University of Groningen

Faculty of Economics and Business

university of groningen / faculty of economics and business

Master thesis in Marketing Analytics and Data Science

The Drivers of Box Office Success: Budget, Star Actors, Reviews and the Moderating Effect of Seasonality across movie genres

---

Angeliki Loulo

S5670101

a.loulo@student.rug.nl

First Supervisor:

Dr. Evert de Haan

Second Supervisor:

Dr. A. (Alec) Minnema

January 17, 2025

Abstract

This study investigates the factors influencing box office performance with a focus on the interaction of budget, star actors, consumer reviews, and professional reviews with seasonality (holidays vs. non-holiday periods) across popular movie genres such as drama, adventure, action, comedy, thriller, horror, and science fiction and sequel movies. By analyzing data from 745 movies released between 2000 and 2023, the research employs Moderated Multiple Regression analysis to explore how these variables impact global box office revenues. Results show that budget has a consistently significant positive impact on revenue, the impact of star actors and professional reviews varies depending on the timing of the release. Surprisingly, professional reviews often have a negative effect during non-holiday periods, and star actors are less effective during crowded holiday seasons, but their presence in sequel movies might impact revenue positively. These findings underline the importance of tailoring release strategies to seasonal trends and genre-specific preferences, providing valuable insights for filmmakers, marketers, and distributors aiming to maximize box office performance.

Keywords*: Box Office Performance, Budget, Star Actors, Consumer Reviews, Professional Reviews, Seasonality, Movie Genres, Moderated Multiple Regression*

Acknowledgments

This thesis journey has been both challenging and rewarding. Along the way, I faced many obstacles that at times felt overwhelming, but I gave my utmost effort to see it through to the end. Each obstacle taught me the importance of perseverance, resilience, and believing in myself.

I am profoundly grateful to my supervisor, Dr. Evert de Haan, for his advice and motivation to keep moving forward even during the most difficulties. His guidance kept me grounded and encouraged me to stay on track and complete this work.

I am also truly grateful to my family and friends, who always believed in me and reminded me that I could do it. Their support and their constant encouragement gave me the strength to finalize this thesis.

Contents

1.Introduction 7
1.1 Relevance of the Topic 7
1.2 Research Objective 8
1.3 Structure Overview 9
2. Literature Review 10
2.1 The Complexity of Movie Success 10
2.2 Defining movie success: Commercial vs Critical Outcomes 11
2.3 Bridging gaps in understanding box office success across genres 13
2.4 WOM & Consumer Reviews 14
2.5 Professional reviews 17
2.6 Budget 18
2.7 Star Actors 20
2.8 Conceptual Framework 21
3. Data Preparation & Description 23
3.1 Data Preparation 23
3.2 Data Collection 23
3.3 Data Integration 24
3.3.1 Control Variables 25
3.3.2 Moderator Variable 26
3.3.3 Star actors 28
3.4 Missing Values 29
3.5 Outliers 31
3.6 Data Description 34
4. Methodology 37

4.1 Explanatory Model ... 37
4.1.1 Model Specification ... 37
4.2 Model Estimations & Assumptions ... 42
4.2.1 Heteroscedasticity ... 42
4.2.2 Multicollinearity ... 43
4.2.3 Non- Normality ... 44
4.2.4 Linearity ... 44
4.2.5 Assessment of Influential Data Points ... 45
4.3 Usage of A.I. ... 46
5. Results ... 47
5.1 Holiday Model ... 47
5.2 Non – Holiday Model ... 48
5.3 Baseline Model ... 49
5.4 Combined Model ... 51
5.5 Combined Model OLS & Ridge ... 52
6. Discussion ... 55
6.1 Hypothesis Testing & Findings ... **55**
6.2 Managerial Implications ... **57**
6.3 Limitations & Future Research ... **58**
7. Conclusion ... 60
Bibliography ... 61
Appendix ... 71
A) Data Preparation ... 71
A1) Movie Titles ... 71
A2) Sequel Movies ... 97
A3) Star Actors ... 99
A4) Correlation Analysis (pre & post imputation) ... 100

B) Data Description 101
C) Assumption Checks 108
Heteroscedasticity 108
Multicollinearity 109
Non- Normality 111
Linearity 112
Influential outlier detection 113
D) Results 118
Pre-Assumption Stage 118
Mean-Centered Models 122
Hypothesis:(isolated interactions effects) 126
Robust Standard Error Models 130
Ridge Regression Models 134

1.Introduction

1.1 Relevance of the Topic

What do movies and earthquakes have in common? In both cases, only about 20% can truly be considered the major "movers" and “shakers” (Scott, 2019), while the remaining 80% are merely minor tremors (Bhattacharya, K. et al., 2009). In the entertainment industry, this shows how movie success is defined, with most revenue coming from a few blockbuster movies, according to (De Vany, 2004), follows a Pareto distribution where 20% of the film generates 80% of revenue.

This study finds both alignment with and divergence from prior research. Consistent with (Liu, 2006), consumer reviews have more significant influence during less competitive non-holiday periods, though their overall effect was weaker than expected. Similarly, professional reviews were found to have a limited influence, consistent with (Basuroy S. et al., 2003), but with a slight negative impact during non-holiday periods, differing from earlier findings. Budget emerged as a key factor, as highlighted by (Scott, 2019), but its reduced effect during holidays suggests that competition from other releases may diminish its impact, making non-holiday periods more favorable for high-budget films. The significant negative interaction between star actors and holidays challenges the assumption of guaranteed success for star-driven films, aligning with (Gunter B, 2018) and (Scott, 2019), who noted that heavy competition during holiday periods reduces their effectiveness. These findings emphasize the role of seasonal effects and market dynamics in shaping box office performance.

However, these studies often assume that success can be forecasted before a movie’s release, ignoring the dynamic and unpredictable nature of the industry. (De Vany, 2004) argued that revenues follow a Pareto distribution, where few movies dominate the box office, making accurate pre-release predictions nearly impossible.

This unpredictability arises from factors such as word-of-mouth buzz, changing audience preferences, and unexpected competition from other movie releases, which can dramatically influence a movie's performance. For instance, a film with little pre-release hype might gain sudden popularity through positive audience reviews after release, while a highly anticipated movie could underperform if it faces competition from another blockbuster released at the same time. These elements emphasize the importance of studying how various factors interact after - release, rather than relying solely on pre-release forecasts.

Given the transformative impact of globalization and technological advancements on the film industry, it is essential to estimate box office performance using fresh and comprehensive data. My study addresses this by analyzing data from 2000 to 2023, capturing the evolving trends and dynamics that shape revenue generation in a globalized market.

My study builds on this prior research but focuses on a critical gap: **seasonality**. Papers like (Pangarker et al., 2013) and (Mirrlees, T. et al., 2013) and (Scott, 2019) noted that movies released during holidays and summer tend to generate higher revenues. However, they did not explore how seasonal timing (holiday and non – holiday) interacts with factors like budget, reviews, or star actors. My study addresses this gap by examining how these variables influence box office performance across different genres during holiday and non-holiday periods.

1.2 Research Objective

The primary objective of this study is to understand how key factors such as **budget, star actors, consumer reviews,** and **professional reviews** interact with seasonality (holiday vs. non-holiday periods) to influence box office performance.

By focusing on popular genres such as drama, horror, adventure, comedy, science fiction, thriller, action, and sequel movies as, the research question guiding this study is: ***“How do key factors such as budget, star actors, consumer reviews, and professional reviews interact with seasonality (holiday vs non – holiday periods) to influence box office performance across specific genres?"***

1.3 Structure Overview

The thesis is structured to offer a logical flow from theory to findings. **Chapter 1)** introduces the topic, explains why it is relevant, and outlines the main objectives and research question of the study. **Chapter 2)** reviews the existing literature, focusing on predictors of box office success and the gaps this study addresses. **Chapter 3)** explains the data sources, preparation, and description, while **Chapter 4)** discusses the methodology, including Moderated Multiple Regression (MMR), to analyze interaction effects. **Chapter 5**) presents the results, highlighting how budget, reviews, and star actors influence performance during holiday and non-holiday periods. **Chapter 6** tests the hypotheses and presents the detailed findings, managerial implications, and the study's limitations while offering directions for future research. Finally, **Chapter 7** concludes the thesis, providing a comprehensive overview of the study’s contributions to the field.

2. Literature Review

2.1 The Complexity of Movie Success

Movies captivate audiences across all walks of life, from timeless animated classics to action-packed franchises. Offering an escape, they make the film industry a vibrant part of our world economy (Michalopoulos et al., 2024).

The Hollywood movies industry produces approximately 600 - 800 movies each year, where 100 – 150 of them come from well–known studios like Disney, Universal, and Warner Bros (Handke, 2021). This means that a significant amount of Hollywood's annual profits comes from these major studios which have the majority high – budget movies that get the most attention from the biggest audiences because of their successful marketing reach and dynamic financial position. On the other hand, the rest of the annual releases are produced by smaller studios that don't have access to high-budget productions (Eliashberg, J. et al., 2006).

