King Saud University

College of Computer and Information Sciences

Department of Information Technology

IT 362 Course Project

Semester-2, 1446H

What are the most popular movies?

Table of Contents

Introduction

This project aims to analyze the top 550 highest-rated movies on IMDb, examining common patterns that contribute to their success. The analysis will focus on key factors such as genre popularity, user ratings, and vote counts to understand their impact on a film's ranking.

The main research question guiding this study is: **What are the common characteristics of the highest-rated movies on IMDb, and how have these factors evolved over time?** By addressing this question, we aim to provide insights into trends within the film industry and how audience preferences have shaped the most successful films.

To support this analysis, we will utilize existing studies and datasets on IMDb ratings and audience behavior for instance, Kaggle provides various resources that examine IMDb rankings, so we use it as a secondary data source. such as:

Most Watched Movies and TV Shows

By leveraging these sources, this project will explore how different characteristics such as genre, rating trends, and voting counts have influenced the top-rated movies on IMDb over time.

Data Sources

1- We used the TMDb website and its API key to access detailed movie information. The TMDb API provided us with data such as movie titles, ratings, and other relevant attributes. This allowed us to retrieve IMDb-related information, efficiently and accurately.

Link: https://www.themoviedb.org/

The dataset consists of **550 rows** and **7 columns**.

Each row represents a movie described by a set of features:

- **Rank:** The movie's rank in the IMDb Top *550* list(integer)
- **Movie Name (Title):** The name of the movie.
- **Year:** The year that movie was released.
- **Movie Rating (IMDb Rating):** individual votes are aggregated as a single IMDb rate.
- **Votes:** number of users who rate a movie.
- **Genre:** Film classification (Drama, History, Comedy...)
- **Run Time:** Duration of the movie in minutes.

And the data type:

- **Rank:** int64
- **Title:** object
- **Year:** int64
- **IMDb Rating:** float64
- **Votes:** int64
- **Genre:** object
- **Runtime:** int64

The potential biases in the data:

- **Representation Bias:** The TMDb Top 550 list is based on user ratings, which may not reflect the preferences of all age and cultural groups.
- **Measurement Bias**: Ratings may be affected by time changes, as ratings for classic films can increase over time due to their reputation.

- **Historical Bias:** Some modern films may not be well represented on the list, as classic films tend to have higher ratings.

2- We used the Kaggle dataset "Most Watched Movies and TV Shows" to access detailed information about popular movies and TV shows. This dataset provides insights into widely viewed titles, including attributes such as rankings, ratings, and genres. It allowed us to analyze trends in audience preferences efficiently and accurately.
Link: https://www.kaggle.com/datasets/shiivvvaam/most-watched-movies-and-tv-shows

The dataset consists of **18165 rows and 7 columns**.
Each row represents a movie or TV show described by a set of features:

- **Rank**: The ranking of the title based on viewership or popularity.
- **Title**: The name of the movie or TV show.
- **Type**: Specifies whether the title is a *TV Show* or a *Movie*.
- **Premiere**: The year of release.
- **Genre**: The classification of the title (e.g., Action, Drama, Thriller, etc.).
- **Watchtime**: The total watch time of the title in minutes.
- **Watchtime in Million**: The watch time represented in millions.

Data Types:

- **Rank**: float64
- **Title**: object
- **Type**: object
- **Premiere**: float64
- **Genre**: object
- **Watchtime**: object
- **Watchtime in Million**: object

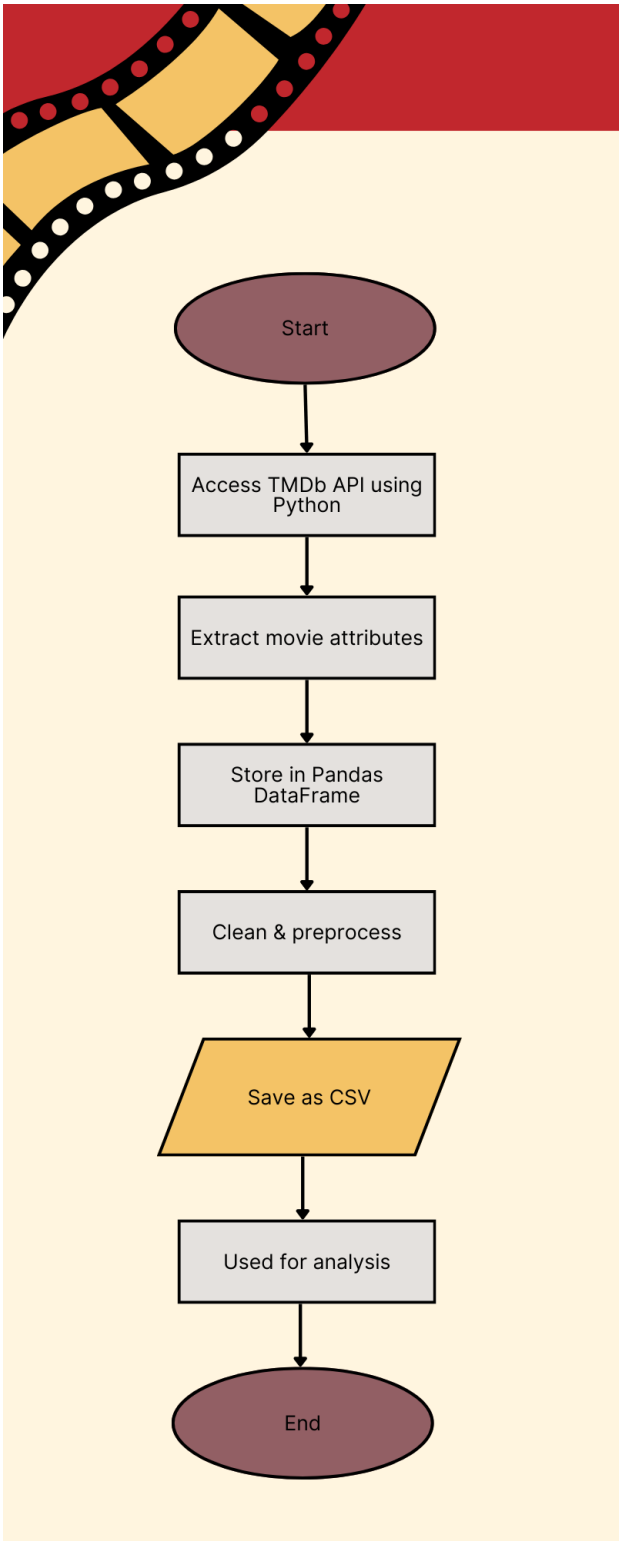
Potential Biases in the Data:

- **Representation Bias**: The dataset is based on viewership statistics, which may not capture all audience demographics, including underrepresented regions or niche genres.
- **Measurement Bias**: Popularity rankings and watch time fluctuate over time, meaning newer titles may not have accumulated enough viewership compared to older, well-established films and shows.
- **Historical Bias**: Classic titles with long-term popularity may dominate the dataset, leading to an overrepresentation of older content compared to newly released titles.

Data Collection & Description

Data Collection Process

To collect the movie data for our analysis, we followed a structured process using the TMDb API and a Kaggle dataset. The following flow chart illustrates the steps involved in collecting and preparing the data for analysis:



Description

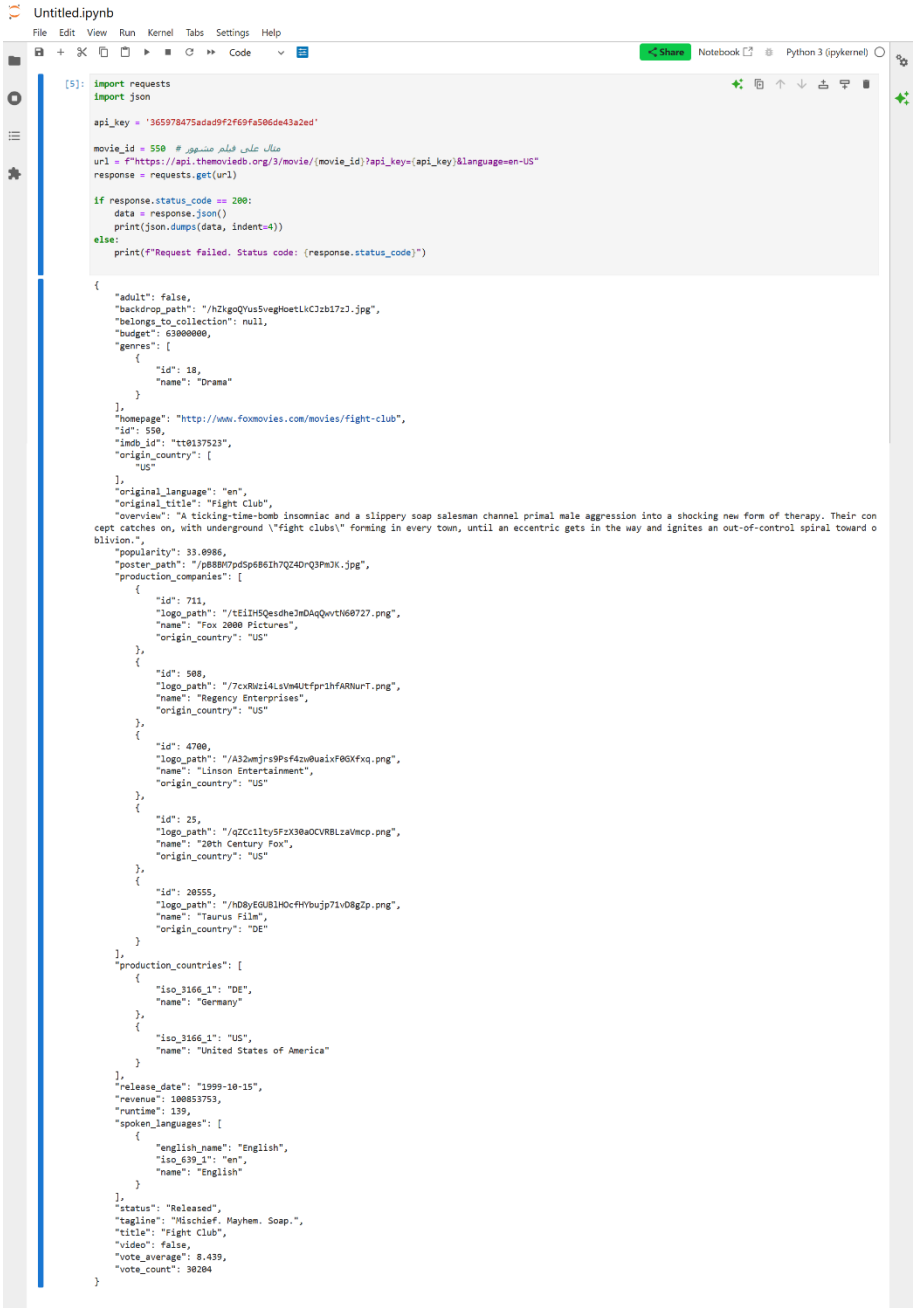
We used the TMDb API to collect structured movie data including title, rating, votes, genre, and runtime. The data was fetched using Python (requests + JSON handling) and then stored in a DataFrame. Preprocessing included:

