coefficient of 0.65. While critics and audiences generally align, their differing priorities—emotional engagement for audiences vs. professional metrics for critics—account for notable variations.

Overall, most films succeed in the 60–80 range, balancing critical acclaim and audience appeal. Outliers underscore the need to consider both perspectives to fully understand film reception.

Genre Analysis of Best Certified Fresh Movies in Theaters

In total, there are 21 different movie genres (shown in Table 5). This variety of genres enables us to explore patterns and trends within each genre more effectively.

The first aspect we focus on is identifying the top genres based on Tomatometer Ratings and Audience Ratings. By calculating the average ratings within each genre, we highlight the top 10 genres for each

rating category (Graph 6). They are Anime & Manga, Sports & Fitness, Documentary, Special Interest, Classics, Musical & Performing Arts, Faith & Spirituality, Television, Art House & International, and Animation.

The graph also shows that the top 10 genres identified by both rating methods are identical. This indicates a strong alignment between critic and audience opinions regarding the highest-rated movies.

Similarly, we select the bottom-ranked genres by Tomatometer Ratings and Audience Ratings (Graph 7). They are Horror, Mystery & Suspense, Cult Movies, Science Fiction & Fantasy, Action & Adventure, Comedy, Kids & Family, Gay & Lesbian, Drama, and Romance.

An interesting phenomenon is that, compared to the graph of the genres with the highest weekly grossing, they show some similarities. This shows that the genres with the highest grossing are actually among those with the lowest ratings. A possible explanation is that those movies focus more on their commercial impacts, which might lead to bad storytelling, which in turn lowers their ratings.

Furthermore, we look into the content ratings within each genre (Graph 8). There are 5 kinds of content rating: G (General Audiences), PG (Parental Guidance Suggested), PG-13 (Parents Strongly Cautioned), R (Restricted), and NR (Not Rated). From the graph, we can see that more mature movie topics like horror movies and Gay & Lesbian movies concentrated more on R rating, while more general topics like Animation movies are generally distributed on G and PG. By looking into these distributions, we can gain more insight into the targeted audience of each kind of movie.

In conclusion, each genre of movie has its unique content distributions, and their genres can also have a significant impact on the overall ratings of the movies.

Sentiment Analysis of Movies Review

Since numbers are easily accessible and straightforward, most movie reports rely on numerical ratings to convey a film's popularity. However, reviews on movie websites often provide deeper insights into the true sentiments of the audience. After all, scoring standards vary from person to person, but the emotions expressed through language are more genuine. Therefore, we conducted a sentiment analysis of reviews collected from movie websites to better understand audience perceptions. The two word clouds in Graph 9 below reflect the most common positive and negative words in all reviews.

We ranked movies within each genre based on their favorable review ratios, calculated as the number of positive reviews divided by the total number of reviews for each movie. The favorable review rate for each genre was then determined by averaging the favorable review ratios of all movies within that genre. The results are summarized in Graph 10 (showing only the top 10 genres with the highest favorable reviews). Compared to the rankings in the previous section, we observe significant differences in both critic reviews and audience reviews, suggesting that numerical ratings alone cannot fully capture audience evaluations.

We further examined the impact of time on these rankings by analyzing reviews from the earliest 5%, 10%, and 20% time periods for each movie. The results showed significant differences compared to those based on all reviews, highlighting the weak correlation between early word-of-mouth and a movie's overall reputation. As for the two outliers observed at the 100% mark in Graph 11, we believe this discrepancy is likely due to an insufficient sample size.(Only 1 “Biography” movie and 2 “War” movies are included in our dataset.)

Machine Learning

This study presents a comprehensive process for predicting audience ratings, including data preparation, feature engineering, model selection, and analysis. In the data preparation phase, irrelevant features like Director and Writer were removed, missing values excluded, and genres simplified to retain only primary genres, dropping rare ones. Content ratings were filtered to include common categories (G, PG, PG-13, R, NR). The dataset was split into training (66%) and testing (34%) sets, with categorical variables one-hot encoded and the target variable label-encoded into two classes: 0 (Spilled) and 1 (Upright). Testing set features were aligned with the training set to ensure consistency.

Feature engineering refined the dataset further. Runtime was discretized into five categories, and new features were derived from the release date, such as decade, month, and day bins. The Genre and Content Rating columns were simplified to major categories, and the final dataset included numerical features (e.g., Runtimebin) and categorical features (e.g., Genre). The target variable was encoded as an ordered factor.

For model selection and training, three models—Support Vector Machines (SVM), k-Nearest Neighbors (kNN), and Random Forest—were trained with 5-fold cross-validation. Hyperparameters like SVM’s penalty parameter (C), kNN’s neighbor count, and Random Forest’s features per split were optimized. Accuracy was the primary evaluation metric on both cross-validation and test sets.

In the results, Random Forest emerged as the best-performing model with high accuracy, precision, and recall for both classes. It achieved the highest AUC (0.71) compared to SVM (0.64) and kNN (0.56). SVM was biased toward the "Upright" class, and kNN lacked discriminatory power. Confusion matrices and ROC curves supported Random Forest’s superior performance.

In conclusion, Random Forest was the most effective model, though the study faced limitations such as moderate overall accuracy and class imbalance. Future work should explore advanced models like XGBoost, feature selection, and techniques like oversampling or weighted loss functions to enhance performance.

Appendix

	Date	Movie	Weekly Gross	Total Gross
1	November 24, 2024	Wicked	$114 million	$114 million
2	November 24, 2024	Gladiator II	$55.5 million	$55.5 million
3	November 24, 2024	Red One	$13.2 million	$52.9 million
4	November 24, 2024	Bonhoeffer: Pastor. Spy. Assassin.	$5.1 million	$5.1 million
5	November 24, 2024	Venom: The Last Dance	$4 million	$133.8 million
6	November 24, 2024	The Best Christmas Pageant Ever	$3.5 million	$25.5 million
7	November 24, 2024	Heretic	$2.23 million	$24.7 million
8	November 24, 2024	The Wild Robot	$2 million	$140.7 million
9	November 24, 2024	Smile 2	$1.11 million	$67.7 million
10	November 24, 2024	A Real Pain	$1.109 million	$4.96 million
11	November 18, 2024	Red One	$32.1 million	$32.1 million
12	November 18, 2024	Venom: The Last Dance	$7.3 million	$127.5 million
13	November 18, 2024	The Best Christmas Pageant Ever	$5.2 million	$19.8 million
14	November 18, 2024	Heretic	$5.1 million	$20.4 million
15	November 18, 2024	The Wild Robot	$4.2 million	$137.6 million
16	November 18, 2024	Smile 2	$2.9 million	$65.6 million
17	November 18, 2024	Conclave	$2.8 million	$26.5 million
18	November 18, 2024	Hello, Love, Again	$2.32 million	$2.32 million
19	November 18, 2024	A Real Pain	$2.22 million	$2.96 million
20	November 18, 2024	Anora	$1.76 million	$10.42 million

