Mr. Avisek Regmi
Avenue de la Foretaille 27b
Chambesy, 1292
E-mail: avisek.regmi@gmail.com

CAS Applied Data Science Final Project

Reel Insights: A Comprehensive Analysis of Movie Recommendation Systems

Alphas

2011-2012 TV-14 2 Seasons

A team of individuals gifted with extraordinary neurological abilities is tasked with solving a series of high-profile crimes. Among them are an autistic man whose mind works as fast as a computer and an FBI agent with super strength. More Info

Starring: David Strathairn, Ryan Cartwright
Creators: Zak Penn, Michael Karnow

Date: 29th May, 2024

Table of Contents

Contents

1. Abstract

This project delves into the multifaceted realm of cinema through two distinct but interconnected parts. Firstly, it embarks on a cinematic journey, unraveling the rich tapestry of film history, trivia, and facts by employing data-driven methodologies. Extensive Exploratory Data Analysis (EDA) is conducted on diverse aspects of movie metadata, including revenues, casts, crews, and budgets across different eras. Additionally, predictive models are crafted to forecast movie revenues and success, shedding light on the pivotal features influencing these outcomes. Secondly, the project pivots towards the development and evaluation of movie recommender systems, encompassing Simple Generic, Content-Based, and User-Based Collaborative Filters. These systems are scrutinized through both qualitative and quantitative lenses, unveiling their efficacy in facilitating personalized movie recommendations. While the first section caters to a broad audience within the movie-making industry, offering invaluable insights for streaming providers, producers, and stakeholders, the latter section holds relevance for businesses reliant on recommendation mechanisms such as Amazon, Netflix, and Hotstar. By enhancing revenue generation, user engagement, and overall user experience, these recommendation engines emerge as indispensable tools for navigating the ever-evolving landscape of cinematic consumption.

2. Introduction

In an age where data serves as a guiding beacon across industries, the world of cinema stands no exception. The allure of storytelling through motion pictures has captivated audiences for decades, transcending boundaries of time and space. However, amidst the glitz and glamour lies a realm ripe for exploration through the lens of data analytics. This project endeavors to unravel the enigma of cinema, peering into its past, deciphering its present, and perhaps even forecasting its future.

The project unfolds in two distinct acts, each offering a unique perspective on the cinematic landscape. Act one, aptly titled **"The Story of Film,"** embarks on a historical odyssey, tracing the evolution of cinema through the annals of time. Through meticulous analysis of movie **metadata** spanning **genres, decades,** and **continents**, I aim to unearth hidden patterns, trends, and insights that shape the fabric of cinematic history. From the humble beginnings of silent films to the blockbuster juggernauts of today, every frame tells a story, and every dataset holds a clue.

Act two, **"Movie Recommender Systems,"** shifts focus to the intersection of art and technology, where algorithms wield the power to curate personalized cinematic experiences. In an era inundated with content, the ability to navigate the vast expanse of cinematic offerings becomes paramount. Through the development and evaluation of recommendation engines, I endeavor to streamline this journey, offering tailored suggestions based on user preferences, content similarities, and collaborative filtering techniques. As the digital landscape continues to evolve, so too must my methods of content discovery, ensuring that audiences find not just what they seek, but what they never knew they desired.

As I embark on this cinematic voyage, I invite you to join us in uncovering the hidden gems, decoding the mysteries, and embracing the magic of storytelling through data. From the silver screen to the digital frontier, the journey promises to be as exhilarating as the stories themselves. Welcome to the intersection of art and analytics, where every data point is a plot twist waiting to be discovered.

3. Data

The data utilized in this project has been meticulously curated from two primary sources: The Movie Database (TMDB) and MovieLens, both renowned repositories within the realm of cinematic analysis.

MovieLens, a bastion of film ratings and reviews, offers a treasure trove of insights derived from a vast dataset. Boasting a staggering compilation of 26 million ratings and 750,000 tag applications across 45,000 movies, contributed by a diverse cohort of 270,000 users, it serves as a cornerstone for empirical analysis. Additionally, the inclusion of tag genome data, comprising 12 million relevance scores across 1,100 tags, enriches the dataset, providing nuanced perspectives on movie categorization and user preferences. Notably, a condensed subset of this expansive dataset, featuring 10,000 ratings for 9,000 movies from 700 users, further refines my analytical lens.

Augmenting this rich repository is data sourced from TMDB, facilitated by the provision of TMDB IDs within the MovieLens dataset. Leveraging these unique identifiers, a comprehensive spectrum of metadata, credits, and keywords for all 45,000 movies was meticulously extracted using a custom script interfacing with TMDB's Open API. Initially structured in JSON format, this wealth of information was seamlessly transposed into CSV files, leveraging Python's versatile Pandas Library, thereby facilitating streamlined data manipulation and analysis.

Central to my analytical endeavor are several pivotal files:

1. **movies_metadata.csv**: This repository encapsulates a plethora of metadata sourced from TMDB, encompassing vital attributes such as budget, revenue, release dates, and genres, among others, for over 45,000 movies.
2. **credits.csv**: A comprehensive dossier detailing the credits associated with each movie, ranging from directors and producers to actors and characters, thus providing holistic insights into the creative ensemble behind each cinematic masterpiece.
3. **keywords.csv**: A repository housing an exhaustive catalog of plot keywords intricately linked with each movie, enabling nuanced exploration of thematic elements and narrative motifs.
4. **links_small.csv**: This file serves as a compendium of movies featured within the condensed subset of the Full MovieLens Dataset, facilitating focused analysis on a manageable scale.
5. **Ratings_small.csv**: The cornerstone of my collaborative filtering analysis, this dataset encompasses 100,000 ratings spanning 9,000 movies, contributed by 700 discerning users, thereby serving as the bedrock for personalized recommendation systems.

In amalgamating these diverse datasets, my aim is to unravel the intricacies of cinematic storytelling, decode user preferences, and forge pathways towards enhanced cinematic experiences through data-driven insights and analysis.

4. Data Collection

The data collection process for this project involved sourcing comprehensive datasets from both the MovieLens and TMDB platforms, ensuring a robust foundation for subsequent analysis.

4.1 MovieLens Dataset

The primary dataset was obtained from the GroupLens website, a well-known repository for movie ratings and metadata (https://grouplens.org/datasets/movielens/). The MovieLens Full Dataset, which includes 26 million ratings from 270,000 users across 45,000 movies, served as a pivotal resource. Within this extensive dataset, the file links.csv provided essential TMDB and IMDB IDs for all listed movies, facilitating a seamless integration with supplementary data sources

4.2 TMDB API Integration

To augment the MovieLens data with richer metadata, I registered for an API Key with TMDB, gaining access to detailed information through three critical endpoints: movie details, cast and crew information, and plot keywords. Each endpoint provided unique insights into various facets of the 45,000 movies, necessitating a systematic approach to data extraction.

4.3 Data Scraping and Processing

Given the scope of the data required, I developed three distinct scrapers to interface with each TMDB endpoint. These scrapers were designed to methodically request and retrieve data for all listed movies. However, TMDB's API imposes a rate limit of 40 requests every 10 seconds, which necessitated a strategic approach to avoid exceeding these constraints. As a result, the complete data extraction process spanned an entire day.

Upon retrieval, the data was initially in stringified JSON format. This raw format required substantial processing to render it usable for analysis. Using Python's Pandas library, the JSON data was meticulously converted into structured CSV files, ensuring that the dataset was not only comprehensive but also readily accessible for subsequent exploratory and predictive analyses.

This rigorous data collection process laid the groundwork for an in-depth exploration of movie metadata, enabling a nuanced understanding of the cinematic landscape and facilitating the development of sophisticated recommendation systems.

5. Data Wrangling

5.1 Removing Unnecessary Features

To streamline the dataset and reduce its dimensionality, certain superfluous features were removed. Attributes such as Backdrop Path, Adult, and IMDB ID were deemed non-essential for my analysis and were thus dropped. This step helped in focusing on the most relevant features, enhancing the efficiency of my data processing workflows.

5.2 Cleaning

The cleaning process involved several key steps to handle missing and malformed data:

1. **Handling Missing Values:** Many features contained values of 0, indicating the absence of data. These placeholders were converted to NaN (Not a Number) to accurately represent missing values and facilitate their appropriate handling in subsequent analyses.

2. **Parsing Stringified JSON Objects:** Some features were still in the form of stringified JSON objects. Using Python's ast library, these were converted into Python dictionaries. To further simplify the dataset, these dictionaries were reduced to lists by retaining only the necessary attributes, discarding IDs, timestamps, and other extraneous information.

3. **Exploding DataFrames:** For features such as genres and production countries, where each movie could have multiple entries, the DataFrame was "exploded." This process transformed lists within cells into separate rows, allowing for more granular and comprehensive analysis of these multi-valued attributes.

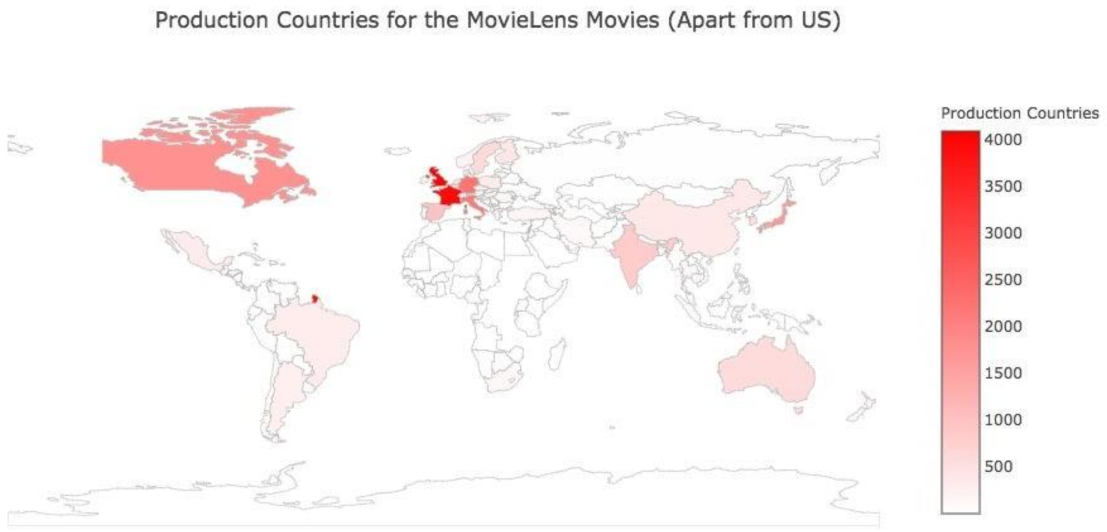
4. **Type Conversion:** To ensure consistency and usability, most features were converted into basic Python data types such as integers, strings, and floats. Dates, initially in string format, were converted into Pandas Datetime objects. From these datetime objects, additional temporal features like month, year, and day of release were extracted, enriching the dataset with more analytical dimensions.

Through these data wrangling steps, the raw datasets were transformed into a clean, structured, and highly functional format. This refined dataset now serves as a robust foundation for subsequent exploratory data analysis and the development of predictive models, enabling deeper insights into the world of cinema.

6. Exploratory Data Analysis and Visualization

In this part, I share the main findings from basic statistics and charts. These findings are central to the first section of my Movie Recommendation System Project, showing interesting patterns and trends in the movie industry.

6.1 Geographical Distribution of Movies


Production Countries for the MovieLens Movies (Apart from US)
Production Countries
4000
3500
3000
2500
2000
1500
1000
500

1. **Dominance of English-Language Films**: The dataset reveals that the majority of movies are produced in the English language and primarily shot in the United States. This underscores Hollywood's significant influence on global cinema.
2. **Prominence of European Cinema**: Europe emerges as a major hub for film production, with the United Kingdom, France, Germany, and Italy ranking among the top five countries. This highlights the rich cinematic heritage and active film industries within these nations.
3. **Leading Asian Markets**: Japan and India are the most prolific Asian countries in terms of movie production. Their vibrant film industries contribute significantly to global cinema, offering diverse and culturally rich storytelling.

6.2 Franchise Movies

1. **Top-Earning Franchises**: The Harry Potter franchise stands out as the most successful movie series, grossing over $7.707 billion from eight films. Close on its heels is the Star Wars franchise, which has garnered $7.403 billion from an equal number of movies.
2. **Remarkable Single-Movie Franchise**: The Avatar collection, despite comprising only one film at present, has amassed nearly $3 billion, making it the highest-grossing single-movie franchise. Nonetheless, the Harry Potter series remains the most lucrative franchise with at least five films.
3. **Largest Franchise by Volume**: The James Bond series is the largest movie franchise in terms of the number of films, boasting over 26 releases. Following at a distance are the Friday the 13th and Pokemon franchises, with 12 and 11 movies respectively.