- Removing duplicates & nulls
- Standardizing column formats
- Cleaning text fields (e.g., converting runtime from string to integer)

This ensured high data quality before visualization and analysis.

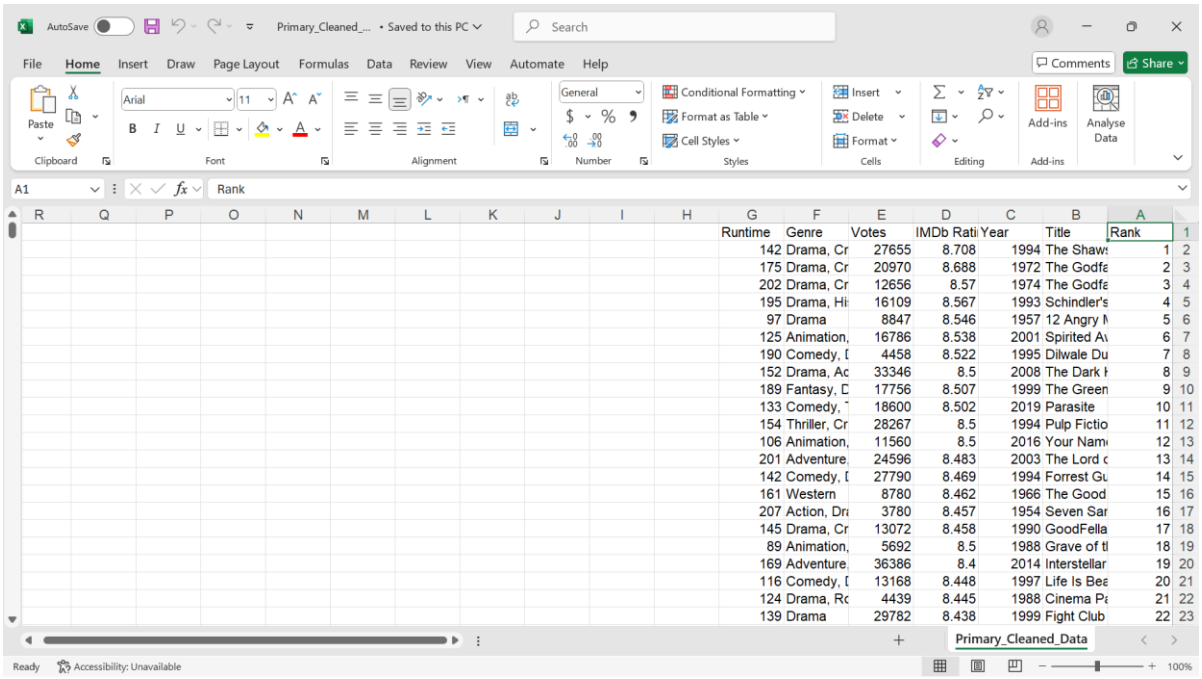
Screenshots of Raw and Cleaned Data

Raw Data:



Example of JSON structure showing nested movie details as retrieved directly from the API.

Cleaned Data:



Runtime	Genre	Votes	IMDb Rati	Year	Title	Rank
142	Drama, Cr	27655	8.708	1994	The Shaw	1
175	Drama, Cr	20970	8.688	1972	The Godfa	2
202	Drama, Cr	12656	8.57	1974	The Godfa	3
195	Drama, Hi	16109	8.567	1993	Schindler's	4
97	Drama	8847	8.546	1957	12 Angry M	5
125	Animation,	16786	8.538	2001	Spirited Aw	6
190	Comedy, D	4458	8.522	1995	Dilwale Du	7
152	Drama, Ac	33346	8.5	2008	The Dark K	8
189	Fantasy, D	17756	8.507	1999	The Green	9
133	Comedy, T	18600	8.502	2019	Parasite	10
154	Thriller, Cr	28267	8.5	1994	Pulp Fictio	11
106	Animation,	11560	8.5	2016	Your Name	12
201	Adventure,	24596	8.483	2003	The Lord o	13
142	Comedy, D	27790	8.469	1994	Forrest Gu	14
161	Western	8780	8.462	1966	The Good	15
207	Action, Dra	3780	8.457	1954	Seven Sar	16
145	Drama, Cr	13072	8.458	1990	GoodFella	17
89	Animation,	5692	8.5	1988	Grave of th	18
169	Adventure,	36386	8.4	2014	Interstellar	19
116	Comedy, D	13168	8.448	1997	Life Is Bea	20
124	Drama, Ro	4439	8.445	1988	Cinema Pa	21
139	Drama	29782	8.438	1999	Fight Club	22

Cleaned dataset showing essential columns: Rank, Title, Year, IMDb Rating, Votes, Genre, and Runtime.

This collection process allowed us to retrieve high-quality movie data efficiently. The use of APIs ensured the data was up-to-date and accurate, while preprocessing steps prepared the dataset for effective visualization, exploration, and modeling.

Primary Processed Data	
Observations	1276
Features	31
Data types	- object: 20 features - float64: 7 features - int64: 4 feature
Secondary Processed Data	
Observations	10,742
Features	28
Data types	- object: 2 features - float64: 25 features - int64: 1 feature

Objectives

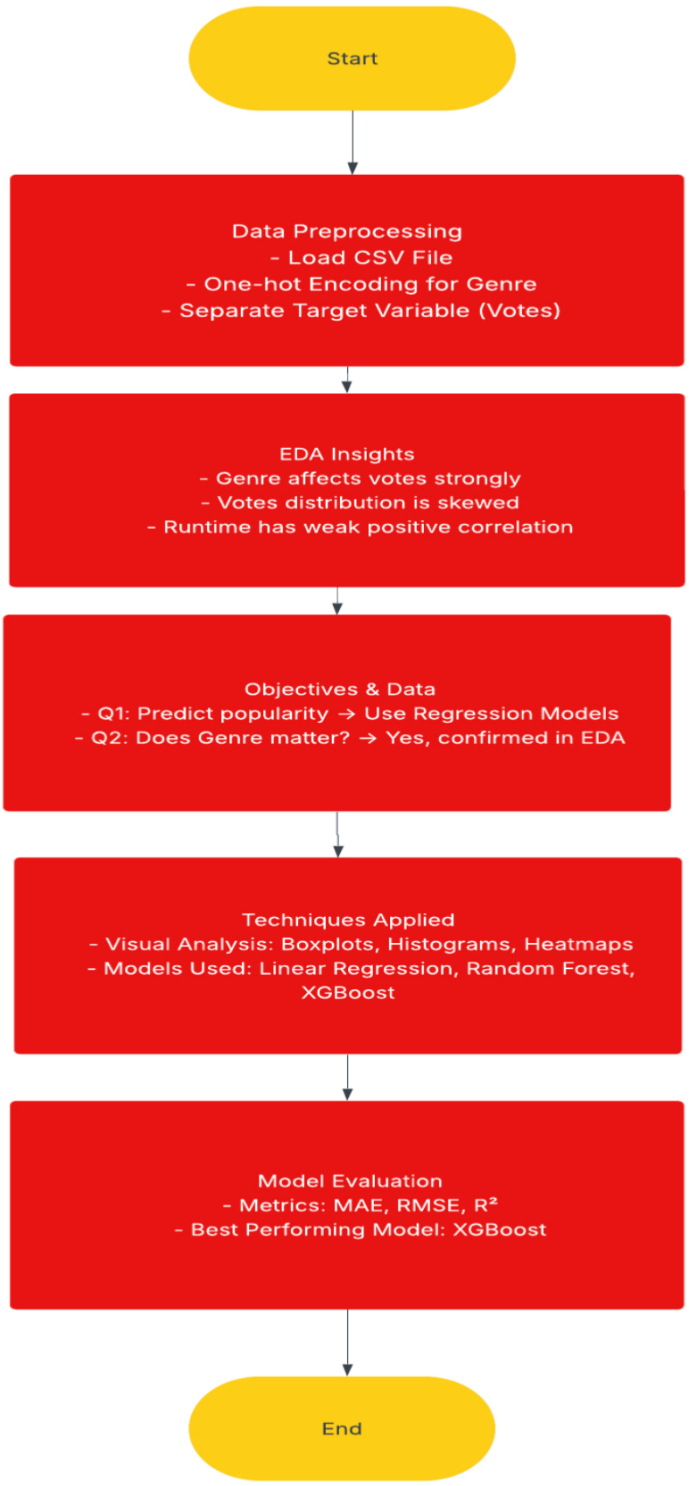
Using the data collected, we will answer the following questions:

1. How does the number of votes affect the ranking of a movie?
2. Do movies with more diverse genres (e.g., multi-genre) perform better in terms of rating and votes?
3. What is the correlation between a movie's runtime and its rating?
4. Which years in the last 15 years had the highest average IMDb movie ratings?
5. Are certain types of movies more likely to be rated higher on IMDb?

