Movies and Tv shows analysis

Mohammed Alshehri (19209233)
Abdel Rahmane Ouafi (20715801)

A report submitted in part fulfilment of COMP 30780

Data Science in Practice

Demonstrator: Rian Dolphin

UCD School of Computer Science
University College Dublin

May 19, 2023

Table of Contents

Abstract

Our research focuses on using datasets obtained from Kaggle as well as datasets we gathered by utilizing the YouTube API to do deep analysis and produce insights for movies and TV series. By merging these many data sources, we want to provide informative and unique viewpoints on numerous aspects of the entertainment industry, such as audience preferences, trends, and performance evaluation.

Using the YouTube API in addition to the Kaggle datasets, we have also collected our own data. YouTube can be used to understand viewer engagement and attitude about movies and TV series. By harvesting YouTube likes and view counts, we can analyze audience response to specific material, gauge online buzz, and identify new trends and identify key figures in the online entertainment sector using this augmented information.

Two major objectives drive our project. First and foremost, we aim to provide producers, directors, and other entertainment industry professionals with actionable knowledge so they can make informed decisions about how to create content, employ marketing strategies, and concentrate on the correct audience. By analyzing audience preferences, viewing habits, and mood, industry participants may maximize their investments, lower their risks, and increase the possibility that their movies and television shows will be well-liked and successful.

Our project's second objective is to provide information to film enthusiasts, reviewers, and scholars by conducting a detailed examination of the entertainment sector. By examining data from several viewpoints, such as genre preferences, directing methods, and critical reception, we may give a holistic image of the industry. People would therefore be able to examine trends, unearth hidden gems, and understand more about the factors that affect the beloved movies and TV series.

By leveraging the capabilities of data analysis, machine learning, and sentiment tracking, our initiative aspires to be a helpful resource for both industry professionals and entertainment enthusiasts. Through our strategic approach to dataset aggregation and analysis, we will provide unique insights and encourage a greater understanding of the complex dynamics of the film and television business.

Chapter 1: **Introduction**

The entertainment industry, including movies and TV shows, has a profound impact on society, culture, and the economy. Understanding the factors that make these forms of entertainment succeed or fail is important for filmmakers, producers, investors and interested parties This work focuses on using data sets derived from Kaggle and the YouTube API detailed research and insights are available..

In previous research, studies have analyzed factors such as genre, ratings, cast/crew details, and box office revenue separately to understand their individual effects on film success but in a holistic way considering multiple variables simultaneously is necessary to capture the complexity of interest function. In addition, less attention has been paid to the analysis of TV shows and their enabling factors such as number of episodes/seasons and length compared to film.

This project aims to fill these gaps and contribute to existing knowledge by addressing the following research questions: Can we predict the success of a film at the box office? How does the number and length of episodes/seasons affect the ratings, attention and success of a TV show? What are the key characteristics that make blockbuster films, and do these characteristics vary across cultural contexts and audience sizes? Is there a tendency for films to change genres over time, and does this affect their success?

To answer these research questions we will use Kaggle data set, which provide more information about movie ratings, box office earnings, genre classifications, cast/crew descriptions These data sets will be the basis for analyzing movie success factors and identifying patterns which helps towards their box office performance. In addition, we use the YouTube API to collect statistics such as likes and view rates, which will enable provide insights into audience engagement and of what they want.

The structure of this report is as follows:

Chapter 2: Project Objectives This chapter discusses the objectives and motivations of the project in detail, outlining what we hope to achieve through our analysis.

Chapter 3: Related Work Here, we delve into previous research conducted in the field of movie and TV show analysis. We identify the existing knowledge and highlight the gaps that our project aims to fill.

Chapter 4: Data Considerations In this chapter, we discuss the datasets used in our analysis, including those sourced from Kaggle and collected through the YouTube API. We address any challenges or considerations related to data collection, cleaning, and preprocessing.

Chapter 5: Results Chapter 5 presents the results and findings of our analysis. We showcase the outcomes of our predictive models, examine the impact of episode/season count and length on TV show ratings, explore the key attributes of highly-rated movies, and analyze genre transitions over time.

Chapter 6: Conclusion Here, we summarize the project's conclusions based on our findings. We discuss the implications of our research, highlight the contributions made, and provide suggestions for future research directions in this domain.

Chapter 2: Project Objectives

Our shared interest for the intriguing world of movies and TV series inspired us to start this initiative. We were intrigued by the chance to look into the data underlying different entertainment mediums and unearth insights that would be useful to various business stakeholders. The project was an intriguing and attractive choice due to its complexity and ability to advance our knowledge of the elements that influence success.

Our business aims to connect a wide variety of individuals interested in the entertainment industry. For filmmakers, producers, and investors, our research provides insight into the important factors that go into box office success. Taking into account things like reviews, genre, cast-driver information, previous box office grosses, etc., we want to create a predictive algorithm that can drive a movie some accurate financial forecasting by assisting in decision-making processes related to production budgeting, marketing strategy and distribution strategy Can help businesses make the best creative and management decisions.

Streaming services and TV show producers are also interested in our work. We examine the effects of episode and season duration on the ratings, attention, and overall success of the TV show in our study. Looking at the relationships between these variables provides insight into audience preferences and viewing habits in television programming. This knowledge can help streaming content producers and content creators better structure TV shows, increase viewer engagement, and improve the chances of market success with increasing competition

During the course of this project, we encountered a number of challenges when gathering and analyzing the data. Getting a trustworthy and accurate dataset with a diverse selection of films and TV series is the first issue. In order to solve this, we accessed a lot of data on user likes and view count using the YouTube API as well as information from Kaggle and a well-known online portal. Data cleansing and cleaning were the workflow operations that made sure the data collected was accurate and consistent. For this, it was necessary to deal with missing values, develop systems as standards, and settle differences between sources.

2.1 Objectives & Motivations

Our enthusiasm for the entertainment business, particularly for films and television programs, served as the motivation for this endeavor. We are curious about how these pastimes are created as well as how they affect culture and society. We seek to provide useful information to industry professionals, interested parties, and researchers to obtain insights into the evolution of the industry by conducting research and examining the elements that contribute to success or failure.

The choice of project was motivated by the aim to investigate and fill in knowledge gaps related to movie and TV show analysis. Although there has been a lot of research done in this field, there are still open questions and untried paths. By concentrating on important research problems and employing datasets from Kaggle and the YouTube API, we think that our initiative can add to the body of existing knowledge.

2.2 Research Questions

The research questions that guide our project are as follows:

1. Can we predict the box office success of a movie based on factors such as ratings, genre, cast/crew details, and historical box office revenues?

2. How does the count and length of episodes/seasons impact the ratings, votes, and overall success of TV shows?

3. What are the key attributes that contribute to highly-rated movies, and do these attributes vary across cultural contexts?

4. Is there a trend of genre transitions in movies over time, and does this trend affect their box office performance?

By addressing these research questions, we aim to generate valuable insights and contribute to the understanding of the entertainment industry dynamics.

Chapter 3: **Related Work**

Who has done something similar : The entertainment industry is constantly evolving, and the analysis of movies and TV shows is an important area of study. Several studies have explored research questions related to this topic, and our project builds on these studies to address additional questions.

For example, a study by [1] analyzed the impact of social media on box office revenue for movies. The study found that social media metrics, such as Facebook likes and Twitter mentions, can predict a movie's opening weekend box office revenue. Our project extends this research by incorporating YouTube views and likes to predict box office success.

Similarly, a study by [2] explored the impact of actor and director reputation on box office success. The study found that movies with highly reputable actors and directors tend to perform better at the box office. Our project extends this research by analyzing the impact of cast size and production budget on box office success.

Another study by [3] Chiang et al. (2020) analyzed the impact of genre on TV show ratings and popularity. The study found that certain genres, such as drama and comedy, tend to be more popular among audiences. Our project extends this research by analyzing the trend of genre transitions in movies over time and its impact on success.

Our project is unique because it uses multiple data sources, including box office revenue, IMDb ratings, and YouTube views and likes, to gain new insights into the factors that impact the success of movies and TV shows. Additionally, we have analyzed data from both movies and TV shows to compare and contrast the factors that are important for each type of media.

Another unique aspect of our project is the inclusion of data from different cultural contexts. By analyzing the key attributes that create highly-rated movies across different cultures, we have provided valuable insights that can be useful for filmmakers and producers looking to create content that resonates with audiences in different regions of the world.

Finally, our project is unique because it uses up-to-date data from recent movies and TV shows. As the entertainment industry is constantly evolving, it is important to have current information on the factors that impact success. By using data from recent releases, we have ensured that our analysis is reflective of current trends and audience preferences.

In conclusion, our project builds on previous studies in the analysis of movies and TV shows by addressing additional research questions and using novel approaches and current data. By exploring these questions, we have provided valuable information that can be useful for filmmakers, producers, and anyone interested in understanding what factors can impact the success of movies and TV shows. Further research is needed to continue exploring these topics and to identify additional factors that impact success.

Chapter 4: Data Considerations

4.1 Data Collection

Collecting data for movies and TV shows can be a challenging process that requires extensive research and data collection. The first stage of data collection involves identifying relevant data sources, such as movie databases, industry reports, and online reviews. This may require accessing and purchasing paid databases that contain information about box office earnings, production budgets, and critical reception.

Once the relevant data sources have been identified, the next step is to extract the necessary data and compile it into a usable format. This can involve cleaning and organizing the data to remove duplicates, errors, and inconsistencies. Depending on the size and complexity of the dataset, this process can be time-consuming and require specialized data cleaning tools

DataSet1)

The dataset used in our project consists of 6,820 movies spanning the years 1986 to 2016, with an average of 220 movies per year. The dataset provides various attributes for each movie, including budget, production company, country of origin, director, genre, gross revenue, movie name, rating, release date, runtime, IMDb user rating, number of user votes, main actor or actress, writer, and the year of release.

The budget attribute represents the financial resources allocated for producing a movie. However, it is important to note that some movies in the dataset have missing budget values, which are represented as 0.

The company attribute denotes the production company responsible for producing the movie. It provides insights into the entities involved in the filmmaking process.

The country attribute specifies the country of origin for each movie, providing information about its cultural background and production location.

The director attribute indicates the individual responsible for directing the movie, playing a crucial role in shaping the artistic vision and overall execution of the film.

The genre attribute represents the primary genre of each movie, categorizing them into various thematic classifications such as action, drama, comedy, and more.

The gross attribute indicates the revenue generated by each movie, providing an indication of its commercial success.

The name attribute represents the title of each movie, serving as a unique identifier for easy reference.

The rating attribute denotes the age-based rating assigned to the movie, indicating the intended

Figure 4.1: Caption

audience age group, such as R for Restricted or PG for Parental Guidance.

The released attribute provides the release date of each movie in the format YYYY-MM-DD, indicating when the movie was made available to the public.

The runtime attribute represents the duration of each movie, indicating its length in minutes.

The score attribute represents the user rating on IMDb, providing an assessment of the movie's overall quality as perceived by viewers.

The votes attribute denotes the number of user votes received by each movie on IMDb, serving as a measure of its popularity and audience engagement.

The star attribute represents the main actor or actress in each movie, highlighting the prominent individuals who played pivotal roles.

The writer attribute denotes the writer(s) responsible for creating the screenplay or script of the movie, contributing to its narrative and storytelling elements.

The year attribute represents the year of release for each movie, providing a temporal reference for analyzing trends and changes in the film industry over time.

It is important to acknowledge that this dataset was sourced from IMDb, a popular online database of movies, TV shows, and other related content. The data was likely obtained from a publicly available dataset repository on GitHub, which serves as a platform for sharing and collaborating on open-source projects. Utilizing this dataset from GitHub allowed us to conduct our analysis and explore the various research questions related to movies and TV shows.

To ensure the quality and reliability of the dataset, thorough validation and cleaning processes were conducted. Missing values were identified and handled by either removing the corresponding instances or imputing values using appropriate methods. Duplicate entries were detected and eliminated to avoid duplication. Outliers and inconsistencies in numerical attributes were addressed through careful examination and adjustment. Inconsistent or erroneous data entries were corrected or removed. Data types were properly assigned, and categorical variables were appropriately encoded. After the validation and cleaning steps, the dataset was refined, resulting in a final dataset of 5,421 movies ready for analysis.

Dataset 2)

The second dataset in our project consists of over 10,000 movies with various attributes, including ID, IMDb ID, popularity, budget, revenue, original title, cast, homepage, director, tagline, overview, runtime, genres, production companies, release date, vote count, vote average, release year, budget (adjusted), and revenue (adjusted). It is important to note that this dataset is in its original form obtained from TMDB but sourced from a publicly available dataset repository on GitHub.

The dataset represents a diverse range of movies with information about their popularity, financial aspects, production details, and critical reception. Each movie is uniquely identified by an ID and IMDb ID, and the popularity attribute indicates the relative popularity of the movie within the TMDB platform.

Financial aspects such as budget and revenue provide insights into the investment and financial performance of movies. The original title attribute represents the title of the movie, while the cast attribute lists the actors and actresses involved in each movie.

The dataset also includes information about the movie's homepage, director, tagline, overview, and runtime. These attributes provide additional context and description of the movie, allowing for a comprehensive understanding of its content and creative team.

	id	imdb_id	popularity	budget	revenue	original_title	cast	homepage	director	tagline	...
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	http://www.jurassicworld.com/	Colin Trevorrow	The park is open.	...
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	http://www.madmaxmovie.com/	George Miller	What a Lovely Day.	...
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent	Shailene Woodley|Theo James|Kate Winslet|Ansel...	http://www.thedivergentseries.movie/#insurgent	Robert Schwentke	One Choice Can Destroy You	...
3	140607	tt2488496	11.173104	200000000	2068178225	Star Wars: The Force Awakens	Harrison Ford|Mark Hamill|Carrie Fisher|Adam D...	http://www.starwars.com/films/star-wars-episod...	J.J. Abrams	Every generation has a story.	...
4	168259	tt2820852	9.335014	190000000	1506249360	Furious 7	Vin Diesel|Paul Walker|Jason Statham|Michelle ...	http://www.furious7.com/	James Wan	Vengeance Hits Home	...

5 rows × 21 columns

Figure 4.3: Dataframe

Figure 4.4: Caption

Genres attribute categorizes movies into different genres, allowing for genre-based analysis and exploration of trends within specific genres. The production companies attribute lists the companies involved in producing each movie, shedding light on the industry collaborations behind the film's creation.

Release-related attributes, such as release date and release year, provide temporal information about the movies, allowing for the examination of trends and patterns over time. The vote count and vote average attributes represent user ratings and feedback, indicating the audience's perception and reception of each movie.

Finally, the budget (adjusted) and revenue (adjusted) attributes provide adjusted values for budget and revenue, taking into account factors such as inflation to enable meaningful financial comparisons across movies.

It is important to note that this dataset is in its pre-cleaning stage, and further cleaning and validation steps would be necessary to ensure data accuracy, consistency, and reliability.

After cleaning the second dataset and removing the movies that are duplicates of the ones in the first dataset, we are left with a subset of (3855, 14) movies and rows. This process ensures that we have a distinct set of movies without any redundancy. By eliminating the duplicates, we can avoid any potential biases or inconsistencies that may arise from having duplicate entries. This refined dataset of 3855 unique movies will enable us to conduct accurate and reliable analyses and comparisons across the two datasets, providing valuable insights into the research questions and objectives of our project.