	Date	Movie	Weekly Gross	Total Gross
215	June 24, 2024	Thelma	$2.2 million	$2.2 million
216	June 24, 2024	The Watchers	$1.9 million	$17.7 million
217	June 24, 2024	Rite Here Rite Now	$1.5 million	$1.5 million
218	June 16, 2024	Inside Out 2	$155 million	$155 million
219	June 16, 2024	Bad Boys: Ride or Die	$33 million	$112.2 million
220	June 16, 2024	Kingdom of the Planet of the Apes	$5.2 million	$157.8 million
221	June 16, 2024	The Garfield Movie	$5 million	$78.5 million
222	June 16, 2024	IF	$3.4 million	$100.9 million
223	June 16, 2024	The Watchers	$3.6 million	$13.6 million
224	June 16, 2024	Furiosa: A Mad Max Saga	$2.4 million	$63.1 million
225	June 16, 2024	The Fall Guy	$1.5 million	$87.9 million
226	June 16, 2024	The Strangers: Chapter 1	$760,000	$33.9 million
227	June 16, 2024	The Lord of the Rings: The Fellowship of the Ring	$633,000	$3.09 million
228	June 10, 2024	Bad Boys: Ride or Die	$56.5 million	$56.5 million
229	June 10, 2024	The Garfield Movie	$10 million	$68.6 million
230	June 10, 2024	IF	$7.8 million	$93.3 million
231	June 10, 2024	The Watchers	$7 million	$7 million
232	June 10, 2024	Kingdom of the Planet of the Apes	$5.4 million	$149.7 million
233	June 10, 2024	Furiosa: A Mad Max Saga	$4.2 million	$58.6 million
234	June 10, 2024	The Fall Guy	$2.5 million	$85 million
235	June 10, 2024	The Lord of the Rings: The Fellowship of the Ring	$2.4 million	$2.4 million
236	June 10, 2024	The Lord of the Rings: The Two Towers	$1.9 million	$1.9 million
237	June 10, 2024	The Strangers: Chapter 1	$1.8 million	$32.1 million

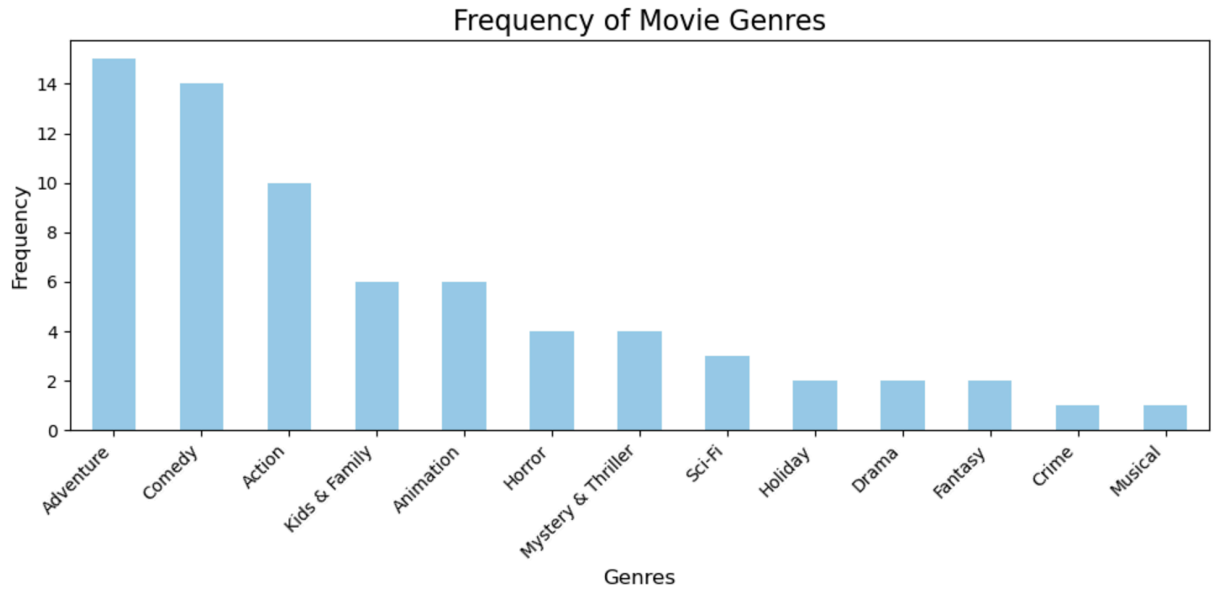
Table 1: Weekly and Total Data of Movies on Weekend Box Office List from June to Dec (2024)

	Date	Movie	Weekly Gross	Total Gross	Weekly Gross (Numeric)
137	August 12, 2024	Deadpool & Wolverine	$54.1 million	$494.3 million	5.41e+07
127	August 18, 2024	Alien: Romulus	$41.5 million	$41.5 million	4.15e+07
118	August 26, 2024	Deadpool & Wolverine	$18.3 million	$577.2 million	1.83e+07
147	August 5, 2024	Deadpool & Wolverine	$97 million	$395.5 million	9.7e+07
197	July 1, 2024	Inside Out 2	$57.4 million	$469.3 million	5.74e+07
177	July 15, 2024	Despicable Me 4	$44.6 million	$211.1 million	4.46e+07
167	July 21, 2024	Twisters	$80.5 million	$80.5 million	8.05e+07
157	July 28, 2024	Deadpool & Wolverine	$205 million	$205 million	2.05e+08
187	July 7, 2024	Despicable Me 4	$75 million	$122.6 million	7.5e+07
227	June 10, 2024	Bad Boys: Ride or Die	$56.5 million	$56.5 million	5.65e+07
217	June 16, 2024	Inside Out 2	$155 million	$155 million	1.55e+08
207	June 24, 2024	Inside Out 2	$100 million	$355.1 million	1e+08
20	November 11, 2024	Venom: The Last Dance	$16.2 million	$114.8 million	1.62e+07
10	November 18, 2024	Red One	$32.1 million	$32.1 million	3.21e+07
0	November 24, 2024	Wicked	$114 million	$114 million	1.14e+08
30	November 4, 2024	Venom: The Last Dance	$26.1 million	$90 million	2.61e+07
60	October 14, 2024	Terrifier 3	$18.8 million	$18.8 million	1.88e+07
50	October 21, 2024	Smile 2	$23 million	$23 million	2.3e+07
41	October 28, 2024	Smile 2	$9,5 million	$40.8 million	9.5e+07
70	October 7, 2024	Joker: Folie à Deux	$37.5 million	$37.5 million	3.75e+07
100	September 16, 2024	Beetlejuice Beetlejuice	$51.6 million	$188 million	5.16e+07
90	September 23, 2024	Beetlejuice Beetlejuice	$26 million	$226.8 million	2.6e+07
80	September 30, 2024	The Wild Robot	$35.7 million	$35.7 million	3.57e+07
110	September 9, 2024	Deadpool & Wolverine	$7.2 million	$614 million	7.2e+06