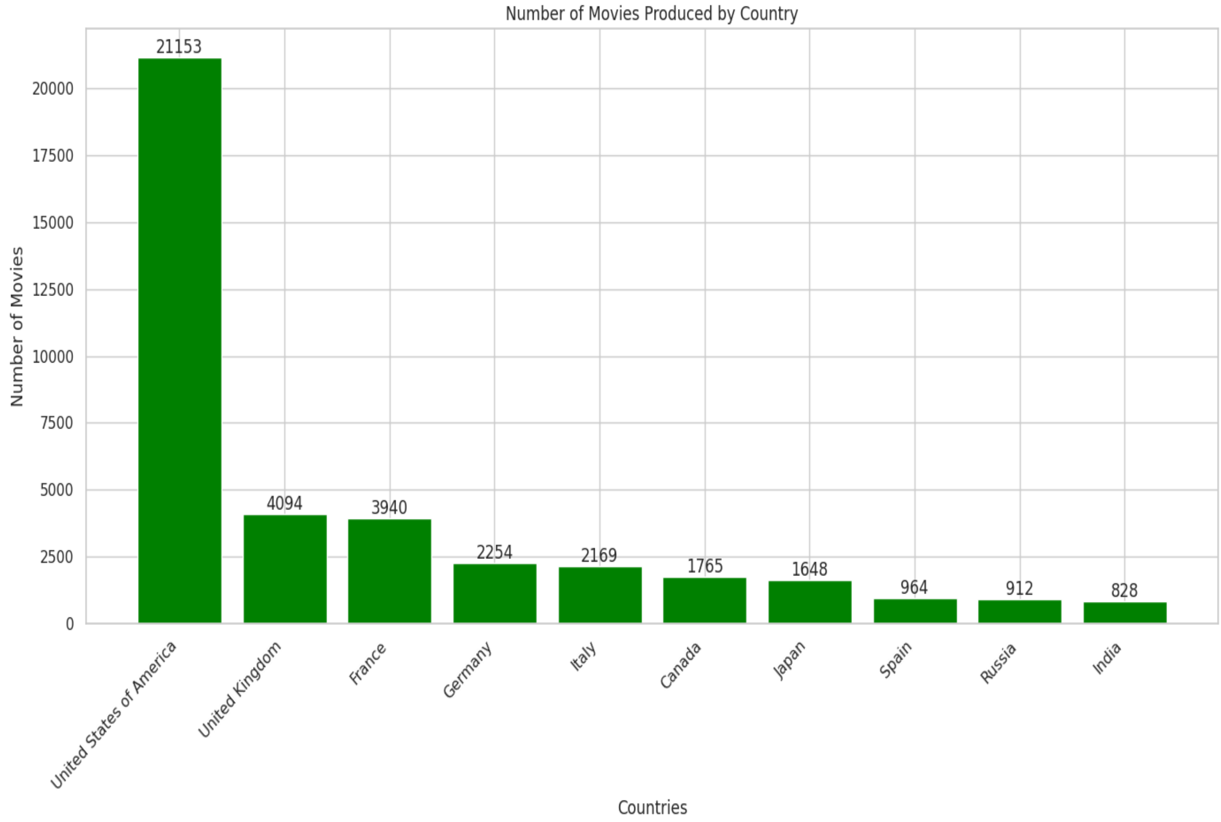
6.3 Production Companies

1. **Top-Grossing Production Companies**: Warner Bros emerges as the highest-earning production company, with a staggering $63.5 billion from nearly 500 movies. Universal Pictures and Paramount Pictures follow, with revenues of $55 billion and $48 billion, respectively.
2. **Most Successful Movies on Average**: Pixar Animation Studios leads in terms of average movie success, attributed to its impressive portfolio of beloved films such as *Up*, *Finding Nemo*, *Inside Out*, *Wall-E*, *Ratatouille*, the *Toy Story* series, and the *Cars* series. Marvel Studios ranks second, with an average gross of $615 million per film, thanks to blockbuster hits like *Iron Man* and *The Avengers*.

7. Data Visualization

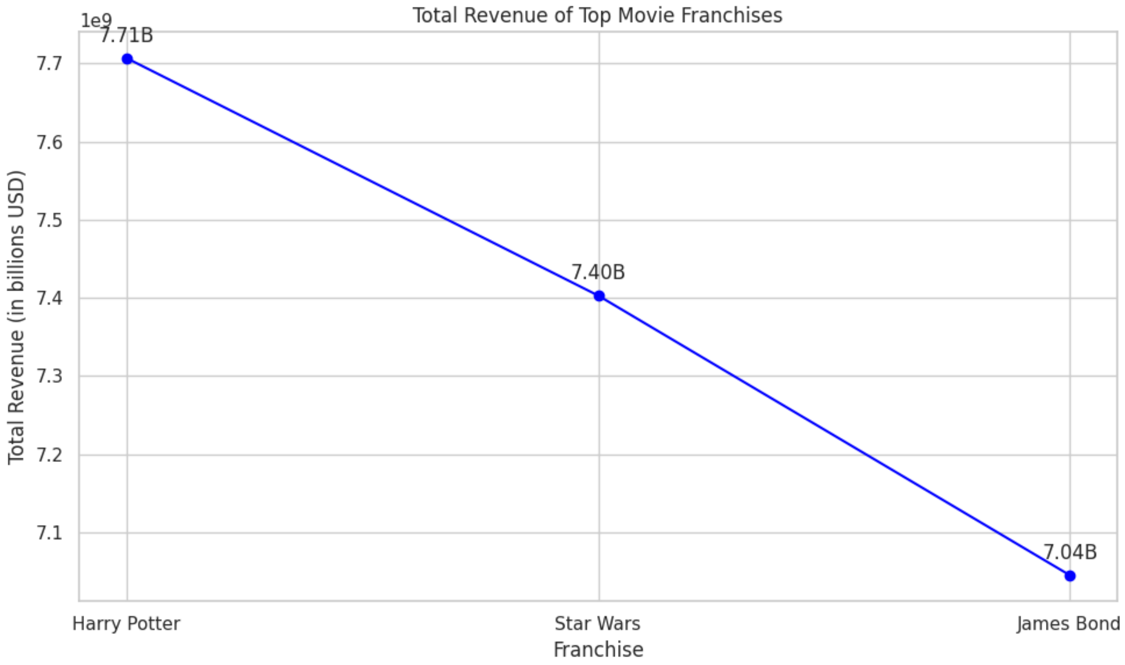
7.1 Geographical Distribution Chart:

The below bar chart is illustrating the number of movies produced in various countries, highlighting the dominance of the United States and notable contributions from European and Asian countries.



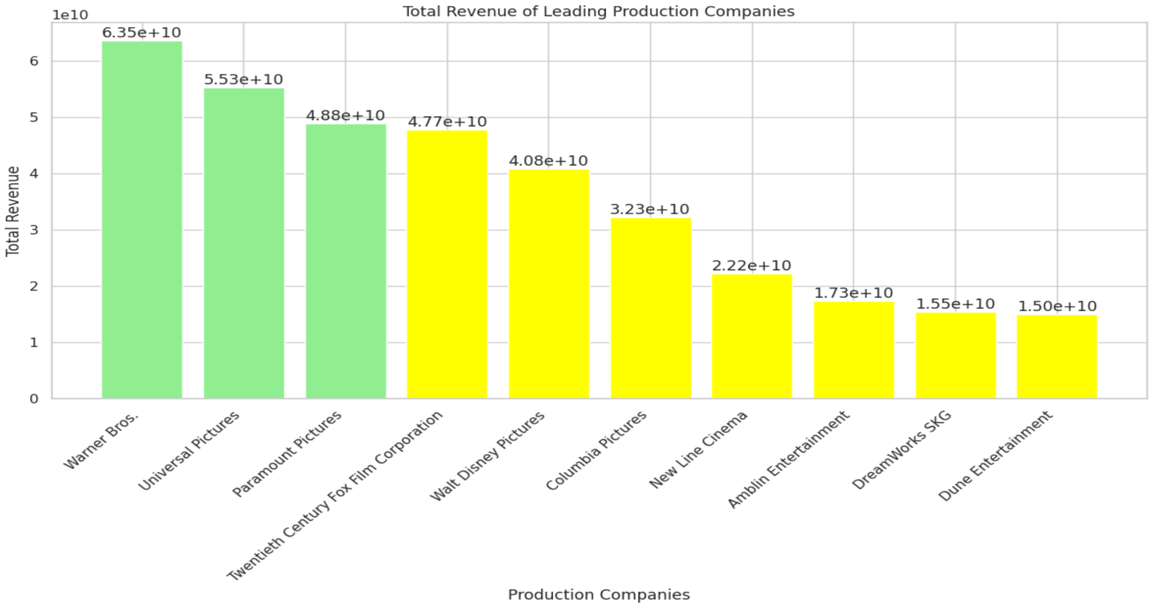
7.2 Franchise Revenue Graph:

A line graph comparing the total revenues of top movie franchises, showcasing the financial success of Harry Potter, Star Wars, and Avatar.



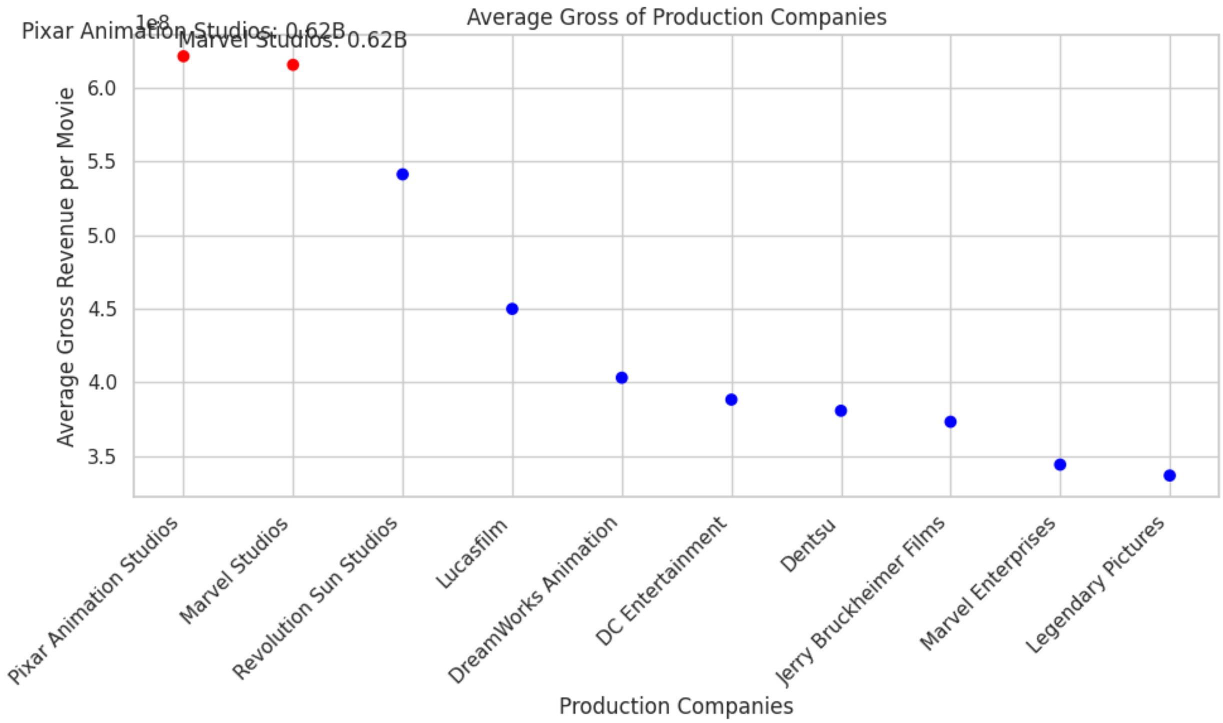
7.3 Production Company Earnings Bar Chart:

Below is a bar chart detailing the revenues of leading production companies, emphasizing the dominance of Warner Bros, Universal Pictures, and Paramount Pictures.



7.4 Average Gross of Production Companies:

Below is a scatter plot displaying the average gross revenue per movie for various production companies, underscoring the exceptional performance of Pixar Animation Studios and Marvel Studios.



These visualizations and analyses provide a comprehensive overview of the movie industry's landscape, revealing the geographical distribution of film production, the financial success of major franchises, and the dominance of top production companies. Through this exploration, I gain valuable insights into the factors driving success in the world of cinema.

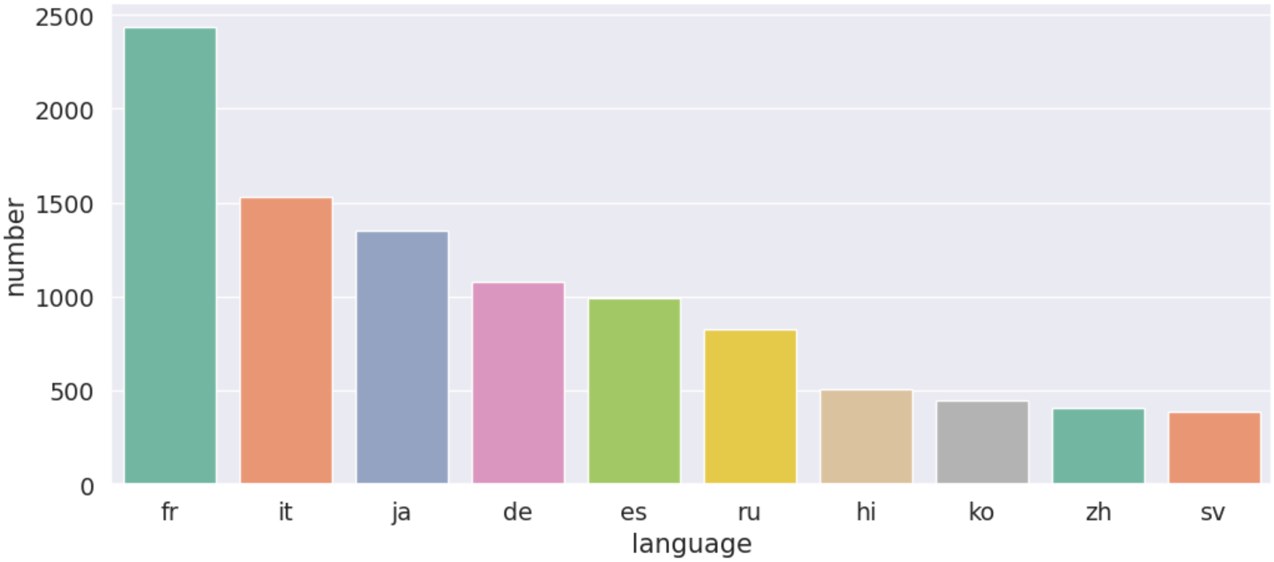
7.5 Movie Title Wordcloud

The image below showcases a wordcloud visualization of movie titles, revealing that **"Love"** stands out as the most frequently used word, emphasizing the dominance of romantic themes in cinema. Additionally, words like **"Girl," "Day,"** and **"Man"** appear prominently, highlighting the recurring focus on personal and everyday stories in movie narratives. This visualization beautifully captures the essence of storytelling in the film industry.