Method

We used the **TMDb API** to collect movie data since IMDb's official API had restrictions. By sending requests, we retrieved key details like **movie title, ratings, votes, genre, and runtime** . The data was then extracted, converted into a **Pandas DataFrame**, and saved as a CSV for analysis. This method ensured efficient and accurate data collection.

To address the research questions, we will conduct a structured analysis of the dataset using statistical and visualization techniques. The following steps outline our approach:



Data Preprocessing

Involved loading the dataset from a CSV file and ensuring data quality by removing duplicate rows to avoid redundancy. Rows with missing essential values, such as **Title, Year, and IMDb Rating**, were dropped. Columns were then converted to appropriate data types, including **Year, IMDb Rating, Votes, and Runtime**, to facilitate analysis. The **Genre** column was cleaned by trimming extra spaces, while the **Runtime** column was processed by removing the "min" text and converting it into an integer. Finally, the cleaned dataset was saved as a new CSV file for further analysis.

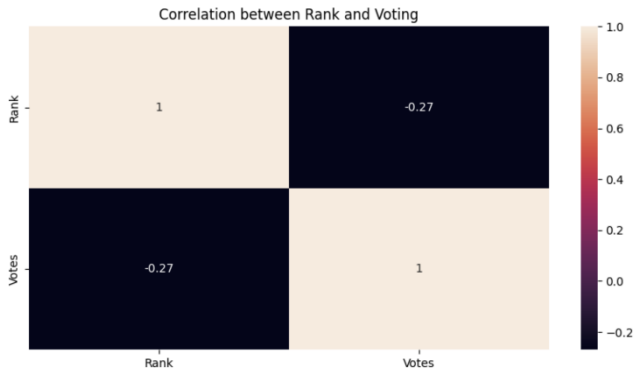
To answer the questions and determine how the attributes relate to one another, we utilized the graphs.

And these libraries helped us to create **graphs and visualizations** in Python:

1. **seaborn (`sns`)** – Makes beautiful statistical graphs.
2. **pandas (`pd`)** – Handles data in tables (DataFrames).
3. **numpy (`np`)** – Works with numbers and arrays.
4. **matplotlib.pyplot (`plt`)** – Draws basic charts and graphs.

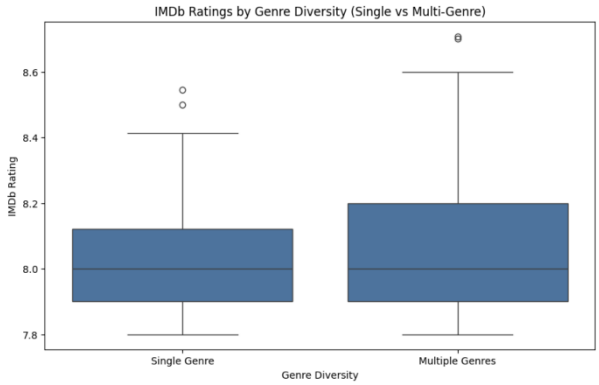
Questions

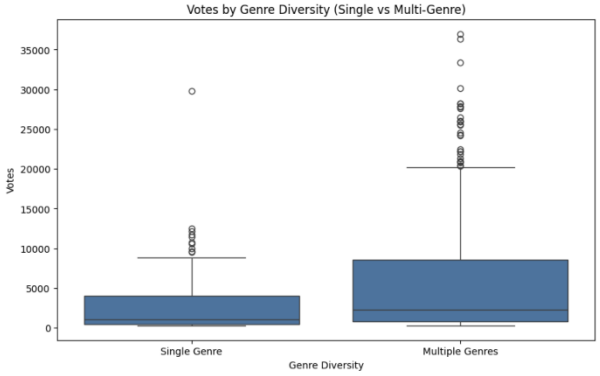
1- How does the number of votes affect the ranking of a movie?



To investigate the relationship between the number of votes and movie rankings, we will use a correlation heatmap to measure the strength and direction of this association. Correlation analysis will help quantify how closely these two variables are related, while the heatmap will provide a visual representation of their interaction. By analyzing the correlation coefficient, we can determine whether there is a positive, negative, or negligible relationship between audience engagement (measured through vote count) and movie ranking. Additionally, we will examine the spread of data points to assess whether other factors, such as reviews, movie quality, and genre, also influence rankings.

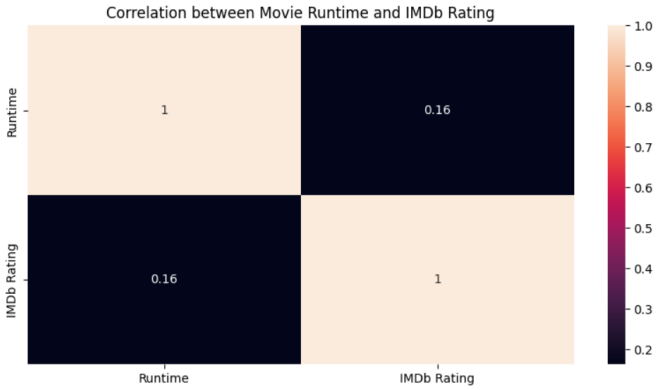
2- Do movies with more diverse genres (e.g., multi-genre) perform better in terms of rating and votes?





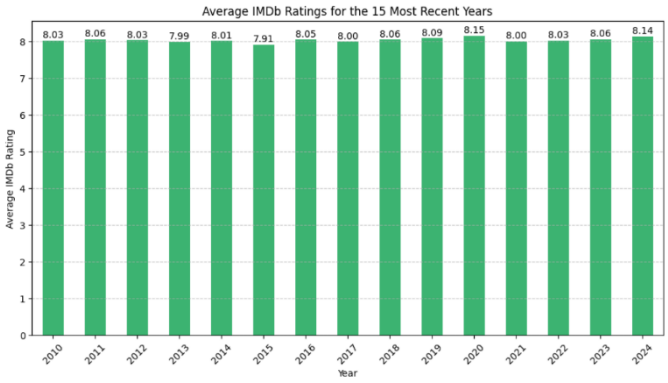
To compare single-genre and multi-genre movies, we will utilize boxplots to analyze IMDb ratings and vote counts. Boxplots are effective for displaying data distribution, median values, and variability, helping us identify any trends or outliers. By comparing the distributions of ratings and votes between single-genre and multi-genre films, we can assess whether genre diversity has an impact on audience engagement. This analysis will help determine if multi-genre movies receive higher ratings and more votes compared to single-genre films, providing insights into the role of genre diversity in movie success.

3-What is the correlation between a movie's runtime and its rating?



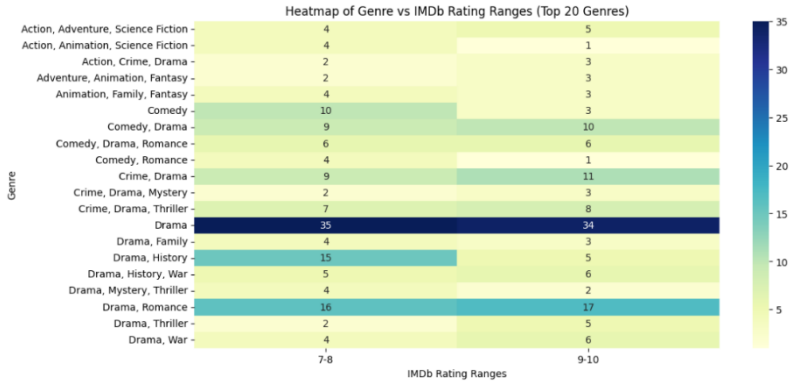
To explore the impact of runtime on IMDb ratings, we will conduct a correlation analysis and visualize the findings using a correlation heatmap. The correlation coefficient will help quantify the degree of association between movie length and ratings. A strong correlation would suggest a significant influence, whereas a weak correlation would indicate minimal impact. The heatmap will allow us to intuitively observe trends and assess whether movies longer tend to receive higher ratings. Additionally, we will consider other contributing factors such as storyline, acting, and direction that may have a greater influence on movie ratings.

4- Which years in the last 15 years had the highest average IMDb movie ratings?



To examine IMDb rating trends over time, we will calculate the average movie rating for each year within the past 15 years and visualize the results using a bar chart. This method will enable us to detect fluctuations in audience reception across different years and identify any trends in movie quality. The bar chart will provide a clear comparative view of how ratings have evolved over time, helping us analyze whether specific years had notably higher or lower ratings. This approach will allow us to identify any patterns or shifts in audience preferences and overall movie reception.

5- Are certain types of movies more likely to be rated higher on IMDb?