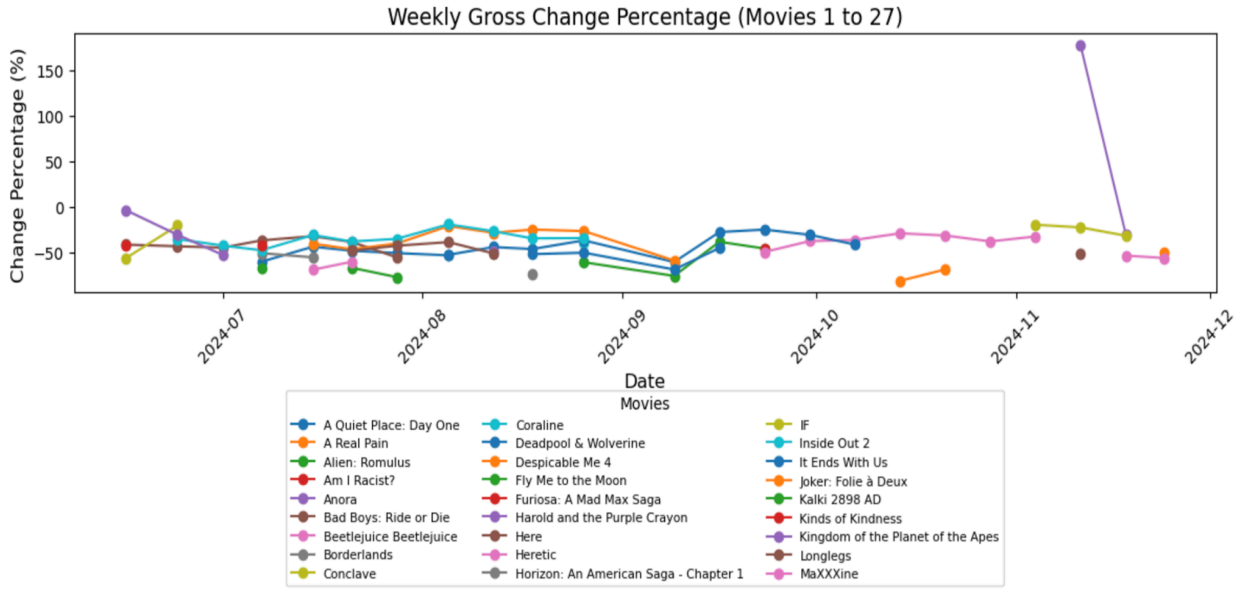
Table 2: Movies with Highest Box Office Each Week

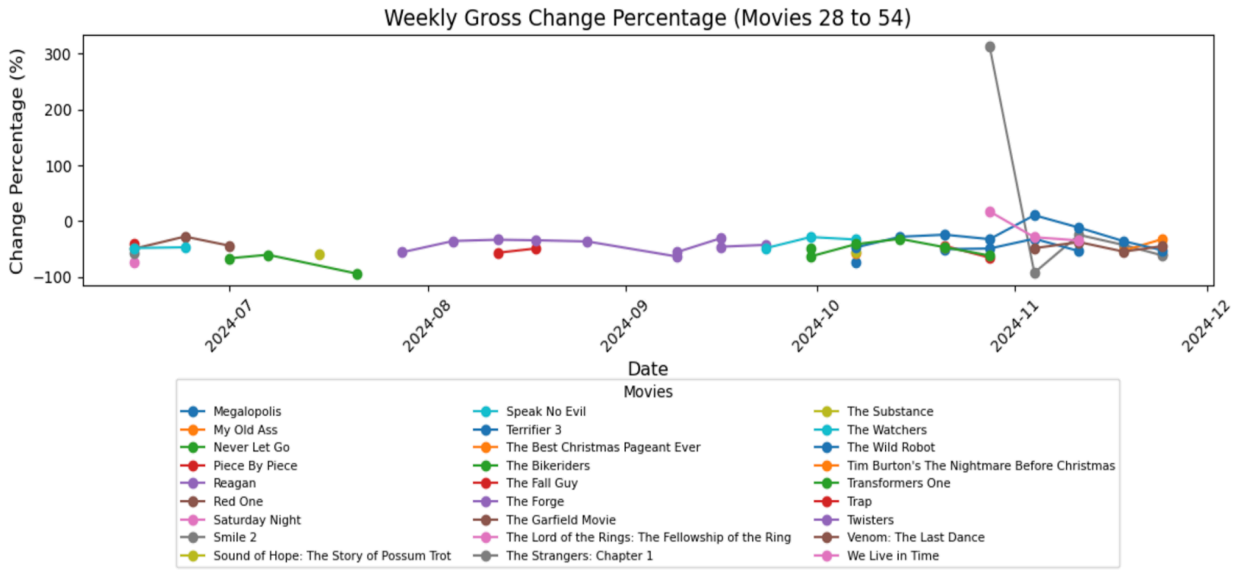
	Date	Movie	Weekly Gross	Total Gross	Genres
0	August 12, 2024	Deadpool & Wolverine	$54.1 million	$494.3 million	Action, Adventure, Comedy
1	August 18, 2024	Alien: Romulus	$41.5 million	$41.5 million	Sci-Fi, Horror
2	August 26, 2024	Deadpool & Wolverine	$18.3 million	$577.2 million	Action, Adventure, Comedy
3	August 5, 2024	Deadpool & Wolverine	$97 million	$395.5 million	Action, Adventure, Comedy
4	July 1, 2024	Inside Out 2	$57.4 million	$469.3 million	Kids & Family, Comedy, Adventure, Animation
5	July 15, 2024	Despicable Me 4	$44.6 million	$211.1 million	Kids & Family, Comedy, Adventure, Animation
6	July 21, 2024	Twisters	$80.5 million	$80.5 million	Action, Adventure, Mystery & Thriller
7	July 28, 2024	Deadpool & Wolverine	$205 million	$205 million	Action, Adventure, Comedy
8	July 7, 2024	Despicable Me 4	$75 million	$122.6 million	Kids & Family, Comedy, Adventure, Animation
9	June 10, 2024	Bad Boys: Ride or Die	$56.5 million	$56.5 million	Action, Comedy
10	June 16, 2024	Inside Out 2	$155 million	$155 million	Kids & Family, Comedy, Adventure, Animation
11	June 24, 2024	Inside Out 2	$100 million	$355.1 million	Kids & Family, Comedy, Adventure, Animation
12	November 11, 2024	Venom: The Last Dance	$16.2 million	$114.8 million	Action, Adventure, Sci-Fi
13	November 18, 2024	Red One	$32.1 million	$32.1 million	Holiday, Action, Adventure, Comedy
14	November 24, 2024	Wicked	$114 million	$114 million	Drama
15	November 4, 2024	Venom: The Last Dance	$26.1 million	$90 million	Action, Adventure, Sci-Fi
16	October 14, 2024	Terrifier 3	$18.8 million	$18.8 million	Holiday, Horror, Mystery & Thriller
17	October 21, 2024	Smile 2	$23 million	$23 million	Horror, Mystery & Thriller
18	October 28, 2024	Smile 2	$9,5 million	$40.8 million	Horror, Mystery & Thriller
19	October 7, 2024	Joker: Folie à Deux	$37.5 million	$37.5 million	Crime, Drama, Musical
20	September 16, 2024	Beetlejuice Beetlejuice	$51.6 million	$188 million	Comedy, Fantasy
21	September 23, 2024	Beetlejuice Beetlejuice	$26 million	$226.8 million	Comedy, Fantasy
22	September 30, 2024	The Wild Robot	$35.7 million	$35.7 million	Kids & Family, Adventure, Animation
23	September 9, 2024	Deadpool & Wolverine	$7.2 million	$614 million	Action, Adventure, Comedy