Love
Man
Girl
Night
Day
Life
Time
Story
Dead
One
Christmas
Little
Movie
World
Death
Last
Black
House
Blood
Boy
Heart
Street
Sister
Zombie
Hour
Monster
America
River
Never
Run
Legend
Journey
Adventure
Prince
Dance
Rise
Queen
Rose
Mother
High
Son
Doctor
Back
Le
Miss
Hell
Midnight
Tale
Take
Terror
Friend
Wild
Dawn
Bride
Sky
Three
Paradise
Lady
Four
Baby
Mystery
American
Will
Beautiful
People
King
New
Evil
Fall
Sweet
Big
Planet
Gun
Eye
Two
Hero
Island
Jack
War
Without
City
Magic
Cat
Meet
Brother
Children
Darkness
Room
Case
Murder
Woman
First
Dark
Father
Living
Die
Space
Ghost
Mountain
Kid
Hill
Go
Kill
God
Land
Crazy
Face
Star
Seven
Bad
Wedding
Crime
Great
Long
Hollywood
Vampire
Killer
Game
Dragon
Beyond
Angel
Water
Battle
Green
Family
Happy
Hunter
Men
El
Wind
Rock
Zero
Mr
U
VIII
Warrior
Master
White
Princess
Kiss
Devil
Five
Thing
Part
Perfect
Let
Cop
Best
Sex
Hand
Dog
Blue
Moon
Stranger
Shadow
Town
Women
Live
Secret
Wife
La
Year
Film
Affair
de
Paris
Home
New York
Dream
Daughter
Good
Lost
Summer
Call
S
End
II
Sun
Young
Earth
Fear
Come
Beast
Song
Fire
Sea
Party
Soul
Red
Way
Lover
Light
Final
Road
West
Country
Heaven
Return

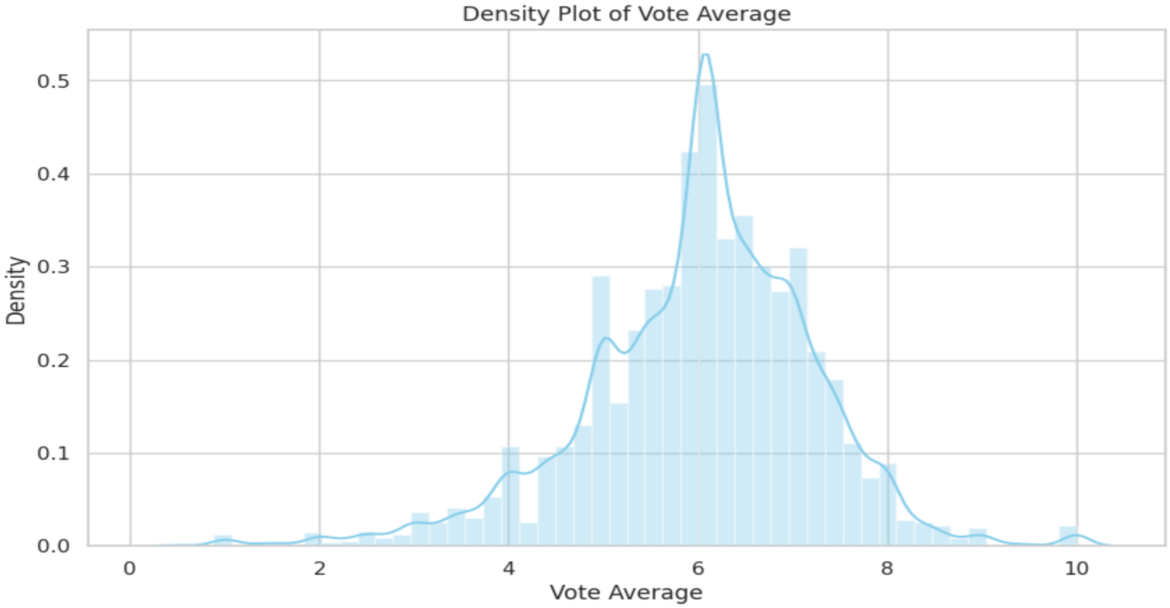
7.6 Original Languages

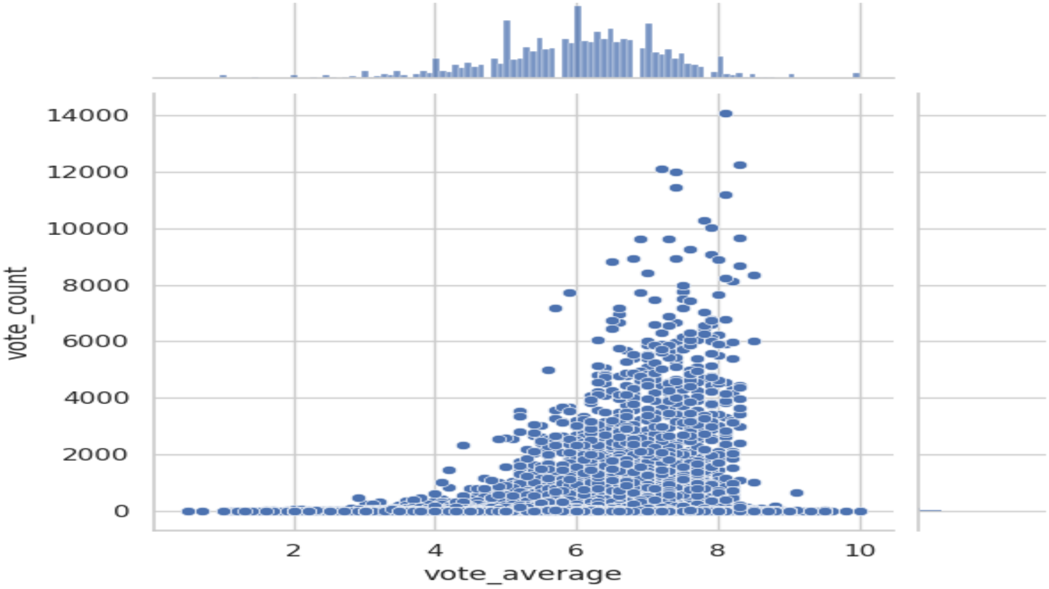
The bar chart below illustrates the diverse linguistic landscape of the movie industry, featuring films in over 93 languages. English-language movies dominate the dataset, underscoring Hollywood's significant influence on global cinema. Following English, French and Italian films are the most prevalent, with Japanese and Hindi leading the representation of Asian languages. This distribution highlights the rich variety of languages that contribute to the world of movies.



7.7 Popularity, Vote Average, and Vote Count

1. **Most Popular Movies**: According to the TMDB Popularity Score, "Minions" tops the list, followed by "Wonder Woman" and "Beauty and the Beast." These popular titles indicate a trend towards family-friendly and female-centric films.
2. **Most Voted On Movies**: Christopher Nolan's films "Inception" and "The Dark Knight" have garnered the highest number of votes, reflecting their widespread acclaim and popularity.
3. **Top-Rated Movies**: "The Shawshank Redemption" and "The Godfather" are the highest-rated movies in the TMDB database, mirroring their positions in IMDB's Top 250 Movies list. Both films boast ratings over 9 on IMDB, compared to 8.5 on TMDB.



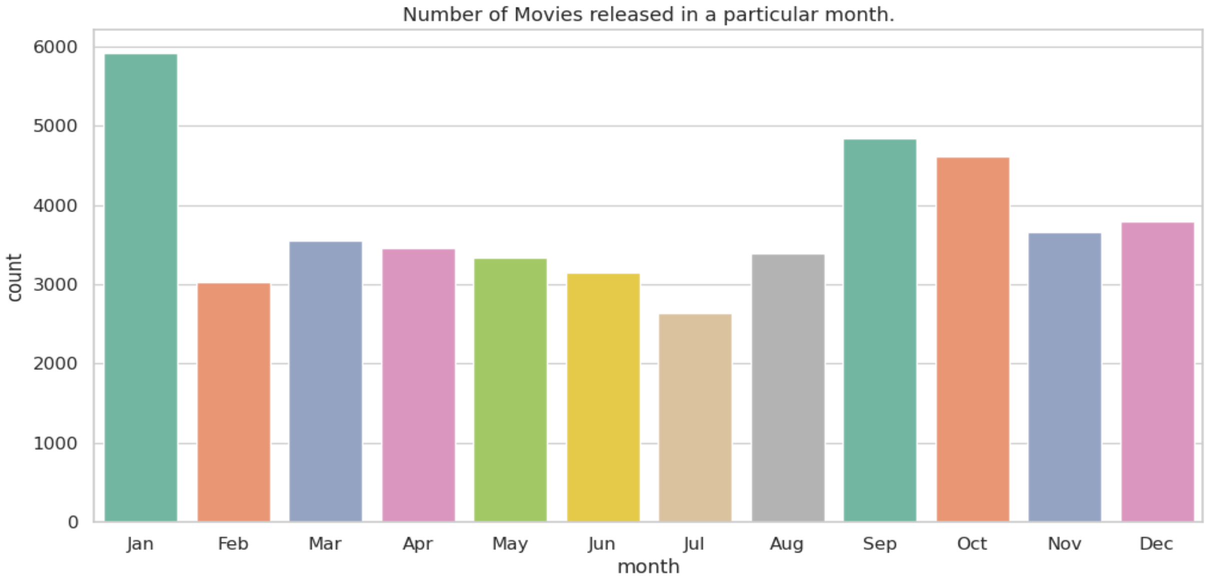


4. **Correlation Insights**: Interestingly, the Pearson correlation coefficient between popularity and vote average is a mere 0.097, indicating no significant correlation. Similarly, there is a minimal correlation between vote count and vote average, suggesting that a high number of votes does not necessarily correlate with better movie ratings.

7.8 Movie Release Dates

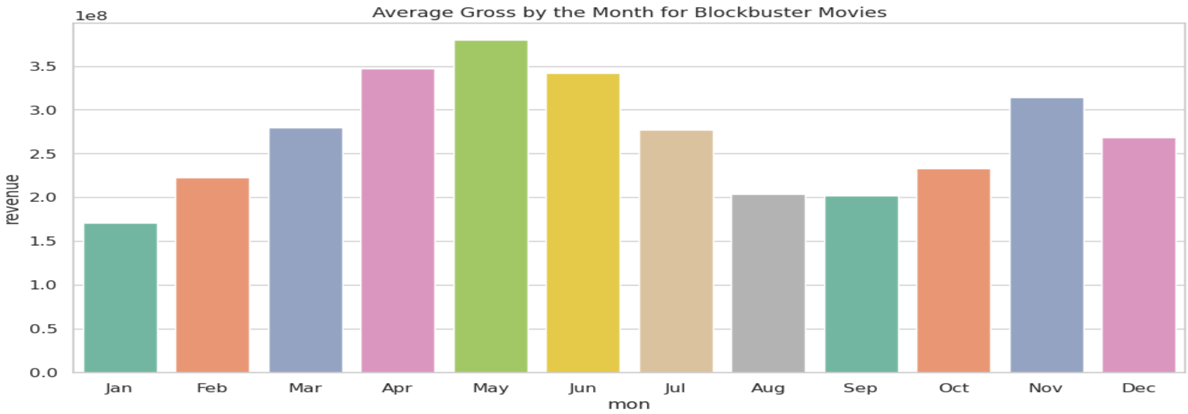
7.8.1 Monthly Trends:

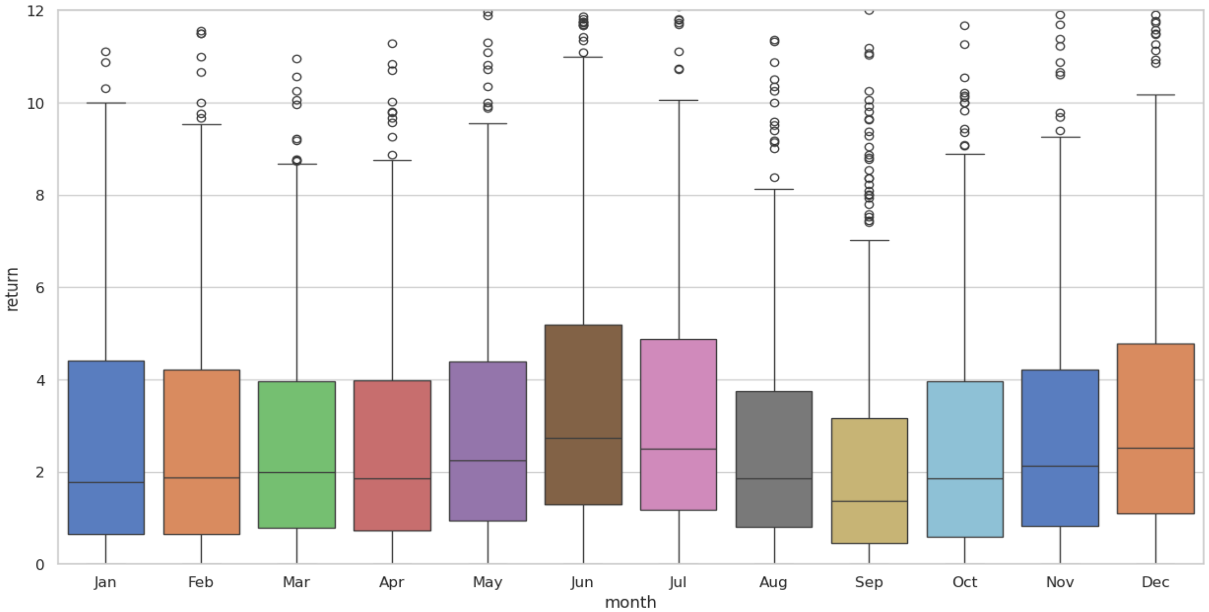
The bar chart below shows January is the most popular month for movie releases, often considered the "dump month" for subpar films. In contrast, April, May, and June see the highest average grosses, attributed to the release of blockbuster movies during the summer vacation period.



7.8.2 Median Returns:

The bar chart below shows movies released in June and July achieve the highest median returns, while September has the lowest. The success of summer releases can be linked to increased audience availability during vacations, whereas September marks the start of the academic year, leading to reduced movie consumption.

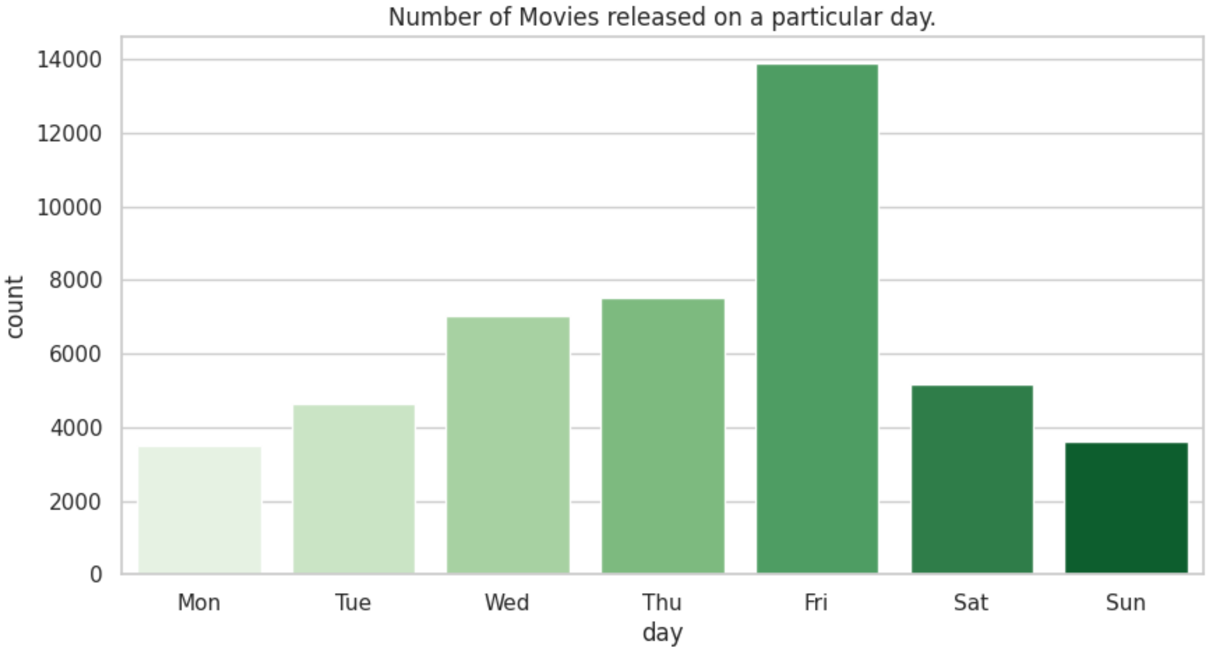




The above box plot shows June and July consistently reign with the highest median returns, contrasting with September, the least successful month by these standards. The success of summer releases in June and July is tied to the vacation season, while September's decline correlates with the onset of the school year, leading to a dip in movie consumption.