The goal of the production of movies is to be aligned with the audience's preferences. People naturally have different tastes when it comes to movies. Some folks go for comedy, others lean towards romance or science fiction, while some might even head to the theater simply because a favorite actor or actress is starring. (Michalopoulos et al., 2024). This creates a universal challenge for the accurate estimation of the successfulness of movies: they are expected to cater to a vast and diverse audience, each with unique preferences. (Ha, 2011). The questions raised are: how successful will a movie be soon, and how to estimate the success of movies when it is a movie a success or a failure? (Ahmed et al., 2019).

2.2 Defining movie success: Commercial vs Critical Outcomes

Definition

The entertainment industry, which has a huge economic and cultural footprint in the world, constantly seeks the key determinant factors that make a movie successful. The term of thriving in the movie industry usually falls into two categories the commercial success, like box office earnings, profits, and extra income from streaming and merchandise (Gao et al., 2019). On the other hand, critical success is about positive professional and consumer reviews in terms of valence and volume rating (quantified through user-rating values, while the total amount of reviews determines volume and prestigious awards. (Duan et al., 2008).

A film can generate impressive box office revenue yet receive mixed reviews or earn critical acclaim while struggling financially. When a movie achieves a strong box office performance and positive critical reception, it reaches a rare level of success. This balance reflects a unique blend of popular appeal and artistic quality, marking the film as impactful on multiple fronts (Gao et al., 2019).

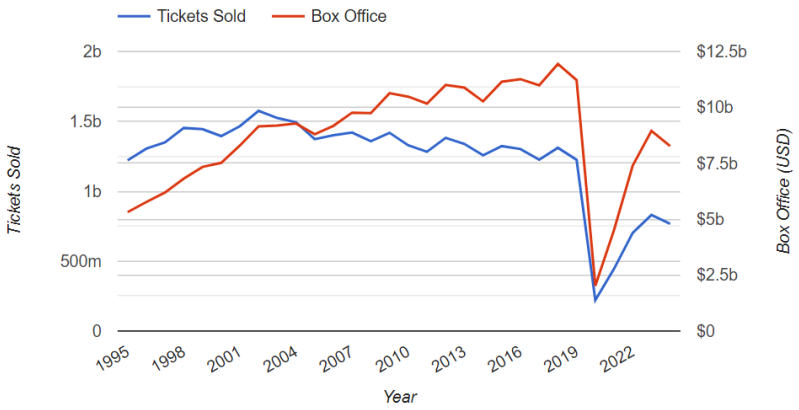


Figure 1) Box office Revenue (1995-2023) The Numbers

According to **Figure 1**, the popular website *The Numbers* focused on providing a wide range of metrics related to the film industry, including box office revenue, ticket sales, and trends over time. The above indicates that in 2018, it recorded the highest box office revenue with the highest sales, with 873 movies released compared to 724 movies released in 2017*(The Numbers).*

Some examples of what we mean by the terms successful and unsuccessful movies are analyzed in this paragraph, which is considered why it is so important to analyze the main factors that drive the increase of the box office revenue and probably how we could predict the future box office performance focusing on the most significant ones. A great example of a movie critics loved, but the impact on box office revenue was negative significant at the box office is *Under the Skin* (2013), directed by Jonathan Glazer and starring Scarlett Johansson. Moviegoers preferred it for its unique style and Scarlett's amazing performance, and it even has a high rating on Rotten Tomatoes. However, despite all the love from critics, it only made about $7.2 million worldwide, which is far less than its $13.3 million budget, making it a financial flop. The movie's box office performance was a surprising flop, with its financial result almost the opposite of its critical audience success. (Internet Movie Database, sd)

Despite receiving a low average volume rating of 42 from nearly 15,000 viewers, placing it among the lowest-rated movies, the 2015 horror movie *The Gallows* turned out to be a surprise box office revenue success. Produced on a tiny budget of just $100,000, it earn $22 million, achieving a remarkable return on investment. Its financial success was impressive, especially given the mixed reception from audiences.

Regarding the above contrasting examples, it is concluded that positive audience ratings and good professional reviews do not always guarantee financial success, nor do low ratings automatically mean a movie won't perform well at the box office. This clarifies that a deeper look at other factors is essential to understanding what truly drives box office success. Factors like budget, genre, seasonal effect, star actors,

sequel status, and marketing efforts can sometimes play a more crucial role in whether a movie becomes a hit or a flop. (Scott, 2019)

For the entertainment industry, understanding these driving factors is incredibly valuable. By analyzing them, producers and distributors can make more intelligent choices, increase the chances of a movie's success, and reduce financial risks. It is not just about learning from past outcomes; it is also about predicting future performance (Gunter, 2018). This kind of insight helps in planning more effective budgets, fine-tuning marketing strategies, and choosing the best release times for distributors but also for supply chain managers where they cooperate with the film industry the merchandise products like DVDs, clothes with movie titles, and specific brands from movies that are hits. Altogether, this approach supports a more profitable, sustainable industry where investments are more likely to pay off (Squire, 2016)

2.3 Bridging gaps in understanding box office success across genres

After reviewing existing literature about which are the most valuable determinant factors that make a movie successful, I concluded that there is plenty of research on this topic but most of it focuses on general key factors like production budgets, release dates, star actors, professional and audience reviews, and many others. Still there are gaps and patterns analyzing movies success. However, these studies often do not explore how these factors work differently depending on the seasonal effect on the movies.

According (Scott, 2019) focused on two models. The one was about the ex-ante model, which included the prerelease factors such as the genres, star actors, seasonal timing (summer/holiday), reviews and sequel movies, and the ex-post model, which included the word of mouth and award nominations. His study showed how the pre–release

factors and after-release factors contributed to the success of movies. The genre is considered a secondary factor and didn't emphasize how these influences might vary between genres like comedy, horror, or drama or a deeper focus on the seasonal effect.

However, (Pangarker et al., 2013) considered the genre as a significant factor in their study, but they concluded that when the genre variable is combined with other variables, its influence decreases in determining box office success. This insight helped me estimate how genres interact with other multiple factors, specifically in terms of the dynamics of those specific genres on box office performance. Moreover, both studies argued that sequels have better performance because they already have a stable audience.

While (Gunter B., 2018) emphasizes the profitability of specific genres like action and horror, it does not consider how these genres interact with other factors like professional and audience reviews, budgets, or seasonality. Additionally, (Lee et al., 2016) study relies more on improving the predictive accuracy through machine learning techniques, including factors such as transmedia storytelling. However, it does not focus deeply on how particular dynamics of seasonality impact specific factors like genres of movies or sequels that might influence box office performance revenues.

2.4 WOM & Consumer Reviews

Word-of-mouth (WOM) is public opinion, and it is one of the most significant factors shaping consumers' perspective in evaluating products. Although, except for the traditional WOM, where the individuals share their opinions and experiences, also the digital version known as electronic word-of-mouth (eWOM) significantly influences consumer choices in the movie industry. This effect is pronounced in the movie industry, as moviegoers often rely on their selections of information released before the movie's first premiere. This refers to often for experience products (movies) with short lifespans. (Kim et al., 2018)

Movies are often called “experience goods” because we cannot judge their quality until we have watched them. Unlike a physical product that we can test or inspect before buying, the actual value of a movie, whether it is the story, the acting, or the production, only becomes clear after we watch it (Eliashberg, J. et al., 1997).

This uncertainty makes it challenging for moviegoers to choose which movies to watch in which season, so they often turn to reviews, ratings, and recommendations for help (Parc J. et al., 2021).

Digital platforms like IMDb, Rotten Tomatoes, and social networks like Facebook, Twitter, and Instagram have become essential hubs for movie discussions and reviews. These spaces not only provide easy access to information but also heighten the impact of social influence. This concept refers to how the attitudes and actions of others shape people’s opinions. A glowing review or enthusiastic post can spark interest in a movie, creating widespread buzz and attracting more viewers. Conversely, harsh critiques or negative comments can deter potential audiences, discouraging them from watching the movie altogether (Chatterjee, P. , 2001).

Social networks and review platforms have become a significant force shaping how people view movies. Positive discussions and good reviews can turn a film into a blockbuster by building excitement and encouraging more people to watch it. On the other hand, negative feedback spreads quickly and can keep audiences away, sometimes even before a movie gets a fair chance. Studios understand this power and use social media for targeted marketing campaigns to generate buzz and guide public perception. This shows that social influence isn’t just about what people say, it’s also a deliberate tool the movie industry uses to drive success (Parc J. et al., 2021).