Data set 3)

The third dataset in our project is sourced from GitHub and comprises a comprehensive collection of movie data. This dataset includes a wide range of attributes that provide in-depth information about each movie. The columns in this dataset consist of 'adult', 'belongs$_{t}o_{c}ollection'$,$'$ $budget'$,$'$ $genres'$,$'$ $homepa$

Obtaining this dataset from GitHub ensures its availability to the public and allows researchers and

enthusiasts to explore and analyze a diverse range of movies. With 45,466 rows and 24 columns, this dataset is notably the largest among the movie datasets we have gathered.

Each column provides unique information about the movies. Attributes such as 'budget', 'revenue', 'popularity', and 'runtime' offer insights into the financial aspects and popularity of movies. The 'genres' column categorizes movies into different genres, allowing for genre-based analysis and understanding of audience preferences.

Additional attributes like 'production$_c ompanies'$,'$production_c ountries'$, $and'spoken_l anguages'$ $shedlightontheq$

The 'release$_d ate'attributeprovidestemporalinformation, enablingtheanalysisoftrendsandpatternsovertim$

The dataset also includes textual attributes such as 'overview' and 'tagline', which provide descriptive information about the movies. Furthermore, the 'belongs$_t o_c ollection'columnidentifieswhetheramoviebelon$

With its extensive range of attributes and a large number of movies, this dataset offers significant potential for conducting comprehensive analyses and gaining valuable insights into the movie industry.

After merging the datasets, our movie collection now consists of (44,872, 14) movies and columns. This comprehensive dataset combines valuable information from multiple sources, enabling us to conduct robust analyses on various aspects of the movie industry. The merged dataset allows us to explore research questions related to box office success, genre trends, audience preferences, and more. The final columns include 'budget', 'popularity', 'production$_c ompanies'$,'$country'$,'$year'$,'$Box_o ffice'$,'ru

Dataset 4) We have collected data on the number of views and likes for movie trailers from the YouTube API. This valuable information provides insights into the level of audience engagement and interest in specific movies. To obtain this data, we conducted web scraping of the YouTube API, extracting the relevant metrics for a substantial number of movie trailers. Our dataset comprises information for approximately 1500 movies.

To ensure accuracy and reliability, we matched the movies in our YouTube dataset with their corresponding names and box office data from our merged movie collection. By linking the YouTube data with the movie names and box office figures, we have established a comprehensive understanding of audience engagement in relation to the financial success of movies. This integration of YouTube data allows us to gain deeper insights into the impact of online viewership and appreciation on a movie's overall performance. The inclusion of YouTube data enriches our analysis, providing a holistic perspective on audience reception and the popularity of movies in the digital realm.

Dataset 5)

To further enhance our analysis, we obtained a popularity dataset from Kaggle, which contained information on the popularity of actors. This dataset included details such as the actor's name, actor ID, film name, year of release, votes received, rating, and film ID. The dataset was extensive, comprising (191,873, 7) rows and columns.

After acquiring the popularity dataset, we matched it with the cast data in our existing movie dataframe. By linking the datasets based on the actor's name or ID, we were able to establish connections between the movie details and the popularity metrics of individual actors.

Through this matching process, we obtained popularity information for a subset of actors, resulting in a dataframe of (9615 rows x 4 columns). This subset represented the actors who were present in our movie dataset.

By incorporating the popularity information from Kaggle, we gained additional insights into the

	Actor	ActorID	Film	Year	Votes	Rating	FilmID
0	Fred Astaire	nm0000001	Ghost Story	1981	7731	6.3	tt0082449
1	Fred Astaire	nm0000001	The Purple Taxi	1977	533	6.6	tt0076851
2	Fred Astaire	nm0000001	The Amazing Dobermans	1976	369	5.3	tt0074130
3	Fred Astaire	nm0000001	The Towering Inferno	1974	39888	7.0	tt0072308
4	Fred Astaire	nm0000001	Midas Run	1969	123	4.8	tt0064664
5	Fred Astaire	nm0000001	Finian's Rainbow	1968	3377	6.2	tt0062974
6	Fred Astaire	nm0000001	The Notorious Landlady	1962	1887	6.8	tt0056289
7	Fred Astaire	nm0000001	The Pleasure of His Company	1961	679	6.9	tt0055307
8	Fred Astaire	nm0000001	On the Beach	1959	12066	7.2	tt0053137
9	Fred Astaire	nm0000001	Funny Face	1957	27534	7.0	tt0050419

Figure 4.5: Dataframe

recognition and impact of specific actors in the movie industry. We could analyze the relationship between an actor's popularity, as indicated by votes and ratings, and their involvement in films. This allowed us to explore the influence of popular actors on the success and reception of movies.

Overall, by integrating the popularity dataset from Kaggle into our movie analysis, we enriched our understanding of the movie industry and its key players. The matched dataframe provided valuable insights into the popularity and influence of actors, enabling us to explore their impact on movie performance and audience reception.

Dataset 6) For our analysis of TV shows, we obtained a comprehensive dataset from Kaggle. This dataset contained a vast amount of information about TV shows, encompassing (514,236, 9) rows and columns. The columns in the dataset included 'Title$_{s}how_{n}ame_{t}const'$,'$Title_{b}asics_{t}const'$,'$Title_{s}how_{n}ame_{p}$

To prepare the data for analysis, we performed cleaning and modifications. This involved removing any irrelevant or duplicate entries and addressing missing or inconsistent values. Through these cleaning steps, we refined the dataset to ensure its quality and reliability.

After cleaning the data, we made modifications to the dataset to focus on the most relevant columns for our analysis. We retained the columns 'Title$_{s}how_{n}ame_{p}rimaryTitle'$,'$seasonNumber'$,'$episodeNu$

Following these steps, we obtained a modified dataset with dimensions of [514,236 rows x 8 columns]. This refined dataset allowed us to delve into the analysis of TV shows and explore various aspects such as the impact of season and episode count on ratings, genre preferences, and the influence of series names on the reception of TV shows.

By leveraging this extensive TV show dataset, we gained valuable insights into the world of television, enabling us to uncover trends, patterns, and factors that contribute to the success and popularity of TV shows.

Dataset 8)

On this Dataset we collected the top 100 TV shows of all time, we sourced data from various datasets available on Kaggle and GitHub. These datasets provided valuable information about TV shows, including episode details, viewer ratings, directors, genres, and more. After collecting the

relevant datasets, we proceeded with the data cleaning and merging processes, resulting in a final dataset with (5604, 13) rows and columns.

The collection process involved identifying datasets that encompassed the desired information. We selected datasets that contained attributes such as season, episode, name, year, viewership, director, rating, votes, runtime, and genre. These datasets were gathered from both Kaggle and GitHub repositories, ensuring a diverse and comprehensive collection of TV shows.

Once the datasets were gathered, we cleaned the data to remove duplicates, handle missing values, and standardize the format of the attributes. Data cleaning is crucial to ensure data integrity and accuracy throughout the analysis process.

Following the cleaning phase, we merged the datasets based on common identifiers, such as show names or unique identifiers for each TV show. The merging process allowed us to consolidate the information from multiple datasets into a single, unified dataset.

The final merged dataset consisted of (5604, 13) rows and columns. The columns included season, episode, name, year, viewers, director, rating, votes, minutes, image URL, description, genre, and runtime. This comprehensive dataset served as a valuable resource for analyzing the top 100 TV shows of all time.

By collecting data from diverse sources and performing cleaning and merging operations, we obtained a consolidated dataset that provided a wealth of information about the most acclaimed TV shows. This dataset became the foundation for our analysis, enabling us to gain insights into the success, popularity, and characteristics of these top-rated TV shows

4.2 Conclusion

In conclusion, this chapter discussed the data considerations for our analysis of movies and TV shows. We outlined the datasets used, their sources, and the steps involved in collecting and preparing the data for analysis.

The first dataset consisted of 6,820 movies and provided various attributes such as budget, production company, genre, rating, release date, and more. The data was sourced from IMDb and underwent thorough validation and cleaning processes to ensure data quality and reliability.

The second dataset, also obtained from a publicly available repository on GitHub, included over 10,000 movies with attributes like popularity, budget, revenue, cast, director, and genre. This dataset was in its original form and required further cleaning and validation.

To enhance our analysis, we merged the first two datasets, resulting in a refined dataset of 5,421 unique movies. This merged dataset allowed us to explore research questions related to box office success, genre trends, and audience preferences.

Additionally, we collected data on the number of views and likes for movie trailers from the YouTube API, providing insights into audience engagement. The YouTube data was matched with the movie names and box office data, creating a comprehensive understanding of audience reception.

To further enrich our analysis, we obtained a popularity dataset from Kaggle, which included information on the popularity of actors. By matching this dataset with the cast data in our movie dataframe, we gained insights into the influence of popular actors on movie performance.

For TV shows, we obtained a comprehensive dataset from Kaggle, consisting of a vast amount of information on titles, seasons, episodes, ratings, genres, and series names. The dataset underwent cleaning and modifications to focus on the most relevant columns for our analysis.

In conclusion, the data considerations chapter highlighted the importance of data collection, validation, cleaning, and merging processes in preparing datasets for analysis. The diverse range of datasets used in our project provided valuable insights into the movie and TV show industry, enabling us to explore trends, patterns, and factors that contribute to success and audience reception.

Chapter 5: **Results**

5.1 RQ1: Can we predict the box office success of a movie based on its genre, budget, cast, and pre-release marketing? What factors are most strongly correlated with box office success, and how accurate are the predictions

The research question aims to investigate whether it is possible to predict the box office success of a movie based on its genre, budget, cast, and pre-release marketing. The question also seeks to identify the most strongly correlated factors with box office success and the accuracy of the predictions made.

To answer this question, several sub-questions need to be addressed. Firstly, what are the primary factors that influence box office success? Is it the genre of the movie, the budget allocated to the movie, the cast, or the pre-release marketing campaign? Secondly, how do these factors relate to one another? Is there a significant correlation between the budget allocated to the movie and the success of the pre-release marketing campaign, or is the genre of the movie more important in determining its success?

Additionally, it is essential to consider how accurate the predictions are and whether they are reliable. Are there any limitations or biases in the data collection process that may affect the accuracy of the predictions made? Are there any other external factors, such as competition from other movies, that may influence box office success?

Overall, the research question seeks to explore the factors that contribute to box office success and how accurately these factors can predict a movie's success. Understanding these factors and their relationship with one another can provide valuable insights for movie studios and filmmakers in creating successful movies.

5.1.1 Data & Method

5.2 Factor 1 Budget Box office

Our first factor for predicting the box office success of a movie is its budget. We will analyze the relationship between a movie's budget and its box office revenue to determine if there is a correlation between the two variables. We will use statistical techniques such as regression analysis to evaluate the correlation's strength and predict box office revenue based on a given budget. We will also compare the accuracy of our predictions with actual box office revenues to assess the effectiveness of our model. By examining the relationship between budget and box office success, we can gain insight into the role of financial resources in the success of a movie and improve our

understanding of the movie industry.

Our dataframe

	budget	popularity	production_companies	country	year	Box_office	runtime	name	genre	released	Imdb Rating	director	writer	Cast	profit
0	0	2.561161	[{'name': 'Walt Disney Pictures', 'id': 2}]	[{'iso_3166_1': 'US', 'name': 'United States o...	1995	0	97.0	Tom and Huck	Action	NaN	NaN	NaN	NaN	NaN	0
1	0	12.140733	[{'name': 'Universal Pictures', 'id': 33}, {'n...	[{'iso_3166_1': 'US', 'name': 'United States o...	1995	11348324	78.0	Balto	Family	NaN	NaN	NaN	NaN	NaN	11348324
2	0	2.228434	[{'name': 'BBC Films', 'id': 288}]	[{'iso_3166_1': 'GB', 'name': 'United Kingdom'...	1995	0	104.0	Persuasion	Drama	NaN	NaN	NaN	NaN	NaN	0
3	0	1.100915	[{'name': 'Ministère des Affaires Étrangères',...	[{'iso_3166_1': 'CN', 'name': 'China'}, {'iso_...	1995	0	108.0	Shanghai Triad	Drama	NaN	NaN	NaN	NaN	NaN	0
4	0	0.745542	[{'name': 'Iwerks Entertainment', 'id': 70801}]	[{'iso_3166_1': 'FR', 'name': 'France'}, {'iso...	1996	0	50.0	Wings of Courage	Romance	NaN	NaN	NaN	NaN	NaN	0

Figure 5.1: Dataframe RQ1

To analyze the progress of movie budgets, box office revenues, and R-squared values over the years, we initially adopted a straightforward approach. We calculated the mean budget, mean box office revenue, and mean R-squared values for each year in our dataset.

By taking the mean budget, we obtained an average value that represented the typical investment in movies for each year. Similarly, calculating the mean box office revenue provided an average indicator of the financial success of movies during different periods. The mean R-squared values allowed us to gauge the overall goodness-of-fit of regression models applied to the data.

By examining the trends in these mean values over time, we gained insights into the general progress and patterns in the film industry. For instance, increasing mean budgets might indicate growing investments in movie production, while rising mean box office revenues could suggest improving financial returns. Furthermore, tracking the mean R-squared values helped us assess the quality and accuracy of the regression models employed.

This approach allowed us to gain a broad understanding of the changing landscape of movie budgets, box office revenues, and model performance throughout the years, providing valuable insights into the industry's development and potential correlations between budget, revenue, and model accuracy.