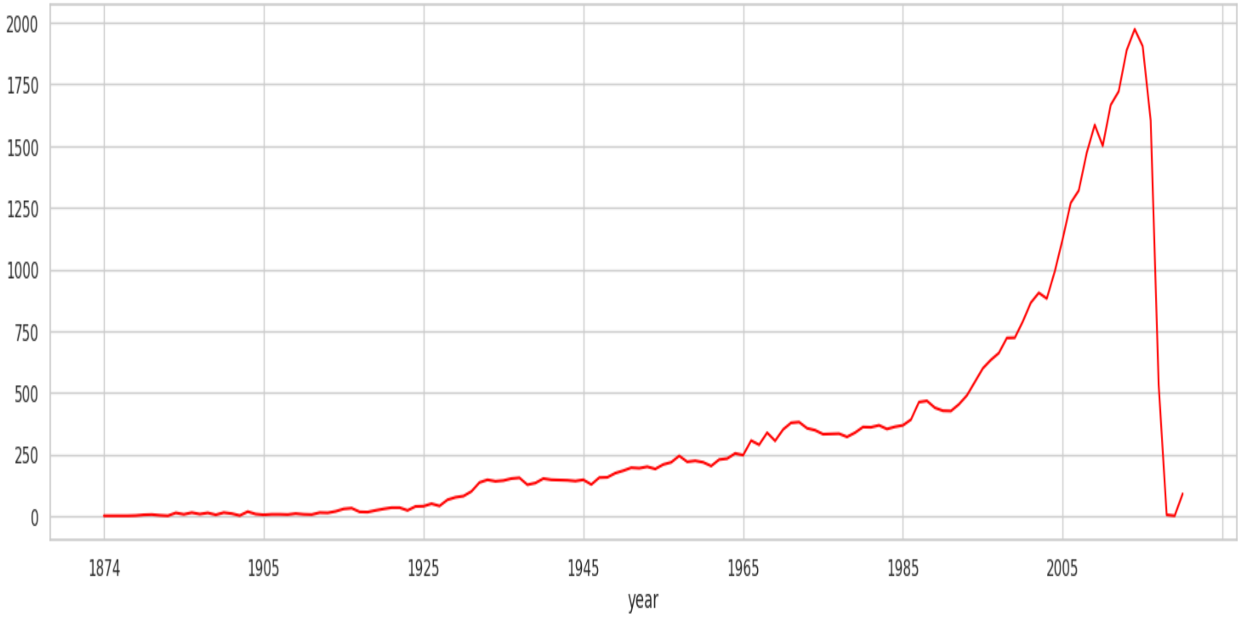
7.9 Day of Release:

The below bar charts shows Friday is the most popular day for movie releases, coinciding with the start of the weekend, while Sunday and Monday are the least popular.



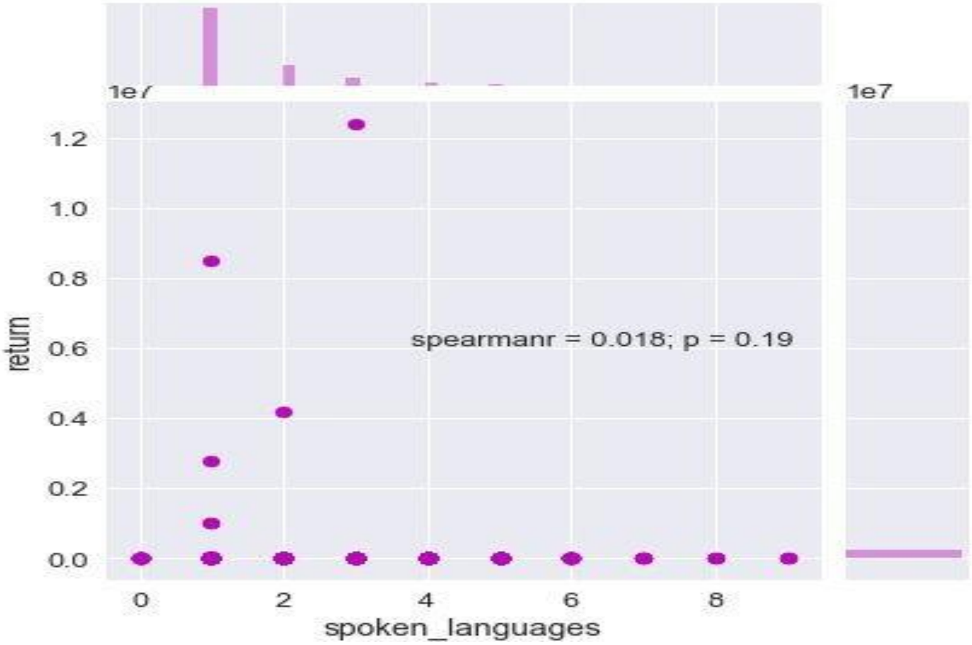
7.10 Historical Movie Release:

The below graph shows the oldest movie in the dataset is "Passage of Venus," a series of photographs from 1874, showcasing early cinematic endeavors.



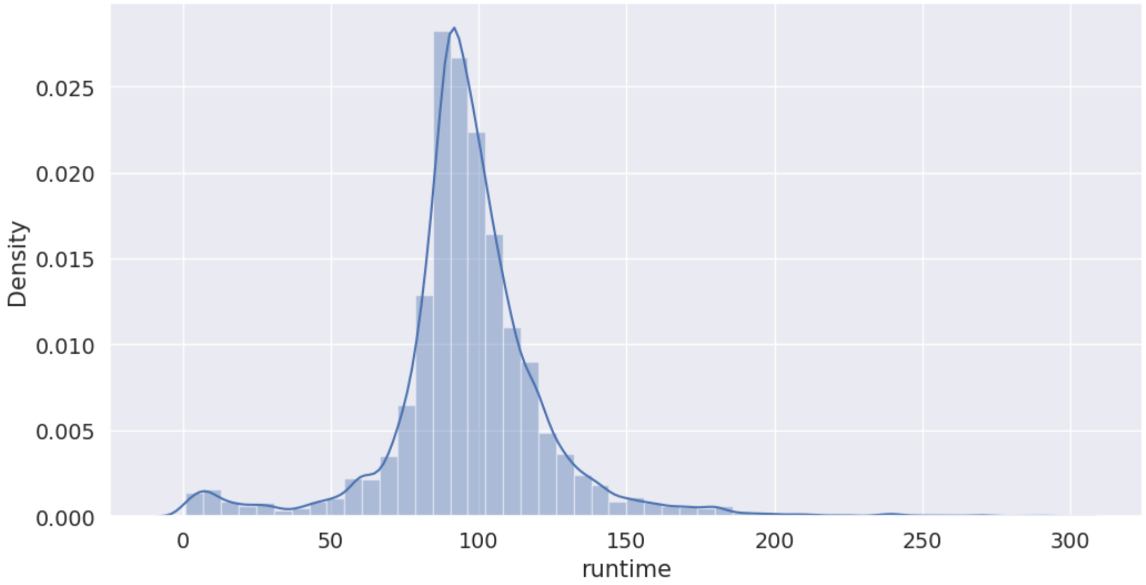
7.11 Spoken Languages

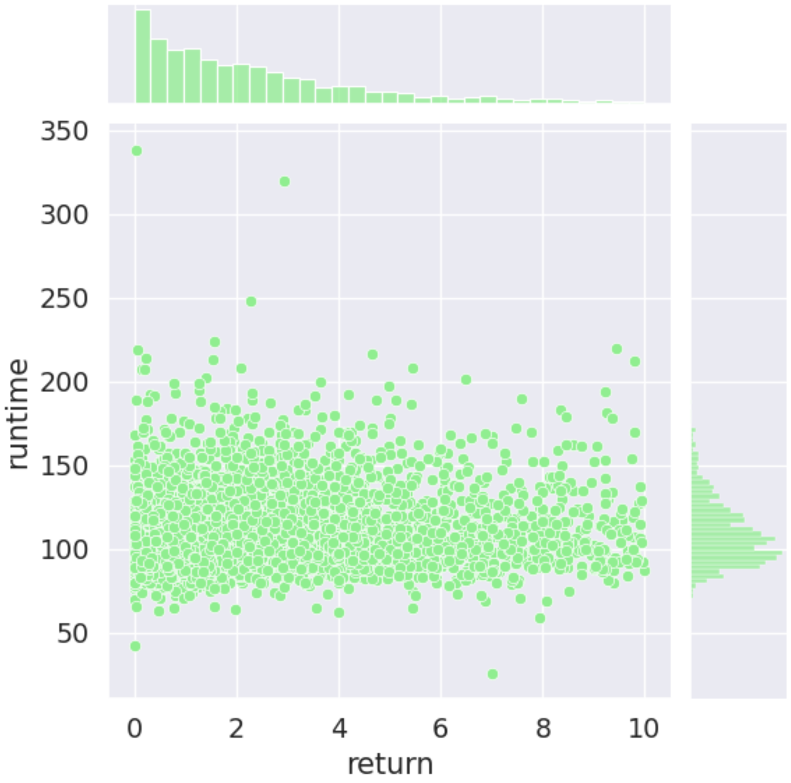
The below graphs shows "Visions of Europe," featuring 25 short films by different directors, stands out for having the most languages. This diversity reflects the collaborative nature of the project. There is no significant correlation between the number of languages spoken in a movie and its financial returns.



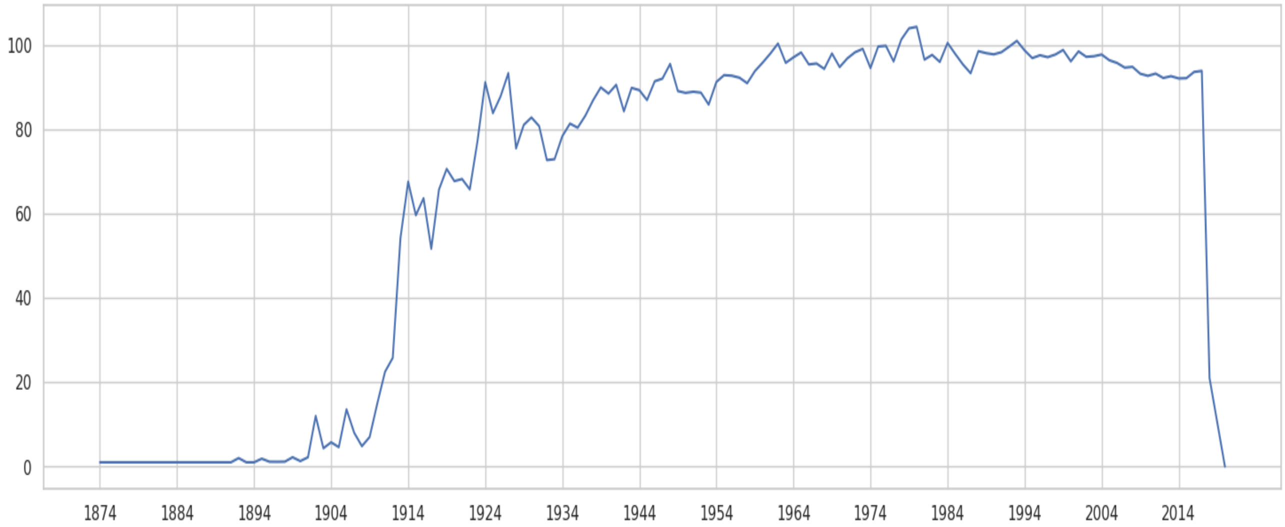
7.12 Runtime

The average movie runtime is approximately 1 hour and 30 minutes. The longest movie in the dataset is 1256 minutes (or 20 hours) long. There is no apparent relationship between runtime and movie success, as the duration does not significantly impact returns.



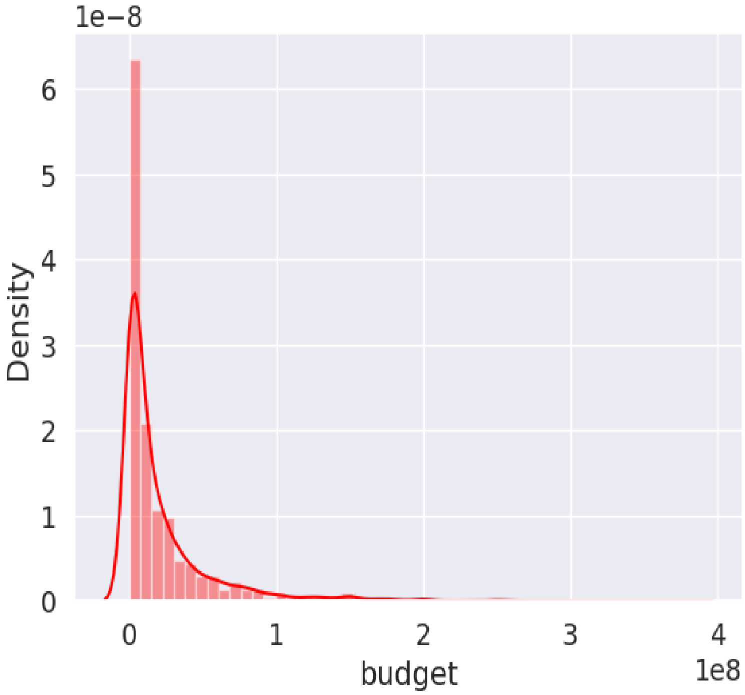


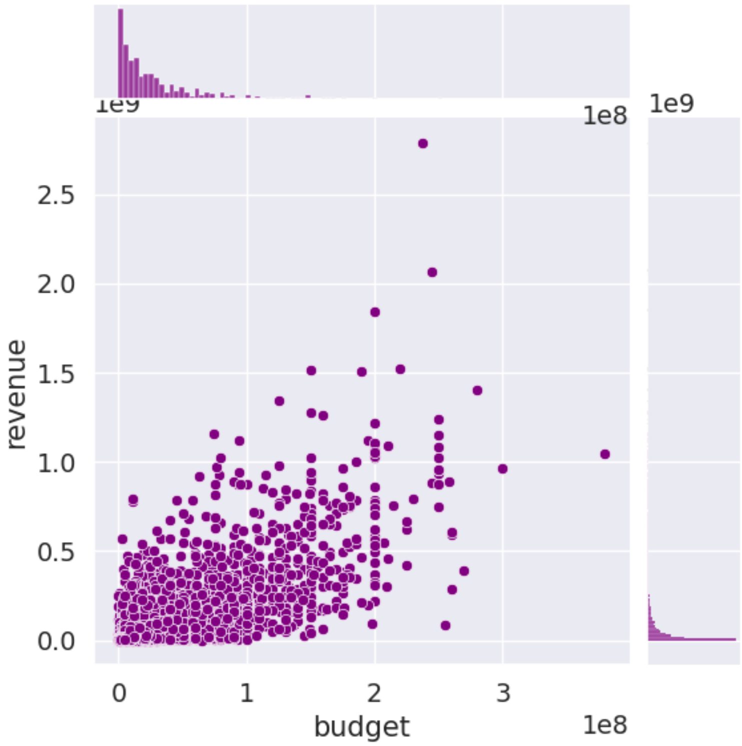
The below graph shows films began reaching the 60-minute mark as early as 1914. By 1924, the standard runtime extended to 90 minutes, a length that has remained relatively constant ever since. This evolution reflects early industry efforts to cater to audience preferences, establishing the 90-minute format as a lasting cinematic standard.