After reviewing multiple studies such as (Cheng, L et al., 2022), (Babic Rosario A. et al., 2016), and (Liu, 2006) analyze the impact of WOM on box office performance, which is used for forecasting sales outcomes and shaping the audience’s behavior. Before analyzing the effect of WOM on box office performance, they categorize the

different aspects of online Word of Mouth into volume, valence, and variance. The volume (frequency) is the total amount of reviews or user interactions. When the volume is high, there is a higher level of audience interaction. While the variance indicates how much disagreement or difference of viewpoints there is among reviewers. When we observe, higher variance may lead to higher anxiety of reviewers and can generate curiosity, especially for experienced products such as movies. For example, (Zimbra et al., 2017) found out that tweets with various perspectives can gain attention depending on timing and the platform. Third, the valence represents the sentiments, and the tone of the reviews, characterized as positively, negatively or neutral. The article explores both positive and negative statements independently, concluding that, while both types generate revenue, positive lines have more significant impact on motivating potential moviegoers.

After analyzing many papers such as the ones above, it is concluded that the number of reviews a movie gets has a significant impact on its box office performance. In many cases, the sheer volume of reviews matters more than the tone of the reviews (valence) (positive or negative) or their variety. This is because the number of reviews reflects how much attention and engagement a movie generates, making it one of the most reliable indicators of its success. A study in *The Electronic Library* found that the total number of reviews a movie receives has the most significant impact on its box office success. While the content of the reviews does play a role, it has more substantial impact on specific genres, making the overall volume of reviews a stronger and more consistent predictor of performance (Li et al., 2018).

High WOM volume relates to popularity, encouraging more individuals to watch the movie, especially in the prerelease and opening week phases. Additionally, action and adventure generate high WOM volume, while R-rated movies attract less WOM. (Liu, 2006).

On the other hand, (Cheng, L et al., 2022) extend this by showing that volume amplifies audience awareness, regardless of sentiment, making it particularly influential for blockbuster genres like action and comedy, while negative reviews influence curiosity in horror and drama. Similarly, (Babic Rosario A. et al., 2016) show that volume consistently outperforms other WOM metrics in driving consumer decisions, suggesting that hedonic products such as movies benefit more from hedonic products than utilitarian ones.

In conclusion, the collective findings of these studies establish a robust foundation for incorporating WOM volume as a key variable in my thesis. By analyzing how WOM volume interacts with other factors such as reviews, budgets, and genre-specific dynamics, my research can contribute to a deeper understanding of movie success, so the following hypothesis is structured:

Hypothesis *1: Consumer reviews positively impact box office performance, with stronger effects during the holiday season compared to non-holiday season.*

2.5 Professional reviews

Professional reviews are different than those of consumers because they rely more on concrete authoritative information, and they are more objective because of their professional status. These reviews significantly impact moviegoers' perceptions, particularly those seeking information before deciding to watch a movie. Academic research has indicated that professional reviews have a significant impact on box office performance, but this depends on which time period is more influential at the production status, post–production status, pre – release or released status of a movie, it depends on the stage of a movie lifecycle (Basuroy S. et al., 2003).

According to (King, T., 2007), professional reviews usually do not correlate with audience reviews because they emphasize more on the filmmaking, technical features such as the cinematography, direction, screenplay structure and areas that are more

relevant to the critical professional perspective or those who are segmented to the niche audience.

On the other hand, the moviegoers are often influenced by emotional impact, entertainment value, or relatability, which may not always align with professionals' opinions. However, it is noticeable to note that the influence of professional reviews varies across different genres and production scales. This is also a further reason why the current study emphasizes the importance of specific genres as a control variables (Gemser, G. et al, 2007).

Nevertheless, the study of (Basuroy S. et al., 2003) highlights that both positive and negative reviews are correlated with weekly box office revenue over eight weeks, indicating that professional reviews can be both as influencers and predictors of box office performance. For instance, when there are negative reviews in the early weeks of the release stage of a movie, the role of professional reviews, if they are negative, is more significant than those of positive reviews.

The study of (Reinstein et al., 2005), which examined the long–term box office performance, indicated that positive professional reviews have a significant positive impact on box office performance, especially movies that are more highly costed which could be defined as niche movies.

To sum up, in terms of all the findings of the ones' above academic studies, this current study ended up with this hypothesis, which supports that:

Hypothesis 2: Professional reviews have a positive impact on box office performance, and this is stronger during the holiday season compared to non-holiday season.

2.6 Budget

The production budget, which represents the total cost of creating a movie, including expenses for the cast, crew, equipment, and post-production, plays a significant role in filmmaking. However, estimating its impact on box office performance can be

challenging for researchers. This is primarily because marketing budgets, often critical to a movie's success, are rarely disclosed and challenging to track.

As a result, researchers often concentrate on the production budget despite it providing only a partial view of the factors influencing a movie's box office performance (Eliashberg et al., 2006).

Although most of the research, such as (Scott, 2019) and earlier studies like (Litman et al., 1983) , have proved that the relationship between budget and box office performance is positive and significant, there is a factor that we should take into consideration which the study of (Scott, 2019) proves it. It mentions that increasing a movie's budget typically leads to higher box office revenue; however, the impact of additional spending diminishes as the budget grows. Once essential investments like star casting, production quality, and marketing are covered, further expenditures tend to produce a smaller return on investments in revenue. This highlights the importance of studios carefully managing budgets to ensure that additional spending delivers meaningful returns.

This is why academic studies typically define budget ranges to align with the economic structure of the film industry. For instance, (Subramaniyaswamy et al., 2017) classified films into three categories based on their production costs based on censoring: low-budget (under $50 million), mid-budget ($50–150 million), and big-budget (over $150 million). This categorization helps to highlight the unique financial dynamics of different types of productions.

Based on these findings, the third hypothesis that this study will assress is: *Hypothesis 3: Budget has a significant positive impact on box office performance, with stronger effects during the holiday season compared to non-holiday season.*

2.7 Star Actors

According to (Gunter B, 2018) and other studies, the impact of star actors is strongly significant and positive on box office performance but complex because it is influenced by different factors such as the various genres of movies or if the movie is a sequel.

Producers often cast famous star actors to make a movie more visible and attract more audiences, leveraging their popularity to generate pre-release buzz and increase box office revenues. Surprisingly, its success is determined by the film's genre and the actor's image, which must match moviegoers' expectations. For example, action and adventure films may benefit more from star power, but horror films may rely more on storylines. (Nelson et al., 2012)

While the presence of star actors often significantly boosts box office revenue due to their ability to attract large audiences, their high salaries can sometimes lead to reduced profits. This occurs when the movies' overall profits are insufficient to offset the substantial cost of hiring the star actor, potentially resulting in diminishing returns on investment. However, during peak release seasons, such as holidays, including a celebrity actor can boost a film's success, exploiting increased audience availability and heightened competition to generate attendance and revenue (Scott, 2019), (Pangarker et al., 2013).

Sequels often rely on the success of their predecessors, and the presence of popular actors in sequels considerably boosts their box office potential. Star actors attract loyal followers and continue the existing audience base from the prior movie (Russo, 2016)

According to those findings, the fourth hypothesis of this study is: *Hypothesis 4: The presence of star actors positively impacts box office performance, with stronger effects during the holiday season compared to the non-holiday season.*

2.8 Conceptual Framework

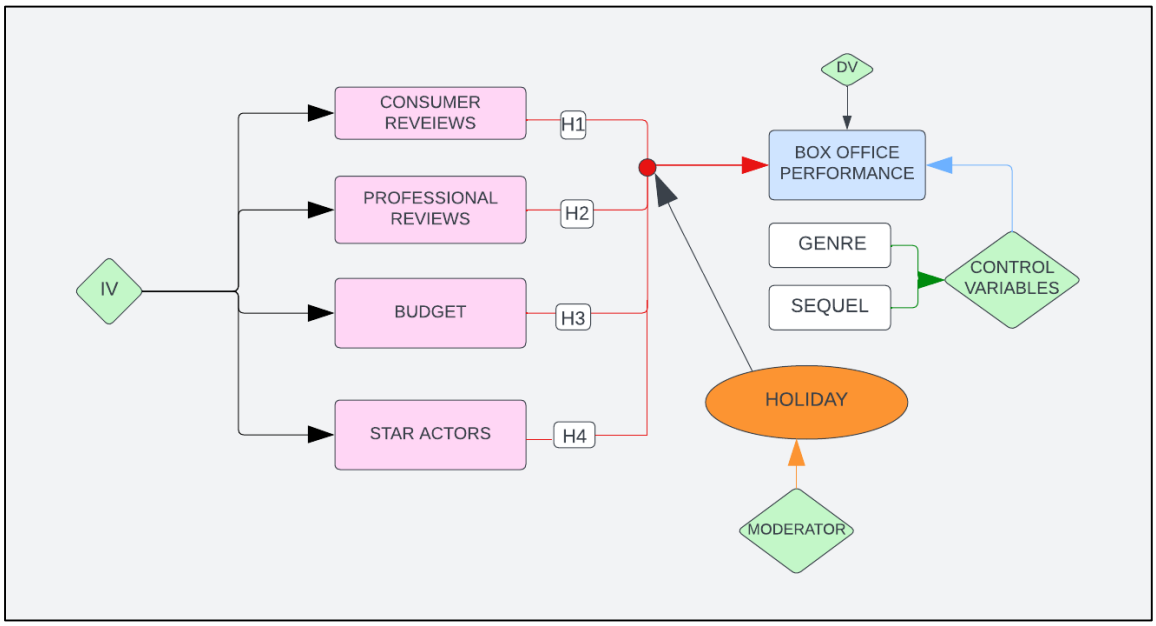


Figure 2) of Conceptual Model made from Lucidspark

The figure of Conceptual Model is created in this study and is designed to visually represent the relationships between key factors influencing box office performance. This model aims to provide a clear structure for testing the research hypotheses outlined in this study while ensuring that the key variables and their interactions are well-defined and logically connected.