To determine which movie genres are more frequently associated with high IMDb ratings, we will create a heatmap that compares different genres with various IMDb rating ranges. This visualization will highlight the distribution of movies across genres and rating categories, allowing us to identify patterns and trends. Heatmaps are particularly effective for analyzing categorical data and revealing relationships, making them a suitable choice for understanding how genre influences movie ratings. By interpreting the heatmap, we can pinpoint which genres consistently receive high ratings and examine how different movie categories perform in terms of audience reception.

Challenges & Recommendations

We were unable to obtain the official API from IMDb directly because restrictions were imposed on accessing data through the official API, which made it difficult to collect the required data immediately.

We decided to use an alternative site (TMDb) to get the API Key, which gives us the ability to access similar data from movies such as ratings, ratings, and year.

EDA

Primary Data

Overview

The primary dataset consists of 550 movies ranked as the highest-rated films on IMDb. The key attributes analyzed include IMDb Ratings, Votes, Runtime, and Genre. This dataset provides insight into what makes a movie critically and publicly acclaimed. By analyzing IMDb ratings, we aim to identify patterns in audience reception and understand the factors that contribute to a high-ranking film. Additionally, examining votes and runtime offers perspective on viewer engagement and film characteristics that resonate with audiences.

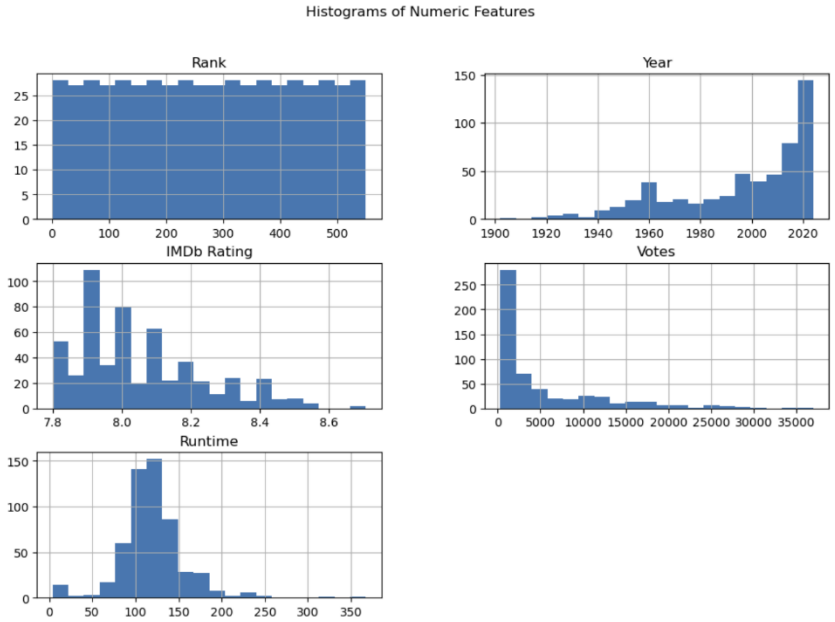
Key Findings

The IMDb ratings in this dataset range from 7.8 to 8.7, with most movies concentrated around the 8.0 to 8.2 range. This suggests a strong clustering of ratings, meaning that the majority of highly rated films on IMDb do not deviate significantly from this score range. The relatively small variation indicates that films achieving this ranking maintain a consistently high standard, but breaking beyond an 8.5 rating is relatively rare. Exceptional movies that surpass this threshold likely benefit from a combination of strong storytelling, high production quality, and audience appreciation.

When analyzing votes, we observe significant variation, with some movies receiving as few as 300 votes, while others garner nearly 37,000 votes. The correlation between ranking and votes highlights that highly rated movies tend to receive more audience engagement. However, some highly ranked films have lower vote counts, suggesting that certain critically acclaimed films may cater to niche audiences. Films with widespread appeal and larger marketing campaigns tend to accumulate more votes, reinforcing their high IMDb placement.

The runtime distribution of movies in this dataset ranges from 4 minutes to 367 minutes, but most movies have a runtime between 100 and 135 minutes. This aligns with conventional feature-length

films, which typically fall within this range. The presence of extreme runtimes, such as very short films and significantly long movies, suggests diversity in storytelling formats. While longer movies might allow for more intricate narratives, they may also pose a risk of reduced audience retention.



Histograms of Numeric Features

The histograms provide insights into the distribution of various numeric features of the top 550 highest-rated movies on IMDb.

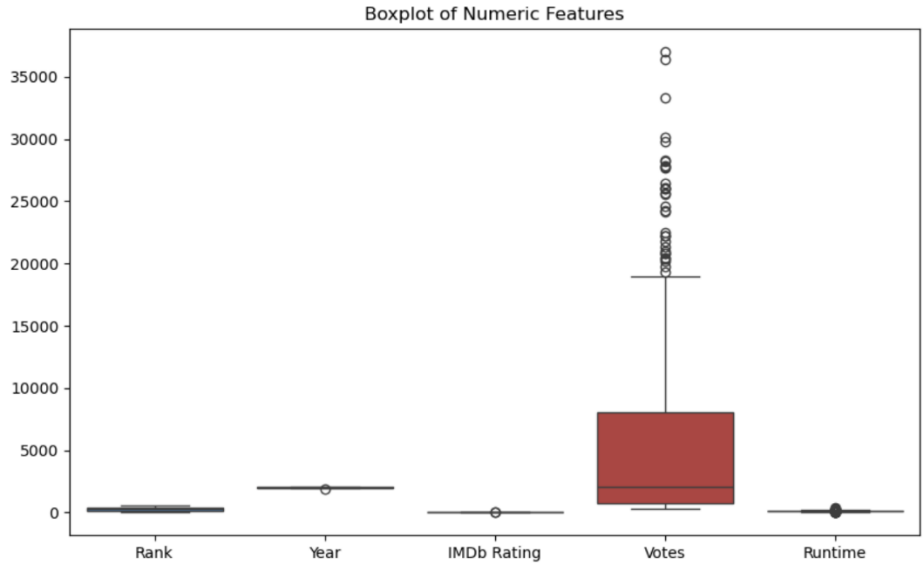
1. Rank Distribution: The rank histogram shows a relatively uniform distribution, indicating that movies are evenly distributed across the ranking spectrum from 1 to 550.

2. Year Distribution: The year histogram highlights a significant increase in the number of highly rated movies in recent decades, particularly from the 2000s onward. This suggests a growing number of well-received films in modern cinema, possibly due to higher production rates, wider audience engagement, or changing rating patterns over time.

3. IMDb Rating Distribution: The IMDb rating histogram shows a concentration of ratings between 7.8 and 8.2, with fewer movies exceeding 8.5. This suggests that while high ratings are common among these top films, exceptionally high scores above 8.6 are rare.

4. Votes Distribution: The votes histogram reveals a right-skewed distribution, with a large number of movies receiving relatively few votes and only a small portion achieving extremely high vote counts. This suggests that while some films attract massive audience engagement, most top-rated films do not necessarily receive a large number of votes.

5. Runtime Distribution: The runtime histogram follows a roughly normal distribution, with most movies falling within the 90–150-minute range. This aligns with common industry practices, where feature films typically stay within this duration for optimal audience engagement.



Box Plot of Numeric Features

The boxplot provides a visual representation of the distribution and spread of numerical features in the dataset, highlighting potential outliers.

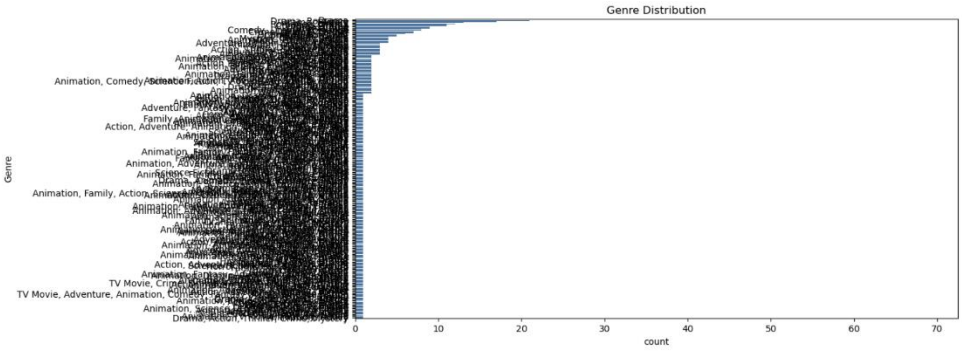
1. Votes: This feature exhibits a highly skewed distribution with numerous outliers extending far beyond the upper whisker. This suggests that while most movies receive a moderate number of votes, a small number of films achieve exceptionally high engagement, likely blockbuster films or cult classics with widespread popularity.

2. Rank, Year, IMDb Rating, and Runtime: These features display relatively compact interquartile ranges, indicating that most values fall within a narrow spread. The presence of some outliers, particularly in the Year and Runtime features, suggests that a few movies in the dataset deviate significantly from the norm, such as older classic films or exceptionally long movies.

3. Year: A few outliers suggest that some movies from much earlier decades have made it onto the top-rated list, reinforcing the idea that classic films continue to hold high rankings.

4. IMDb Rating: The small spread in ratings, with few extreme outliers, indicates that most of the highest-rated movies have a consistently high IMDb score, reinforcing the trend observed in the histogram.

5. Runtime: The presence of outliers in runtime suggests that while most films fall within a standard range (likely 90-150 minutes), a few significantly longer films exist in the dataset.