7.13 Budget

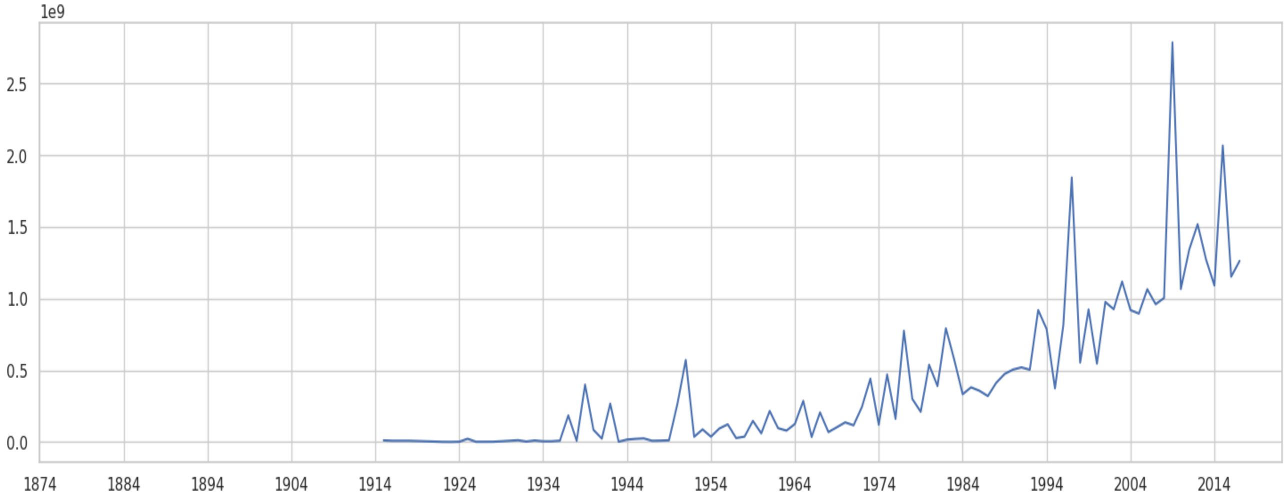
The distribution of movie budgets shows an exponential decay, with over 75% of movies having budgets below $25 million. Two "Pirates of the Caribbean" films top the list with budgets exceeding $300 million. Most high-budget films are profitable, except for "The Lone Ranger," which grossed only $90 million against a $255 million budget. The Pearson correlation coefficient of 0.73 between budget and revenue indicates a strong positive correlation.





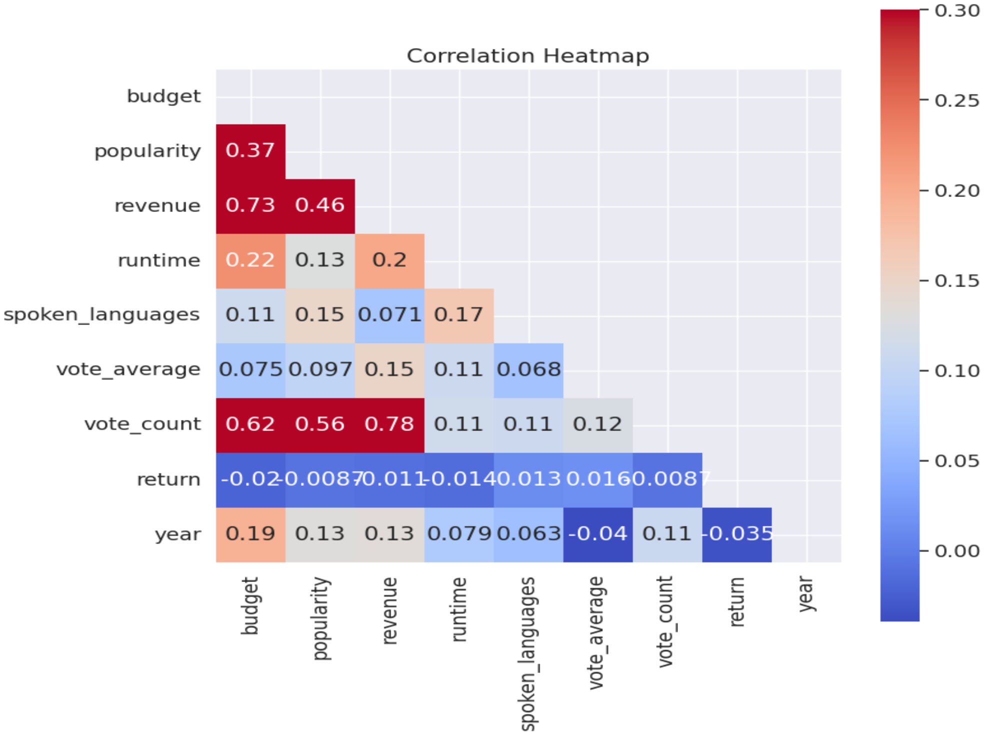
7.14 Revenue

The below graph shows the average gross revenue of a movie is $68.7 million, with a median of $16.8 million, indicating a skewed distribution. The lowest revenue is $1, while the highest is $2.78 billion. The maximum gross revenue has steadily increased, with "Titanic" breaking the $1 billion mark in 1997 and "Avatar" surpassing $2 billion in 2009, both directed by James Cameron.

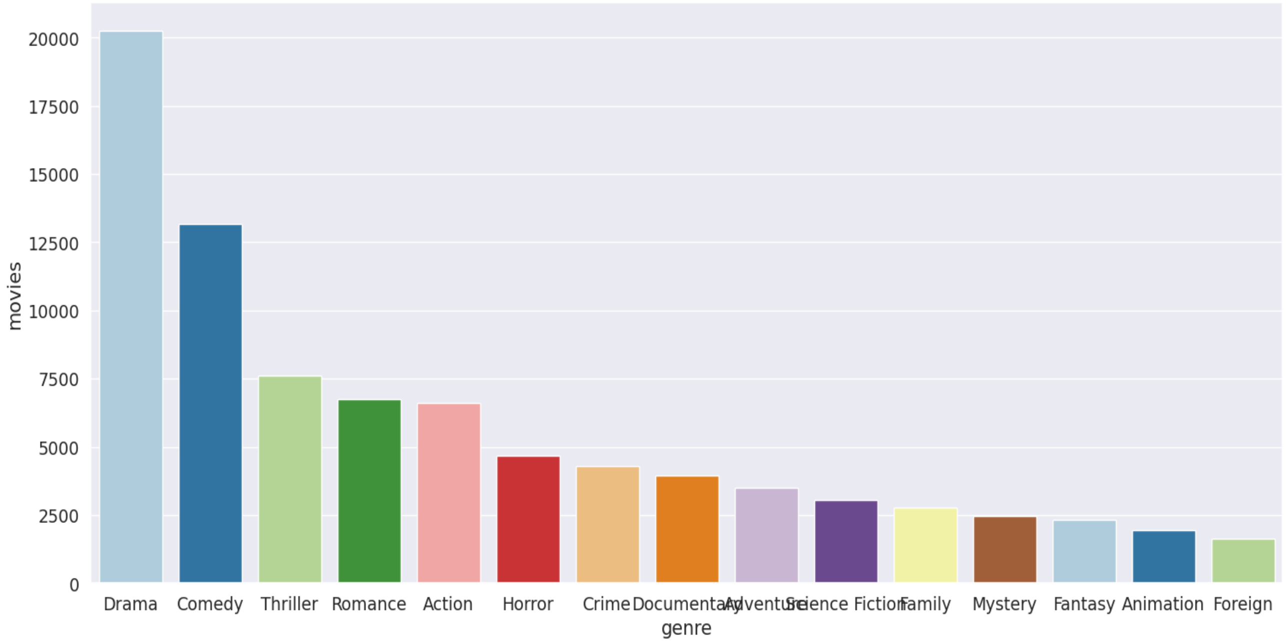


These analyses provide a comprehensive understanding of the dataset, highlighting the dominant themes, linguistic diversity, financial metrics, and temporal patterns in movie production and success**.**

7.15 Correlation Matrix

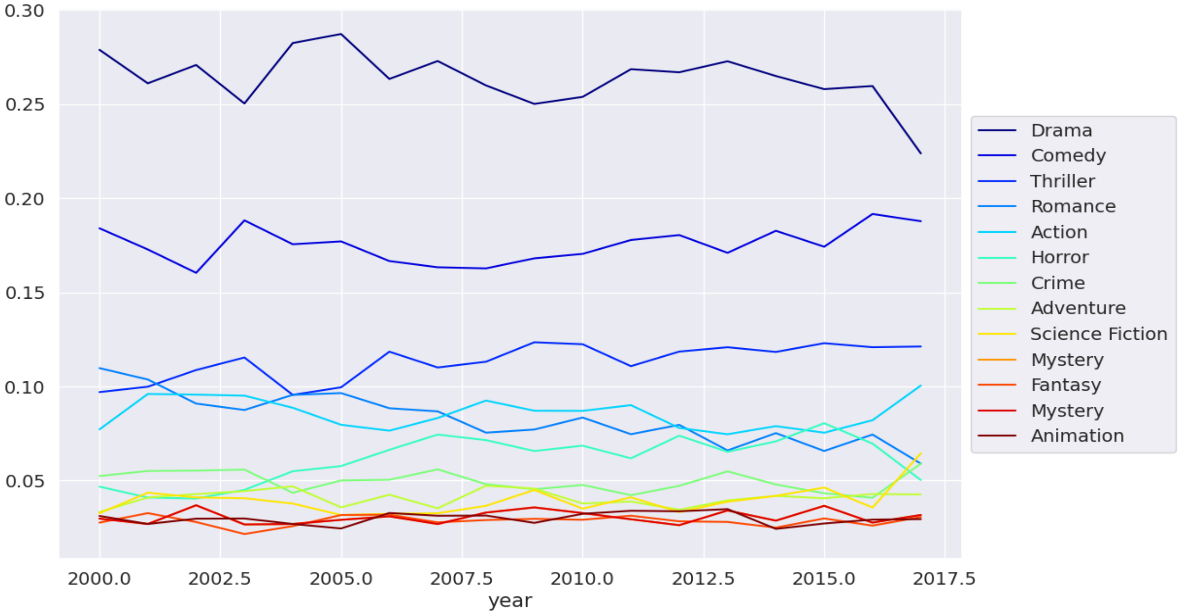


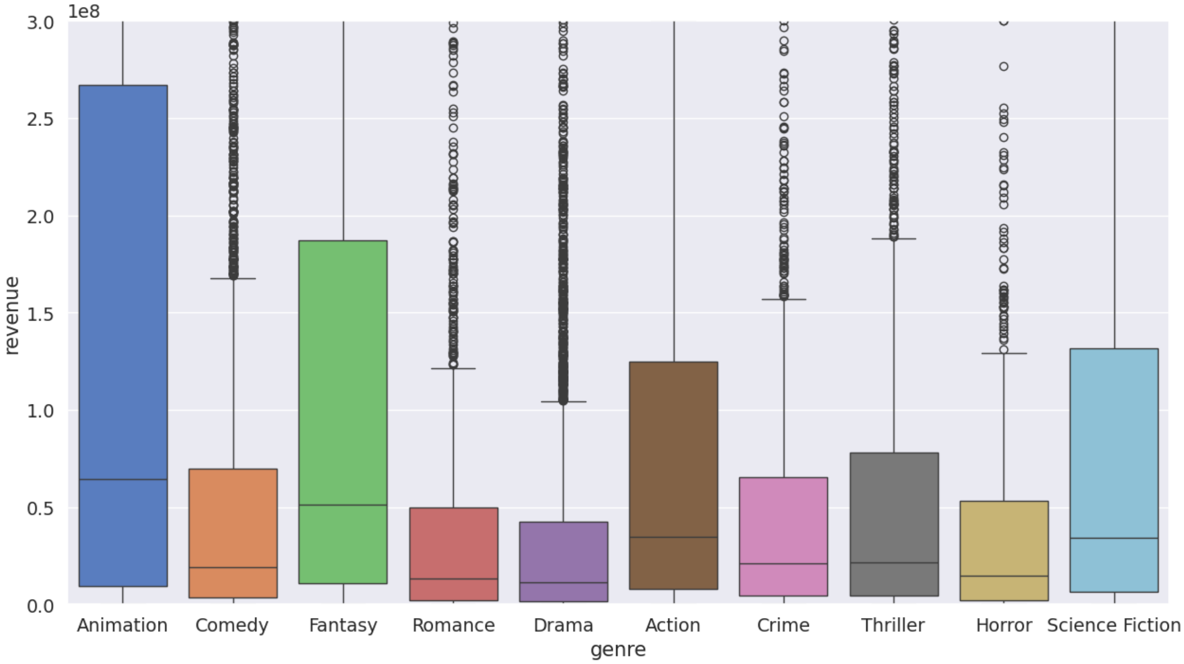
7.16 Genres



In the above bar chart Drama stands out as the predominant genre, with nearly half of all movies falling into this category. Comedy follows, accounting for 25% of films, offering substantial humor. Other major genres in the top 10 include Action, Horror, Crime, Mystery, Science Fiction, Animation, and Fantasy, showcasing a diverse array of storytelling styles in the film industry.