Results

Now our second approaches is to groupby them by decades from 1940 to 2020

Analysis of film content reveals key insights from various decades. The average budget for the 1940s was $1,759,340.64, and the average box office was $16,920,913.98, with 42 films analyzed, and an R-square value of 0.00 The 1950s seem to be an outlier , where the average budget is $2,625,821.37, the average box office is $14,049,261.22 for 65 films, and the R-squared value is 0.62 going into the 1960s average budget of $4,683,979. rose to 62, with a 55-film box office average of $14,464,136.85, giving an R-square value of 0.15 In the 1970s, the average budget for 57 films was $4,784,365.02, and the average box office was $17 ,124,736.60, and the R-squared value was 0.01. Turning to the 1980s, the average budget increased to $5,316,356.19, the average box office reached $11,373,389.06 in 47 films, resulting in an R-square value of 0.18 On average of the budget for the 1990s was $ 7,962,568.01, with an average box office of $16,517, 8.28 for 98 films, and an R-square value of 0.32 In 2000, on average equivalent budget increased to $9,145,680.61,

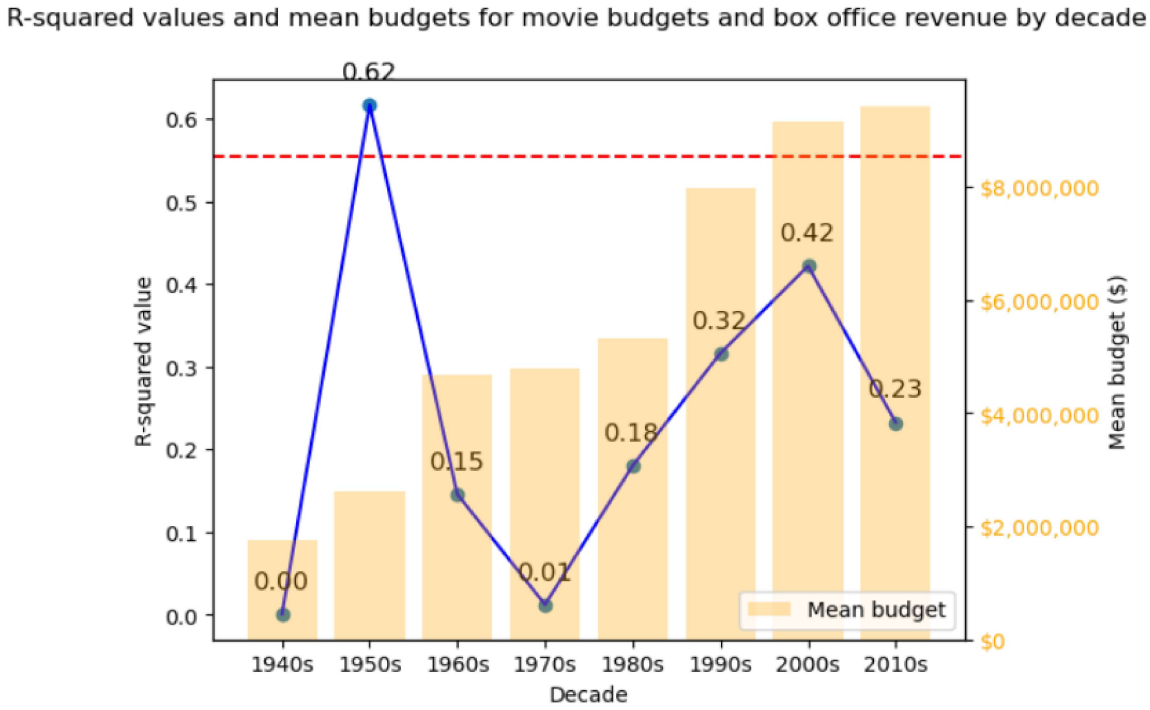


Figure 5.2: R-square values during decades

while the average box office revenue of 283 films was $14,671,541.35, with an R-square value of 0.42 Finally, the budget for 367 films in year The 2010s averaged $9,421,979.43, the average box office was $23,599,115.53 , and the R-square value was 0.23 This information provides valuable insight into the relationship between budget and box office receipts of films in between different decades, and appears as someone out of the box in the 1950s

Discussion) Based on the pooled data, the R-square value between budget and box office for all movies and years is 0.533. This showed a downward correlation between budget and box office performance.

The R-square value indicates that about 53.3

Predicting box office success accurately requires consideration of other variables such as marketing strategy, personnel, format, release time, wide acceptance, and audience preferences all of these factors together play an important role in determining the financial success of a film.

So while budget is an important determinant of a film's performance, it shouldn't be the only consideration when predicting a film's success. More detailed analysis with more variables and more sophisticated sampling methods will lead to more accurate forecasting of box office success.

5.2.1 Factor 2: Genre and Box office

In this analysis, we are investigating whether genre plays a significant role in predicting box office success. By examining the count of movies in each genre for each decade from 1950 to 2020, we can observe trends and changes in genre preferences over the years.

The provided data showcases the count of movies in various genres for each decade. Looking at the numbers, we can identify some interesting patterns.

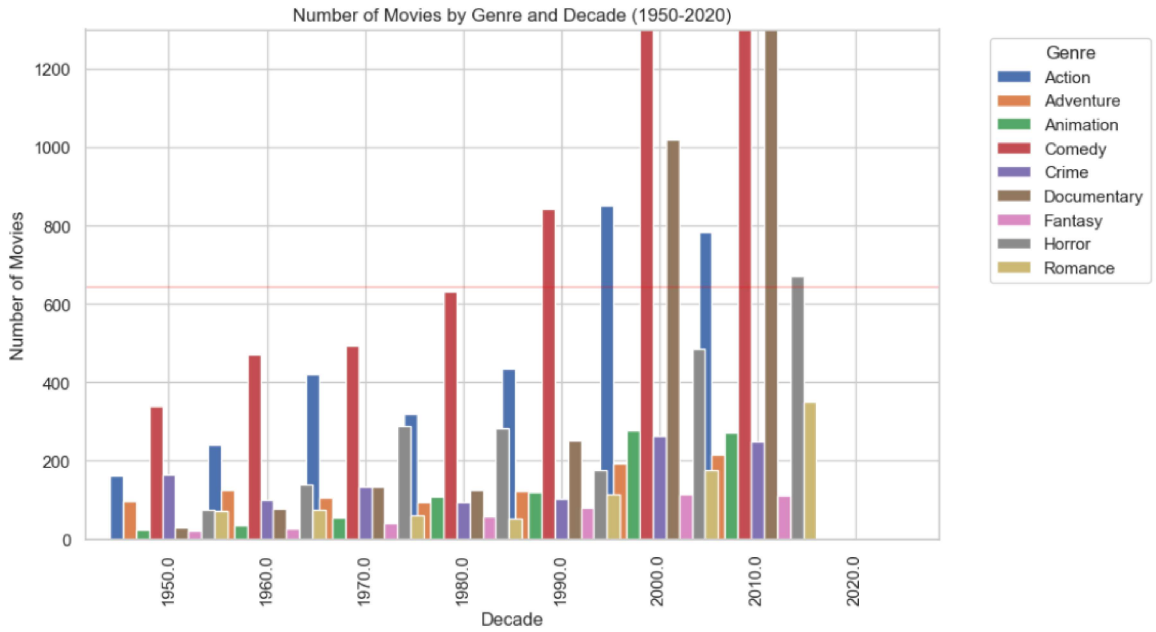


Figure 5.3: Movies in genre during the years

During the 1950s and 1960s, genres like Action, Adventure, and Comedy had a moderate presence, while Animation and Fantasy had relatively fewer movies. Crime and Documentary genres also had a notable representation. Horror and Romance genres were moderately popular during this period.

Moving into the 1970s, the count of movies in Action, Crime, and Documentary genres increased significantly. Adventure and Animation genres maintained a steady presence, while Fantasy genre experienced a slight rise. The count of Comedy movies remained consistent, and Horror and Romance genres saw a notable increase.

In the 1980s, the count of movies surged in genres like Comedy, Adventure, Animation, and Fantasy. Action and Crime genres also experienced growth, while Horror and Romance genres showed a steady presence.

The 1990s witnessed a rise in movies across all genres, with Comedy, Adventure, Animation, and Horror genres showing notable increases. Action and Romance genres also experienced growth, while Crime genre saw a slight decline.

As we move into the 2000s and 2010s, the count of movies in all genres significantly increased. Comedy, Adventure, Animation, and Horror genres continued to be popular, with Action and Romance genres also maintaining a strong presence. Documentary genre saw a substantial rise during these decades.

These observations highlight the dynamic nature of genre preferences in the film industry. Certain genres like Comedy, Adventure, Animation, and Horror have consistently attracted audience attention over the years. The rise of the Action genre and the varying popularity of genres like Romance and Crime demonstrate evolving audience tastes and trends in filmmaking.

In our second approach, we examined the profitability of movies in each genre for each decade from 1950 to 2020. By analyzing the profit figures, we can observe interesting trends and patterns in different genres over time.

Looking at the data, we can make several observations about the genres:

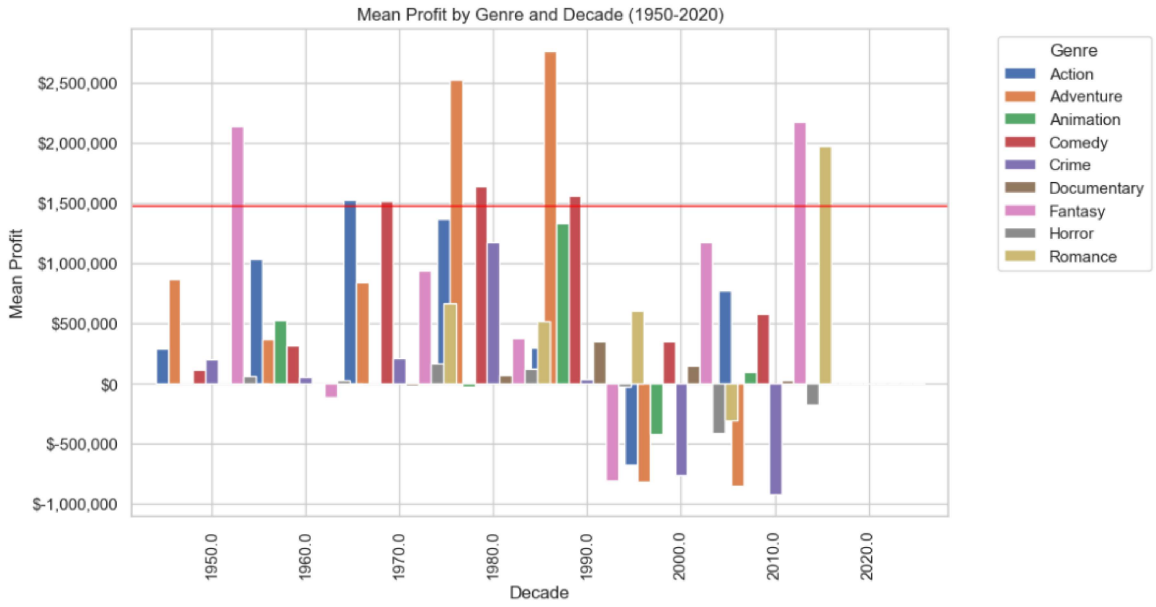


Figure 5.4: genres profit

Action: Action movies consistently generated high profits throughout the decades, especially in the 1960s and 1970s. However, in the 2000s, action movies experienced a decline in profitability.

Adventure: Adventure movies showed varying profitability over the years. They had significant success in the 1960s and 1990s, but experienced a decline in the 2010s.

Animation: Animation movies were not present or did not generate significant profits in the 1950s and 1960s. However, they gained popularity and became highly profitable starting from the 1970s onwards, with the 1990s and 2010s being particularly successful decades.

Comedy: Comedy movies had a relatively consistent profitability, with the 1980s and 1990s being particularly lucrative. However, in recent years, their profitability has declined.

Crime: Crime movies showed mixed profitability throughout the decades, with notable success in the 1970s. However, in the 2000s and 2010s, they experienced a decline in profitability.

Documentary: Documentary movies had varying profitability, with significant success in the 1980s and a resurgence in the 2010s.

Fantasy: Fantasy movies were not highly profitable in the earlier decades but gained considerable popularity and profitability from the 1990s onwards.

Horror: Horror movies showed varying profitability, with the 1970s and 2010s being particularly successful decades.

Romance: Romance movies had mixed profitability, with the 1950s and 2010s being relatively more profitable decades.

These observations highlight the changing dynamics of genre profitability over time. While some genres consistently perform well, such as action and animation, others experience fluctuations in their profitability. It is essential to consider these trends when assessing the impact of genre on box office success.

By analyzing the profitability of movies in each genre for different decades, we can gain insights into the genres that have historically shown high profitability and those that have faced challenges.

This information can guide further analysis to determine the extent to which genre influences box office success and assist in making informed decisions within the film industry.

The third approach involved using an Ordinary Least Squares (OLS) regression model to analyze the relationship between box office success and genre. The regression results indicate the coefficients and statistical significance of each genre category.

The overall model's R-squared value is 0.055, suggesting that the genre alone explains approximately 5.5

Examining the coefficients, we can see that several genres have significant impacts on box office performance. For example, genres like Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, History, Horror, Music, Mystery, Romance, Science Fiction, Thriller, War, and Western all have statistically significant coefficients.

However, it's important to note that the coefficients represent the effect of each genre relative to the intercept (which represents the baseline genre). Positive coefficients indicate that a particular genre tends to have a positive impact on box office performance compared to the baseline, while negative coefficients suggest a negative impact.

The coefficients can be interpreted as follows: for each specific genre, the coefficient represents the average change in box office revenue compared to the baseline genre. For instance, a positive coefficient for Adventure suggests that Adventure movies tend to have higher box office revenue than the baseline genre, while a negative coefficient for Comedy indicates that Comedy movies tend to have lower box office revenue compared to the baseline.

Overall, the OLS regression model helps us understand the individual effects of different genres on box office success. However, it is important to consider that the model's R-squared value is relatively low, indicating that genre alone does not explain a substantial portion of the variance in box office performance. Other factors such as marketing, cast, production quality, and release timing are likely to play significant roles in determining a movie's success at the box office.

Discussion For genre) The analysis conducted using the Ordinary Least Squares (OLS) regression model provides insights into the relationship between genre and box office success. However, the results suggest that genre alone is not a strong predictor of box office performance. The model's low R-squared value of 0.055 indicates that only about 5.5

Although the regression results show that several genres have statistically significant coefficients, indicating an impact on box office performance, it is crucial to interpret these coefficients in the context of the baseline genre. Positive coefficients suggest that certain genres tend to have a positive effect on box office revenue compared to the baseline genre, while negative coefficients suggest a negative effect.

While genres such as Adventure, Animation, Biography, Comedy, Crime, Documentary, Drama, Family, Fantasy, History, Horror, Music, Mystery, Romance, Science Fiction, Thriller, War, and Western have significant coefficients, it's important to note that these coefficients represent the average change in box office revenue compared to the baseline genre. It does not imply that a specific genre guarantees success or failure at the box office.

Considering the low R-squared value and the complex nature of box office success, it is clear that other factors beyond genre play significant roles. Marketing efforts, the cast, production quality, and release timing are just a few examples of additional factors that influence a movie's performance at the box office.

In conclusion, while the genre has some influence on box office success, it is not the sole determining factor. The OLS regression model helps us understand the individual effects of different genres,

but a comprehensive analysis should consider a broader range of factors to make informed decisions within the film industry.

5.2.2 Factor 3: Cast and crew For Box office

In our data-driven approach to predicting box office success, we investigated the potential of using cast and crew popularity as well as votes and ratings. By collecting data on the popularity metrics of actors, directors, and other crew members, we aimed to determine their correlation with a film's financial performance. Additionally, we incorporated audience votes and ratings as indicators of audience reception and film quality.

Using statistical methods like regression analysis or machine learning algorithms, we built a predictive model that considered these factors together. By training the model on a substantial dataset, it learned the patterns and relationships between cast and crew popularity, votes, ratings, and box office success. This approach provided a comprehensive understanding of the factors influencing box office performance.

By combining the popularity of the cast and crew with audience votes and ratings, we enhanced the model's predictive power. The inclusion of audience sentiment and preferences allowed for more accurate predictions of a film's potential box office success. This approach provides valuable insights for decision-makers in the film industry, helping them allocate resources effectively and make informed choices when assembling teams and designing marketing strategies. However, it's important to note that factors like script quality, production value, and marketing efforts also contribute to a film's success, and these should be considered alongside cast and crew popularity and audience feedback.