Table 3: Movies with Highest Box Office Each Week Attached with Genres



Graph 1: Frequency of Movie Genres



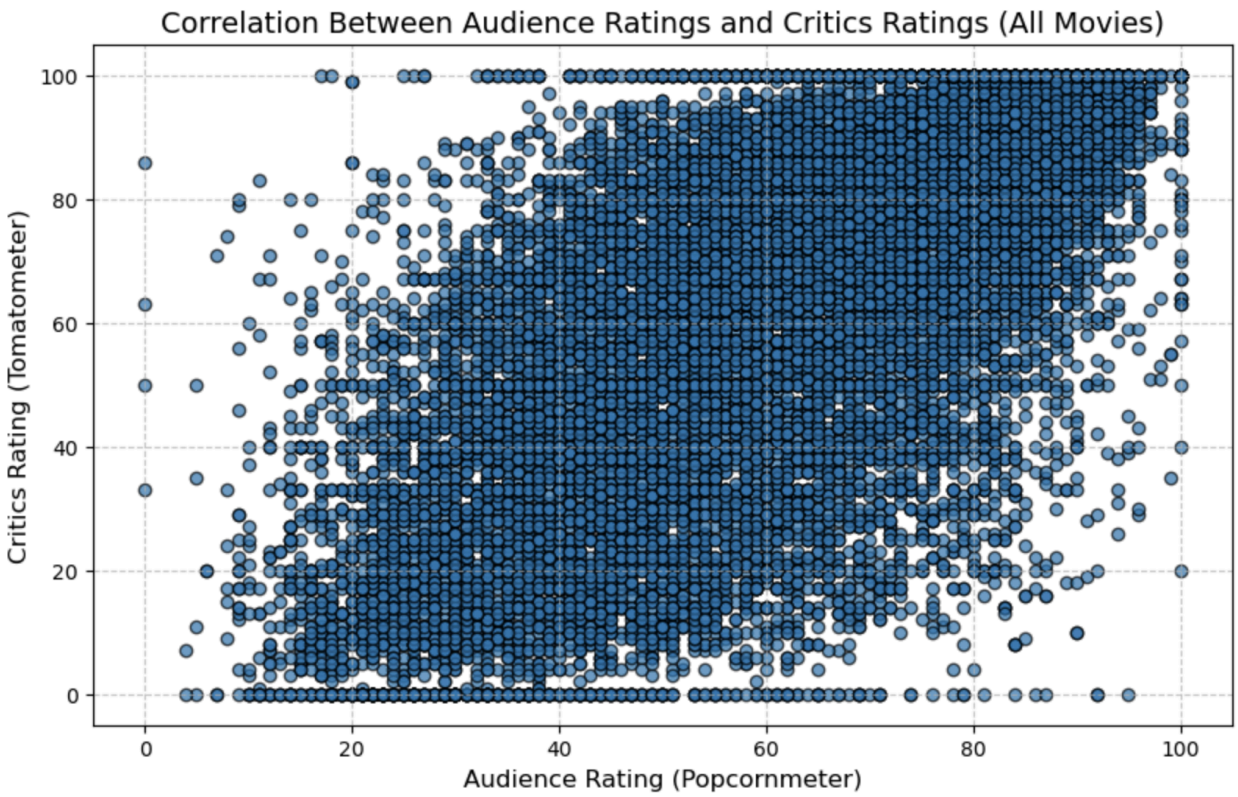


Graph 2: Frequency of Movie Genres

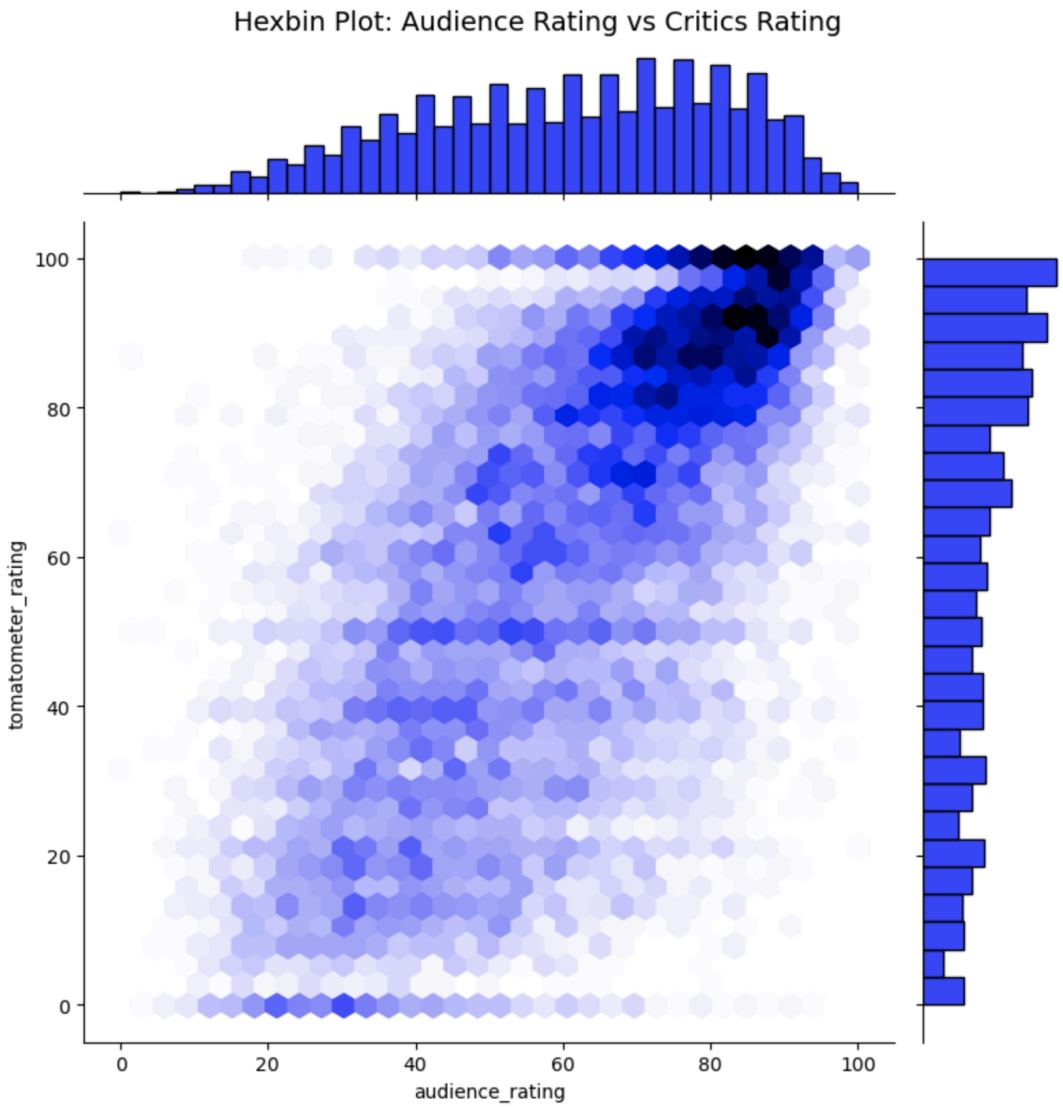
Movies with the Highest Weekly Gross Volatility:

	Movie	Volatility (Standard Deviation)
34	Smile 2	166.792765