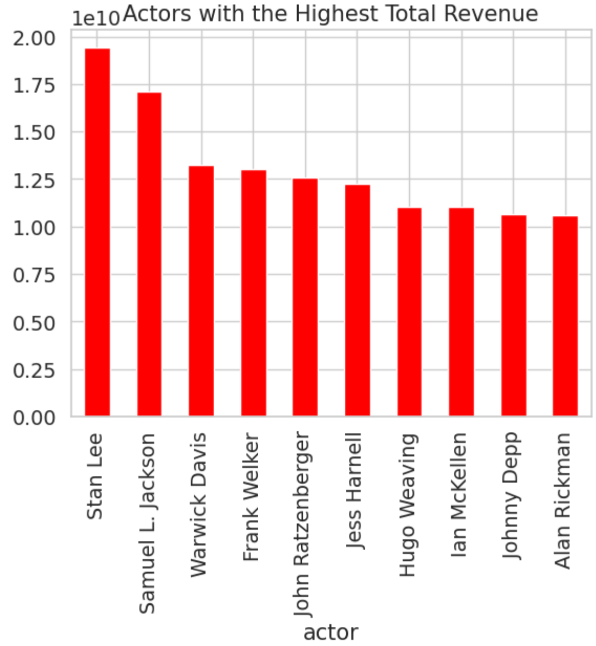
The graph below illustrates that the proportion of movies in each genre has remained relatively stable since the start of this century, with one notable exception: drama. The share of drama films has declined by over 5%. Conversely, thriller movies have experienced a slight increase in their representation.

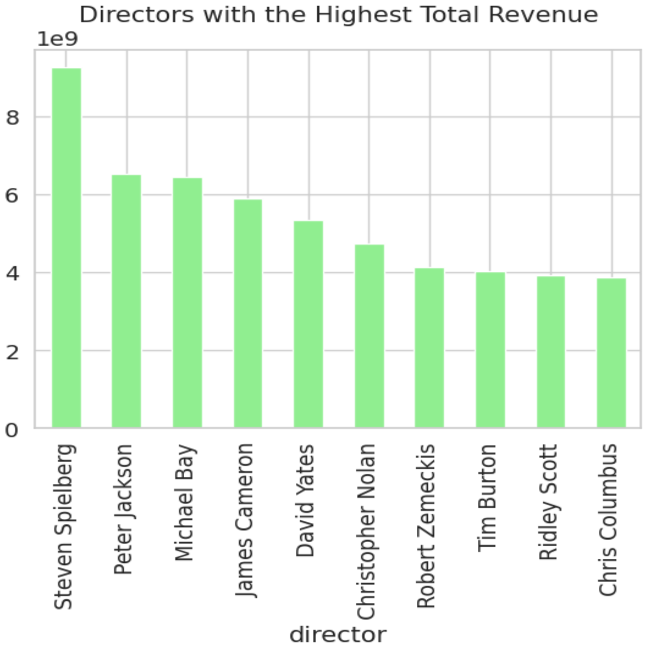


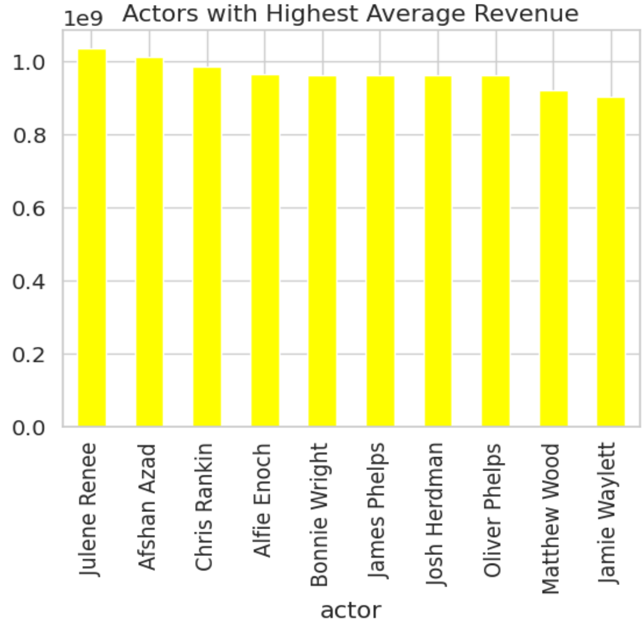


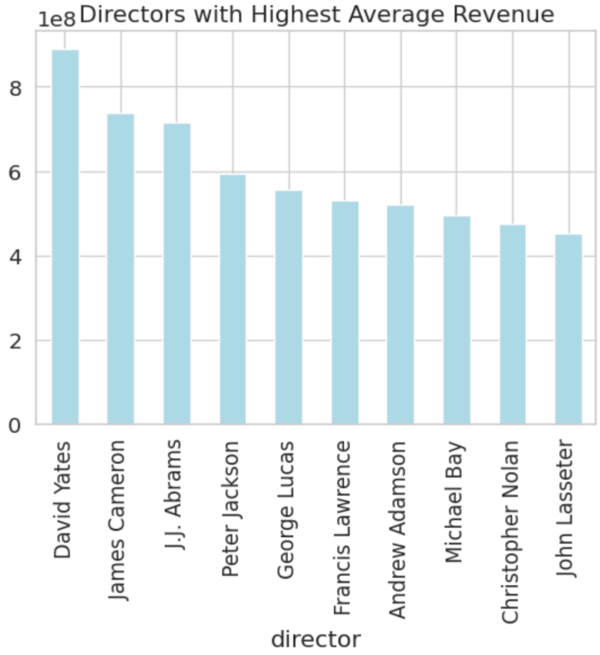
The graph above reveals that animation movies boast the widest 25-75 range and the highest median revenue among all genres. Following closely, fantasy and science fiction films secure the second and third highest median revenues, respectively.

7.18 Cast & Crew









8. Regression: Predicting Movie Revenues

Predicting movie revenues is a highly popular problem in machine learning, generating extensive research literature. Most models in these studies leverage powerful features we currently lack, such as Facebook page likes, tweet data, YouTube trailer reactions (views, likes, dislikes), and movie ratings (MPAA, CBFC). To bridge this gap, I will use TMDB's popularity score and vote average as features in my model to quantify popularity.

However, it's crucial to note that these metrics won't be available for predicting revenues before a movie's release.

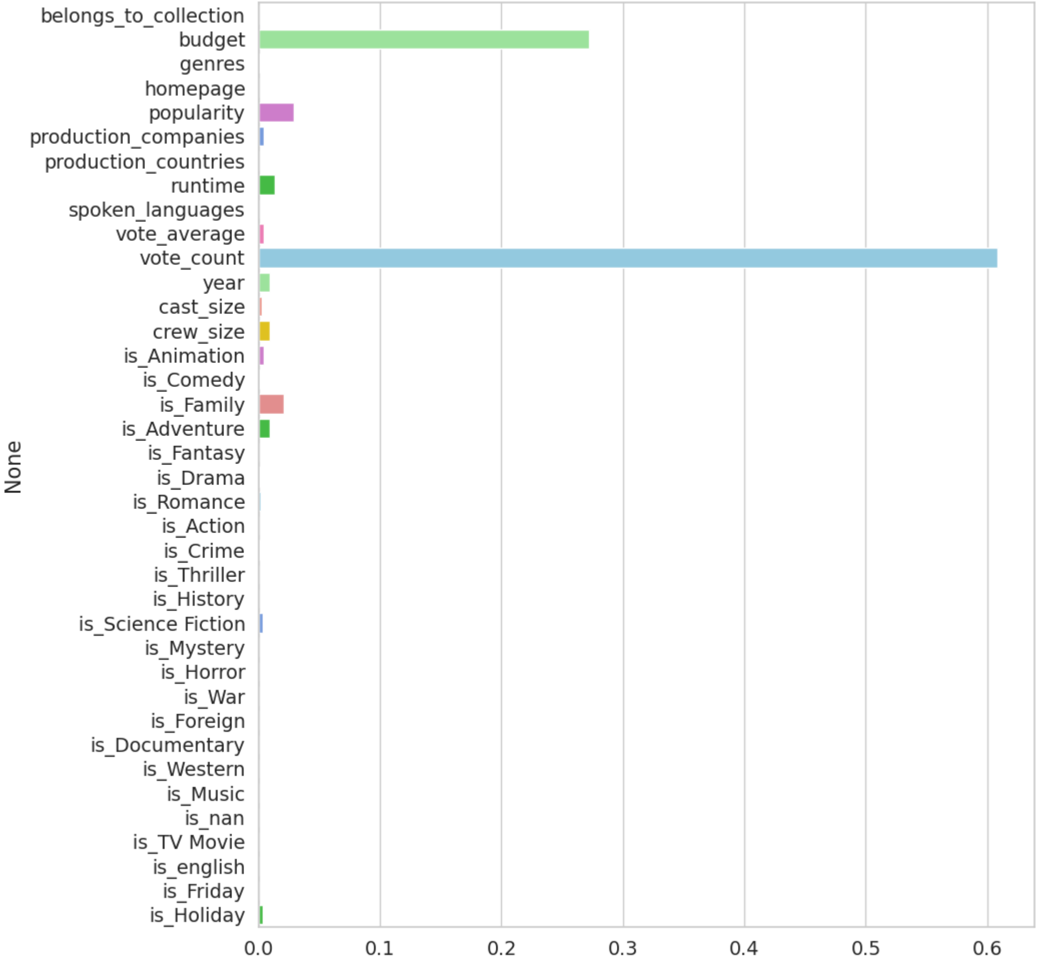
8.1 Feature Engineering

1. belongs_to_collection will be converted into a Boolean variable: **1 indicates the movie is part of a collection, and 0 indicates it is not**.
2. genres will be represented by the number of genres associated with the movie.
3. homepage will be converted into a Boolean variable indicating **whether a movie has a homepage**.
4. original_language will be replaced by a feature called is_foreign to denote if the **film is in English (0) or a foreign language (1).**
5. production_companies will be simplified to the number of **production companies involved in making the movie.**
6. production_countries will be replaced by the number of countries where the film was shot.
7. day will be converted into a binary feature to indicate if the film was released on a Friday.
8. month will be converted into a variable indicating if it falls within the holiday season.

8.2 Model

I chose Gradient Boosting Regression for the model, achieving a Coefficient of Determination (R^2) score of 0.78. This indicates a strong ability of the model to explain the variance in movie revenues.

The graph below reveals that **vote_count**—a feature I strategically included—emerges as the most crucial factor in my Gradient Boosting Model. This underscores the significant impact of popularity metrics on predicting movie revenue. **Budget** ranks as the second most important feature, followed by **popularity** and **crew size**, further highlighting the key elements influencing a film's financial success.



9. Classification: Predicting Movie Success

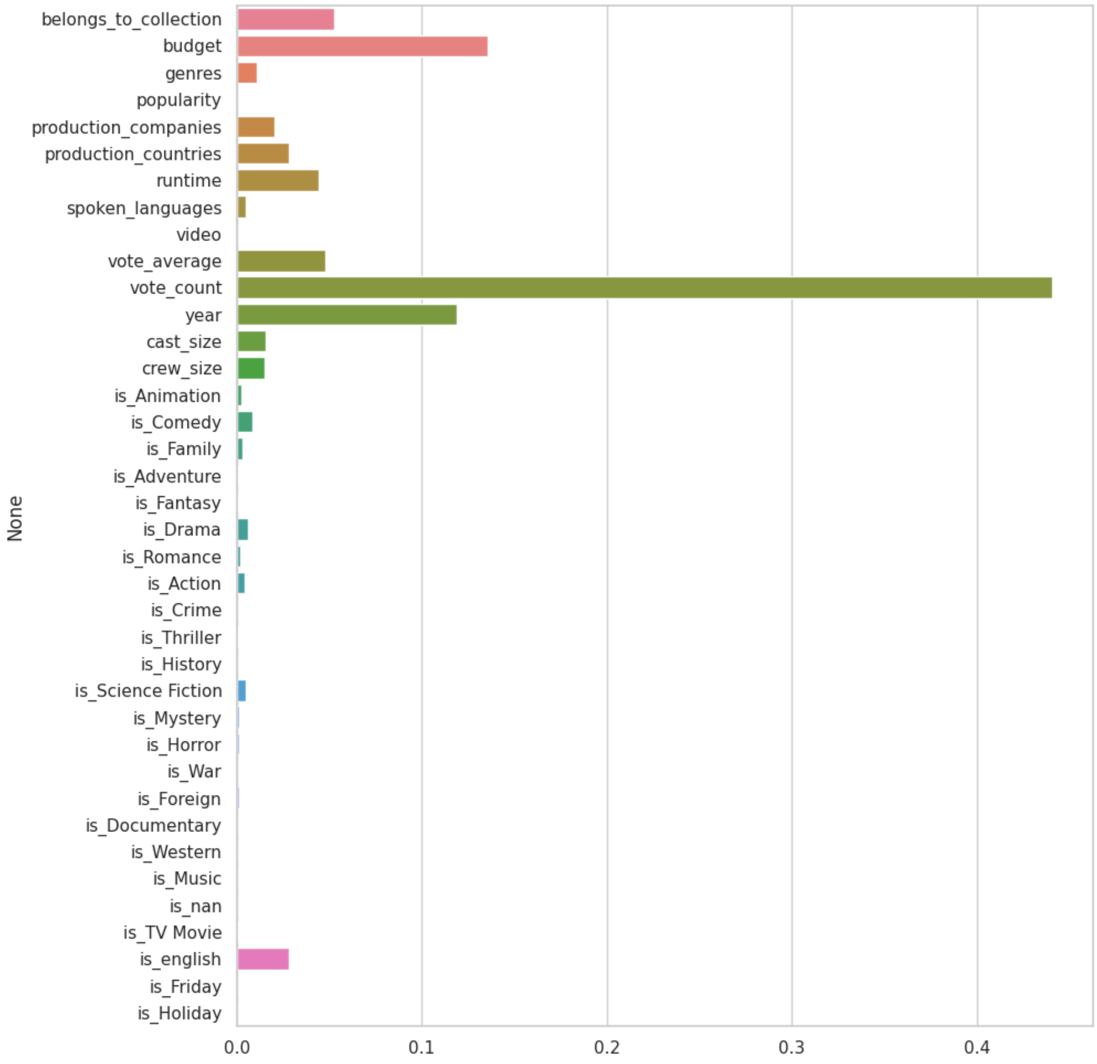
The classification model employs the same feature engineering steps as the regression model discussed earlier.