The ***independent variables*** in the conceptual model include ***consumer reviews***, ***professional reviews, budget***, and ***star actors*****.** These variables are selected based on their importance in literature and relevance to understanding box office success.

Regarding ***consumer and professional reviews***, the focus is on ***volume***, which means the total number of reviews a movie receives. This isn't just a measure of how many people are talking about the movie, but a reflection of how much attention the movie is getting.

The more reviews there are, the greater the audience's engagement and curiosity, making the volume a strong indicator of a movie's overall reach and impact.

Star actors play a significant role in a movie's success even after its release. Their presence often keeps audiences interested and drives people to theaters based on their performance and popularity. A well-known actor can enhance a movie's appeal and sustain audience engagement over time, making their contribution vital to a film's post-release box office performance.

The ***dependent variable, box office performance***, is the central focus of this research. It is defined as the total revenue generated by a movie during its theatrical release.

Seasonality is an essential factor depicted in the conceptual model, and it has been written as a Holiday, focusing on how the timing of a movie's release during holidays like Christmas, Thanksgiving, Easter, Halloween, New Year's Eve, Valentine's Day, or the summer, affects its box office performance. These times are significant because people are likelier to go out, spend time on entertainment, and watch movies with family and friends, creating an ideal opportunity for films to perform better.

The diagram also considers ***genres*** **l**ike **drama, adventure, action, sci-fi, thriller, comedy,** and **horror** because different types of movies appeal to various audiences and perform uniquely at the box office. It also includes ***sequel status*****,** as sequels often benefit from prior audience familiarity and established interest. These factors are included to isolate better and understand the specific impact of the main variables on a movie's box office performance.

3. Data Preparation & Description

3.1 Data Preparation

At this chapter will be described the process that the data were structured, the sources of the data, the way that the variables of the dataset were created and structured, cleaned and steps of creating the main final dataset for analysis. This process is critical because it ensures data quality, identifying and correcting errors like missing values, duplicates, outliers, or other inconsistencies that might skew the accuracy of analysis results. Otherwise, unhandled outliers or missing values in the dataset could distort the analysis of factors like the impact of professional reviews or seasonality on box office performance. (Max Kuhn et al., 2013).

Moreover, a well-prepared and described dataset ensures that the data is aligned with the research objectives and hypotheses of the study. It allows us to focus on relevant variables and relationships that directly address the research question with clarity and transparency. (S. B. Kotsiantis et al., 2006)

3.2 Data Collection

The data collection was made from two main data sources widely recognized in the entertainment industry and available publicly: the **IMDb (Internet Movie Database)** dataset and the **Rotten Tomatoes** dataset. Both datasets are extracted from Kaggle.com, which is a well-known platform where data scientists and researchers share datasets for various projects. By accessing their datasets through Kaggle, I could leverage reliable, publicly available data that aligns perfectly with this study's research objectives, which are relevant to understanding the factors that drive movie success.

Additionally, IMDb and Rotten Tomatoes are industry-standard sources of movie data. The origins of these datasets were carefully reviewed to ensure their accuracy and reliability.

It was confirmed that both were created using web scraping techniques, with data directly collected from the official IMDb and Rotten Tomatoes websites. This verification provided confidence that the datasets were sourced from legitimate platforms and contained trustworthy information for analysis.

The **IMDb (Internet Movie Database)** dataset includes detailed data on movie titles, release dates, budgets, box office revenues, and audience ratings. For this research, the IMDb dataset provided essential insights into the financial aspects of movies and audience reception. Its inclusion was critical for understanding how variables like budget and revenue contribute to box office success.

On the other hand, the **Rotten Tomatoes** dataset complements the IMDb dataset by focusing on reviews and ratings from both professionals and audiences. Rotten Tomatoes is well-known for its “Tomato meter” score, which aggregates professional critic reviews, and its audience score, including the public opinion. Additionally, it includes information on the number of reviews and the volume providing an indicator of engagement and interest. This dataset was invaluable for examining the influence of consumer and professional reviews on box office performance.

3.3 Data Integration

The previously mentioned IMDB dataset contains 10,178 rows and 12 columns are the titles of the movies, the release date of the movies, the score, which is the volume of the consumer reviews as numerical values, the overview, which is the descriptions or summaries of the movies, the crew which includes the actors of the movies, the titles of the movies. Categorical variables include the language of the movies, the country of productions and the genre of the movies. While the budget of movies and the revenue is about the worldwide box office revenue as continues numerical values. Also, it is chosen for the worldwide box office revenue instead of the domestic because it depicts

a better picture of movies' success around the world, and it's a more complete and realistic way to measure its true impact.

The Rotten Tomatoes dataset includes 143,258 rows of movies and contains 16 columns of variables, specifically the Tomato Meter, which is numerical and describes the volume of professional reviews. It also contains the budget, revenues, and genres as the IMDB dataset.

The final dataset was created by merging the two datasets based on the same movie titles and release dates. Before merging them, I verified that both datasets have standardized column names and formats and no duplicates.

The data were collected from 2000 to 2023, including more recent data compared to the studies that were discussed in the literature review. The merged movie dataset includes 778 movies with the same titles and release dates from 2000 to 2023.

The dataset initially included 778 movies but was refined to 745 ***(Appendix A1)*** after filtering out movies from Asia, specifically countries such as China, India, and Thailand, and genres such as biography and western. This adjustment was made to simplify the analysis, particularly when considering the holiday season as a moderator.

3.3.1 Control Variables

3.3.1.1 Genre

The genre categorical variable is defined as a dummy variable (1 if the movie is in the specific genre or 0 if it is not), and it is estimated with the IF Excel formula. The movies are grouped by genre as drama with 254 movies, horror with 167 movies, action with 155 movies, comedy with 177 movies, thriller with 311 movies, sci-fi with 53 movies, and adventure with 90 movies.

Using dummy variables as control variables is a common approach in regression analysis to identify the impact of specific categories while accounting for their influence on the outcome being studied. Dummy variables make it possible to compare how

different categories (such as a movie genre) affect the results, using one category as a baseline for comparison (Wooldridge J. M., 2013).

Genre	Number of Movies
Thriller	311
Drama	254
Comedy	177
Horror	167
Action	155
Adventure	90
Sci-Fi	53

Table 1: Number of movies by Genre

3.3.1.2 Sequels

To identify sequels in the dataset, I created a dummy variable called "Sequel," assigning a value of 1 for sequels and 0 for non-sequels. I used a systematic approach to detect sequels by searching for common indicators in movie titles, such as numbers (e.g., "2," "3") and keywords like "Part," "Returns," or "Rebirth." After this automated process, I manually reviewed the remaining movies to ensure accuracy. Out of the 745 movies in the dataset, 89 were identified as sequels. This approach aligns with the (Russo, 2016) findings, which highlight the significance of sequels in box office success due to their existing audience base. ***(Appendix A2)***

3.3.2 Moderator Variable

To integrate seasonality into the analysis, I created several variables to capture key holiday periods and seasonal effects across different regions. A dummy variable named **holiday** was introduced to account for the general influence of holidays on box office performance.