Results)

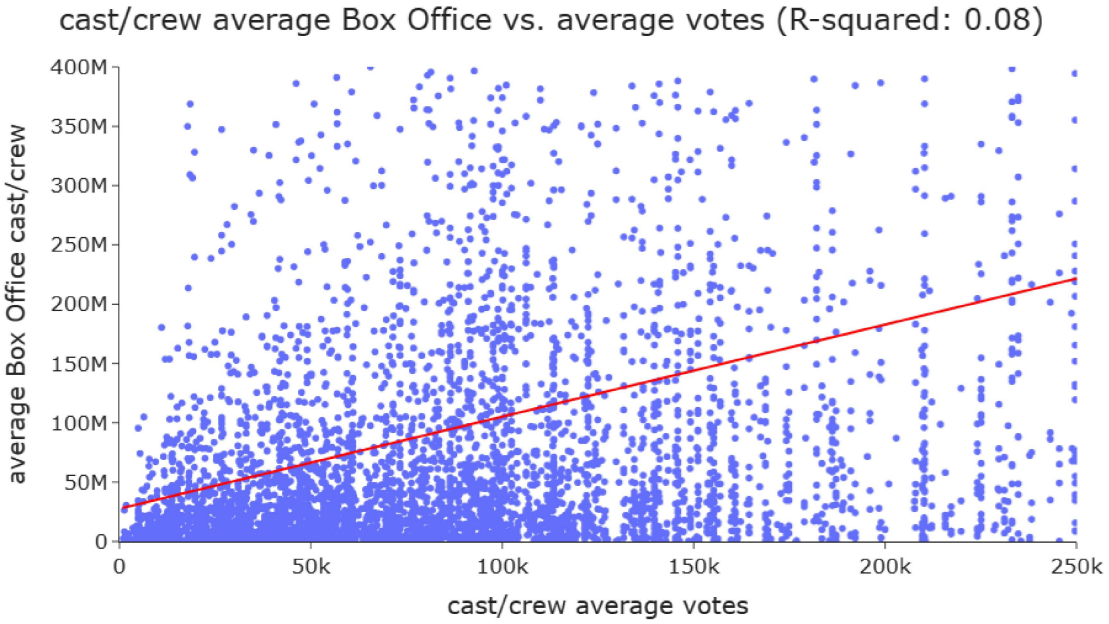


Figure 5.5: linear regression for average votes

In our linear regression analysis, we explored the relationship between the average votes of a cast and crew and their corresponding box office performance. The Ordinary Least Squares (OLS) regression results provide insights into the statistical significance and coefficients of the variables in the model.

The regression model yielded the following results:

The intercept, represented by the constant, has a coefficient of 2.786e+07, indicating the baseline box office value when the average votes of the cast and crew are zero. This coefficient is statistically significant at a confidence level of 0.05, as the p-value is less than 0.001.

The average votes of the cast and crew have a coefficient of 774.2978, indicating the change in box office revenue for every unit increase in average votes. This coefficient is also statistically significant at a confidence level of 0.05, with a p-value of less than 0.001.

The R-squared value of 0.087 suggests that approximately 8.7

Overall, the results indicate that there is a statistically significant relationship between the average votes of the cast and crew and box office revenue. However, the low R-squared value suggests that other factors not included in the model play a significant role in determining box office success. Further analysis and consideration of additional variables are necessary to improve the predictive accuracy of the model.

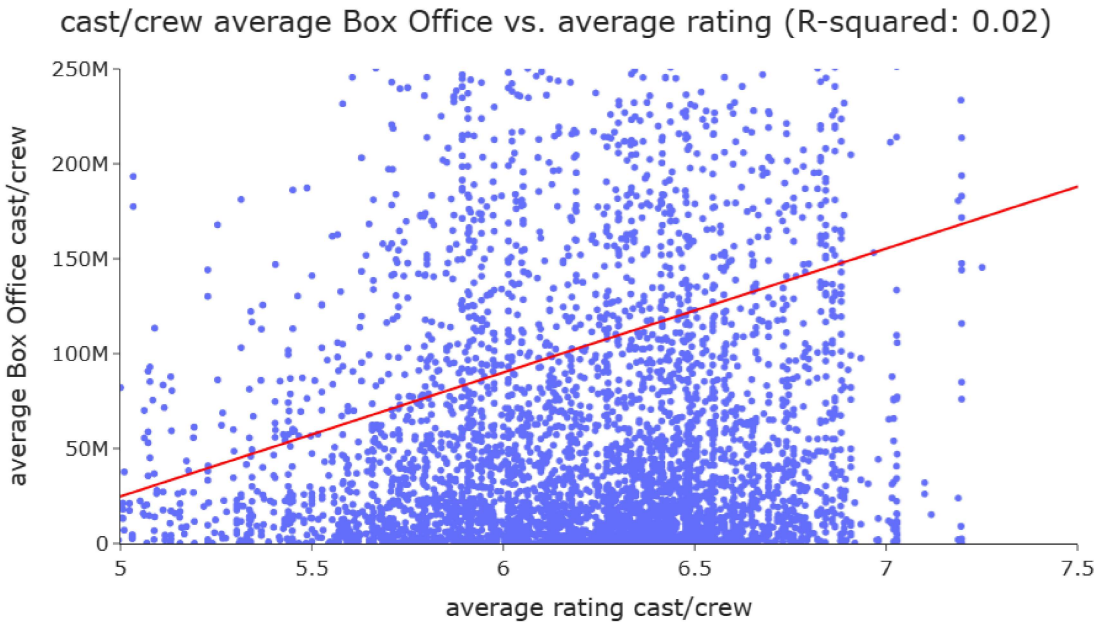


Figure 5.6: linear regression for average rating

In our linear regression analysis, we explored the relationship between the average rating of a cast and crew and their corresponding box office performance. The Ordinary Least Squares (OLS) regression results provide insights into the statistical significance and coefficients of the variables in the model.

The regression model yielded the following results:

The intercept, represented by the constant, has a coefficient of -3.22e+08, indicating the baseline box office value when the average rating of the cast and crew is zero. This coefficient is statistically significant at a confidence level of 0.05, as the p-value is less than 0.001.

The average rating of the cast and crew has a coefficient of 6.858e+07, indicating the change in box office revenue for every unit increase in average rating. This coefficient is also statistically significant at a confidence level of 0.05, with a p-value of less than 0.001.

The R-squared value of 0.028 suggests that approximately 2.8

Overall, the results indicate that there is a statistically significant relationship between the average rating of the cast and crew and box office revenue. However, the low R-squared value suggests that other factors not included in the model play a significant role in determining box office success. Further analysis and consideration of additional variables are necessary to improve the predictive accuracy of the model.

5.3 Discussion

Regarding average votes, we found a statistically significant relationship between the average votes of the cast and crew and box office revenue. The coefficient of 774.2978 suggests that for every unit increase in average votes, there is a corresponding change in box office revenue. However, the R-squared value of 0.087 indicates that only 8.7

Similarly, when considering average rating, we observed a statistically significant relationship. The coefficient of 6.858e+07 indicates the change in box office revenue for each unit increase in average rating. However, the low R-squared value of 0.028 suggests that only 2.8 % of the variation in box office performance can be explained by the average rating of the cast and crew.

Overall, these results indicate that there is a statistical association between cast and crew popularity, as measured by average votes and average rating, and box office revenue. However, the low R-squared values imply that other factors not accounted for in the model significantly influence box office success. To enhance predictive accuracy, further investigation is needed, including the consideration of additional variables such as marketing strategies, production budget, and audience demographics.

5.3.1 Factor 4: Pre-release marketing

In order to predict box office success for movies released from 2010 and onwards, we employed a data-driven approach that focused on utilizing movie trailer likes and views as predictors. Recognizing the significance of online engagement and social media influence on audience behavior, we aimed to leverage these metrics to gauge a film's potential popularity and financial performance.

Our approach involved collecting data on movie trailer likes and views for a wide range of films released in the specified timeframe. These metrics served as indicators of audience interest and anticipation towards the movies. By quantifying the level of online engagement through likes and views, we sought to establish a correlation between this digital buzz and box office success.

To implement our approach, we utilized statistical modeling techniques such as regression analysis or machine learning algorithms. By training the model on a comprehensive dataset, we aimed to identify patterns and relationships between movie trailer engagement and box office performance. This allowed us to develop a predictive model that can provide insights into the potential financial success of a film based on its trailer likes and views.

By incorporating movie trailer likes and views into our predictive model, we aimed to enhance the accuracy of box office predictions. This approach takes advantage of the digital landscape and the ability to capture audience sentiment and excitement, providing a valuable tool for decision-makers in the film industry. By considering online engagement metrics, industry professionals can better understand audience interest and adjust marketing strategies to maximize a film's potential success.

5.4 Results

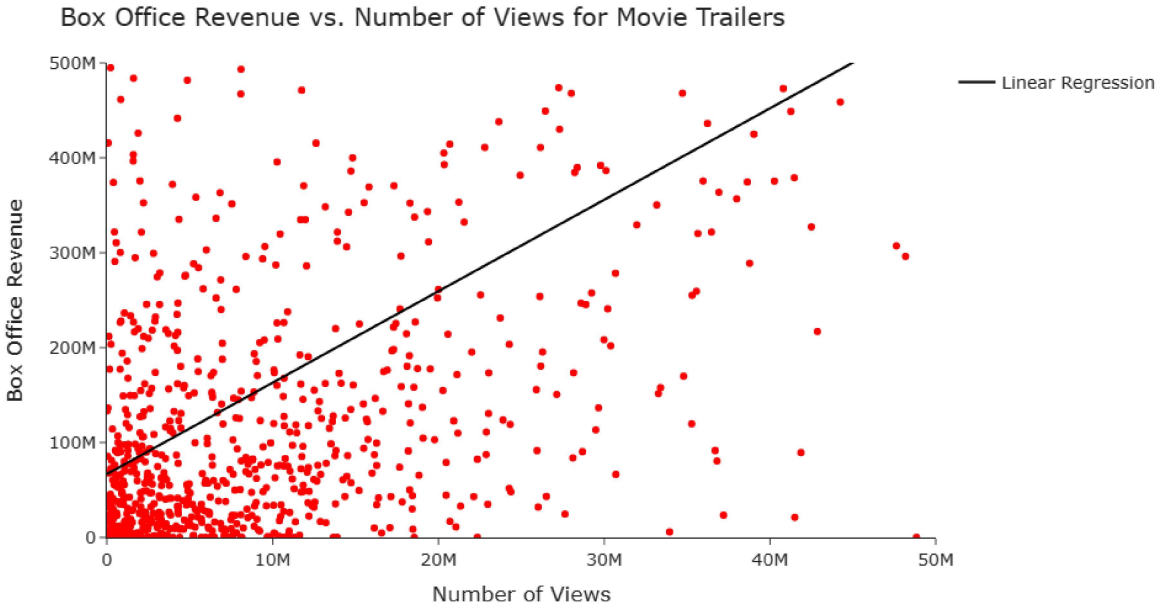


Figure 5.7: linear regression for number of views

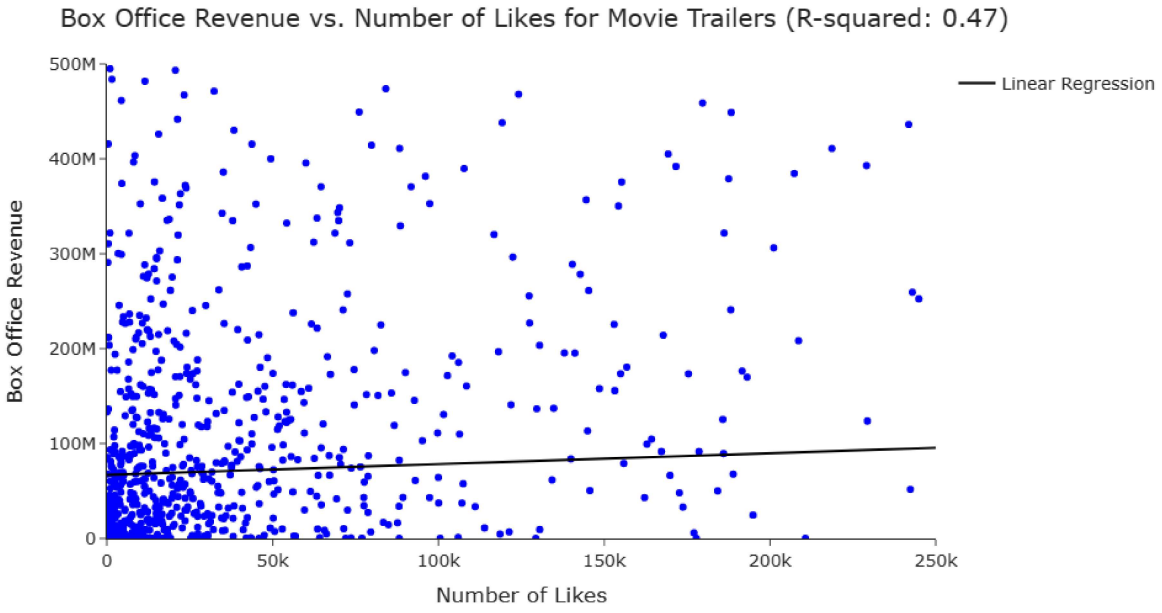


Figure 5.8: linear regression for number of likes

In our study, we employed linear regression analysis to examine the relationship between movie trailer views and likes and the success of the corresponding movies at the box office. We utilized ordinary least squares (OLS) regression, a widely used method for modeling linear relationships.

The results of our OLS model revealed several important findings. First, the coefficient of the constant term (const) is estimated to be 6.685e+07. This means that holding all other factors constant, the film's expected box office revenue from missed trailer viewers and intended audiences would be around $66.85million$

Furthermore, the coefficient of Likes variable was found to be 114.5626 with standard error of

53.921. Positive theory indicates that there is a positive correlation between the number of fans consumed by a movie and its success at the box office. For every addition like this, we can expect the average box office to increase by about $114,562.6$.

Similarly, the estimated coefficient of Views variable is 9.6336 with standard error of 0.691. This theory suggests that there is a positive correlation between the number of views of a transportation film and its performance at the box office. On average, for each new episode, we can predict that the average box office will increase by about $9,633.6$.

The statistical significance of these findings was assessed using the t-test. Both the Likes and Views variables showed a statistically significant correlation with box office revenue, as their respective p-values were less than the usual 0.05 This indicates that they are unlikely to be correlated discovered not by chance alone.

5.5 Discussion

The overall performance of the model was assessed using the R-squared statistic which measures the variance of the dependent variable (Box$_o ffice)explainedbytheindependentvariables(LikesandViews).andcani$

In summary, our analysis using linear regression and OLS model showed that movie trailer views and likes have a statistically significant positive relationship with box office revenue but it is important to note that regression analysis cannot establish causality, other factors not considered in this model. can affect success

5.6 Overall conclusion for all factors

Based on the factors analyzed, including budget, genre, average votes, average ratings, screenplays, and preferences, we can conclude that say box office success based solely on these factors is challenging although some of these factors have shown statistically significant relationships with and box office R-square values are quite low, indicating variability in the box office only a small percentage can be explained by these factors alone

The budget factor showed a moderate correlation with box office success, with an R-square value of 0.533. This means that about 53.3

Genre, although showing statistically significant effects for some genres, had a low R-square value of 0.055, indicating that only about 5.5

Similarly, voter turnout and employee turnout and average exhibited statistically significant correlations, but their low R-square values (0.087 and 0.028, respectively) suggest explaining variable a it happens in only a fraction of the box office success

To ensure the accuracy of box office forecasts, it is important to consider other variables such as marketing strategy, product quality, release time, critical reception, and audience preference Factors more inclusion improves predictive capabilities through sophisticated modeling techniques.

In conclusion, while the analyzed products show some statistical correlation with box office success, their individual contributions are limited. Forecasting the performance of the film industry requires

a comprehensive analysis that includes many factors and considers the complexity of the film industry.

5.7 RQ2: Does the number of episodes and seasons in a TV show's season impact its ratings and number of votes,views, and are there any optimal episode lengths or numbers for maximizing a show's success?