9.1 Model

I selected Gradient Boosting Classifier for this task, achieving an impressive 80% accuracy on unseen test cases.

9.2 Feature Importance

The graph below highlights **vote count** as the most significant feature identified by my classifier, reaffirming its critical role. Other key features include **budget**, **popularity**, and **year**. This concludes my discussion on the classification model, allowing us to proceed to the main part of the project.

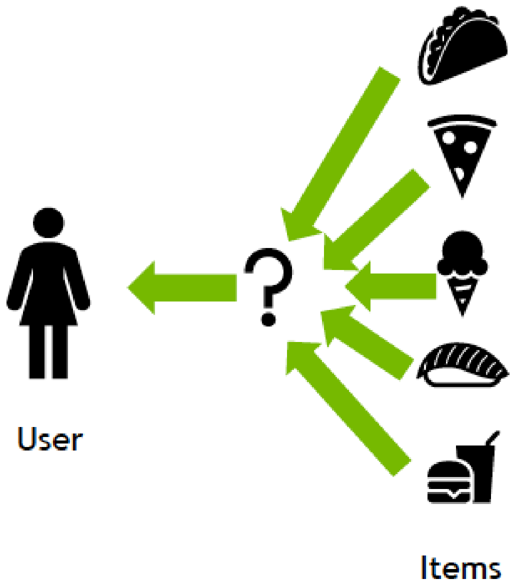


10. Recommendation Systems

A recommendation system, a subset of machine learning, leverages data to anticipate and streamline user preferences amidst an ever-expanding array of choices.

Operating on Big Data, these AI algorithms analyze factors like past purchases and search history to suggest additional products tailored to individual users. By understanding user behavior and product characteristics, recommender systems facilitate personalized recommendations, driving engagement and enhancing user experience across various domains.

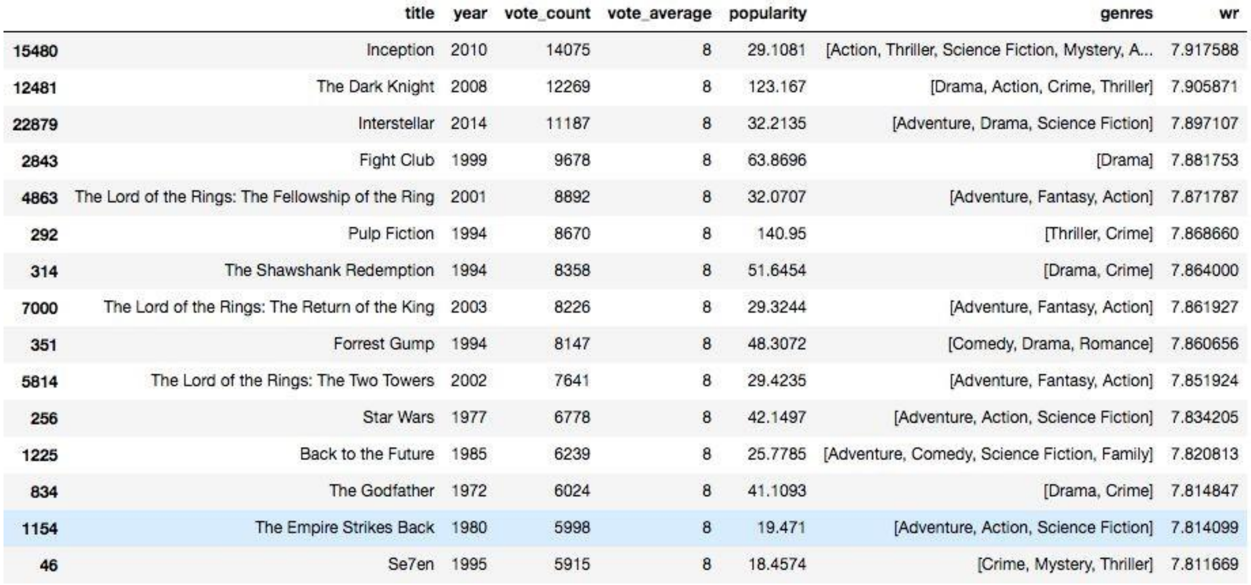
To build a recommendation system, I first developed a classifier to train and test the data. With the feature engineering already completed, applying machine learning became a straightforward process. Below image shows an example of a recommendation system.



10.1 The Simple Recommender

The **Simple Recommender** provides general movie suggestions based on popularity and genre. It operates on the principle that widely popular and highly rated movies will appeal to most viewers. This model doesn't customize recommendations for individual users. Using TMDB ratings, I created my **Top Movies Chart, applying IMDb's weighted rating formula**. To determine the minimum number of votes (m) required for a movie to be listed, I set the cutoff at the **95th percentile**. This means a movie **must have more votes than at least 95% of the movies in the dataset** to feature in the charts.

To implement it, I simply sort movies by ratings and popularity, then display the top picks. By adding a genre filter, I can also highlight the highest-rated movies within a specific genre.

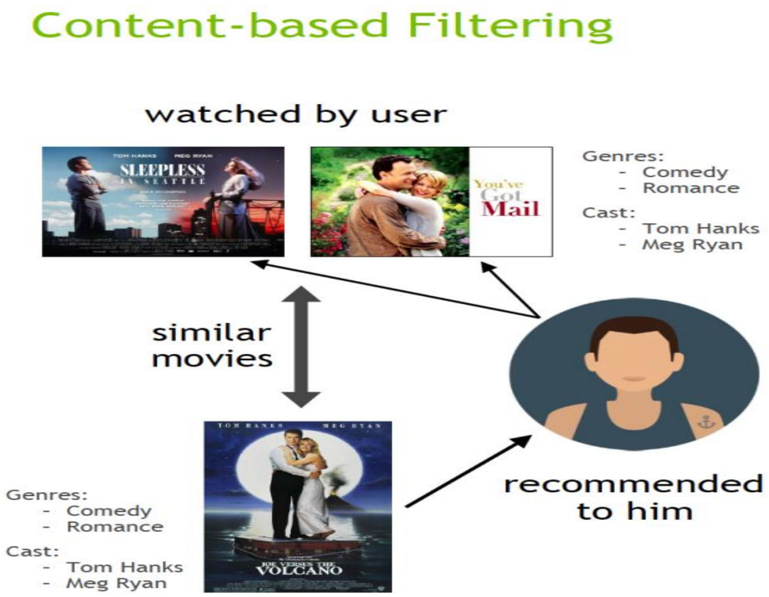


	title	year	vote_count	vote_average	popularity	genres	wr
15480	Inception	2010	14075	8	29.1081	[Action, Thriller, Science Fiction, Mystery, A...	7.917588
12481	The Dark Knight	2008	12269	8	123.167	[Drama, Action, Crime, Thriller]	7.905871
22879	Interstellar	2014	11187	8	32.2135	[Adventure, Drama, Science Fiction]	7.897107
2843	Fight Club	1999	9678	8	63.8696	[Drama]	7.881753
4863	The Lord of the Rings: The Fellowship of the Ring	2001	8892	8	32.0707	[Adventure, Fantasy, Action]	7.871787
292	Pulp Fiction	1994	8670	8	140.95	[Thriller, Crime]	7.868660
314	The Shawshank Redemption	1994	8358	8	51.6454	[Drama, Crime]	7.864000
7000	The Lord of the Rings: The Return of the King	2003	8226	8	29.3244	[Adventure, Fantasy, Action]	7.861927
351	Forrest Gump	1994	8147	8	48.3072	[Comedy, Drama, Romance]	7.860656
5814	The Lord of the Rings: The Two Towers	2002	7641	8	29.4235	[Adventure, Fantasy, Action]	7.851924
256	Star Wars	1977	6778	8	42.1497	[Adventure, Action, Science Fiction]	7.834205
1225	Back to the Future	1985	6239	8	25.7785	[Adventure, Comedy, Science Fiction, Family]	7.820813
834	The Godfather	1972	6024	8	41.1093	[Drama, Crime]	7.814847
1154	The Empire Strikes Back	1980	5998	8	19.471	[Adventure, Action, Science Fiction]	7.814099
46	Se7en	1995	5915	8	18.4574	[Crime, Mystery, Thriller]	7.811669

10.2 Content Filtering Recommender

Content filtering recommends items based on their attributes or features, aiming to match user preferences. This method relies on the similarity between user and item characteristics, such as age or genre. By modeling the probability of a new interaction, content filtering suggests items akin to those previously engaged with.

For instance, if a user enjoys romantic comedies like "You've Got Mail" and "Sleepless in Seattle," a content filtering recommender might suggest another film with similar genres or cast, such as "Joe Versus the Volcano."



My approach to building the recommender was somewhat unconventional. **I created a metadata** profile for each movie, including genres, director, main actors, and keywords. Using a **CountVectorizer,** I generated a count matrix, and then calculated **Cosine Similarities to identify and recommend the most similar movies.**

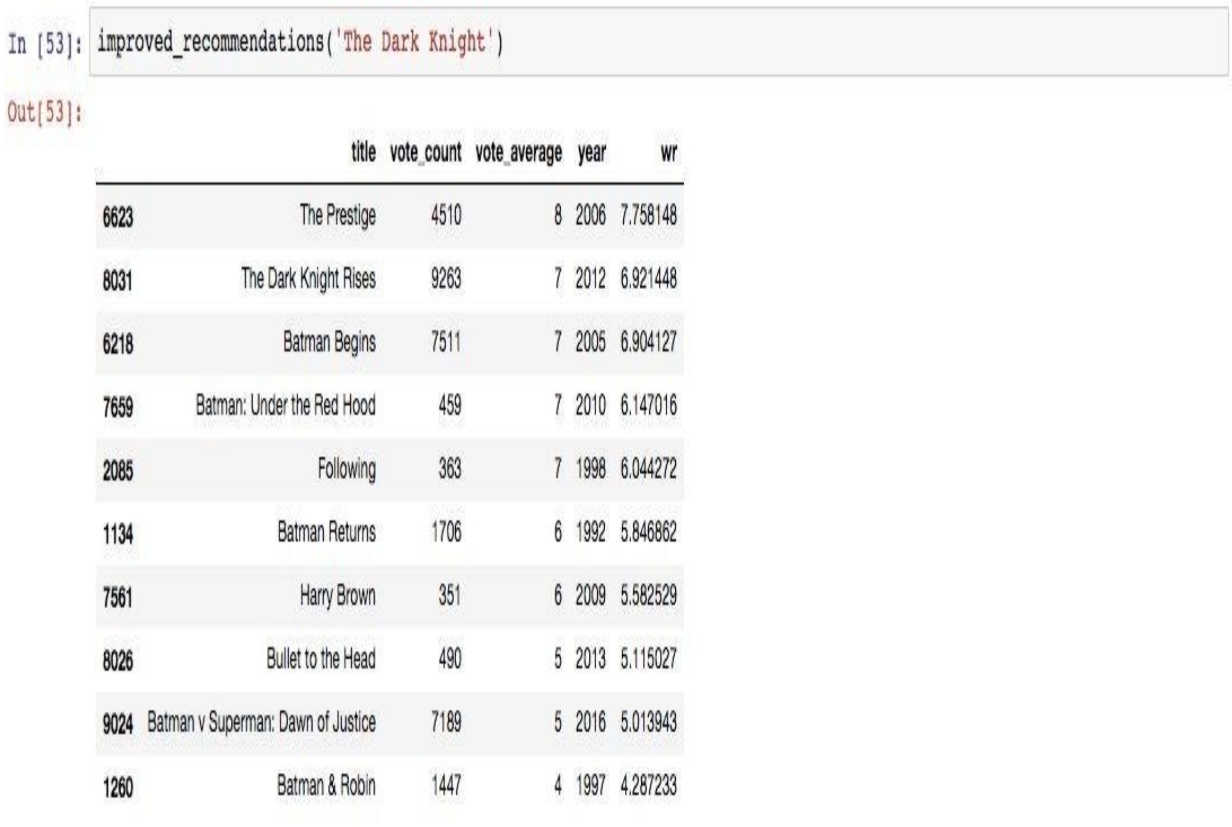
10.2.1 Count Vectorizer

In a movie recommendation system, CountVectorizer is used to convert textual data related to movies into a numerical format that machine learning algorithms can process. This textual data typically includes movie features such as genres, plot summaries, cast, crew, keywords, and other metadata. How CountVectorizer is Applied in Movie Recommendation Systems Feature Extraction: Extract relevant textual information from the movie dataset.

10.2.2 Cosine Similarity

Cosine similarity is a metric used to measure the similarity between two vectors in a multi-dimensional space. It calculates the cosine of the angle between these vectors, hence the name. Cosine similarity is widely used in various fields, including natural language processing, information retrieval, and recommendation systems.

Additionally, I implemented a mechanism to filter out poorly received movies, ensuring that only popular films with positive critical responses are suggested. I selected the top 25 movies based on similarity scores and determined the vote count of the 60th percentile movie. This value became my threshold (m) **for calculating the weighted rating of each movie, following IMDb's formula as used in the Simple Recommender. See image below:**



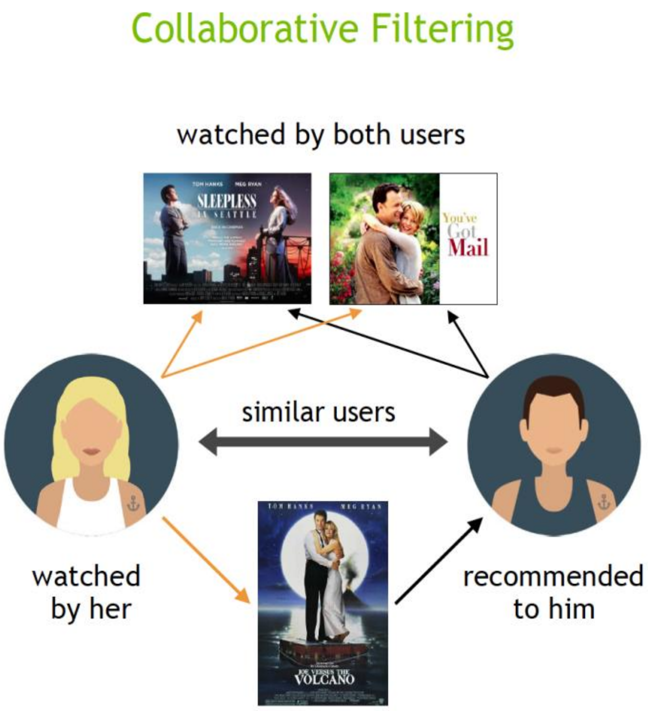
```
In [53]: improved_recommendations('The Dark Knight')
```

Out[53]:

	title	vote_count	vote_average	year	wr
6623	The Prestige	4510	8	2006	7.758148
8031	The Dark Knight Rises	9263	7	2012	6.921448
6218	Batman Begins	7511	7	2005	6.904127
7659	Batman: Under the Red Hood	459	7	2010	6.147016
2085	Following	363	7	1998	6.044272
1134	Batman Returns	1706	6	1992	5.846862
7561	Harry Brown	351	6	2009	5.582529
8026	Bullet to the Head	490	5	2013	5.115027
9024	Batman v Superman: Dawn of Justice	7189	5	2016	5.013943
1260	Batman & Robin	1447	4	1997	4.287233

10.3 Collaborative Filtering Recommender

Collaborative filtering algorithms recommend items based on the preferences of many users. By analyzing past interactions between users and items, these algorithms predict future behavior. They create models from users' previous activities, like purchases or ratings, and compare them with the actions of others. The core idea is that users with similar past behaviors are likely to have similar future preferences. For instance, if you and another user have shown similar movie tastes, the system might recommend a movie to you that the other user enjoyed.