Additionally, more specific variables were created to identify distinct holiday periods and seasons such as *Christmas, New Year's Eve*, *Summe*r, *Thanksgiving*, *Easter*, and *Halloween* also was created *Valentine's Day,* but it includes only 4 movies that were released that date, so it will not taken into account much in the study. The holiday dummy variables are created at specific dates for the countries, for example, Australia has different summer dates than the other countries ***(Table 2).***

Including the seasonal dummy variables in regression models ensures more precise estimates and a clearer understanding of the relationships between variables, as releasing movies during holidays and peak times often attracts larger audiences and boosts box office earnings. (Thomas, J. et al., 1971)

The formula for including seasonal dummy variables in a regression model is as follows (Berry et al., 1985)*:*

$$Yt = \beta_0 + \beta_1 D_1 + \beta_2 D \ldots. \beta_n - D_n - 1 + \varepsilon t$$

Where:

- *Yt: Dependent variable (e.g., box office revenue)*
- *Di: Seasonal dummy variables (e.g., Christmas, New Year's Eve, Summer, Thanksgiving, Easter, Halloween)*
- *β0: Intercept*
- *εt: Error term*

Holiday/Season	*Dates*	*Countries*
Christmas	Dec 1 – Dec 26	US, CA, GB, IE, AU, FR
New Year's Eve	Dec 27 – Jan 5	US, CA, GB, IE, AU, FR
Thanksgiving (US)	Nov 20 – Nov 30	US
Thanksgiving (Canada)	Oct 1 – Oct 15	Canada
Halloween	Oct 15 – Nov 1	US, CA, GB, IE, AU, FR
Easter (2000-2023)	Varies by year (see detailed ranges)	US, CA, GB, IE, AU, FR
Summer (Northern Hemisphere)	June, July, August	US, CA, GB, IE, FR
Summer (Southern Hemisphere)	Dec, Jan, Feb	AU

Table 2: Holiday Dates

3.3.3 Star actors

The **star actor** variable is a dummy variable set to 1 if at least one of the 150 selected top actors appears in the movie's cast of the crew variable. This extensive list was created to accurately identify star actors and ensure adequate coverage across a wide range of movies. A smaller list resulted in fewer identified star actors, limiting the scope of the analysis. By including 150 actors, 228 movies were determined to feature star actors, providing a strong foundation for analyzing the influence of star power on box office success. This approach aligns with (Scott, 2019) findings, who chose the star actors from *Forbes* based on publicly available data, which includes all the famous highly paid actors. ***(Appendix A3)***

3.4 Missing Values

In the dataset, the **tomato** variable had approximately 61 missing values, which is 8% of the data, in the ***(Figure 3).*** To address this, I applied the **Multiple Imputation by Chained Equations (MICE)** method, which is suitable under the assumption of **Missing at Random (MAR).**

The tomato_missing dummy variable was created to identify which variables are related to the tomato missingness variable with correlation analysis ***(Appendix A4: Table 1).*** The correlation analysis showed that the tomato missingness was moderately associated with other variables like *score (-0.2598)*, which means that higher score values are associated with fewer missing tomato values.

The imputation process used regression models to predict missing values based on variables such as **score**, **revenue**, and **budget_x**, preserving the relationships within the dataset. The imputed values, averaging approximately 44.95 ***(Figure 4)*** for missing tomato entries, reflect the average prediction from regression models during the MICE process. This method ensures that the imputed values maintain consistency with the dataset and reduce the risk of bias caused by missing information. Additionally, the distribution of the **tomato** variable after imputation closely matches the original, indicating the effectiveness of the imputation process.

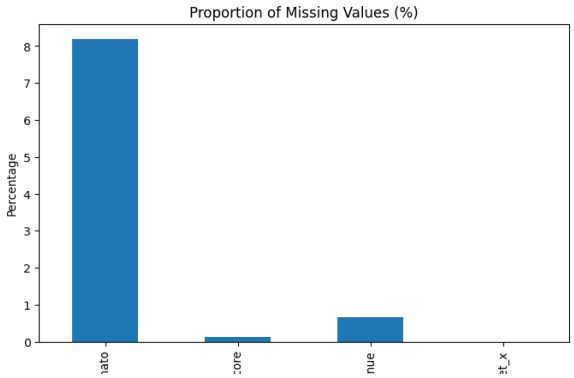


Figure 3) Proportion of Missing Values (%) per tomato, revenue, budget, score

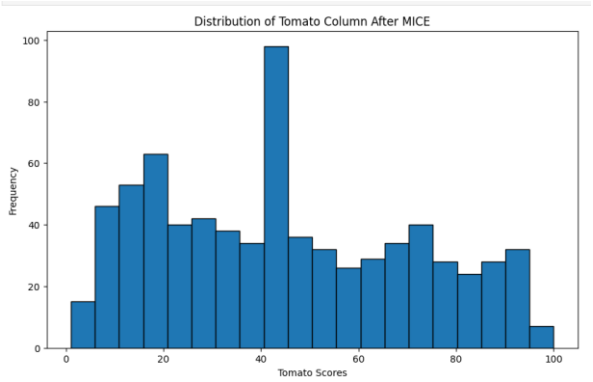


Figure 4) Tomato After MICE imputation

The **revenue** column in the dataset has only 5 missing values, which is just a small portion (0.67%) of the data. Even though it's a tiny amount, I addressed it to keep the analysis accurate. Since revenue numbers are very skewed **(Figure 5),** I used the median value, **46,611,048.5**, because it gives a better idea of typical revenue without being thrown off by extremely high values. This way, the missing data won't affect the overall results.

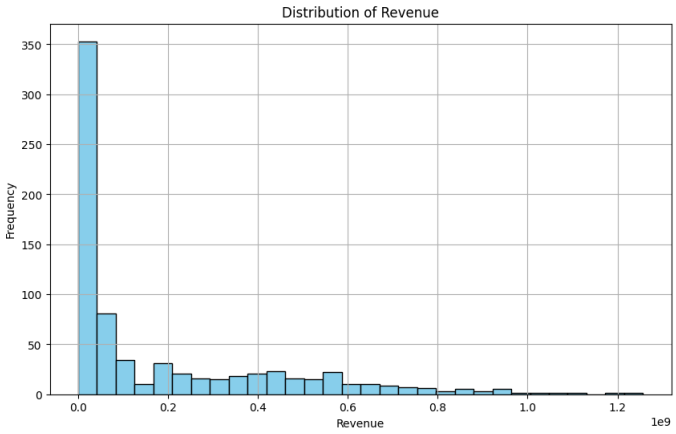


Figure 5) distribution of revenue after imputation

Variable	Missing Values
tomato	61
score	1
revenue	5
budget_x	0

Table 3) Missing Values 1

The **score** variable, the consumer reviews, has only 1 missing value, which I chose to retain in the dataset and replace with the mean, which is 60.88, so at 61 to ensure the completeness of the analysis without dropping any data.

According **to (Appendix A4: Figure 1),** after imputation the correlation matrix shows some interesting patterns. For example, movies with higher budgets tend to make more money, as seen in the strong connection between budget and revenue (0.68). Similarly, professional reviews (tomato) and audience scores (score) are somewhat aligned (0.44), showing that critics and viewers often share similar opinions. On the other hand, the weaker correlations between the genres and revenue, for example the adventure genre and revenue (0.16), suggest that the success of a movie isn't heavily tied to its genre.

3.5 Outliers

Outliers are unusual data points that are very different from the rest of the data and deviate from the stable pattern of distribution of the dataset. An outlier is higher or lower, specifically more than 1.5 times the interquartile range above the upper quartile or 1.5 times below the lower quartile. It's important to find them before analyzing the data because they can change the results, make models less accurate, and lead to wrong conclusions. At this point, the outliers are analyzed to ensure reliable estimations and conclusions of the analysis (Charu C. Aggarwal, 2017).

Variables	Outliers	IQR	Z-Score
revenue	25	3.36%	1.48%
budget	10	1.34%	0.94%
score	7	0.94%	0.40%
tomato	0	0	0

Table 4: Outlier Detection Summary

Regarding the **score** variable **(volume of consumer reviews)** with the IQR method, 7 outliers were detected with a percentage of 0.94 % < 5% from the IQR method, while the Z-Score method flagged 0.40% ***(Table 4),*** suggesting a low proportion of outliers that likely have minimal impact on the overall analysis. The boxplot for the **score (volume of consumer reviews)** of the **(Figure 6)** score variable shows that most of the data is centered between approximately 50 and 70 with the **median approximately 60.** Most of the outliers being presented at the lower and 1 at the upper end of the range at the value of 87 score.

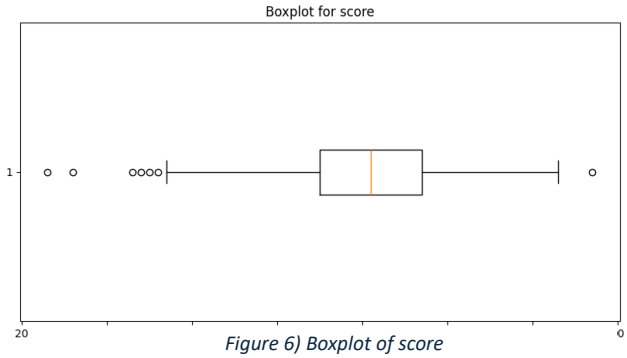


Figure 6) Boxplot of score

These outliers are values that fall outside the typical range but provide important information about movies that performed significantly better or worse than average. Keeping these outliers in the analysis is crucial as they can reveal valuable insights into extreme cases.