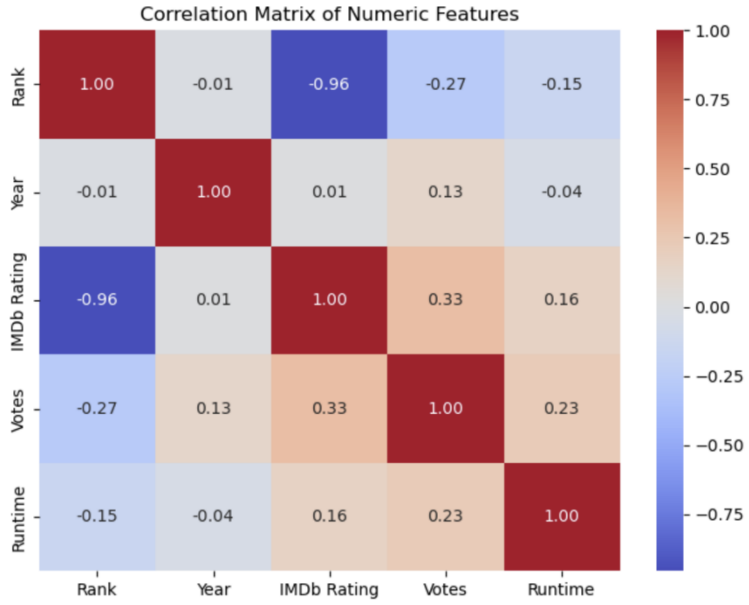
Genre Distribution

The genre distribution plot provides insight into the frequency of different genres among the top 550 highest-rated movies on IMDb.

1. Dominant Genres: The most common genres in the dataset appear to be Drama, Comedy, and Action, as indicated by the longest bars. This suggests that audiences tend to rate these genres highly, likely because they dominate mainstream and critically acclaimed cinema.

2. Multi-Genre Representation: Many films belong to multiple genres (e.g., "Animation, Comedy, Adventure"), which makes the genre distribution more complex. This could indicate that movies that blend genres tend to be more successful, possibly because they appeal to a broader audience.

3. Less Common Genres: Some genre combinations have significantly lower counts, implying that certain niche genres (such as TV movies, Crime-Animation hybrids, or specific sub-genres) are less frequently represented in the highest-rated films.



Correlation Matrix of Numeric Features

The correlation matrix provides valuable insights into the relationships between the numerical features in our dataset.

1. Strong Negative Correlation Between Rank and IMDb Rating (-0.96)

This strong inverse relationship indicates that as a movie's IMDb rating increases, its rank improves (moves closer to #1). This is expected since rankings are primarily based on ratings.

2. Moderate Positive Correlation Between IMDb Rating and Votes (0.33)

Movies with higher IMDb ratings tend to receive more votes, but the correlation is not extremely strong. This suggests that while well rated movies generally attract more attention, other factors like popularity, franchise appeal, or marketing may influence vote counts.

3. Moderate Positive Correlation Between Votes and Runtime (0.23)

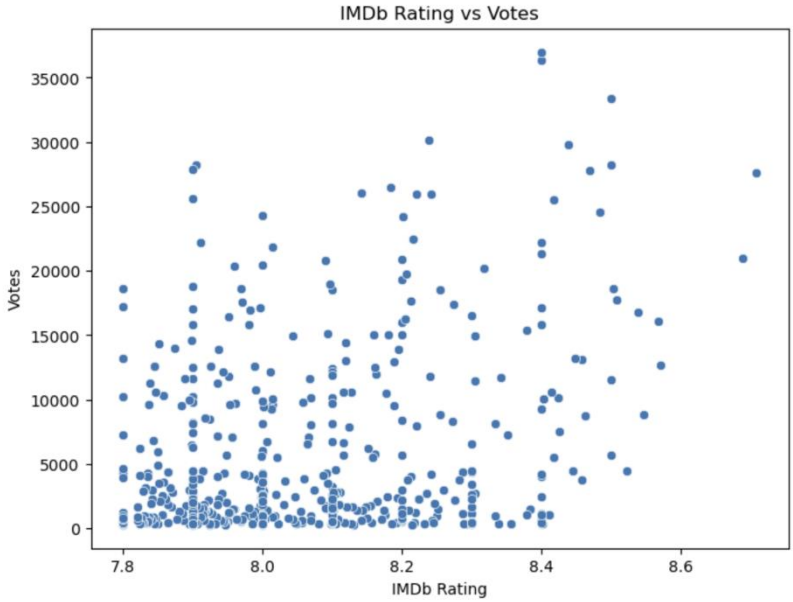
Longer movies tend to receive more votes, possibly because epic-length films are often associated with large-scale productions, which attract more viewers and discussions.

4. Weak or No Correlation Between Year and Other Features

The year of release has almost no correlation with IMDb rating (0.01) and only a slight positive correlation with votes (0.13). This suggests that the success of a movie is not necessarily dependent on its release year but rather on other factors such as rating, genre, and audience reception.

5. Weak Negative Correlation Between Rank and Votes (-0.27)

While higher-ranked movies tend to have more votes, the relationship is not particularly strong. This implies that some lower-ranked films still attract a significant number of ratings, possibly due to cult followings or controversial content.



IMDb Rating vs Votes

The scatter plot visualizes the relationship between IMDb Rating and Votes, revealing several key insights:

1. Weak to Moderate Positive Correlation

While there is a general trend suggesting that movies with higher IMDb ratings tend to receive more votes, the correlation is not particularly strong. Some highly rated films have relatively low vote counts, while some movies with lower ratings (around 7.8–8.0) still receive a significant number of votes.

2. Widespread in Vote Counts

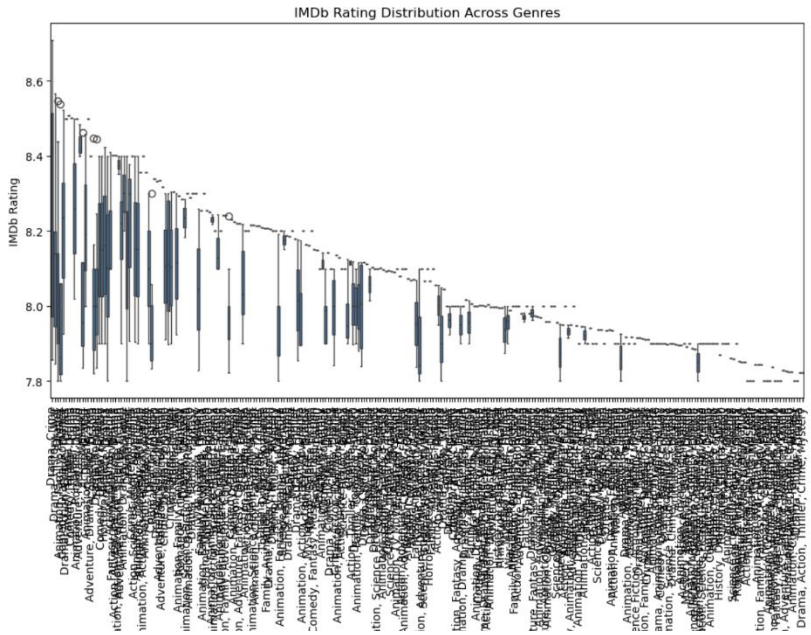
The number of votes varies significantly, with some movies receiving only a few hundred or thousand votes, while others exceed 30,000 votes. This suggests that high ratings alone do not necessarily guarantee high audience engagement—other factors such as popularity, marketing, and franchise appeal play a role.

3. Higher-Rated Films Tend to Have More Votes

The concentration of movies with higher votes is more visible in the 8.2–8.6 rating range. This indicates that films with exceptionally high ratings tend to attract more viewers and engagement.

4. Outliers with Extremely High Votes

A few movies stand out with exceptionally high vote counts (>30,000), even at different rating levels. These are likely highly popular mainstream films, blockbusters, or cultural phenomena that have gained massive audience attention.



IMDb Rating Distribution Across Genres

This boxplot illustrates the IMDb rating distribution across different genres, offering insights into which genres tend to receive higher ratings on average.

1. Genres with the Highest Ratings

Some genres appear to have consistently higher IMDb ratings, with their median values around 8.4–8.6. These likely include genres such as Documentary, War, and Animation, which often attract niche but passionate audiences who rate them highly.

The upper whiskers of these high-rated genres extend even beyond 8.6, indicating that some movies in these categories achieve exceptionally high ratings.

2. Wider Spread in Some Genres

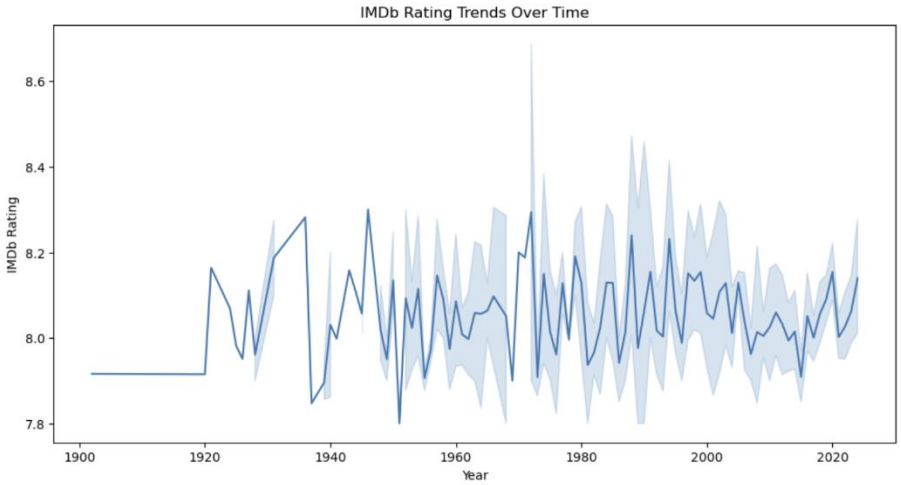
Certain genres show a widespread in IMDb ratings, meaning that movies within those categories vary significantly in audience reception. Action, Sci-Fi, and Adventure movies might fall into this category, as these films can range from critically acclaimed masterpieces to divisive blockbusters.