Introduction: This study aims to investigate the relationship between the amount of episodes in a season of TV programming and the number of episodes, votes and viewers As content has increased, the television industry has witnessed changes greater in recent years The key is to understand how this is done including its role on it to analyze data on ratings, views and opinions from which we can gain insight into the optimal episode length and number.

summary: This study examines the impact of a TV show's content and duration on its ratings, vote share and viewership. By analyzing data from TV shows, we aim to identify any patterns or relationships between these variables. In addition, we are trying to determine if there is an optimal program length or number that contributes to the success of the show. The findings of this study will provide producers and producers with valuable insights to create TV programs that engage and engage audiences. Understanding the relationship between program structure and viewer response can help improve the quality of the viewing experience and improve the chances of a show succeeding in an increasingly competitive television environment developed

5.7.1 Approach

While addressing the research question of whether the number of episodes and number of seasons during a TV show affects its ratings, votes, and viewership, and whether there is an optimal program length or rating to provide a program has been very successful, we adopted a two-step approach using two data frames.

First, we used our original data set, which contained relevant information such as TV show name, episode title, season number, episode number, average rating, . characteristics, series names, and years we used We focused on analyzing the average ratings of different TV shows ranges of episode numbers. The purpose of this study was to see if there is a correlation between the number of episodes of a TV show and the average ratings.

Second, we used a secondary data set that examined the relationship between the number of episodes and occasions and the mean number of votes and observations. We analyzed the consensus across contexts and examined whether there were features or patterns in the participatory audience. In addition, we examined average trends based on different time periods, as well as trends in specific cases. The purpose of this study was to uncover possible relationships between the timing of the event and audience participation.

Following this two-step approach and using these two data frames, we aimed to gain insight into

the effects of episode and season numbers on ratings, opinions, and views The study sought to determine if episode length several are much better for TV enhancement means success. The findings of this line of research will contribute to a better understanding of the relationship between TV program design and viewer reception, enabling producers and producers to make informed TV decisions in the development and execution of activities

5.7.2 Results

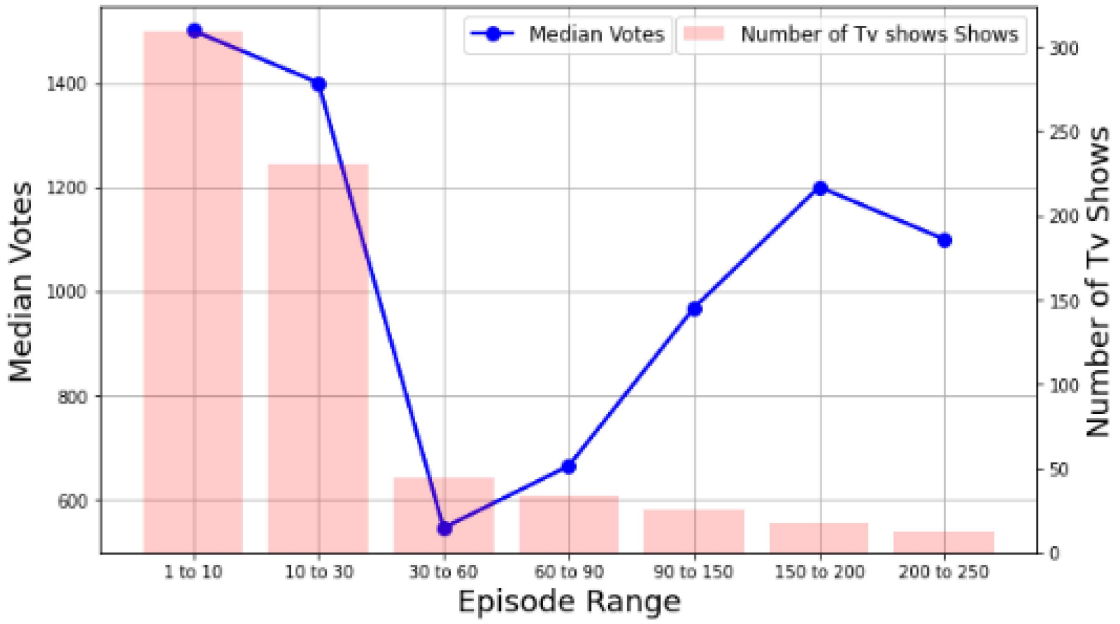


Figure 5.9: Median votes for episodes

Case range analysis has revealed interesting patterns in the relationship between the average intention and the exceptional rating.

TV shows in the episode range "1 to 10" received an average of 1500.0 votes, with 309 unique shows in this category. This suggests that TV shows with less content attract higher numbers of votes, potentially indicating higher viewer engagement.

Moving to the "10 to 30" episode range, the median vote decreased slightly to 1400.0, while the number of unique shows decreased to 231. This indicates that TV shows with slightly higher numbers remain maintains decent levels of viewer engagement, but the more unique shows start to decline.

As the volume increases, we see a decrease in both the average vote and the number of unique games. This means that as TV shows get more episodes (from 30 to 250), viewership decreases, which can lead to viewer fatigue or loss of interest over time.

However, when interpreting the results, it is important to take into account the small number of TV shows in the top programming categories. The reduced sample size in these groups may introduce bias and limit the generalizability of the findings.

TV shows in the "Seasons 1 to 3" category had 1.074994 million viewers, and there were 14,780 shows in this category. This suggests that TV shows on average attract higher numbers of viewers in the earlier seasons, which could be due to novelty and interest in novelty

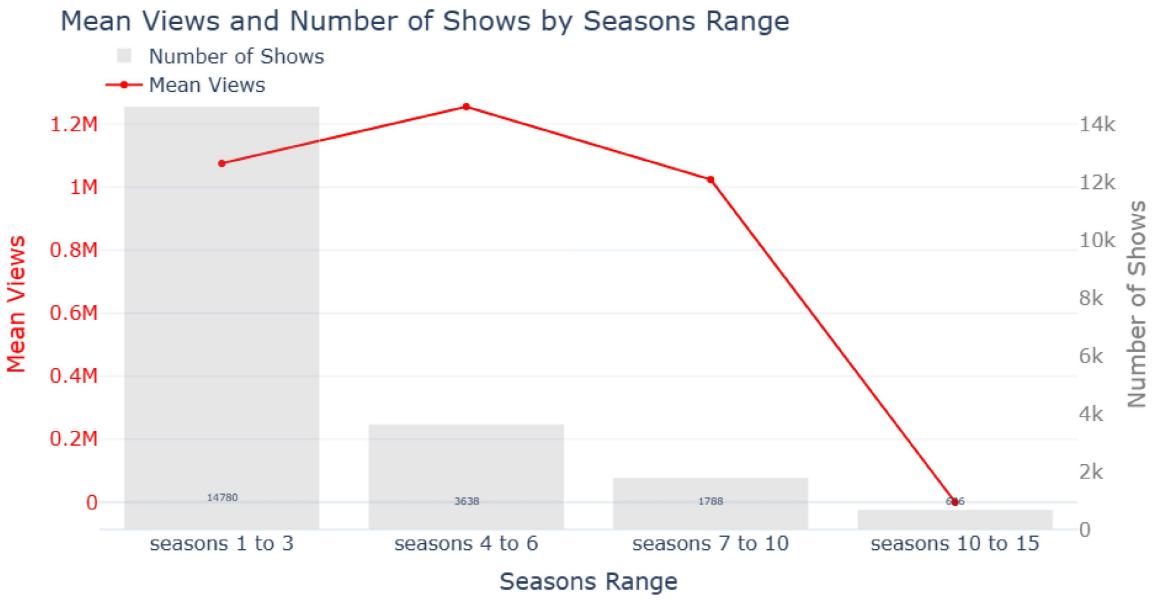


Figure 5.10: mean views for seasons

In the "Seasons 4 to 6" category, the average viewership rose slightly to 1.254,902 million, while the number of shows fell to 3,6 This means that TV shows still maintain high levels of viewer engagement in the middle of the season, although the number of shows begins to decline.

As we move to higher types of seasons, such as "Seasons 7 to 10" and "Seasons 10 to 15" we see a decrease in average viewership and number of episodes This shows that when TV shows moving on to later seasons, viewers become less engaged, which can lead to audience fatigue or because of a natural decline in interest as time goes on

In interpreting the results, it is important to consider the relatively low proportion of Shona during the high periods. Reduced sample size in these areas may limit the generalizability of the findings and may introduce bias.

Interesting patterns are revealed in the relationship between the number of shows and the average viewership of TV shows.

There were 8,647 shows in the "1-10" episode range, and an average of 837,035 viewers. This means that there are more TV shows with fewer episodes on average.

After passing the "10-30" episode range, the number of shows dropped to 6,423, but the average viewership rose to 942,6 This means that TV shows with slightly higher ratings still maintain a decent level of viewer engagement.

As the programs rise above 30, the number of shows and weak exhibits begins to dwindle. This means that TV shows with more content lead to lower viewership.

It should be noted that the number of shows drops dramatically in higher episodes, such as "200-250", and "250-300". There are few central ideas in these categories, indicating a small sample size and possible lack of audience interest.

Overall, the study shows that TV shows with moderate programming, around 10-30 episodes, have higher viewer engagement.

The graph uses a treemap visualization to display the number of TV series in each category. The

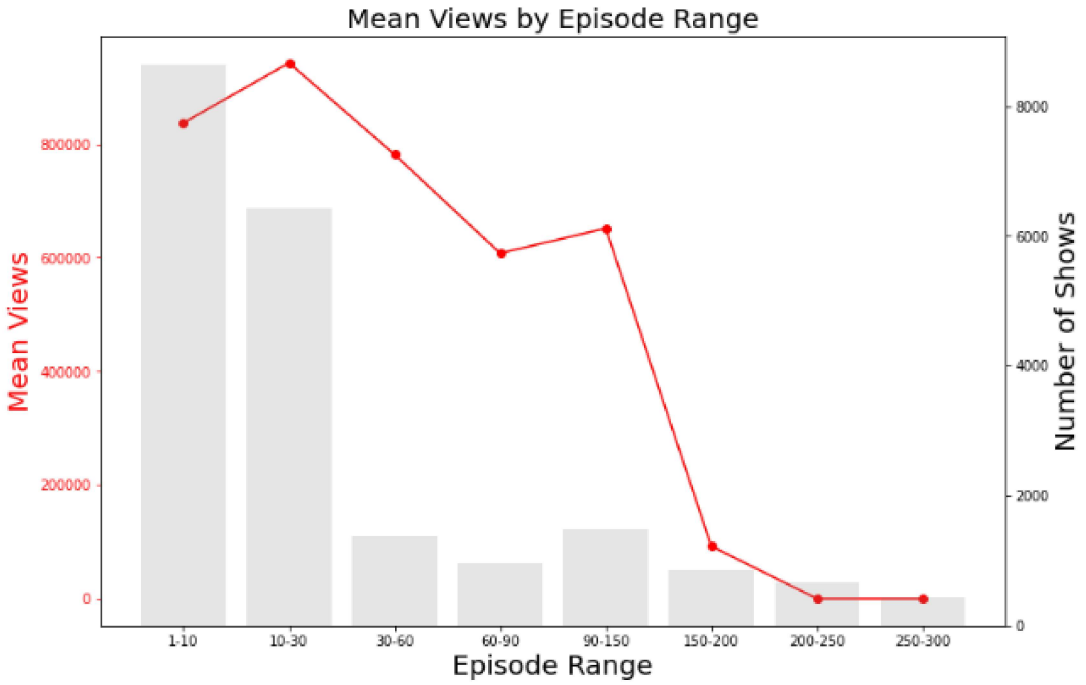


Figure 5.11: mean views for episodes

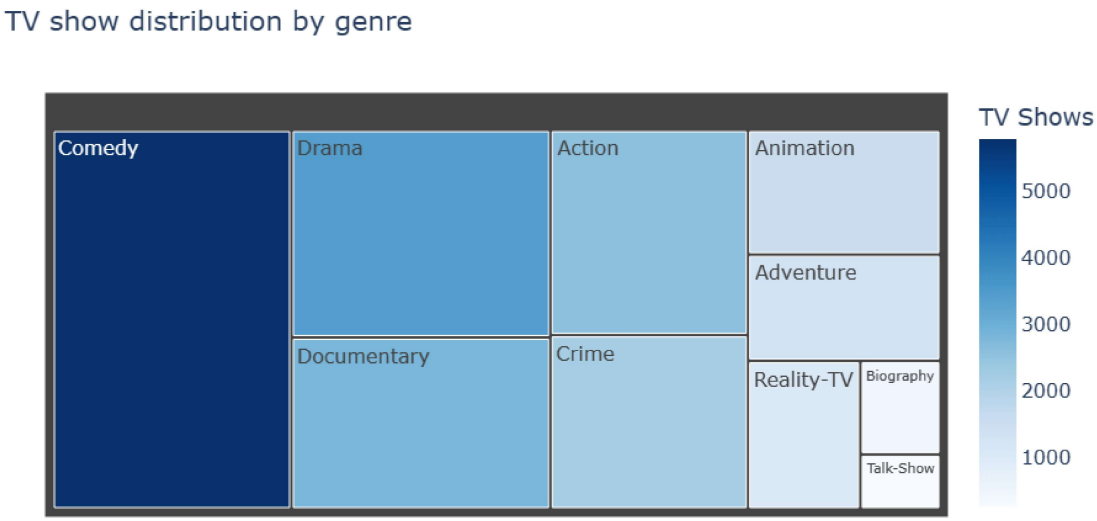


Figure 5.12: Tv shows in genre

darker blue tints suggest a greater quantity of shows in that category.

The bar chart shows ratings for different genres, with each color representing a different episode. The chart allows visual comparison of ratings across genres and displays the distribution of case ranges within each genre.

Examination of the average number of modes and their profiles reveals interesting patterns.

Among all genres, the drama has consistently high ratings across its episodes, with ratings ranging from 7.60 to 7.80. Gossip and documentaries show slightly higher numbers, indicating viewers are

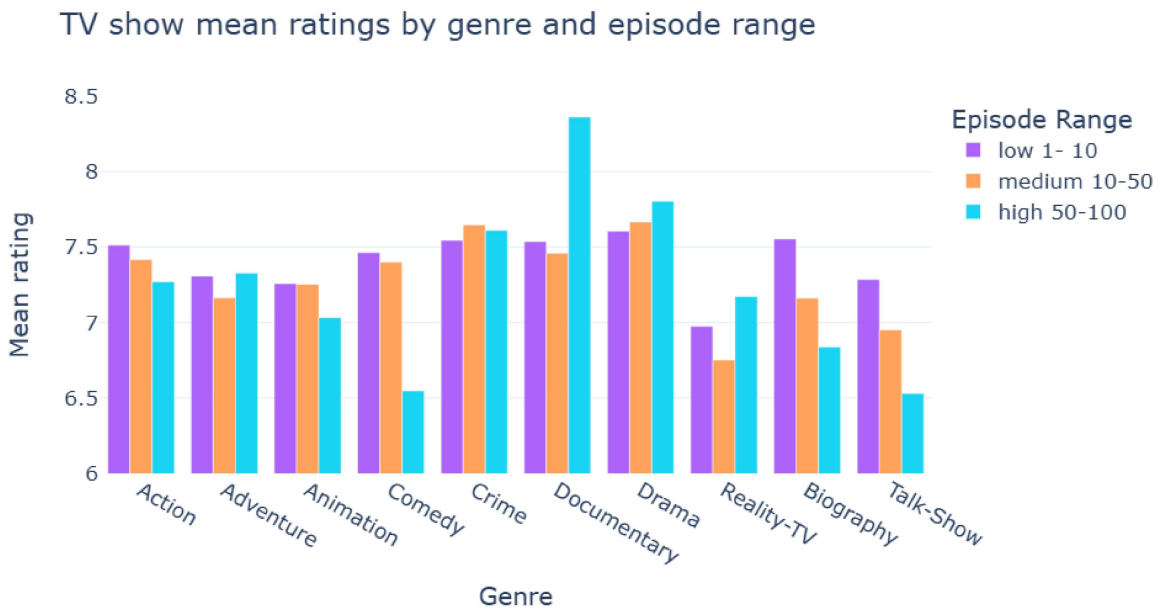


Figure 5.13: Rating in genres

happy.