The **budget** variable has 10 outliers ***(Table 4)*** and are depicted in the boxplot ***(Figure 7)*** 200 million with a median roughly **0.4×10^8** (or **40 million),** which is the orange line within the box represents, indicating the middle value of the budget distribution, while the edges of the box represent the **first quartile (Q1)** and **third quartile (Q3)**, which capture the middle 50% of the data. Using the IQR method, **1.34%** of the data were identified as outliers, while the Z-Score method flagged **0.94%**. These high-budget outliers, often tied to blockbuster productions, are retained in the analysis as they provide critical insights into how large budgets impact box office performance.

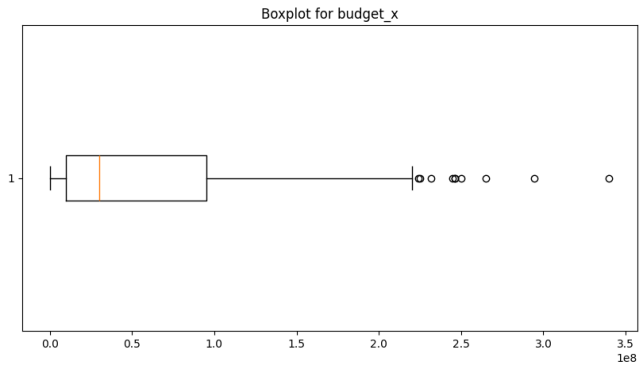


Figure 7) Boxplot of budget

The **tomato** variable the **(volume of professional reviews)** has no outliers **(*Table 4)*** because there are no data points beyond the whiskers, also the median value of professional reviews is around 50 ***(Figure 8).***

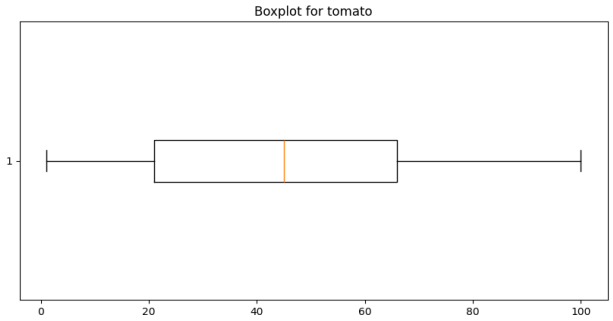


Figure 8) Boxplot of tomato

Regarding the **revenue** variable, 25 outliers are detected ***(Table 4),*** with a median of approximately **0.05×10^8** (or **5 million***), (figure 9).* Most of the revenue values are concentrated upper approximately **0.8 billion**, several movies have exceptionally high revenues exceeding this range. Moreover, the outliers are about 3.36% with the IQR method and 1.48% with the Z-score method, both remain below 5% ***(Figure 9).*** Although the revenue has quite a lot more outliers than the other variables, still it is reasonable to keep the outliers at the analysis because it might provide valuable insights about blockbuster movies.

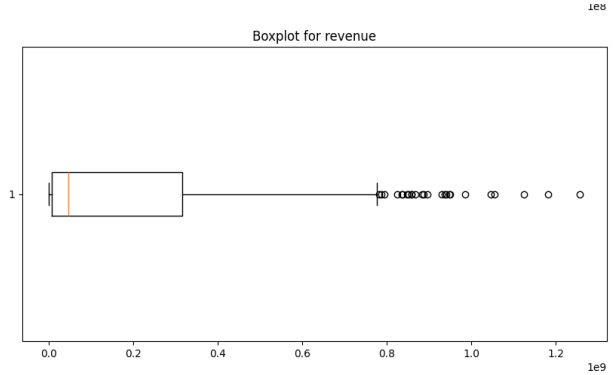


Figure 9) Boxplot of revenue

3.6 Data Description

The table of descriptive statistics ***(Table 5)*** shows that the average movie **revenue** is approximately $182 million, but with a significant variation, as indicated by the high standard deviation of $247 million. This highlights the large gap between blockbuster movies and less profitable ones, which is also confirmed based on ***(Appendix B: Figure 2)*** that it is right skewed. Similarly, the average movie **budget** is around $56 million based on ***(Table 5),*** with most movies having smaller budgets and only a few with very high budgets ***(Appendix B: Figure 1).*** The table of descriptive statistics further confirms this variability, showing the minimum and maximum values for revenue and budget to be $43 and $1.25 billion, and $85,645 and $340 million.

The descriptive statistics ***(Table 5)*** *and* visualizations **(Appendix B: *Figures 4, 11, 15, and 16)*** from data description appendixes highlight key trends in movie reviews. **Professional reviews** (mean = 45, SD = 26) are highly variable, while consumer reviews are more consistent and favorable, with a mean of 61 (SD = 9). **(Appendix B: *Figure 4)*** shows consumer' ratings clustering higher than professional scores. Seasonal releases (***Appendix B: Figure 11)*** slightly improve reviews, particularly during holidays. Genre-wise, drama, adventure, and sci-fi perform best in both professional and consumer ratings ***(Appendix B: Figures 15 and 16),*** while horror scores are lowest.

(Appendix B: Figure 3) reveals that genres such as adventure, action, and sci-fi dominate in box office revenue, with adventure movies showing the highest median revenue and significant variability, while comedy and horror genres consistently generate lower revenues. This suggests that action-packed and visually engaging genres are more lucrative compared to lighter or niche genres.

(Appendix B: Figure 5) highlights the influence of **star actors** which accounts for 30% of the movies in the dataset according to the table of descriptive statistics and **sequels** which accounts for 12% of movies on **revenue**. Movies featuring star actors that generally earn higher revenues compared to those without, while sequels consistently outperform non-sequels. Interestingly, the combination of sequels and star actors results in the highest revenue range, underlining the commercial advantage of leveraging established brands and talent.

Seasonal timing also impacts revenue, as seen in **(Appendix B: *Figures 6 and 8),*** October emerges as the most profitable release month ***(Figure 8)***, likely due to strategic positioning for holiday periods, while the overall distribution of movie releases peaks during late summer and early autumn ***(Figure 6).*** This reinforces the significance of targeting audience availability during peak seasons for maximizing box office returns.

The **holiday** variable indicates that 41% of movies were released during holiday periods, which is further emphasized in ***(Appendix B: Figure 18),*** showing summer as the most popular season for releases, followed by Christmas and Halloween. Additionally, **(Appendix B: *Figure 13)*** shows that during the holiday period, the presence of star actors slightly increases the revenue. This aligns with strategic efforts to capitalize on higher audience availability during holidays. ***(Appendix B: Figure 7)*** shows a consistent increase in the number of movie releases over the years, with a noticeable spike after 2020.

This is complemented by ***(Appendix B: Figure 14),*** which highlights a sharp increase in average revenue after 2015, suggesting that recent movies are more frequent and more lucrative, potentially due to larger budgets.

Variables	Data Description	N	Mean	Standard Deviation	Minimum	Maximum
title	Title of movie i	745				
Dependent Variable						
revenue	Global Box office revenue for movie I in USD$	745	182,042,633	247,492,045	43	1,256,887,580
Independent Variables						
budget_x	Production budget of movie I (in USD)	745	56,253,180	59,110,483	85,645	340,000,000
tomato	Average professional review score (1-100), Rotten Tomatoes	745	44.9557047	25.9434232	1	100
score	Average consumer reveiws (1-100) IMDB	745	60.88590604	8.976170815	23	87
star_actor	Indicator for whether the movie I features a star actor (1 = Yes, 0 = No)	745	0.30738255		0	1
Moderator Variable						
Holiday	Indicator for whether the movie I was released during a holiday season (1 = Yes, 0 = No)	745	0.413422819		0	1
Control Variables						
sequel	Indicator for whether the movie I is a sequel (1 = Yes, 0 = No)	745	0.119463087		0	1
Genre	Indicator for where the movie I belongs to the genres : (drama, comedy, sci-fi, adventure, thriller, action, horror)	745				

Table 5) Descriptive Statistic

4. Methodology

This study aims to analyze with the use of Moderated Multiple Regression (MMR) the impact of key factors such as budget, star actors, and consumer and professional reviews on box office performance, moderated by seasonality (holiday and non–holiday seasons). The MMR is analyzed in Python and R language. This chapter starts with the explanatory model which describes the algorithms of the modelsused in this study, then model specification, model estimations and validations.