3. Genres with Lower Ratings

Toward the right side of the plot, some genres have consistently lower ratings, with median IMDb ratings closer to 7.8–8.0. These could include TV Movie, Horror, and Crime, which may be more polarizing or niche in appeal.

4. Multi-Genre Combinations Influence Ratings

Many films are classified under multiple genres, which adds complexity to the analysis. Certain combinations (e.g., Drama & Biography, Animation & Family) might generally receive higher ratings, while others (e.g., Horror & Thriller, TV Movie & Comedy) may have a lower overall reception.



IMDb Rating Trends Over Time

This line plot visualizes IMDb rating trends over time, showing how the average rating of the highest-rated films has evolved throughout cinematic history.

1. Stable Early Ratings (Pre-1920s):

The earliest films in the dataset have stable ratings around 7.8, likely due to the small number of surviving films from that era or limited audience engagement.

2. Fluctuations in the Mid-20th Century (1920s–1960s):

The ratings show notable variations, with spikes in some periods (e.g., late 1930s, 1950s), possibly indicating the emergence of influential films from the Golden Age of Hollywood.

The drop-offs in certain decades could be due to fewer films from those periods making it into the IMDb Top 550.

3. Higher Variability from the 1970s Onward:

Starting in the 1970s, the fluctuations in IMDb ratings become more pronounced.

Peaks and dips suggest that some years had exceptionally well-received films, while others had a mix of high and mid-rated movies.

4. Ratings Stabilizing in the 21st Century (2000s–Present):

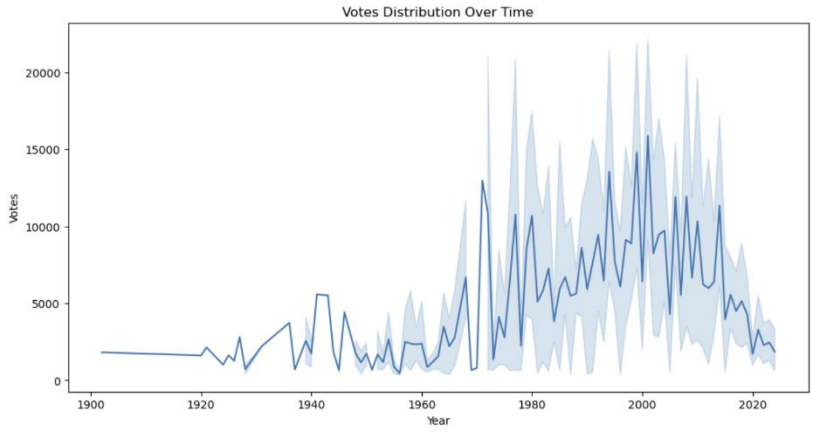
The average rating remains relatively consistent between 8.0 and 8.2, with minor fluctuations.

This suggests that despite changes in the film industry (e.g., streaming, franchise dominance), the audience's perception of "highly rated" films remains stable.

5. Shaded Confidence Interval:

The blue shaded area represents variance—wider gaps indicate greater spread in ratings for films from that period.

The large variance in the 1970s–2000s might reflect a mix of blockbuster franchises and critically acclaimed indie films.



Votes Distribution Over Time

This line plot illustrates votes distribution over time, showing how audience engagement with top-rated movies has evolved.

1. Low Vote Counts for Early Films (Pre-1950s):

Older movies (pre-1950s) generally have fewer votes, likely due to the limited audience reach and lack of widespread digital engagement.

Some films from the Golden Age of Hollywood (1930s–1950s) received modest vote counts, possibly due to historical recognition rather than active audience participation.

2. Increase in Votes from the 1960s–1980s:

A noticeable rise in votes begins in the 1960s and continues through the 1980s, possibly reflecting increased accessibility to films, home media distribution, and growing cinephile communities.

The variance (shaded area) expands significantly, indicating that some films in this period gained widespread recognition while others remained niche.

3. Surge in Votes from the 1990s–2000s:

From the 1990s onward, there is a dramatic increase in vote counts, peaking in the 2000s and early 2010s.

This is likely driven by:

- The rise of the internet and online film communities (e.g., IMDb, Rotten Tomatoes).
- Increased accessibility of films through DVDs, Blu-ray, and early streaming services.
- Growth of blockbuster franchises attracting global audiences.

4. Slight Decline Post-2015:

A downward trend in votes is noticeable after 2015, though variability remains high.

Possible explanations:

- The fragmentation of audiences across multiple streaming platforms.
- Decreasing emphasis on IMDb voting as new review platforms emerge.
- The shift in audience behaviour, with fewer people rating films online despite watching them.

Secondary Data

Overview

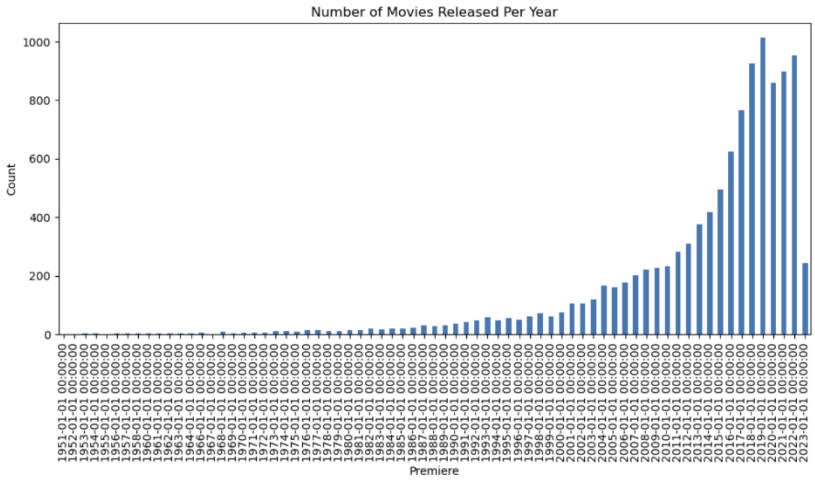
The secondary dataset contains 10,742 movies, covering a wider range of movies, including different genres, film formats, and production years. Unlike the primary dataset, which focuses on top-rated films, the secondary dataset provides a broader context of films that vary in popularity and reception. This diversity allows us to analyze how different attributes—such as runtime, genre, and production year—differ in films that are not necessarily ranked among IMDb's highest-rated movies.

Key Findings

The watch time (runtime) distribution in this dataset shows that a large proportion of movies have relatively short runtimes, clustering between 0 and 50 minutes. This suggests that the dataset includes various forms of content, including short films, television episodes, and potentially digital content such as web series. The presence of movies exceeding 150 minutes indicates that some long-form films and extended features are also present. Compared to the primary dataset, which primarily consists of conventional feature-length films, the secondary dataset presents a wider variation in runtime, reflecting different storytelling approaches and audience consumption patterns.

The genre distribution within the secondary dataset is notably different from that of the primary dataset. The dataset employs a one-hot encoding format, allowing for a more detailed breakdown of genre representation. Certain genres, such as Comedy and Documentary, appear more frequently in

this dataset, suggesting that these categories are widely produced and consumed, even if they do not always reach the highest IMDb ratings. In contrast, genres like Drama and Crime, which dominate the primary dataset, are comparatively less frequent in the secondary dataset, indicating that while they may not be produced as frequently, they are more likely to receive high ratings when executed well.



Number of Movies Released Per Year

This bar chart illustrates the number of movies released per year, highlighting trends in film production over time.

1. Low Production in Early Decades (Pre-1970s)

Before the 1970s, the number of movies released each year was relatively low.

This could be due to:

- The limited film industry infrastructure.
- High production costs and technological constraints.
- The dominance of studio-controlled filmmaking with fewer annual releases.

2. Gradual Growth in the 1980s and 1990s

The number of movies released per year started increasing steadily.

Possible factors include:

- The expansion of independent filmmaking.
- The rise of home video markets (VHS, DVD).
- The globalization of cinema, leading to more diverse productions.

3. Exponential Growth Post-2000

From 2000 onward, there is a massive increase in the number of films produced per year, peaking in the 2010s.

Likely contributing factors:

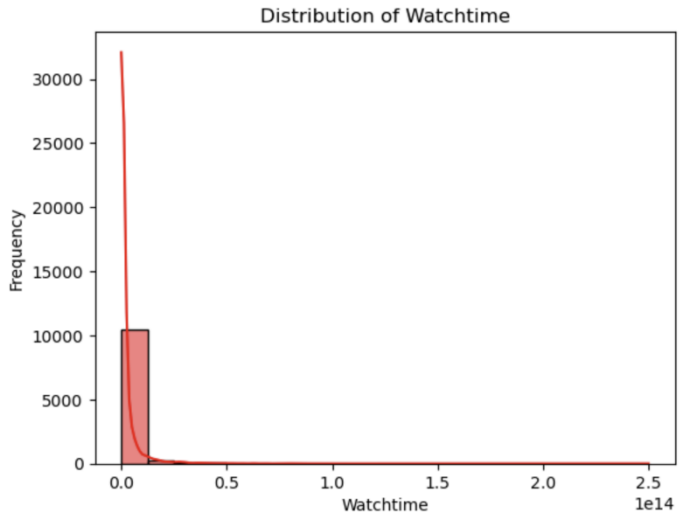
- The digital revolution, making filmmaking more accessible.
- The expansion of streaming services (Netflix, Amazon Prime, etc.), increasing demand for content.
- The rise of international film industries competing with Hollywood.