4	Anora	146.644367
53	We Live in Time	28.530000
18	IF	25.329994
24	Kingdom of the Planet of the Apes	24.580425
47	The Wild Robot	20.018035
39	The Bikeriders	18.070132
41	The Forge	17.939971
2	Alien: Romulus	16.639908
30	Piece By Piece	14.799621

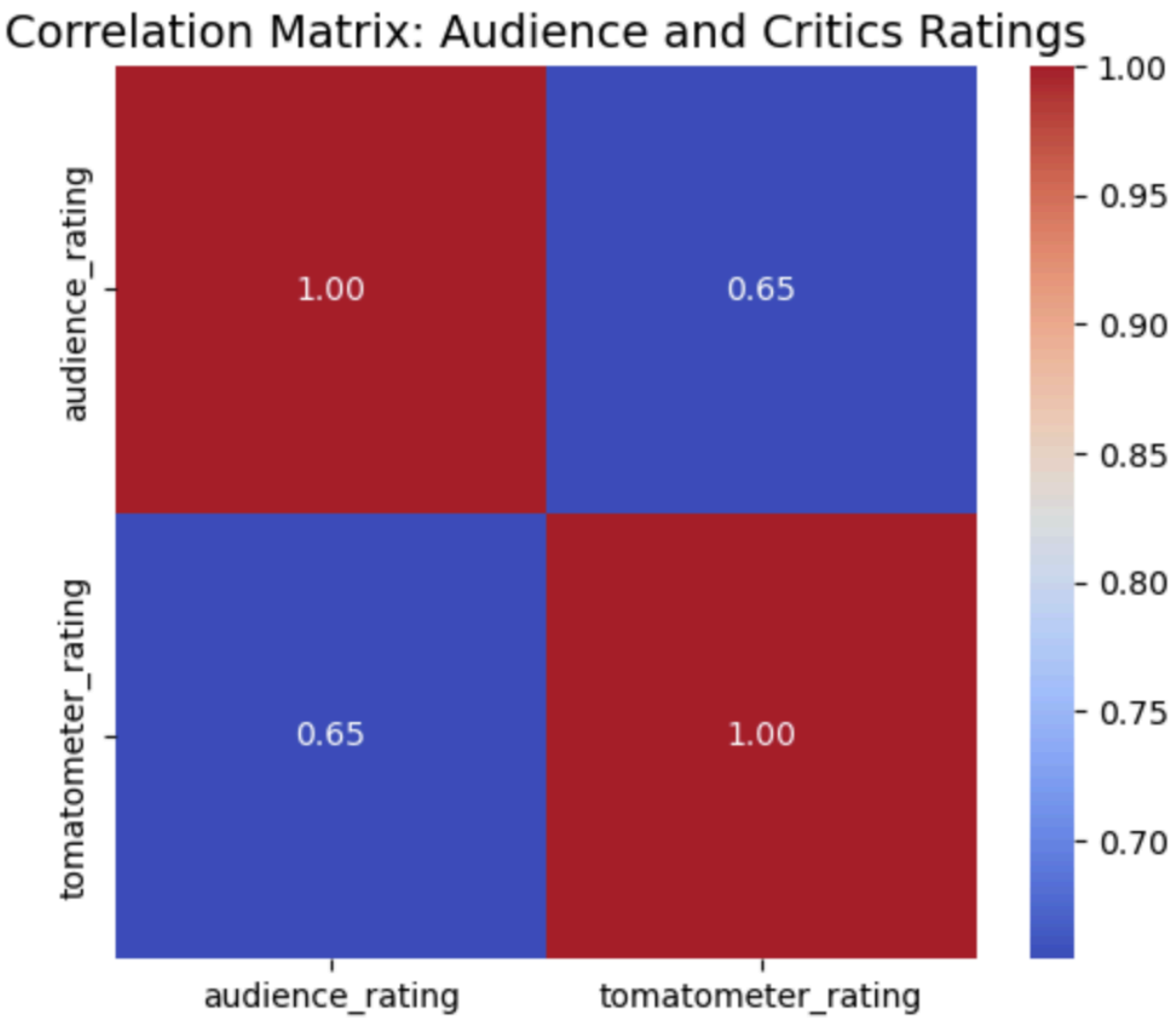
Table 4: Top 10 Movies with the Highest Weekly Gross Volatility