4.1 Explanatory Model

Multiple Linear Regression with interaction terms are used as the primary statistical approach in this study, which investigates linear relationships between dependent and multiple independent variables, allowing for the assessment of direct effects and interaction terms. The inclusion of interaction terms helps the model to capture how the impact of key predictors on dependent variable changes depending on the moderator variable, for the current study, how the impact of predictors on box office performance changes based on seasonality (holiday vs non-holiday periods). (Hayes, A. et al., 2009)

The general MMR formula is depicted as: (Berry et al., 1985).

$$Y_i = \alpha + \beta_x X_i + \beta_z Z_i + \beta_{xz} X_i Z_i + \varepsilon_i$$

4.1.1 Model Specification

The models that are estimated for this study are the Holiday Model, the Non–Holiday Model, the Baseline Model, which are estimated by Multiple Linear Regressions, and the Combined Model by Multiple Moderator Regression, which will be described.

Before analyzing them, it is noteworthy to mention that the variables of revenue and budget are log-transformed before starting the analysis because from the ***(Appendix B: Figures 1,2)***. The figures show that the revenue and budget are not normally distributed but skewed. (John Paul et al., 2017).

The **Holiday Model** estimates the impacts of the main predictors (score, which is the volume of consumer reviews, tomato, which is the volume of professional reviews, budget, and the dummy variable star_actor (famous actors' presence) during the holiday periods, (Holiday = 1).

The formula is:

$$log\_revenue_i = \beta_0 + \beta_1 log\_budget_i + \beta_2 scores_i + \beta_3 tomato_i + \beta_4 star\_actor_i + \beta_5 (Holiday = 1)_i + \beta_6 sequel_i + \beta_7 action_i + \beta_8 comedy_i + \beta_9 drama_i + \beta_{10} horror_i + \beta_{11} adventure_i + \beta_{12} thriller_i + \beta_{13} sci\text{-}fi_i + \epsilon_i$$

Where:

- β_0:the intercept.
- β_1: log-transformed budget of movie i
- β_2: volume of consumer reviews (from IMDB) scores of movie i
- β_3: volume of professional reviews (Tomato) score of movie i
- β_4: the presence of star actors (1, 0 no star actors) for movie i
- β_5: movies i released in holiday periods
- β_6: dummy variable for sequels (1=sequel, 0=non-sequel) for movie i
- β_7,β_8,β_9,β_{10}, β_{11}, β_{12}, β_{13},= coefficients for genre dummy variables (action, comedy, drama, horror, thriller and sci-fi). The reference category others_genres is included in this model.
- ϵ_i= the error term for movie i

The **Non–Holiday Model** estimates the effects of predictors for movies released during non – holiday periods (Holiday = 0), so the formula will be the same as the Holiday model but with Holiday = 0, which represents the movies that are released during the non–holiday period

Where:

$log\_revenue_i =$

$$\beta_0 + \beta_1 log\_budget_i + \beta_2 scores_i + \beta_3 tomato_i + \beta_4 star\_actor_i + \beta_5 (Holiday = 0)_i + \beta_6 sequel_i + \beta_7 action_i + \beta_8 comedy_i + \beta_9 drama_i + \beta_{10} horror_{i+} \beta_{11} adventure_i + \beta_{12} thriller_i + \beta_{13} sci\text{-}fi_i + \epsilon_i$$

The **Baseline Model** includes only the direct effects of the independent and control variables on the dependent variable, excluding the moderator variable Holiday and non – Holiday**.** The formula is:

$log\_revenue_i =$

$$\beta_0 + \beta_1 log\_budget_i + \beta_2 scores_i + \beta_3 tomato_i + \beta_4 star\_actor_i + \beta_5 sequel_i + \beta_6 action_i + \beta_7 comedy_i + \beta_8 drama_i + \beta_9 horror_{i+} \beta_{10} adventure_i + \beta_{11} thriller_i + \beta_{12} sci\text{-}fi_i + \epsilon_i$$

Combined Model: This model entails interaction terms of independent variables (budget, star_actor, score, tomato) with the moderator variable Holiday (for holiday and non-holiday periods), investigating how seasonality moderates the relationships between predictors and box office performance. The formula is:

$log\_revenue_i =$

$$\beta_0 + \beta_1 log\_budget_i + \beta_2 scores_i + \beta_3 tomato_i + \beta_4 star\_actor_i + \beta_5 ((Holiday = 1) * log\_budget)_i + \beta_6 ((Holiday=1) * scores)_i + \beta_7 ((Holiday=1)) * tomato)i + \beta_8 ((Holiday=1) * star\_actor)_i + \beta_9 sequel_i + \beta_{10} action_i + \beta_{11} comedy_i + \beta_{12} drama_i + \beta_{13} horror_i + \beta_{14} adventure_i + \beta_{15} thriller_i + \beta_{16} sci\text{-}fi_i + \beta_{17} Holiday_i + \epsilon_i$$

Where:

- $\beta_5((Holiday=1) * log\_budget)_i$: shows the interaction effects between the log – transformed budget of movie i and holiday variable which shows what is the impact of budget on box office performance revenue differently during the holiday periods
- $\beta_6 ((Holiday=1) * scores)_I$: shows the interaction effect between the volume of consumer reviews of movie i and the holiday variable, this coefficient represents what is the impact of consumer reviews on box office performance revenue during holidays compared to non – holiday periods.
- $\beta_7 ((Holiday=1) * tomato)i$: represents the interaction effect between the volume of professional reviews of movie i and the holiday variable which shows how the impact of professional reviews on box office performance revenue differs during holiday periods compared to non – holiday periods.
- $\beta_8 ((Holiday=1) * star\_actor)_i$: examines how the presence of star actors in movie I impacts box office revenue during holiday periods compared to non – holiday periods.

Isolated Interaction Effects

The **isolated interaction effects** estimate how the main key factors (score, tomato, log_budget, and star_actor) interact with the moderator variable (Holiday) independently. These models are created to test whether the four hypotheses structured to this study are supported or not.

Isolated Interaction Effect of Score (volume of consumer reviews) and Holiday: examines if consumer reviews have a stronger or weaker effect on box office revenue during holiday periods.

The formula is:

$$log\_revenue_i = \beta_0 + \beta_1 scores_i + \beta_2 sequel_i + \beta_3 action_i + \beta_4 comedy_i + \beta_5 drama_i + \beta_6 horror_i + \beta_7 adventure_i + \beta_8 thriller_i + \beta_9 sci\text{-}fi_i + \beta_{10} (Holiday * scores)_i + \beta_{11} (Holiday_i = 1) + \epsilon_i$$

Isolated Interaction Effects of Budget and Holiday estimates the interaction effect of the budget of a movie i during holiday periods compared to non – holiday periods, on box office revenue. The formula is: where Holiday = 1

$$log\_revenue_i = \beta_0 + \beta_1 log\_budget_i + \beta_2 sequel_i + \beta_3 action_i + \beta_4 comedy_i + \beta_5 drama_i + \beta_6 horror_i + \beta_7 adventure_i + \beta_8 thriller_i + \beta_9 sci\text{-}fi_i + \beta_{10} (Holiday * log\_budget)_i + \beta_{11} (Holiday_i = 1) + \epsilon_i$$

Isolated Interaction Effect of Tomato (professional reviews) and Holiday: examines if the effect of the volume of professional reviews of a movie i varies between holiday and non – holiday periods. The formula is:

$$log\_revenue_i = \beta_0 + \beta_3 tomato_i + \beta_5 sequel_i + \beta_6 action_i + \beta_7 comedy_i + \beta_8 drama_i + \beta_9 horror_i + \beta_{10} adventure_i + \beta_{11} thriller_i + \beta_{12} sci\text{-}fi_i + \beta_{13} (Holiday * tomato)_i + \beta_{14} (Holiday_i = 1) + \epsilon_i$$

Isolated Interaction Effect of star actors and Holiday period: examines which is the effect of star actors that are presented in the movie i, on box office revenue during holiday periods.

$$log\_revenue_i = \beta_0 + \beta_4 star\_actor_i + \beta_5 sequel_i + \beta_6 action_i + \beta_7 comedy_i + \beta_8 drama_i + \beta_9 horror_i + \beta_{10} adventure_i + \beta_{11} thriller_i + \beta_{12} sci\text{-}fi_i + \beta_{13} (Holiday * star\_actor)_i + \beta_{14} (Holiday_i = 1) + \epsilon_i$$

4.2 Model Estimations & Assumptions

The coefficients of the MMR models are estimated to be using Ordinary Least Squares (OLS), a widely used method in multivariate analysis. The main goal of the OLS is to minimize the sum of squared residuals between the observed and predicted values, ending up with unbiased estimates. (Dismuke, C. et al., 2006)

Its simplicity, interpretability and accuracy make it ideal for examining complex interactions, such as how factors like budget and reviews are moderated by seasonality to influence box office performance. To ensure robust results, assumptions of heteroscedasticity, multicollinearity, non – normality, linearity, and assessment of influential data points (Wooldridge J. M., 2013).