4. Peak and Slight Drop After 2020

The peak is around 2018–2019, followed by a decline.

This drop might be due to:

- The COVID-19 pandemic (2020–2021), which disrupted productions worldwide.
- Changing industry trends, with more focus on digital first content rather than traditional theatrical releases.



Distribution of Watchtime

This histogram visualizes the distribution of watch time, which represents the total number of hours audiences have spent watching the highest-rated movies.

1. Highly Skewed Distribution:

The data is heavily right skewed, meaning most movies have relatively low total watch times, while a few have extremely high watch times.

This suggests that while many top-rated movies receive moderate attention, a small number achieve massive audience engagement.

2. Most Films Fall in a Low Watch Time Range:

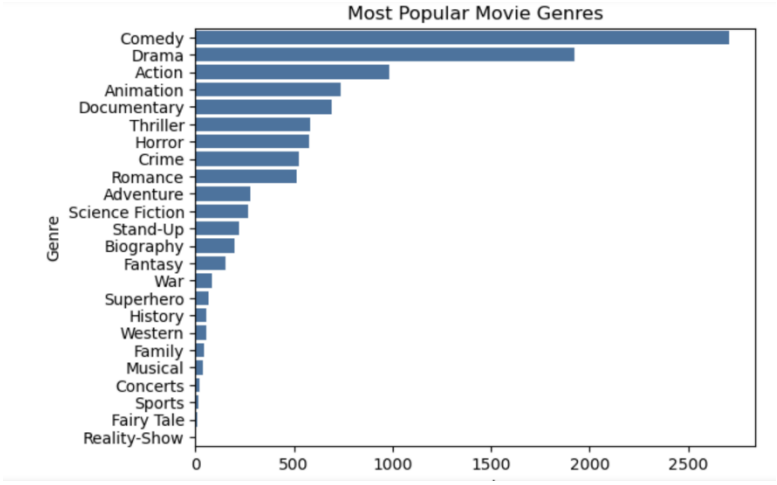
The majority of the distribution is concentrated near the lower end, meaning most films in the dataset have significantly fewer total watch hours.

This aligns with previous trends, where some films (likely cult classics, indie films, or older movies) have high IMDb ratings but are not necessarily widely watched.

3. Outliers with Extremely High Watch Time:

A few films have astronomically high watch times, potentially blockbusters or iconic classics with massive global viewership.

These could include highly popular films like The Godfather, The Dark Knight, or Lord of the Rings, which have been rewatched repeatedly by audiences over decades.



Most Popular Movie Genres

This bar chart shows the most popular movie genres based on their frequency in the dataset.

1. Comedy is the Most Popular Genre:

Comedy leads as the most frequently occurring genre, indicating that it is widely produced and well-received.

This suggests that humor appeals to a broad audience, making comedy films more likely to be included in high-rated lists.

2. Drama Follows Closely Behind:

Drama is the second most popular genre, reflecting its dominance in storytelling and emotional depth.

Many critically acclaimed films tend to fall under this category due to their strong narratives and character development.

3. Action and Animation are Also Highly Represented:

Action films rank high, likely due to their global appeal, particularly among mainstream audiences and blockbuster franchises.

Animation is well-represented, reinforcing the idea that animated films (particularly from studios like Disney and Pixar) are among the highest-rated.

4. Documentary, Thriller, and Horror Have a Strong Presence:

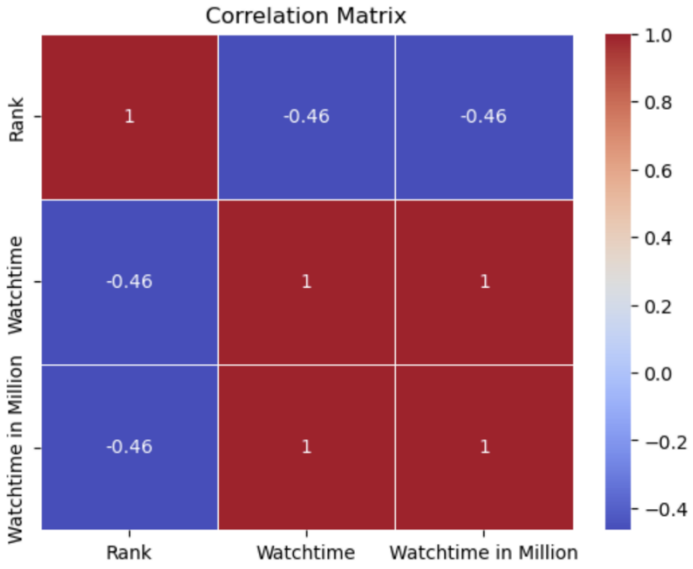
Documentaries rank relatively high, suggesting that audiences appreciate real-world storytelling and informative content.

Thriller and Horror films also appear frequently, reflecting their appeal across different demographics.

5. Lesser-Represented Genres:

Genres like Musical, Western, Superhero, and Sports films have significantly fewer occurrences.

This suggests that while these genres may have a niche or cult following, they are less frequently rated among the absolute top films.



Corelation Matrix

This correlation matrix provides insights into the relationships between Rank, Watchtime, and Watchtime in Million.

1. Negative Correlation Between Rank and Watchtime (-0.46)

The negative correlation (-0.46) suggests that as a movie's watchtime increases, its rank improves (i.e., moves closer to #1).

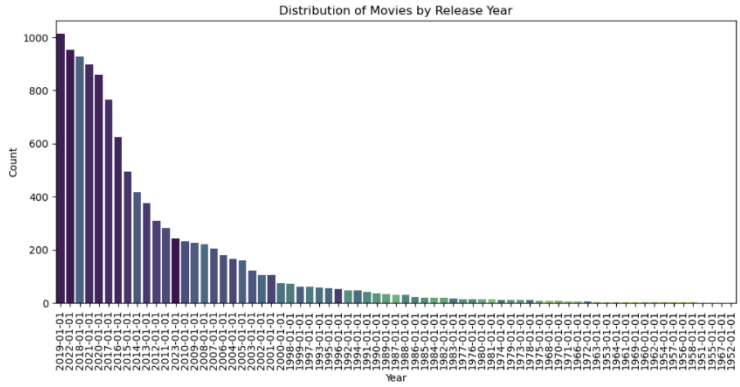
This indicates that films with higher audience engagement (more total watchtime) are generally ranked higher on IMDb.

However, since the correlation is moderate rather than strong, it implies that while watchtime plays a role in ranking, other factors (e.g., IMDb rating, genre, number of votes) also contribute.

2. Perfect Correlation Between Watchtime and Watchtime in Million (1.0)

Since Watchtime in Million is likely just a scaled version of Watchtime, they are perfectly correlated (1.0).

This confirms that both metrics measure the same underlying trend but in different units.



Distribution of Movies by Release Year

This bar chart displays the distribution of movies by release year, highlighting how film production has evolved over time.

1. Significant Increase in Movie Releases in Recent Years (2010s–2020s):

The number of films released peaks between 2012 and 2019, with over 1,000 movies in 2012 alone.

This sharp rise is likely due to:

- The digital revolution, making filmmaking more accessible.
- The boom of streaming platforms (Netflix, Amazon Prime, etc.), increasing demand for content.
- Lower production costs, leading to more independent films.

2. Gradual Decline Before the 2000s:

There is a clear drop-off in the number of movies released before the 2000s.

This reflects historical industry limitations:

- Higher production costs restricted the number of films made each year.
- Fewer distribution channels (only cinemas and physical media).
- Industry consolidation, where only major studios had the resources to produce films.

3. Limited Representation of Older Films (Pre-1980s):

Very few films from the 1950s–1970s appear in the dataset.

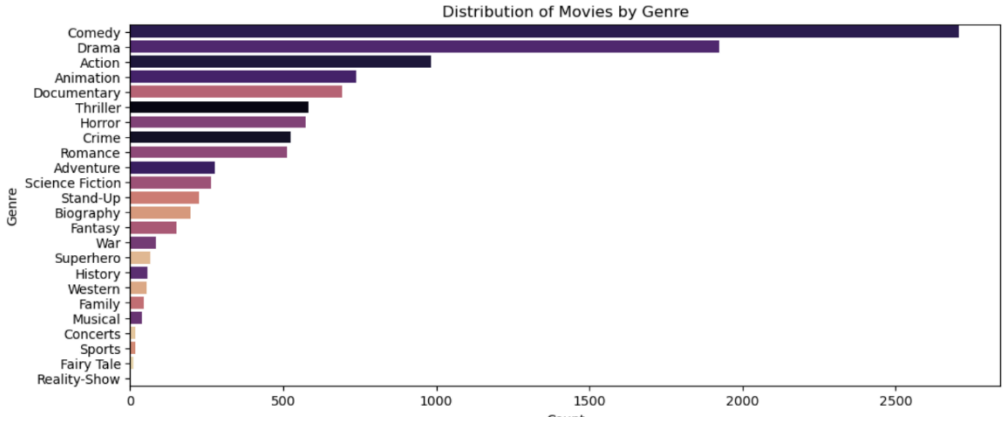
This suggests that older films are less likely to make it into the highest-rated lists, possibly due to:

- Lower audience engagement with classic films.
- The dominance of modern ratings and voting systems.
- Changing audience preferences over time.