Graph 3: Correlation between Audience Ratings and Critics Ratings



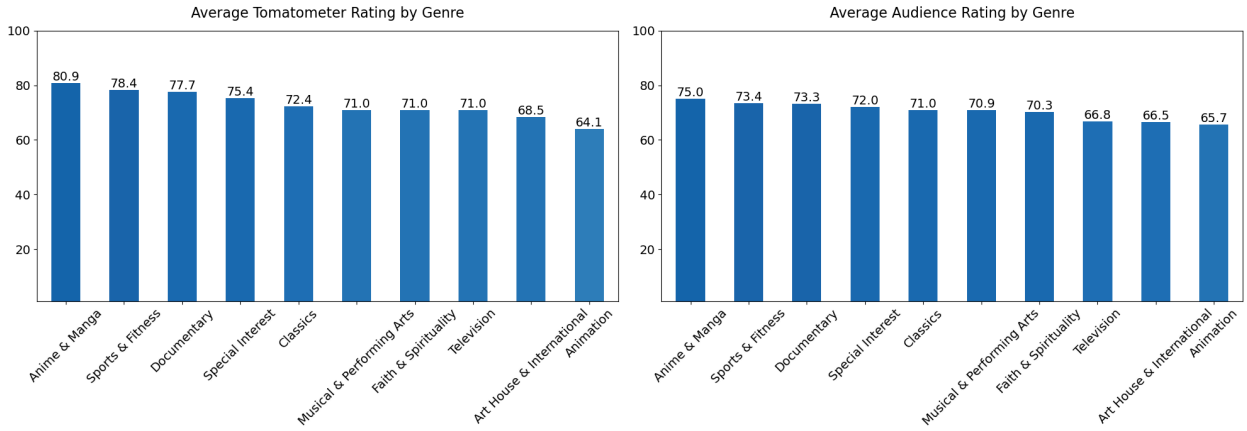
Graph 4: Hexbin Plot



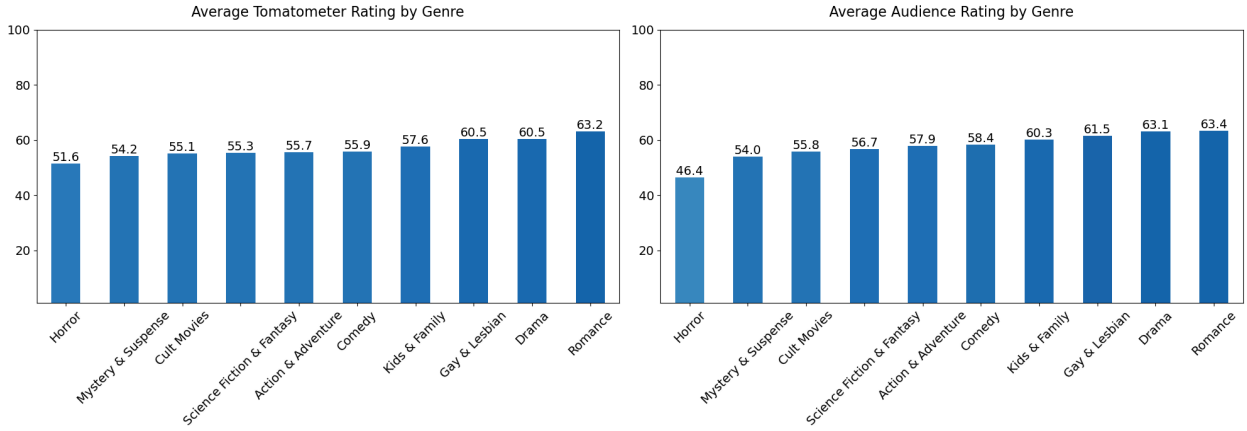
Graph 5: Correlation Matrix

Action & Adventure	Horror
Animation	Kids & Family
Anime & Manga	Musical & Performing Arts
Art House & International	Mystery & Suspense
Classics	Romance
Comedy	Science Fiction & Fantasy
Cult Movies	Special Interest
Documentary	Sports & Fitness
Drama	Television
Faith & Spirituality	Western
Gay & Lesbian	