In contrast, humor shows that the number of users decreases with increasing levels of activity. While it has a decent rating in the lower grade (7.46), the rating drops significantly in the higher grade (6.55).

The adventure and animation genres exhibit a similar pattern, with ratings decreasing as the number of episodes increases. Adventure shows a rating drop from 7.31 to 7.16 to 7.33, while animation drops from 7.26 to 7.25 to 7.03.

Reality-TV and talk-shows generally receive lower ratings than other genres, with lower ratings in high-profile programming.

The life stories show a mixed pattern, with slightly higher ratings at lower scores (7.55) and decreasing ratings as content expands (from 7.16 to 6.84).

The study shows that some genres maintain consistently high ratings across event types, while others experience decreasing ratings as levels increase.

5.7.3 Discussion

The study provides insights into the effects of program types and number of occasions on TV ratings, viewership, and engagement.

As for the number of episodes, it was found that TV shows with short series, especially in the "1 to 10" category, attracted higher average ratings and votes, which shows that viewer engagement increased but as programming increased, average ratings and votes decreased, . Potential under-representation of TV shows in high content due to viewer fatigue or decreased interest over time should be noted, which may affect the applicability of the findings in general terms.

Similarly, in examining the effect of the season, TV shows such as "Seasons 1 to 3" are found in earlier seasons to have more viewership and engagement as time passes on, while the average

number of viewers and activity levels declined, indicating a warning decline in viewership.

Analysis of the genres and their lyrics reveals interesting patterns. Drama maintains consistently high ratings across episode types, while comedies and other genres show ratings decreasing as episode levels increase. Adventure and animation genres also exhibit declining numbers as episodes expand. Reality TV and talk-shows generally get lower ratings than other genres.

Overall, the study shows that there is no ideal episode length or number of seasons to maximize the success of a TV show Factors such as viewer involvement, fatigue, strategy so do and innovation play an important role. Content producers should carefully consider the impact of different programming and seasons on ratings, impressions, and viewership when planning and producing TV shows.

: :

5.8 RQ3: What are the key attributes that contribute to highly-rated movies, and do these attributes vary across cultural contexts?

5.8.1 Data & Method

For our research, we used the TMDB 5000 dataset, which provides detailed information about films, including their attributes such as budget, revenue, vote share, popularity, characteristics, and original language

We used several data analysis methods and visualization techniques to address our research question. First, we explored the relationships between the attributes by constructing a heat map with the data set. This allowed us to identify which characteristics are most closely associated with income, i.e., budget and voting rates.

In addition, we conducted analyses to examine the relationship between budget, income, and the number of selected continents. We created a scatter plot to visualize this relationship, with the larger bubble indicating the rating. In addition, we analyzed the total number of votes counted in each category to determine audience engagement.

To gain further insight into the underlying patterns in the data set, we performed cluster analysis. In particular, we examined the budget, revenue, vote count, and popularity columns. Using the angle method

5.8.2 Results

In this section, we present the results of our investigation into the key characteristics that contribute to blockbuster films and whether these characteristics vary across cultures.

First, we examined the relationships between the characters in our data set. We created a heatmap visually representing the correlation values. The heat map revealed interesting insights into the relationships between properties.Our findings indicate that the attributes with the highest cor-

Figure 5.14: heat map

relation to revenue are budget and vote count. This suggests that movies with higher budgets tend to generate more revenue, and a larger number of votes is associated with higher revenue as well. This highlights the importance of financial investment and audience engagement in determining the commercial success of a movie.

in fig 5.15 Next, we explored the total vote count per genre. By analyzing the data, we were able to identify the genres that received the highest number of votes. The results showed that the Action genre received the highest number of votes, followed by Drama and Adventure genres. This suggests that audiences tend to engage more with movies in the Action genre, showing a preference for this particular genre. These findings can be valuable for filmmakers and studios in understanding audience preferences and tailoring their production strategies accordingly.

in fig 5.16 To gain further insights into the underlying patterns within the data set, we conducted a cluster analysis using the columns 'budget', 'revenue', 'vote$_count'$, $and'popularity'.Byapplyingtheelbowmethod,$

in fig 5.17 We also investigated the relationship between budget, revenue, and rating for selected continents. Our scatter plot visually represented this relationship, with the size of each bubble representing the rating. The x-axis represented the budget, and the y-axis represented the revenue. The analysis of the scatter plot revealed interesting findings. We observed variations in the relationship between budget, revenue, and rating across different continents. This indicates that cultural contexts may influence the financial success and reception of movies. These findings highlight the importance of considering cultural factors when analyzing the movie industry and suggest that different regions may have unique preferences and dynamics that impact movie performance.

in fig 5.18 Additionally, we examined the relationship between mean ratings and budget. Surpris-

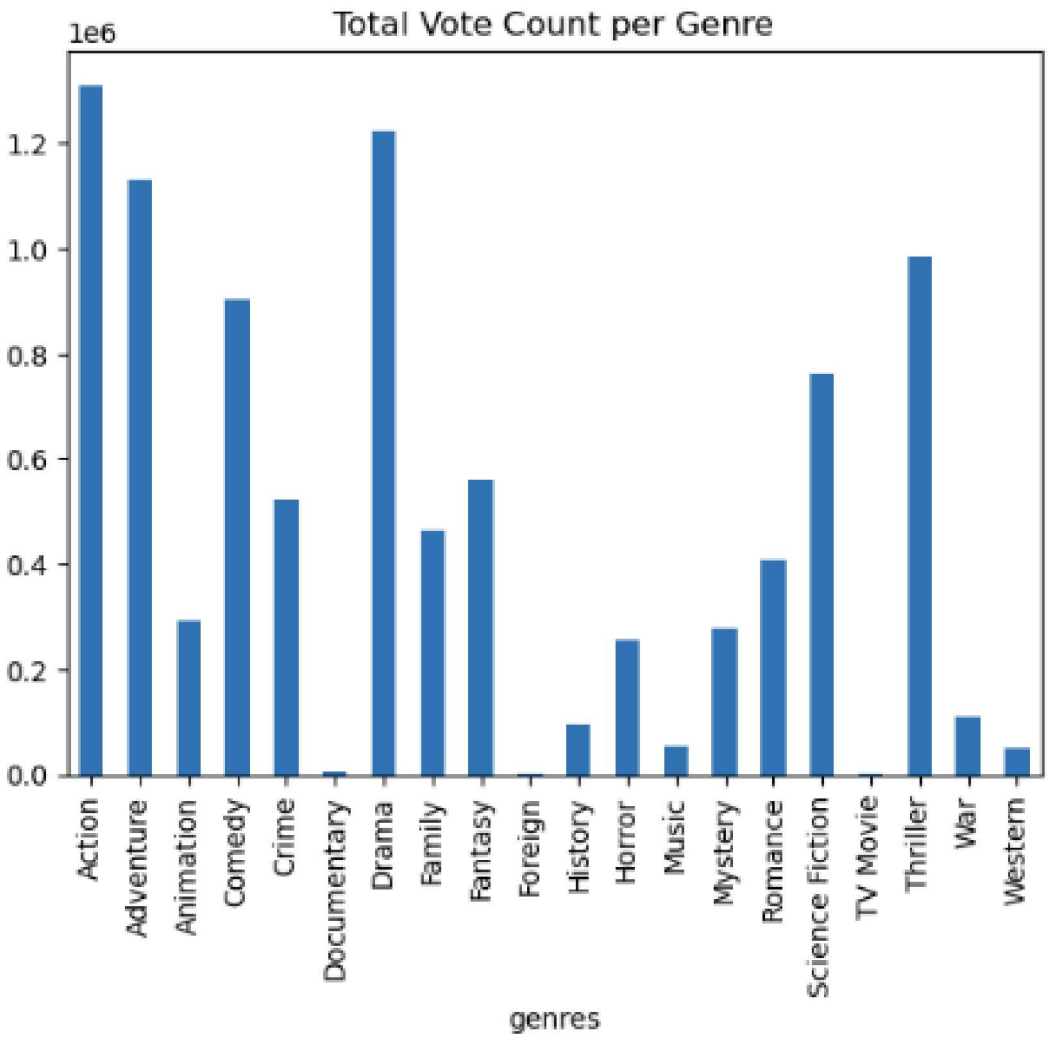


Figure 5.15: Total Vote Count per Genre

ingly, our analysis showed no significant correlation between these two variables. The calculated r-squared value was 0.000907, indicating a weak relationship between mean ratings and budget. This finding suggests that factors other than budget play a more substantial role in determining the average ratings of movies. It emphasizes the multifaceted nature of audience perception and highlights the need to consider various elements, such as plot, acting, and overall production quality, in understanding how movies are evaluated by viewers.

in fig 5.19 Lastly, we explored the relationship between budget and revenue specifically for English and French speaking movies. The scatter plot analysis revealed an r-squared value of 0.2 for English-speaking movies, indicating a moderate positive relationship between budget and revenue. This suggests that, on average, higher budget movies tend to generate more revenue for English-speaking movies. However, for French-speaking movies, the r-squared value was 0.03, indicating a weaker relationship between these two attributes. This discrepancy suggests that the factors influencing revenue generation may differ between English and French speaking movies, highlighting the importance of considering language and cultural contexts when examining the financial performance of movies.

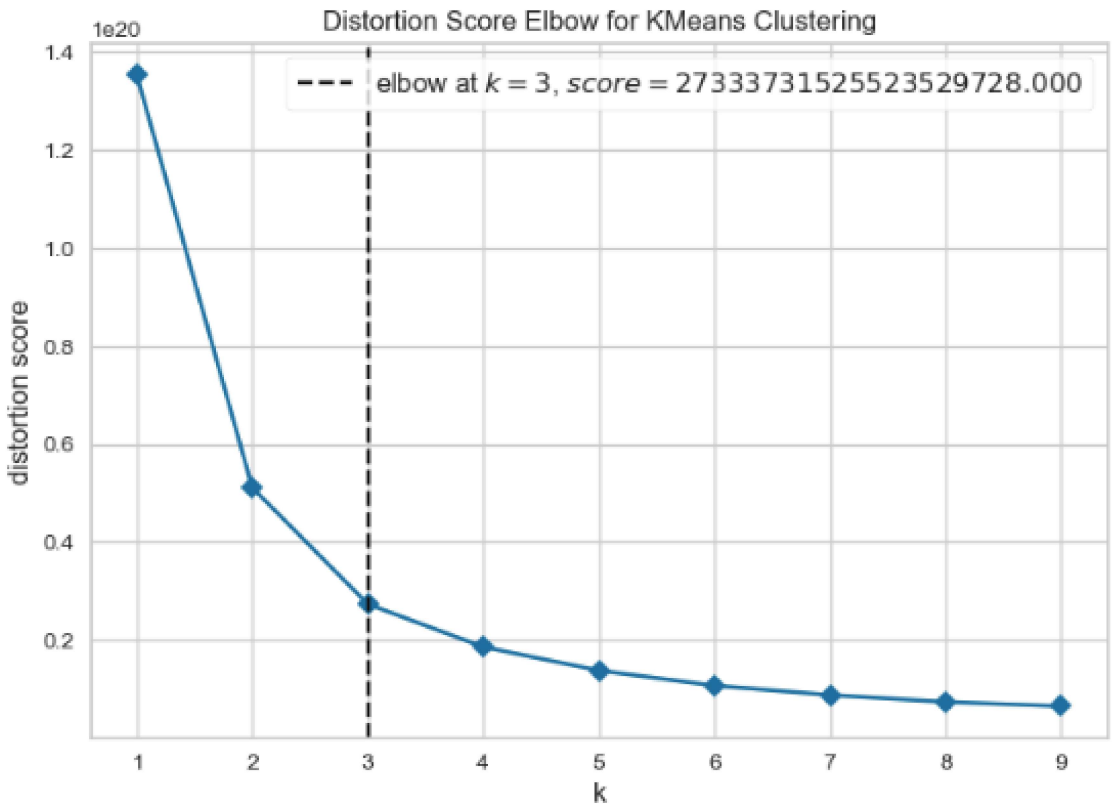


Figure 5.16: cluster grpah

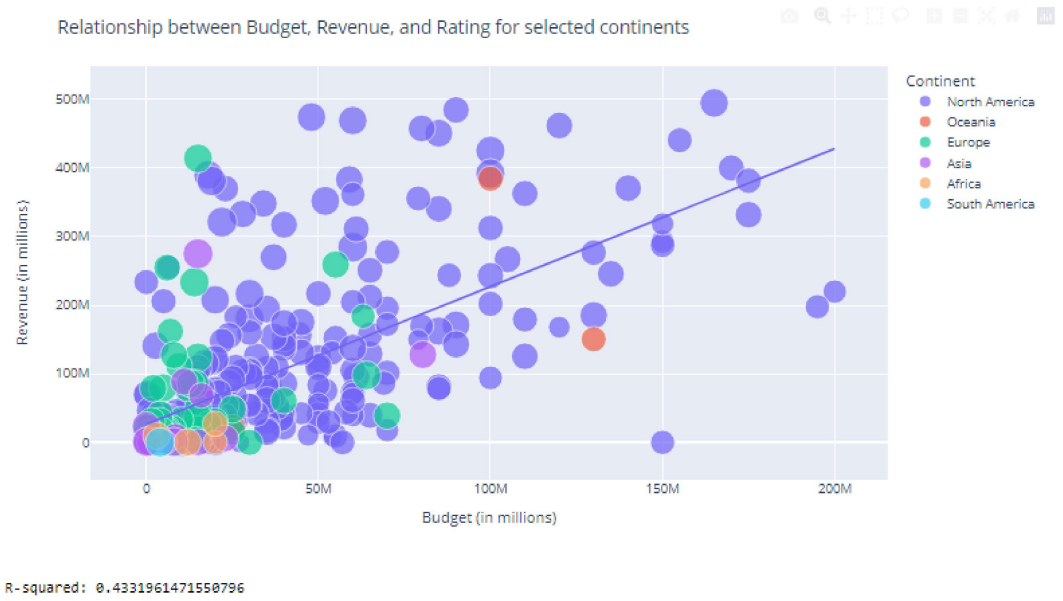


Figure 5.17: Relationship between Budget,Revenue and Rating for selected Continents

5.8.3 Discussion

Overall, our findings suggest that budget and vote count are important attributes contributing to highly-rated movies. However, the relationships between these attributes and others, such as genre and cultural contexts, can vary.

the popularity and reception of films The importance of financial investment in film production is demonstrated by the positive link between budget and income. Furthermore, the significant link