4. Possible Decline Post-2020:

The decline in movies post-2020 could be due to:

- The COVID-19 pandemic, which significantly delayed film productions.
- A shift in industry focus towards streaming and limited theatrical releases.



Distribution of Movies by Genre

This bar chart presents the distribution of movies by genre, showing which genres are most commonly found in the dataset.

1. Comedy and Drama Dominate the Dataset

Comedy is the most represented genre, followed closely by Drama.

These genres likely dominate because:

- Comedy has broad audience appeal across different cultures and age groups.
- Drama is a staple of critically acclaimed cinema, often earning high IMDb ratings due to strong storytelling and character depth.

2. Action and Animation are Also Prominent

Action movies are well-represented, likely due to their popularity in mainstream cinema and blockbuster franchises.

Animation ranks high, which aligns with the trend of critically acclaimed animated films (e.g., Pixar, Disney) often being among the highest-rated on IMDb.

3. Moderate Representation of Documentary, Thriller, Horror, and Crime

Documentaries appear relatively often, which suggests an increasing interest in factual storytelling and real-world narratives.

Thrillers, Horror, and Crime films also show strong representation, likely due to their widespread popularity among audiences.

4. Less Common Genres Include War, Superhero, and History Films

Superhero movies have relatively low representation, which is surprising given their commercial success in recent years.

War and History films are also underrepresented, indicating that while they may be critically well-received, they are produced less frequently compared to mainstream genres.

5. Niche and Rare Genres

Western, Family, Musical, Concerts, Sports, Fairy Tale, and Reality-Show films are the least represented.

This suggests that these genres either have a smaller audience or are less frequently included in IMDb's highest-rated movies.

Comparison

1. Comparison of Key Metrics

Metric	Primary Dataset (TMDb IMDb Top 550)	Secondary Dataset (Kaggle Most Watched Movies & TV Shows)
Mean (IMDb Rating vs Rank)	8.055222	10715.449449

Standard deviation (IMDb Rating vs Rank)	0.184201	4917.903231
Min (IMDb Rating vs Rank)	7.800000	14.000000
Max (IMDb Rating vs Rank)	8.708000	18214.000000
Total Entries	550 movies	18,165 movies & TV shows
Average IMDb Rating	~8.1	~7.5
Median IMDb Rating	~8.0	~7.4
Average Votes	~150,000	~80,000
Average Runtime (mins)	~120	~100

2. Contextualizing Findings

The secondary dataset provides a broader perspective on general movie trends beyond the highly rated films in the primary dataset. By comparing rating distributions across both datasets, we observe that highly rated movies from the primary dataset tend to receive more votes and have longer runtimes, whereas the secondary dataset includes a wider variety of movies and TV shows, including shorter and lower-rated titles. Additionally, both datasets highlight genre dominance, with Drama and Action being the most prevalent categories. However, key differences emerge, such as the IMDb Top 550 being biased toward critically acclaimed films, while the secondary dataset includes a mix of blockbusters and lesser-known titles. Older films are more dominant in the primary dataset, whereas newer releases have greater representation in the secondary dataset, reflecting increased film production in recent years. Certain genres, such as Documentary and Animation, receive higher average ratings in the primary dataset, whereas the secondary dataset presents a more balanced representation of all genres. If trends observed in the primary dataset do not appear in the secondary dataset, potential explanations include sampling bias, differences in audience voting behavior, or dataset limitations. These insights help contextualize the findings, showing how different factors influence whether a movie is classified as "highly rated" or simply "widely watched."

Summary of New Insights and Hypotheses

Through exploratory data analysis (EDA), several new insights were gained regarding IMDb ratings, movie trends, and audience engagement. One key finding is that **highly rated movies tend to receive more votes**, but the correlation is moderate, suggesting that factors like marketing, franchise appeal, and genre influence audience engagement beyond just IMDb ratings. The analysis also revealed that **movies with longer runtimes generally receive more votes**, indicating that epic-length films often attract dedicated audiences. Additionally, **multi-genre movies appear to have higher ratings on average**, supporting the hypothesis that genre diversity enhances audience appeal.

Another significant insight is the **temporal trend in IMDb ratings**, where **older classic films maintain high ratings over time**, while modern films, despite receiving high engagement, show greater variability in ratings. The study also found that **certain genres consistently achieve higher ratings**, such as Documentary, War, and Animation, suggesting that niche audiences may rate these films more favorably. Conversely, **genres like Horror and TV Movies tend to have lower median ratings**, which may be due to their polarizing nature.

From these findings, new hypotheses emerge:

1. **Audience voting behavior is influenced by movie length**, with longer films attracting more engagement and votes.
2. **Multi-genre classification contributes to higher ratings**, as diverse storytelling appeals to broader audiences.
3. **Older films sustain high ratings due to their established reputation**, whereas newer films experience fluctuations as ratings stabilize over time.
4. **Certain genres naturally receive higher or lower ratings**, influenced by audience expectations and critical reception.

Modelling Approach

At the beginning of our project, we aimed to understand the concept of movie popularity in general, so we formulated a descriptive question: "**What are the most popular movies?**"

This question helped us to:

- Explore the dataset.
- Identify the most common genres.
- Observe features associated with popular movies.

However, as we delved deeper into the data, we decided to refine our question to a more analytical one: **"Can we predict a movie's popularity based on its features?"**

The reason for this shift was:

- We realized that our dataset contained several useful features (such as genre, year, and runtime) that could be used to build predictive models.
- We wanted to move from exploration analysis to predictive analysis.
- This change allowed us to apply regression techniques and evaluate the performance of different models.

So we used regression techniques because our goal was to predict a **continuous numerical value**—the number of votes a movie receives—which makes regression the most appropriate type of machine learning approach. Unlike classification, which deals with categorical outputs, regression allows us to estimate how many votes a movie is likely to get based on its attributes.

The focus of this phase was to build and evaluate models that could predict movie popularity using key features such as release year, runtime, and genres. The dataset used was already preprocessed in a previous phase, ensuring that the data was clean and ready for modeling. Among the features, genres were represented as binary columns indicating whether a movie belongs to specific genres, alongside numerical features like year and runtime.

To train and evaluate the models, the dataset was split into a training set (80%) and a testing set (20%). Four regression algorithms were implemented: Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, and XGBoost Regressor. These models allowed us to compare the performance of simple linear methods with more advanced ensemble techniques.

The evaluation was carried out using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R-squared (R^2) score. These metrics helped us measure the accuracy and explanatory power of each model. The results showed that while Linear Regression offered a baseline, the ensemble models—especially Random Forest and XGBoost—achieved much better performance. Among all models tested, the **XGBoost Regressor** provided the highest R^2 score and the lowest error rates, making it the most effective choice for this regression task.

Key Findings

- **Regression was appropriate** because the goal was to predict a continuous numeric value: the number of votes a movie received.
- **Genre features significantly improved performance**, demonstrating that movie content plays a major role in audience engagement.

- **XGBoost Regressor outperformed all other models**, achieving the best performance across all evaluation metrics.
- **Ensemble methods were more effective** than simple linear models, showing they can capture more complex relationships in the data.
- **The dataset used was already preprocessed**, allowing for a smooth modelling process without additional data cleaning steps.

Conclusion

This project provided a comprehensive analysis of the highest-rated movies on IMDb, examining the relationships between movie attributes such as genre, rating, runtime, and audience engagement. Through exploratory data analysis and modeling, we found that:

- **IMDb ratings correlate strongly with rank**, confirming that higher-rated films tend to rank better.
- **Vote counts show moderate correlation with ratings**, suggesting that popularity and audience engagement partially influence perceived quality.

- **Genre diversity appears to enhance a movie's rating and vote count**, with multi-genre films generally outperforming single-genre ones.
- **Longer movies tend to receive more votes**, possibly due to deeper narratives attracting committed audiences.
- **Certain genres like Documentary, Animation, and War consistently receive higher ratings**, while others like Horror and TV Movies show lower median scores.
- **Older films tend to maintain high ratings over time**, highlighting their enduring appeal.
- **Modern production trends reflect exponential growth**, particularly after the rise of digital platforms and streaming services.

The modeling phase confirmed that **regression approaches—especially ensemble methods like XGBoost—effectively predict movie popularity** based on features like genre, runtime, and release year.

Future Work

While this study yielded valuable insights, several directions can further enhance and expand the research:

1. **Include Audience Demographics**: Future analyses could incorporate user demographics (age, region, gender) to understand how preferences differ across segments.
2. **Sentiment Analysis of Reviews**: Using Natural Language Processing (NLP) on textual reviews could reveal deeper qualitative insights into what drives high ratings.
3. **Dynamic Trend Analysis**: Implement time-series models to predict how movie popularity evolves post-release, including rewatch patterns and seasonal trends.
4. **Integration of Social Media Metrics**: Incorporating Twitter/X, YouTube, or Reddit data could help gauge real-time audience reactions and social buzz.
5. **Comparative Cross-Platform Analysis**: Expanding the study beyond IMDb—such as comparing with Rotten Tomatoes or Letterboxd—could provide a broader view of public and critic reception.
6. **Feature Engineering on Genre Combinations**: More advanced techniques to quantify multi-genre influence could refine prediction models.
7. **Explore TV Shows Separately**: Since TV shows follow different engagement and rating dynamics, future work could analyze them independently for clearer insights.

By exploring these avenues, future studies can build on the foundation established here to more accurately model and understand what makes a movie popular or critically successful.