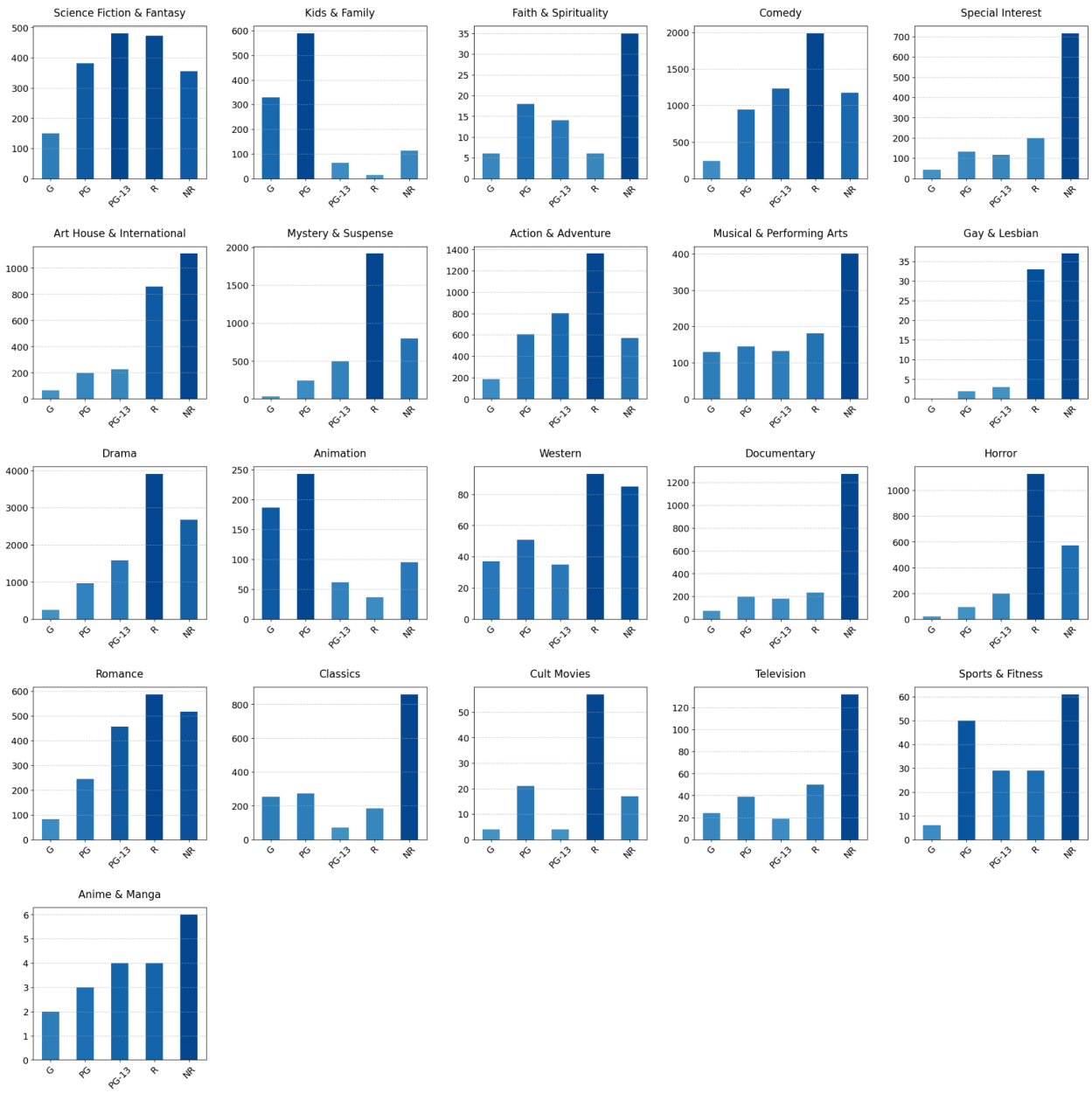
Table 5: 21 Different Genres



Graph 6: Highest Rated Genres

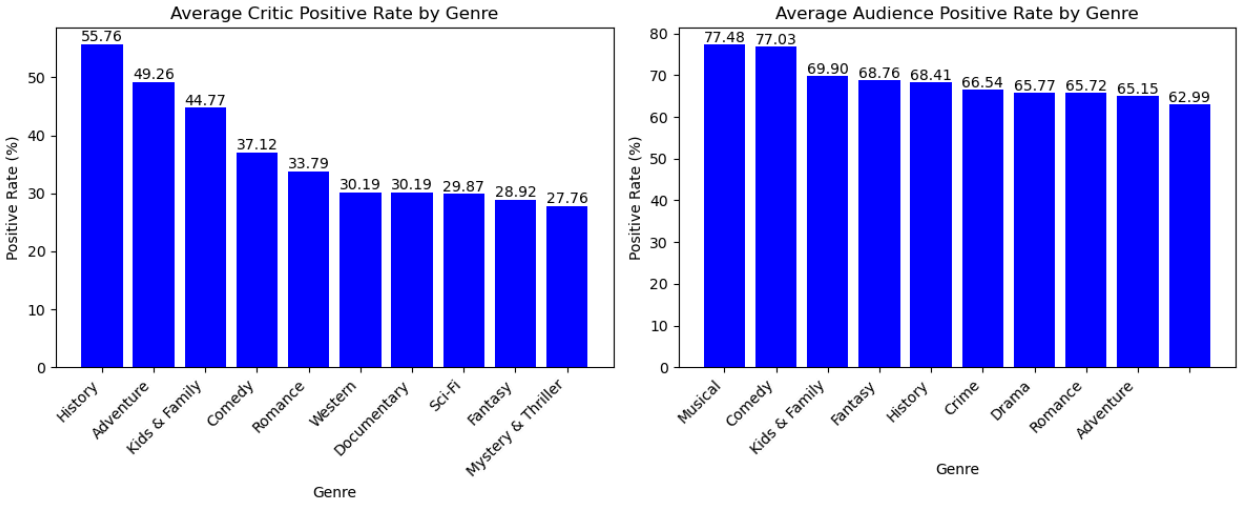


Graph 7: Lowest Rated Genres

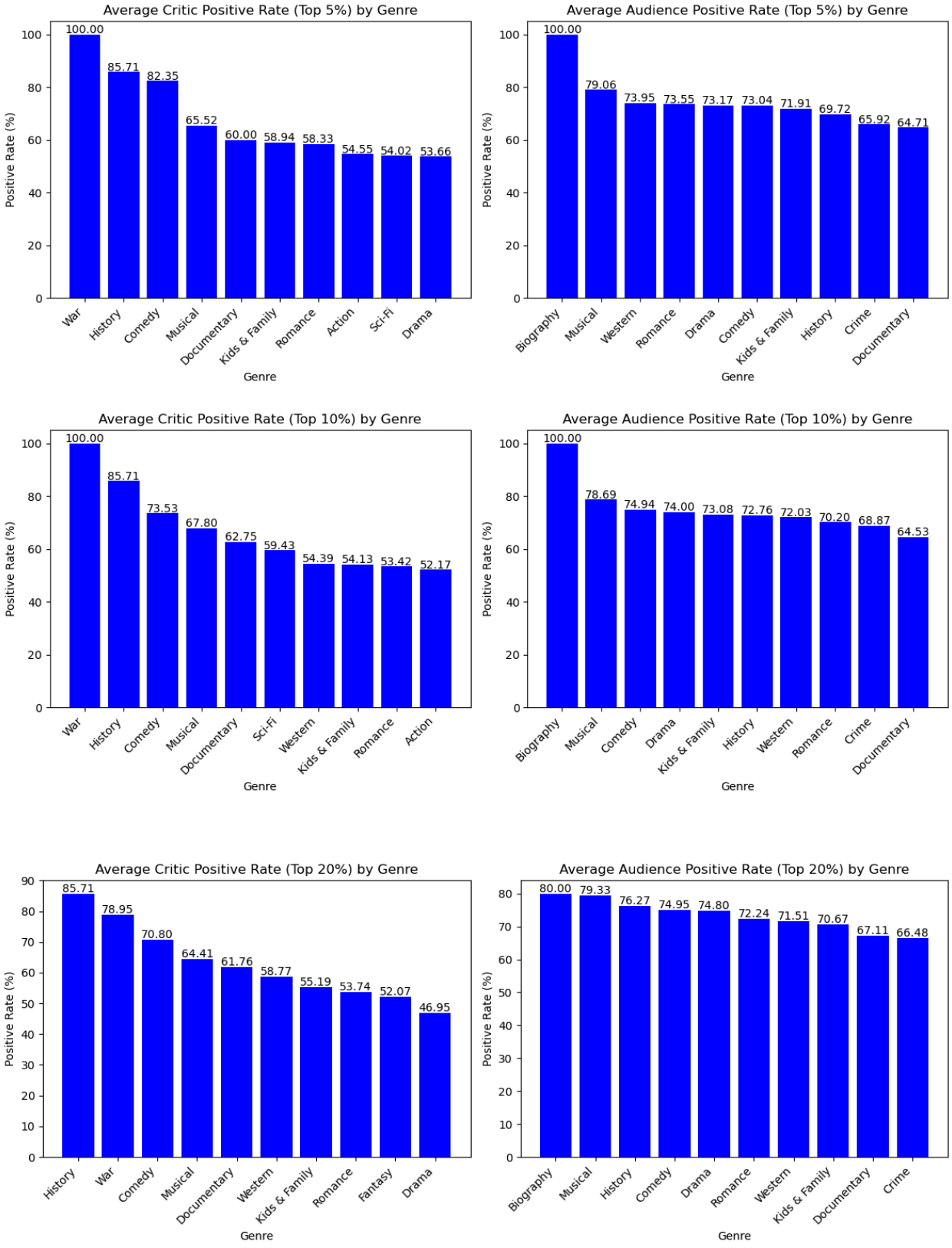


Graph 8: Content Ratings within Each Genre

Graph 9: Positive and Negative Word Cloud



Graph 10: Average Critic or Audience Positive Rate by Genre



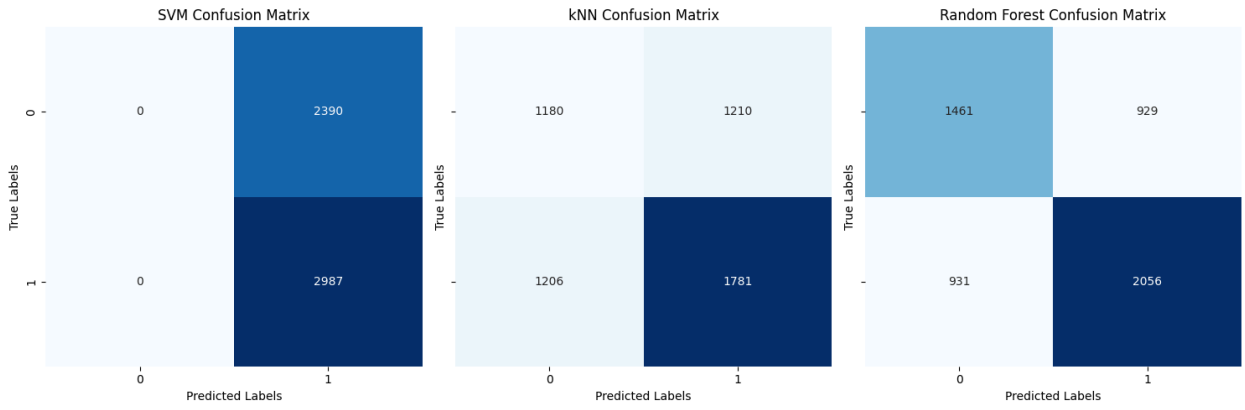
Graph 11: Average Critic or Audience Positive Rate by Genre(Earliest 5%/10%/20%)

Feature Name	Type	Description
Genre	Categorical	The primary genre of the movie (e.g., Action, Comedy).
Content	Categorical	Motion Picture Association (MPA) rating (e.g., G, PG, R).