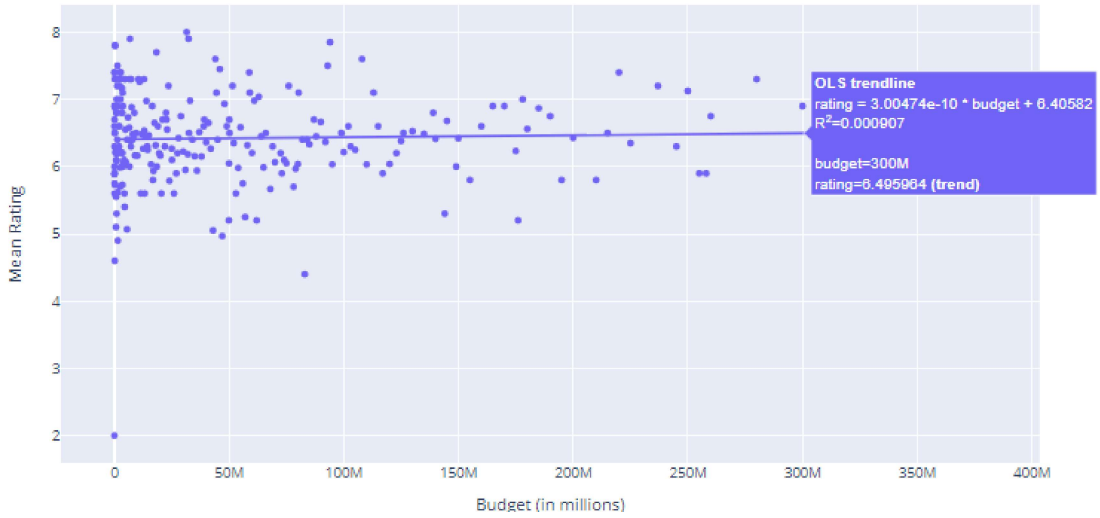


Figure 5.18: Relatiohsip between Mean Ratings and Budget

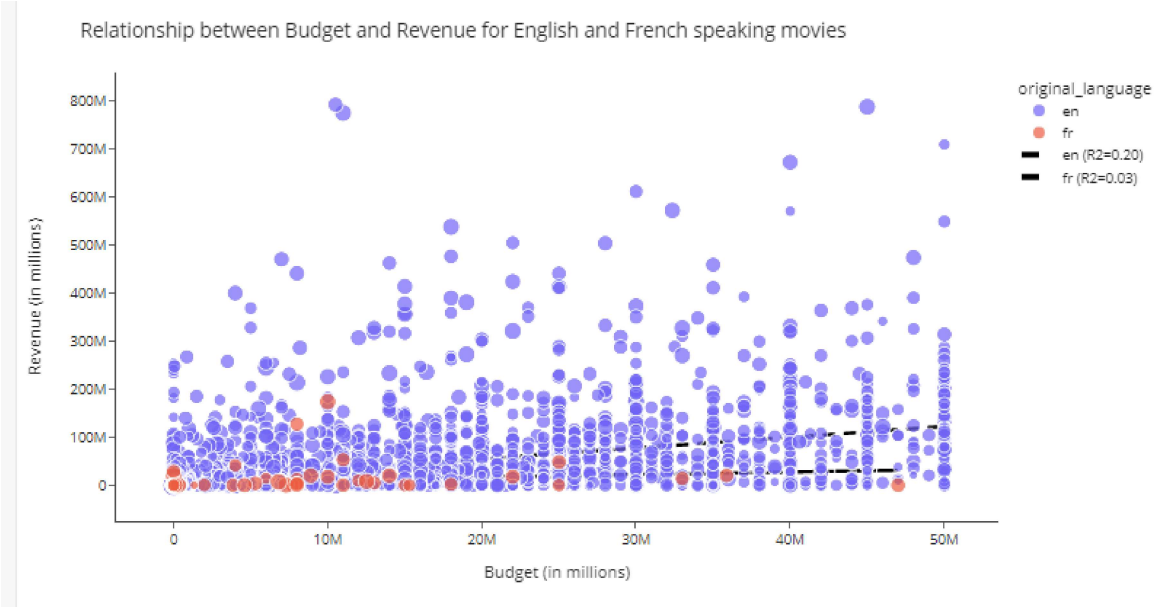


Figure 5.19: Relationship between Budget and revenue for English and French speaking Movies

between vote totals and revenue implies that audience participation and word-of-mouth are key factors in a film's success. However, it is crucial to take into account the impact of additional elements, such as genre, cultural preferences, and regional contexts, which may affect how well a film does financially and in terms of ratings. These difficulties highlight the requirement for a thorough comprehension of the multiple nature of film success, taking into account many factors and their interactions within distinct cultural and contextual contexts. Filmmakers, studios, and other industry professionals can modify their plans and make educated decisions by understanding these complexities.

Furthermore, one interesting finding from our analysis was the divergence between movie success in terms of revenue and ratings. Our visualization, where the bubble size represented the rating, revealed that highly-rated movies did not always correspond to high revenue. This suggests that the financial success of a movie does not solely rely on its critical acclaim or positive audience reception. Other factors such as marketing strategies, target audience preferences, and market competition may significantly influence revenue generation. This finding underscores the importance of considering both financial and qualitative aspects when evaluating the overall success and

impact of a movie. Filmmakers and industry stakeholders should recognize that a high rating does not guarantee commercial success, and a comprehensive understanding of the complex dynamics between revenue and ratings is crucial in making informed decisions within the film industry.

5.9 RQ4: Is there a trend of genre transitions in movies over time, and does this trend affect their box office performance?

5.9.1 Data & Method

In this section, we provide an overview of the data and methodology employed in our analysis of the trend of genre transitions in movies over time and its impact on box office performance.

For our research, we utilized the TMDB 5000 dataset, which is a comprehensive collection of movie data that includes various attributes such as genre, release date, rating, revenue, and budget. This dataset provided us with a rich and diverse set of information necessary to investigate the relationship between genre transitions and box office performance.

To begin our analysis, we first pre-processed the dataset. We converted the release date column into a datetime format, allowing us to extract the release year. This step was crucial in filtering the data to focus on movies released within our specified time frame of 1980 to 2020.

Next, we performed several exploratory data analysis techniques to uncover trends and insights. One of our key analyses involved examining the frequency of genre transitions over time. By grouping the movies based on their release year, we calculated the occurrence of different numbers of genres within each year. This analysis enabled us to visualize the evolving patterns of genre combinations and transitions throughout the selected time period. We created a line plot with different markers to illustrate these trends and identify any emerging patterns.

Additionally, we investigated the relationship between the number of genres in a movie and its box office performance. To accomplish this, we calculated the average rating for movies grouped by the number of genres they encompassed. This analysis provided us with insights into how the number of genres influences the audience's perception of a movie's quality. We created a scatter plot with a trend line to visually represent this relationship and further examine any correlations between the number of genres and box office performance.

Our methodology allowed us to gain a comprehensive understanding of the trends in genre transitions over time and their potential impact on box office performance. By leveraging the TMDB 5000 dataset and employing various data analysis techniques, we were able to extract valuable insights and shed light on the dynamics of genre transitions in the film industry.

5.9.2 Results

In our analysis of genre transitions over time, we found an intriguing pattern. The line plot depicting the frequency of genre transitions revealed that movies with a single genre were the most prevalent over the years, while movies with multiple genres experienced a gradual increase in frequency. This suggests a growing trend of incorporating multiple genres into movies, with

combinations of 2 to 5 genres being the most common.

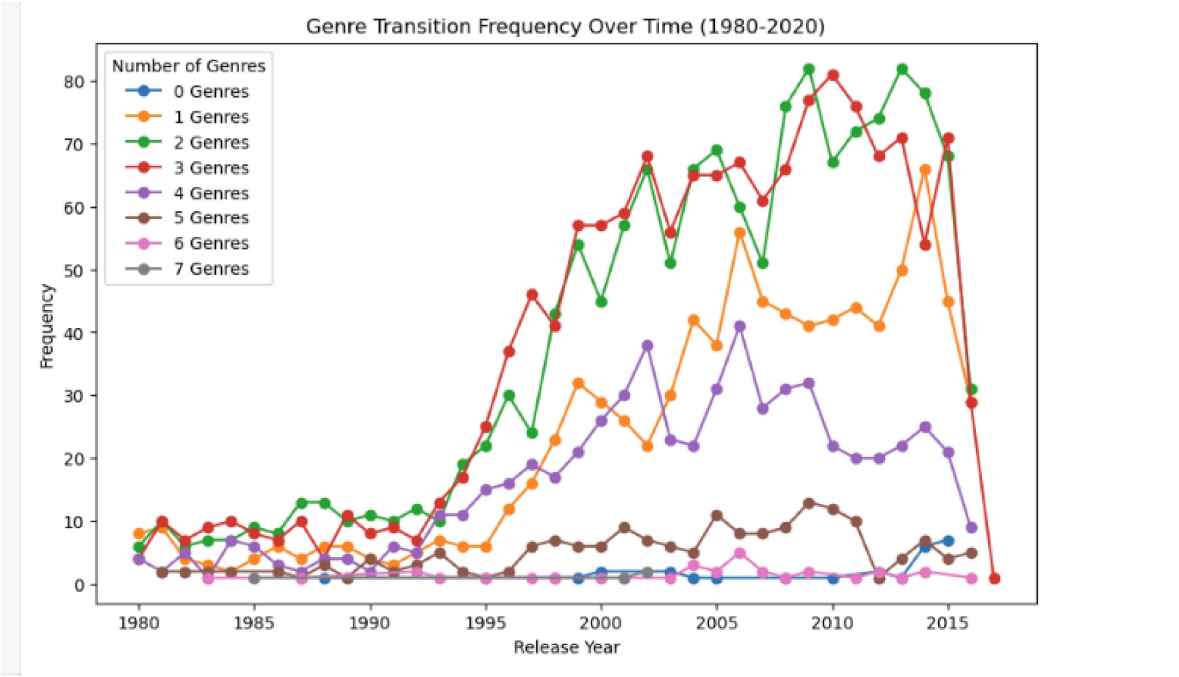


Figure 5.20: genre combinations over the years

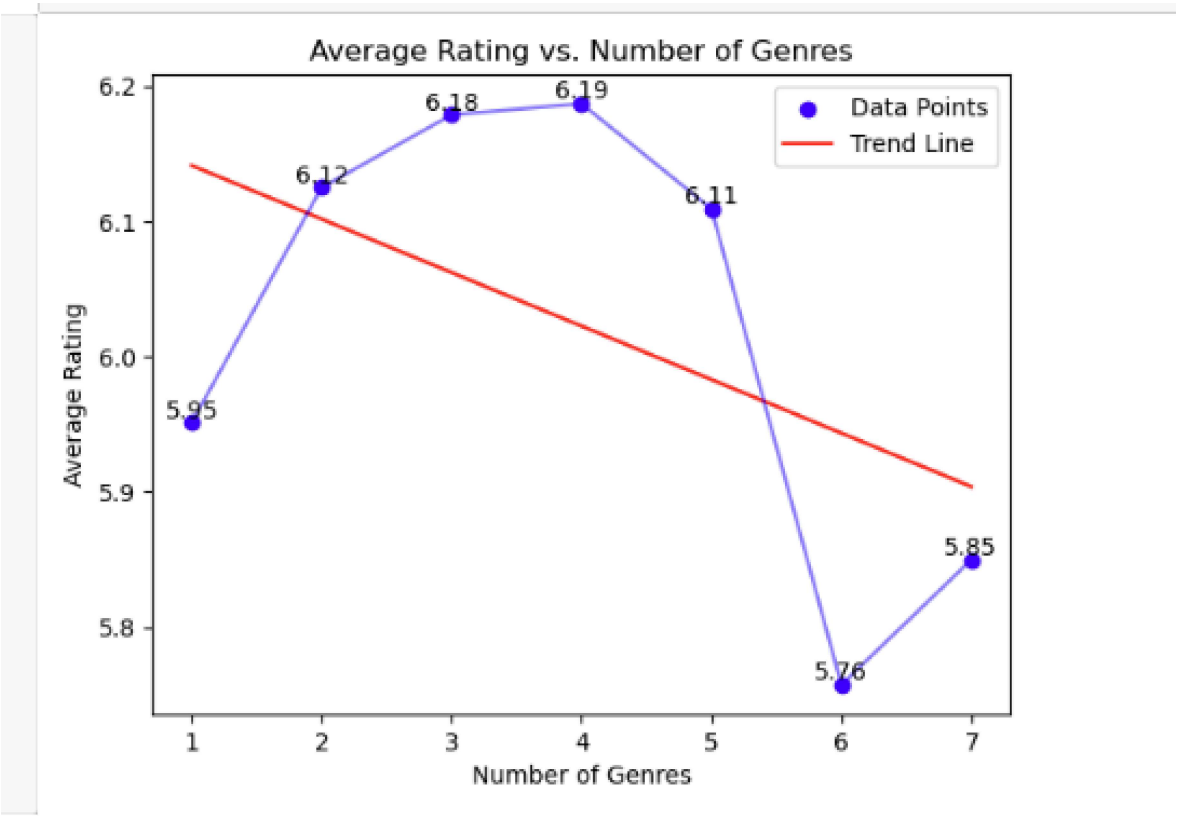
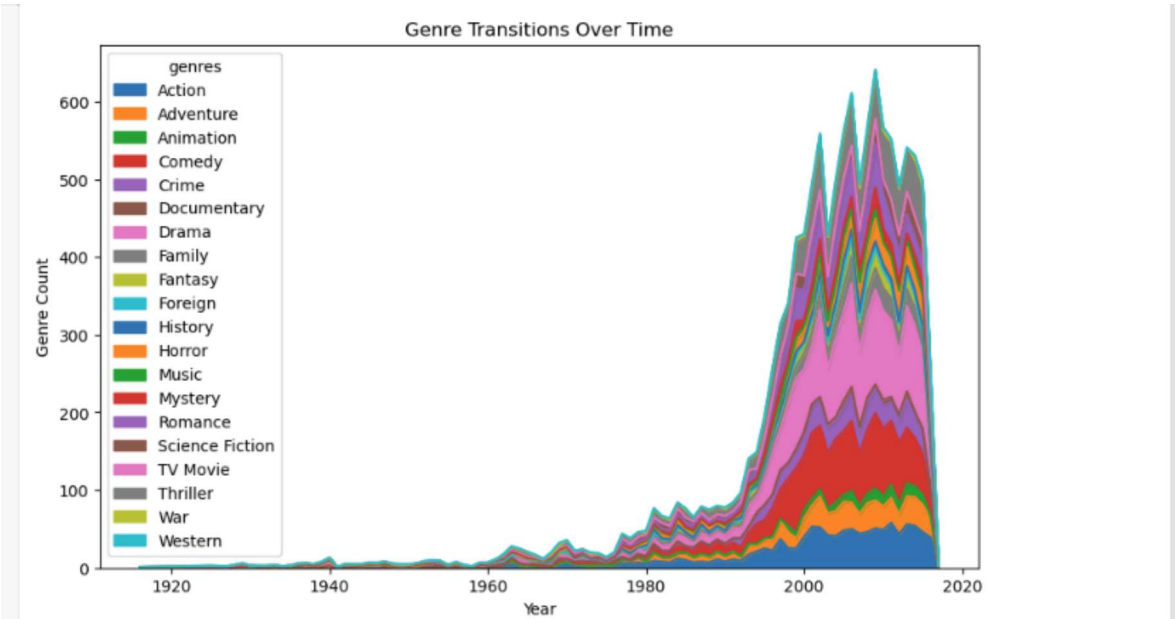