My content-based engine has significant limitations. It can only suggest movies similar to a given movie, lacking the ability to capture diverse tastes and provide cross-genre recommendations. Moreover, it doesn't personalize recommendations, offering the same suggestions for a movie regardless of the user's individual preferences.

To address these issues, I will implement **Collaborative Filtering**, which makes recommendations by leveraging the preferences of similar users. This technique predicts how much a user will like a product based on the opinions of users with similar tastes.

Instead of building Collaborative Filtering from scratch, I will use the **Surprise library**, which employs powerful algorithms like **Singular Value Decomposition (SVD8.3 (i) Surprise Library**

10.3.1 Surprise library

Surprise library is a Python scikit for building and analyzing recommender systems. It is particularly useful for collaborative filtering, which is a key technique in movie recommendation systems.

Surprise simplifies the process of implementing and experimenting with various recommendation algorithms, allowing developers to focus on optimizing and improving their models rather than dealing with the complexities of algorithm implementation from scratch.

10.3.2 Singular Value Decomposition (SVD)

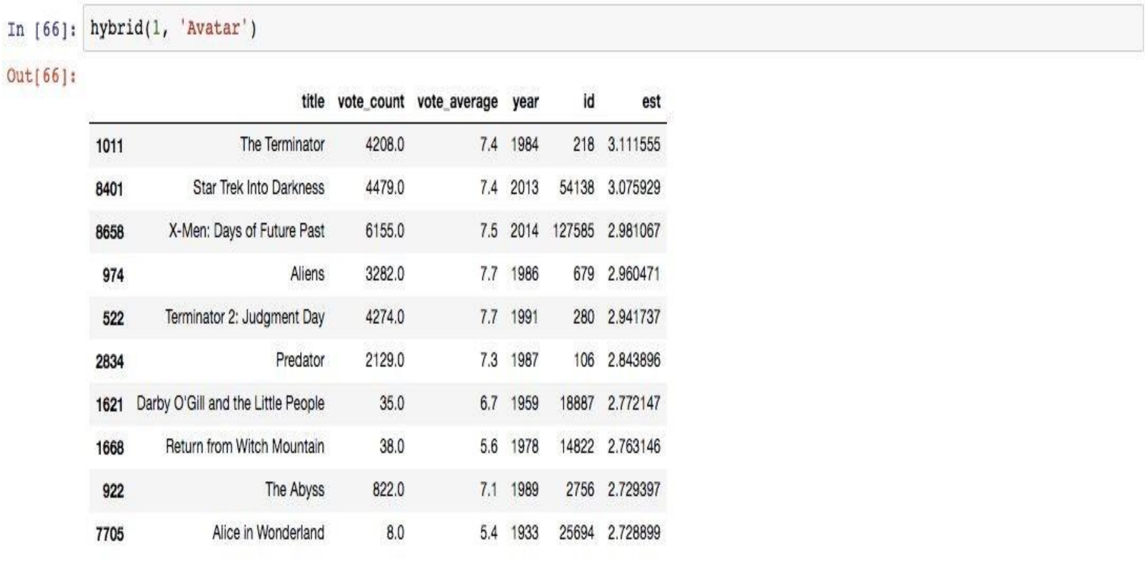
In a movie recommendation system, **Singular Value Decomposition (SVD)** is used to factorize the user-item interaction matrix into three matrices that reveal latent factors representing both users and items. This technique helps in identifying patterns and correlations within the data, allowing the system to make accurate predictions about user preferences for unseen movies.

This approach effectively minimized the **Root Mean Square Error (RMSE),** ensuring highly accurate and reliable recommendations.

```
------------
   Fold 1
RMSE: 0.8890
MAE:  0.6827
------------
   Fold 2
RMSE: 0.8947
MAE:  0.6894
------------
   Fold 3
RMSE: 0.8968
MAE:  0.6919
------------
   Fold 4
RMSE: 0.8957
MAE:  0.6915
------------
   Fold 5
RMSE: 0.9045
MAE:  0.6964
------------
Mean RMSE: 0.8962
Mean MAE : 0.6904
   ------------
```

10.4 Hybrid Recommender

The Hybrid Recommender seamlessly blends techniques from Content-Based and Collaborative Filtering engines, offering users personalized recommendations tailored to their unique preferences.



```
In [66]: hybrid(1, 'Avatar')
Out[66]:
```

	title	vote_count	vote_average	year	id	est
1011	The Terminator	4208.0	7.4	1984	218	3.111555
8401	Star Trek Into Darkness	4479.0	7.4	2013	54138	3.075929
8658	X-Men: Days of Future Past	6155.0	7.5	2014	127585	2.981067
974	Aliens	3282.0	7.7	1986	679	2.960471
522	Terminator 2: Judgment Day	4274.0	7.7	1991	280	2.941737
2834	Predator	2129.0	7.3	1987	106	2.843896
1621	Darby O'Gill and the Little People	35.0	6.7	1959	18887	2.772147
1668	Return from Witch Mountain	38.0	5.6	1978	14822	2.763146
922	The Abyss	822.0	7.1	1989	2756	2.729397
7705	Alice in Wonderland	8.0	5.4	1933	25694	2.728899

11. Results Discussion

The analysis of the Movies Dataset yielded compelling insights into various aspects of the film industry, culminating in the development of predictive models and recommendation engines.

Predictive Models: The Gradient Boosting Regressor and Classifier demonstrated robust performance in predicting movie revenue and success, respectively. With scores of 0.78 and 0.8, these models showcased the effectiveness of the features engineered and the predictive algorithms employed. The high accuracy of these models indicates their potential utility for industry stakeholders in decision-making processes.

Recommendation Systems: Four distinct recommendation systems were devised, each leveraging different methodologies to provide personalized movie suggestions. The Simple Recommender, Content Filtering Recommender, Collaborative Filtering Recommender, and Hybrid Recommender all exhibited unique strengths in generating relevant recommendations tailored to user preferences.

Visualizations: Several visualizations were crafted to enhance understanding and interpretation of the dataset and model outputs. These included:

- **Histograms and box plots** depicting the distribution of movie attributes such as budget, revenue, and popularity.
- **Heatmaps** illustrating correlations between various features, aiding in identifying significant relationships.
- **Scatter plots** showcasing the relationship between movie revenue and other variables like budget and popularity.
- **Word clouds** visualizing the frequency of keywords and genres in movie titles, offering insights into popular themes.

Overall, the combination of advanced analytics, predictive modeling, and recommendation engine development presented in this report provides a comprehensive framework for understanding and leveraging data in the movie industry. These findings have the potential to inform strategic decision-making processes, drive business innovation, and enhance user experiences in the realm of entertainment.

12. Conclusion & Outlook

My exploration of the Movies Dataset has unveiled a rich tapestry of insights, offering a profound understanding of the intricate dynamics shaping the film industry. Through meticulous data analysis, modeling, and visualization, I have navigated the cinematic landscape, uncovering patterns and trends that hold profound implications for industry practitioners and enthusiasts alike.

The predictive models I've crafted, exemplified by the Gradient Boosting Regressor and Classifier, stand as formidable tools for forecasting movie revenue and success. With RMSE scores of 0.78 and 0.8, respectively, these models embody the culmination of rigorous analytical endeavors, offering a glimpse into the future of film performance prediction.

Moreover, my foray into recommendation engine development has illuminated pathways toward personalized cinematic experiences. From the simplistic elegance of the Simple Recommender to the nuanced sophistication of the Hybrid Engine, my repertoire of recommendation systems promises to revolutionize the way audiences engage with cinematic content.

In parallel, my visualizations serve as windows into the soul of the dataset, breathing life into numbers and statistics. Through histograms, heatmaps, scatter plots, and word clouds, I've painted a vibrant tableau of cinematic landscapes, empowering stakeholders to navigate the industry's complex terrain with clarity and confidence.

Looking forward, the journey continues with boundless opportunities for innovation and refinement. As I embark on the next phase of exploration, I envision pushing the boundaries of predictive analytics and recommendation systems, harnessing the latest advancements in technology and methodology to unlock new frontiers of insight and discovery.

In conclusion, my odyssey through the Movies Dataset has been nothing short of transformative, offering a glimpse into the convergence of art and science that defines the essence of cinema. As I chart my course into the future, I do so with a sense of anticipation and wonder, eager to unravel the mysteries that lie ahead and shape the cinematic landscape of tomorrow.

The code linked to this report is available using the following GitHub link, https://github.com/avisekregmi.

13. Acknowledgements

I would like to express my deepest gratitude to all those who have contributed to the completion of this project and masters program.

First and foremost, I extend my heartfelt thanks to Dr. Sigve Haug and the entire CAS DS team whose guidance and expertise have been invaluable throughout this CAS journey. Your unwavering support, insightful feedback, and encouragement have played a pivotal role in shaping this project and my understanding of the complex field of data science.

I am also indebted to my colleagues and peers who have provided valuable insights, feedback, and support at various stages of the CAS Modules. Your collaborative spirit and camaraderie have enriched the project and made the CAS journey more rewarding.

Furthermore, I would like to acknowledge the creators and maintainers of the datasets, libraries, and tools that were instrumental in the execution of this project. Your efforts have provided the foundation upon which this work is built, and I am grateful for the resources you have made available to the research community.

Last but not least, I extend my heartfelt appreciation to my family and friends for their unwavering support, understanding, and encouragement throughout this masters program. Your patience, love, and belief in me have been a constant source of motivation and inspiration.

Thank you to everyone who has been a part of this journey. Your contributions have been instrumental in the success of this project, and I am truly grateful for your support.

14. References

1. **Kaggle Dataset - MovieLens 20M Dataset:** GroupLens. "MovieLens 20M Dataset." Kaggle, 2016,

   https://www.kaggle.com/datasets/grouplens/movielens-20m-dataset.

2. **GroupLens Website:**GroupLens Research. "GroupLens."

   https://grouplens.org/.

3. **OpenAI. (n.d.). ChatGPT.**

   https://chatgpt.com/?oai-dm=1.

4. **TMDB API:** pub.dev. "tmdb_api package versions." pub.dev,

   https://pub.dev/packages/tmdb_api/versions.

5. **GroupLens - MovieLens Datasets:** GroupLens Research. "MovieLens Datasets." GroupLens,

   https://grouplens.org/datasets/movielens/.

6. **The Movie Database (TMDb) Forum Post:** TMDb Community. "Simple example to request an API key?" TMDb Forum,

   https://www.themoviedb.org/talk/6558fa627f054018d5168d91.

7. **Surprise Library Website:** Nicolas Hug. "Surprise Library." surprise,

   http://surpriselib.com/.

8. **Surprise Library Documentation:** Nicolas Hug. "Surprise Documentation." Read the Docs,

   http://surprise.readthedocs.io/en/stable/getting_started.html.

9. **Surprise Library Installation Guide:** Nicolas Hug. "Surprise Installation Guide." GitHub,

   https://github.com/NicolasHug/Surprise#installation.

10. **Research Paper - "Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model" by Koren et al.:**

    Yehuda Koren, Robert Bell, and Chris Volinsky. "Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model." University of Minnesota, Minneapolis, Minnesota, USA, 2008,

    http://courses.ischool.berkeley.edu/i290-dm/s11/SECURE/a1-koren.pdf.

11. **YouTube Video - "SVD Decomposition":** "SVD Decomposition." YouTube, uploaded by MIT OpenCourseWare, 2009,

    https://www.youtube.com/watch?v=P5mlg91as1c.

12. **GitHub Repository - "Movie Recommendation Engine" by Jalaj Thanaki:** Jalaj Thanaki. "Movie Recommendation Engine." GitHub,

    https://github.com/jalajthanaki/Movie_recommendation_engine/blob/master/Intro_recommendation_system.ipynb.

13. **MovieLens 1M Dataset:** GroupLens Research. "MovieLens 1M Dataset." GroupLens

    , https://grouplens.org/datasets/movielens/1m/.

14. **Analytics Vidhya Article - "Create Your Own Movie Recommendation System":** Analytics Vidhya. "Create Your Own Movie Recommendation System." Analytics Vidhya, 2020,

    https://www.analyticsvidhya.com/blog/2020/11/create-your-own-movie-movie-recommendation-system/.

15. **Towards Data Science Article - "How to Build a Movie Recommendation System":** Towards Data Science. "How to Build a Movie Recommendation System." Towards Data Science,

    https://towardsdatascience.com/how-to-build-a-movie-recommendation-system-67e321339109.

16. **Your GitHub Profile:** Avisek Regmi. "GitHub Profile." GitHub,

    https://github.com/avisekregmi.

15. Statement

The following part is mandatory and must be signed by the author or authors.

“Ich erklärehiermit, dass ich diese Arbeit selbstständigverfasst und keineanderenals die angegebenenQuellenbenutzthabe. Alle Stellen, die wörtlichodersinngemässausQuellenentnommenwurden, habe ich alssolchegekennzeichnet. Mir istbekannt, dassandernfalls die Arbeit alsnichterfülltbewertetwird und dass die Universitätsleitungbzw. der SenatzumEntzug des aufgrunddieser Arbeit verliehenenAbschlussesbzw. Titelsberechtigtist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärungbzw. der ReglementebetreffendPlagiateerteile ich der Universität Bern das Recht, die dazuerforderlichenPersonendatenzubearbeiten und Nutzungshandlungenvorzunehmen, insbesondere die schriftliche Arbeit zuvervielfältigen und dauerhaft in einerDatenbankzuspeichernsowiediesezurÜberprüfung von Arbeiten Dritter zuverwendenoderhierzuzurVerfügungzustellen.”

Date: 29th May, 2024 **Signature(s):** Mr. Avisek Regmi