Runtimebin	Ordered Factor	Binned runtime categories (e.g., 1–60 min, 61–120 min).
Rdecade	Categorical	The release decade of the movie (e.g., 1980s, 1990s).
Rmonth	Ordered Factor	Month of release (e.g., January, February).
Rdaybin	Categorical	Binned day of release (Early Month, Mid-Month, Late Month).
AStatus	Binary (Target)	Audience rating status: Upright (1) or Spilled (0).

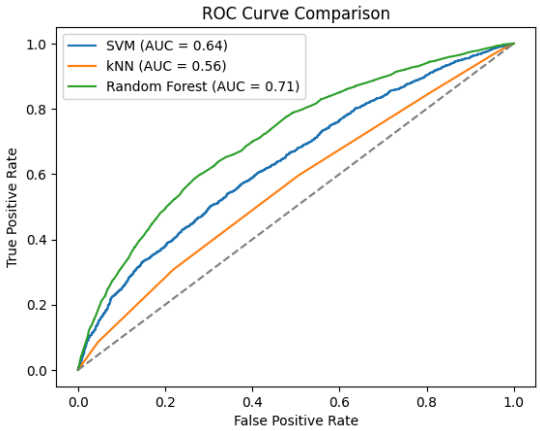
Table 6: Summary of Dataset Feature

Model	CV Accuracy (%)	Test Accuracy (%)
SVM	54.30	55.55
kNN	54.82	55.07
Random Forest	65.34	65.41

Table 7: Comparison of Model Accuracy



Graph 12: Matrix Confusion Comparison



Graph 13: ROC Curve Comparison