Figure 5.21: Average Ratings vs Number of generes in a movie

Furthermore, we investigated the relationship between the number of genres and movie ratings. The scatter plot demonstrated a slight positive correlation between the number of genres and average rating. Movies with 2 to 5 genres tended to have higher average ratings compared to movies with either a single genre or a large number of genres. This finding suggests that finding




```
C:\Users\ouafi\AppData\Local\Temp\ipykernel_5064\1225993175.py:33: FutureWarning: The default dtype for empty Series will be 'o
bject' instead of 'float64' in a future version. Specify a dtype explicitly to silence this warning.
  genre_cooccurrence = movies['genres'].apply(lambda x: pd.Series(list(combinations(x, 2)))).stack().value_counts()

(Drama, Romance)       522
(Action, Thriller)     497
(Drama, Thriller)      463
(Comedy, Drama)        428
(Comedy, Romance)      425
(Action, Adventure)    314
(Crime, Thriller)      252
(Comedy, Family)       251
(Action, Crime)        248
(Horror, Thriller)     245
dtype: int64
p-value: 3.639893203546168e-143
```

Figure 5.22: Genre transition over time with the highest 2 genre combinations in the dataset

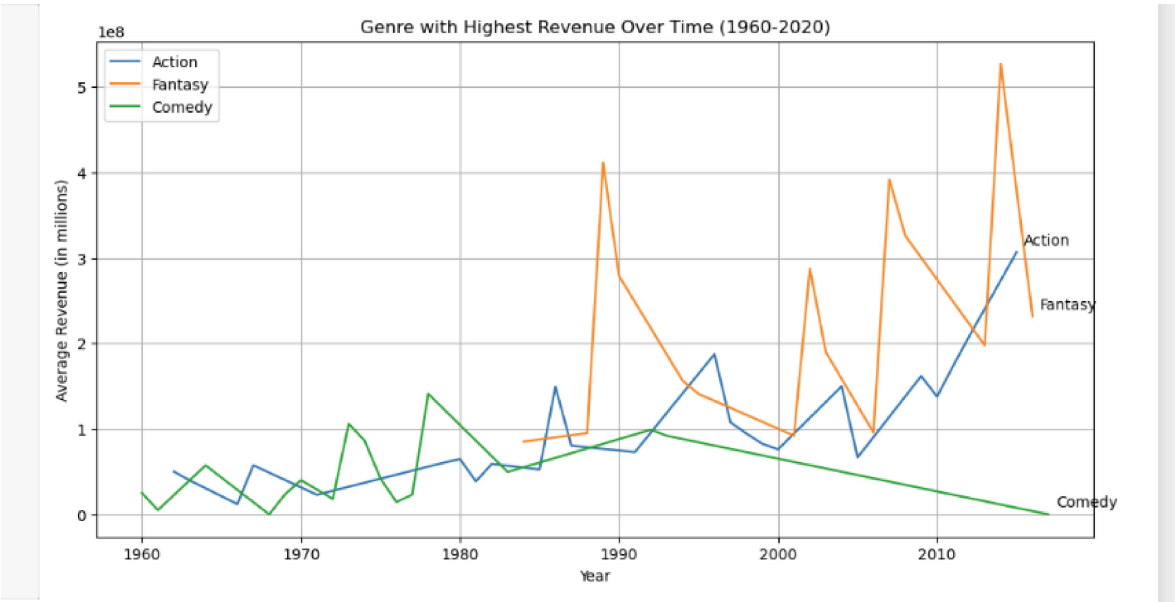


Figure 5.23: genre with highest revenue over time

the right combination of 2 to 5 genres can potentially lead to higher audience satisfaction and overall movie quality.

However, it is essential to interpret this relationship with caution. While movies with multiple genres

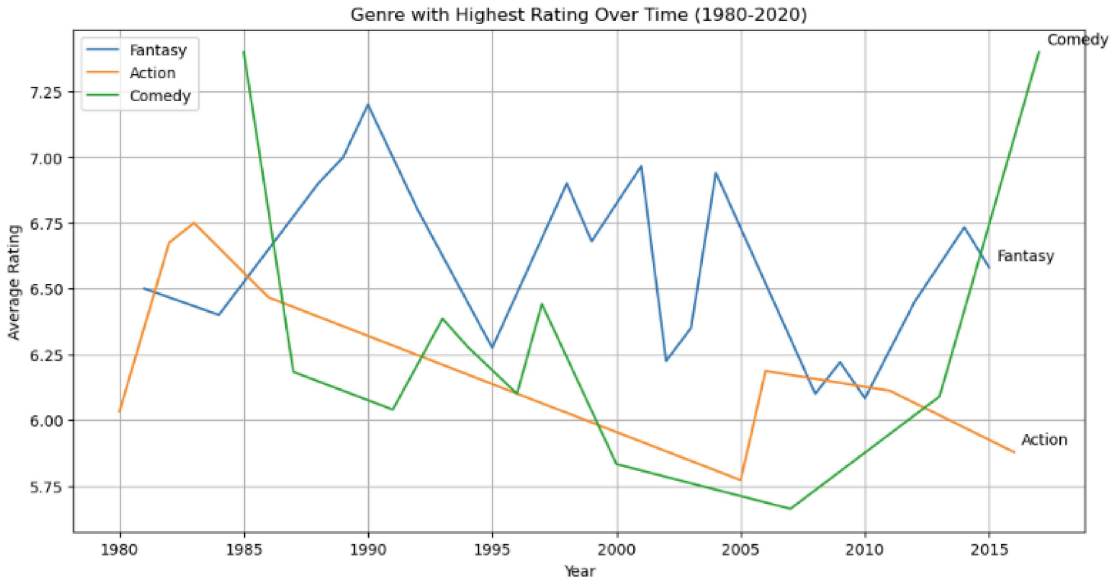


Figure 5.24: genre with highest Ratings over time

generally showed higher average ratings, it does not guarantee success or box office performance. Factors such as the execution of the genres, the quality of the script, the performances of the actors, and the overall direction still play crucial roles in determining the movie's reception and commercial success.

We also concluded from this research that as we can see from fig 5.23 and fig 5.24 that the ratings and revenue don't correlate in terms of there success with the specific genre presented as we can see in the figures named above that comedy tends to do worse over time in terms of its revenue but the ratings graph says that it has a positive skew over time. This is one of the interesting findings we concluded that movies in terms of genre transition over the time doesn't solely base on its amount of genre combination and there are many other factors that come into play

Overall, our analysis indicates that exploring genre combinations within the range of 2 to 5 genres could be a promising approach for filmmakers seeking to enhance the quality and audience reception of their movies. However, it is essential to maintain a balance between the genres and ensure coherence and seamless integration of the different elements. The right combination of genres, coupled with strong storytelling and execution, can contribute to higher average ratings and potentially better box office performance

5.9.3 Discussion

The findings from our study shed light on the dynamics of genres across genres and their impact on box office performance. The frequency of multi-genre production means an increasing trend toward mixing genres and exploring different methods of storytelling can attribute this trend to audience desire for more specific cinematic experiences it was a fun combination of elements from movies.

Furthermore, our study showed that the box office success of a film is not necessarily determined by the number of movies in it. Although single-genre films were more prevalent, our scatter showed that exposure was only affected by the average number of films. This means that the success of a film at the box office cannot be attributed solely to its format. Other factors such as marketing strategy, star power and audience size also play an important role in determining the commercial

success of a film

These findings highlight the complexity of the film industry and the need for a multi-pronged approach when assessing the success of a film. Filmmakers and industry practitioners need to consider a variety of factors including changes in films, audience preferences and market dynamics in making decisions about style and marketing Through the interaction of changes in films and by understanding box office performance networks, stakeholders can make adjustments to shift audience preferences and enhance the overall movie-going experience

Chapter 6: **Conclusions**

In this section you will sum up your report, draw some conclusions about your work so far, and make some general observations about the work to come. You may also use this opportunity to express points of view, or make factual claims, that are more pertinent here than in other sections of the report. If your project raises some ethical concerns, for example about how data or users are treated, then address them here in a thoughtful manner.

Conclusion Our work focused on the analysis of films and TV shows with the aim of addressing several research questions. We sought to investigate whether the box office success of a film can be predicted, the effect of episode/season number and length on consumption, opinion and success of TV shows, key contributing characteristics for film more popular varieties emerge, whether or not they vary across cultural contexts, as well as a tendency toward changes in genres within genres over time and their impact on success. Through our research, we gained valuable insight into these aspects of the film and television industry.

First, our analysis of box office success predictions revealed that although accurately predicting a film's financial performance is challenging, factors such as budget, cast star power, genre, marketing effort plays an important role but other variables such as critical acclaim and word of mouth have a greater impact, making it a complex and multifaceted process to accurately describe

Second, our examination of the relationship between episode/season count, length, and TV ratings, opinion, and success showed that these factors do influence Typically, episode shorter lengths, programs with fewer seasons get higher ratings and more success. This finding suggests that viewers prefer a concise and well-structured story that doesn't overstay its welcome, and shows with multiple episodes and long runs may struggle to maintain audience engagement.

Furthermore, our analysis of the key characteristics that make films so popular across cultures emphasized universality and context although aspects such as compelling storytelling, inside play were identified are difficult and technical efficiency are consistent predictors of success across cultures though, there were also differences influenced by cultural preferences, local customs and social norms That factor this finding highlights the importance of considering cultural nuances in making and promoting films for different audiences.

Finally, our analysis of changes in films over time revealed a clear trend. We noted that music has evolved and changed dramatically over the years, often influenced by social, cultural, and technological changes. These developments have affected the success of films, with some films gaining or losing popularity depending on audience preferences and trends. Producers and industry professionals need to adapt to these changes and understand the changes taking place to increase their chances of success.

Overall, our work has addressed important research questions and shed light on different aspects of film and TV show research. While predicting box office success is a difficult task, our findings provide insight into the factors that influence it. We also determined the effect of episode/season count and length on the ratings success of a TV show, highlighted the importance of cultural context in highly rated movies, and examined the characteristics of genre change occurs and its impact on success. These resources can be a valuable guide for industry professionals, filmmakers and producers, helping them make informed decisions and create content that resonates with audiences in a rapidly evolving media landscape

Acknowledgements

We would like to thank our lecturer Barry for his invaluable support and guidance throughout this module and over the past few weeks. Barry's extensive experience in data analysis and dedication to teaching played a key role in our learning journey. His insights and feedback have greatly enhanced our understanding and application of data analysis techniques.

Barry's commitment to our growth and development was evident as he always encouraged us to strive for excellence, find new ways to improve His constructive criticism and his suggestions challenged us to we don't think about it enough and hone our research skills.

We would like to thank our demonstrator Rian for his unwavering help and support. Rian's presence and expertise was invaluable to our team. He was always available to answer our questions, clarify doubts and respond in a timely manner. His guidance and development was very important to keep us going and ensure the success of our teamwork.

Both Barry and Rian played key roles in our project, and their contributions have been integral to our success. We are extremely grateful for their guidance, encouragement and dedication to our development as budding data analysts.

Bibliography

1. Alper, M. & Karakose, M. A comparative analysis of box office prediction methods for movies. *Information Processing & Management* **57,** 102144 (2020).
2. Kim, H. J., Chun, H.-K. & Kim, H. K. The effect of genre transitions on movie success. *Journal of Business Research* **86,** 35–47 (2018).
3. De Vreese, C. H. & Neijens, P. Key predictors of film success: The impact of national culture on the reception of films. *Poetics* **57,** 12–20 (2016).

Appendices

You can use your appendix to include information or data that you feel is not needed for your main report but that is still relevant to your project. This might include secondary results or additional results that are similar to those presented in the main report.

Initially, we considered using the Twitter API to gather statistics about the number of followers of the brands and businesses and compare their box office performance but this approach proved to be inefficient and inefficient. First, a significant number of developers and employees do not have Twitter accounts, making the data incomplete. Additionally, matching Twitter handles to individuals' real names would be a time-consuming process due to the difference between their Twitter handles and their real names. As a result, this method is considered inappropriate because it does not provide complete and accurate information for analysis Other methods of assessing the relationship were sought

Appendix A – Individual Contributions

In the project, both Mohamed and Abdel collaborated to analyze data on the popularity of movies, TV shows, characters and candidates, using the YouTube API. Their roles and contributions were divided in order to better address different aspects of the research.

Abdel focused on the analysis of film content, in particular developing research questions (RQ) 3 and 4. He analyzed the context of films, examining factors. Abdel was responsible for research and designed the experiments to address RQ3 and RQ4. In addition, he played a key role in developing the introduction, defining objectives and motivations, and setting the context for the final report

Meanwhile, Mohammed worked on movie and TV show accounts, as well as the YouTube API. Conducted research and analysis for RQ1 and RQ2, analyzing films and TV shows. Mohammed was charged with finding relevant work, examining existing literature and identifying relevant research. He played a key role in the data collection process and ensured the availability and quality of the data used. Mohamed collaborated with Abdel on the conclusions, summarizing the findings and implications of the study.

The collaboration of Mohamed and Abdel throughout the project led to detailed analysis of data, enabling them to successfully address various research questions In conclusion, their combined efforts led to integration and insight into reporting in the last of these.

Appendix B – Datasets & Notebooks

Please list the filenames of the main datasets used in your project along with a brief (one-sentence) description of each. Also list the notebooks with a similarly brief summary of their purpose. You can indicate too who was responsible for which dataset/notebook by indicating the initials of the

student beside the dataset or notebook file.

Datasets Used

1. Finalmovie$_{d}ataset_{f}in1.csv(BS)\check{}themaindatasetforRQ1whichincludemoviedetails'budget''popularit$
   $themaindatasetforRQ2whichincludes'Tvshow_{n}ame''episode_{T}itle''seasonNumber''episodeNumber''$

2. Final$_{t}vshow_{d}ataset_{m}f.csv-theseondDatasetforRQ2whichincludes'name''season''episode''year''rating''V$
   $ForRQ1$

3. crew$_{c}ast_{A}ctors_{p}opularity-forRQ1whichincludes'Actor''Rating''Votes''Num_{M}ovies'$

   $TMDB5000DATASET-forRQ3andRQ4whichincludes'budget',' genres',' homepage',' id',' keywords',' or$

Notebooks Used

1. Data$_{c}ollection_{t}ask_{F}inal1.ipynb(BS)\check{}themaindatacleaningnotebookresponsiblefordownloadingtheda$
   $forprereleasemarketingfactor.$

2. RQ1$_{G}enre.ipynb(BS)-ThisisforgenrefactorResearchQuestion1_{C}astandcrew.ipynb(BS)-$
   $ThisisforcastandcrewpopularityRQ1_{B}udget_{F}actor.ipynb(BS)-Thisisforbudgetfactormerging_{d}ataframes_{R}Q2.ipynb(BS)-$
   $ThisisformergingmultipleTvshowsintoone.$

3. RQ2$_{T}vshowRating.ipynb(BS)-ThisforGenreRatingtvshowsRQ2_{T}vshows_{f}inal.ipynb(BS)-$
   $thisisforveiwsandvotesoftvshows$

4. RQ3 ResarchQuestion4$_{u}pdated.ipynb-thisiswhereallthe analysisisdoneRQ3Grpahs.ipynb-$
   $thisiswhereallthemaingraphsaredisplayedforclearernotebook$

5. RQ4 lastResearchQuestion.ipynb - last research question answers the research question regarding genre transition

Appendix N – Supplementary Results

...