4.2.1 Heteroscedasticity

An important assumption in regression analysis is that the variance of the error terms for each predictor variable should have a constant variance, a condition known as homoscedasticity (Rosopa P. J. et al., 2013). However, the residual plots for the **Holiday, Non-Holiday, Combined,** and **Baseline models** reveal that this assumption is violated, as the spread of residuals varies across the predicted values. This indicates heteroscedasticity, which can make the parameter estimates less reliable and potentially impact the validity of the model's conclusions ***(Appendix C.***

The **Breusch-Pagan Test** is a statistical test that checks for heteroscedasticity in a regression model, which is the unequal spread of residuals across the predicted values of the dependent variable. The hypothesis that estimates are null hypothesis **(**Homoscedasticity: constant variance of residuals) and the alternative hypothesis: (heteroscedasticity: variance of residuals depends on predictors) (Breusch, T. S., 1979).

For all models (**Holiday, Non-Holiday, Combined, Baseline**), the **p-values** are extremely low ($p<0.001$), indicating significant evidence of heteroscedasticity.

Breusch–Pagan Holiday Model	Breusch–Pagan Non – Holiday Model
p-value: 0.0002	p-value: $4.08e^{07}$

Table 4) Breusch – Pagan test p-values

Breusch–Pagan Combined Model	Breusch–Pagan Baseline Model
p-value: $4.92e^{11}$	p-value: $3.03e^{12}$

Table 5) Breusch – Pagan test p-values

4.2.2 Multicollinearity

Multicollinearity usually happens in a multiple regression model when two or more predictor variables are highly correlated with one another and share a similar relationship with the outcome variable (Franke, 2010). The multicollinearity assumption was evaluated using VIF values and correlation matrices across all models. A VIF between 5 and 10 indicates high correlation that might be a significant constraint (Noora Shrestha, 2020). Both the **Non-Holiday Model** and the **Baseline Model** demonstrated no significant multicollinearity concerns, as all VIF values were below 2. This indicates that the predictors in these models are independent and reliables. However, the **Holiday Model** and **Combined Model** presented higher VIF values for the interaction terms, such as *holiday_budget* (VIF: 131.29) and *holiday_score* (VIF: 61.28), ***(Appendix C: Figure 8)***. Additionally, the correlation matrices revealed strong relationships between some interaction terms and main predictors.

These findings suggest that the **Holiday** and **Combined Models** may produce fewer stable results unless adjustments, like centering the interaction terms, are made. On the other hand, the **Non-Holiday** and **Baseline Models** appear more robust and provide a solid foundation for further analysis ***(Appendix C).***

4.2.3 Non- Normality

The normality assumption checks whether the residuals (errors) of the regression model follow a normal distribution which ensures valid statistical estimations (Jorge I. et al., 2015). Regarding the **Q-Q plots *(Appendix C*)**, indicate approximate normality for most residuals in the central, although deviations are observed in the tails, it is concluded as a mild non-normality. This was confirmed by the **Shapiro-Wilk test** which is a statistical test, and the null hypothesis of it is that the residuals (error terms) should follow a normal distribution while the alternative one, the opposite argumentation supports (Pek, J. et al., 2018).

Shapiro-Wilk Holiday Model	Shapiro-Wilk test Non – Holiday Model
P-value: 0.004, W-statisic= 0.9371	p-value: 0.0008 W-statisic= 0.9898

Table 6) Shapiro – Wilk test p-values

Shapiro-Wilk Combined Model	Shapiro-Wilk test Baseline Model
P-value: 0.004 W-statisic= 0.9371	p-value: 0.0008 W-statisic= 0.9898

Table 7) Shapiro - Wilk test p-values

According to ***(Table 6,7)*** the results of the Shapiro Wilk tests for all the models are highly significant which reject the null hypothesis of normality, but this contradicts with the W- statistic as the values are close to 1 which means that is closer to normality (Wilk M. B. et al., 1965) In conclusion, the models have mild normality which is acceptable as the linear regressions are resilient to violations of normality.

4.2.4 Linearity

The connection between the independent and dependent variables must be linear. If this requirement is not satisfied, the results may be misinterpreted.

To ensure this condition is met, the linearity assumption is evaluated by examining **scatterplots** and conducting the **Rainbow Test** which the null hypothesis assumes that the relationship between the independent and dependent variables are linear (Casson et al., 2014) . The results show that the relationship between the independent variables and the dependent variable is mostly linear for **the Non-Holiday Model, Baseline Model,** and **Combined Model**, as the **Rainbow Test** shows high p-values. Although the **Holiday Model** had a slightly lower p-value, the overall findings suggest that the models generally meet the linearity assumption and can be used for further analysis***. (Appendix C).***

Rainbow Test Holiday Model	Raibow Test Non – Holiday Model
P-value: 0.004, W-statisic= 0.9371	p-value: 0.0008 W-statisic= 0.9898

Table 8) Rainbow Test p-values

Rainbow Test Combined Model	Rainbow Test Baseline Model
P-value: 0.004 W-statisic= 0.9371	p-value: 0.0008 W-statisic= 0.9898

Table 9) Rainbow Test p-values

4.2.5 Assessment of Influential Data Points

After deciding to retain the outliers in the dataset, it became essential to evaluate whether these outliers significantly impacted the regression results. To investigate this, **Cook's Distance and Residuals vs. Leverage plots** were created for both models, highlighting the presence of potential influential data points**. *(Appendix C).***

Across all models (**Non-Holiday, Holiday, Combined** and **Baseline),** the **Cook's Distance values** mostly fall below the threshold, indicating that most data points have

little impact on the regression results (Cook, 1979). Similarly, the Residuals vs. Leverage plots ***(Appendix C)*** show that most data points are within the acceptable range for both residuals and leverage (Fitrianto A. et al., 2022). These findings support the decision to keep all observations in the analysis, as removing them is unlikely to significantly affect the outcomes. This also highlights the models' ability to handle variability within the dataset effectively.

4.3 Usage of A.I.

Throughout this thesis, I utilized AI tools to enhance efficiency and quality of the analysis. ChatGPT-4 (https://openai.com/index/gpt-4/) was instrumental in refining the clarity of my writing and, most importantly, in supporting my coding tasks in Python. It helped for debug errors quickly, saving significant time and ensuring smooth progress during data analysis. Additionally, I used Perplexity AI (https://www.perplexity.ai/) and Consensus (https://consensus.app/search/)to efficiently provided relevant academic papers, enabling a comprehensive and up-to-date literature review. These AI tools facilitated my workflow, allowing me to focus more on critical analysis and interpretation, ultimately contributing to the overall quality of the thesis.

5. Results

This chapter presents the outcomes of regression analyses to understand how the budget, star actors, consumer reviews, and professional reviews influence box office performance moderated by seasonality (holiday vs. non-holiday). Four main models, Non-Holiday, Holiday, Baseline, and Combined, were estimated across multiple stages**: pre-assumption OLS regressions**, **mean-centered interaction models**, **robust standard error corrections**, and **Ridge regression** for multicollinearity adjustments.

5.1 Holiday Model

The Holiday Model examines which factors influence the box office revenue during the holiday season, the movies released during peak holiday periods (Holiday = 1) are 308. **R-squared = 0.334,** explains the variance in box office revenue during holiday. Also, the **Adjusted R-squared = 0.306**, explains the variance of the model, after adjustments and it is slightly lower and close to R squared which means a slight penalty for including variables it does not make the model more complex. The model seems to be statistically significant as the **F-statistics = 12.30 (p <0.001)** meaning the variables included in the model collectively influence revenue during holidays. For comparison of the model with the other estimated models, **AIC = 1385,52** and **BIC = 1424.01** are useful metrics and lower values of them means a better fit of the model. ***(Appendix D, Tables 2,6,14).***

For the Holiday Model with ridge regression, the **Mean Squared Error (MSE) is 4.84**, which shows how far off the model's predictions are from the actual box office revenues. A smaller MSE means the model is doing a better job at making accurate predictions. ***(Appendix D, Table18)***

The **log_budget** with coefficient 0.477 seems to be statistically significant ($p<0.001$) which means for every 1% increase in the movies' budget, box office revenue

increases by approximately 4.77% during the holiday season. Also, the **holiday_star** actor with coefficient -0.430 ($p<0.01$) means that when the star actors are included in the movies negatively impacts revenue during holidays. Moreover, the **action genre** with a coefficient of -0.622 and ($p<0.1$, marginal significance), means that they do not perform better during holidays, also the **horror genre** with a coefficient of 0.656 ($p<0.1$, marginal significance) seems to perform well during holidays. On the other hand, regarding the non-significant predictors, **tomato (professional reviews)** with a coefficient near–zero means that do not impact revenue during holidays also same for **score (consumer reviews**). ***(Appendix: D, Tables 2,6,14).***

The results above seemed to remain consistent through the four stages that were mentioned the **before assumptions checks (raw model)**, **after assumptions checks (mean – centered)** and **robust standard errors (adjusting for violations like heteroscedasticity),** and **ridge regression (facing multicollinearity**), that means that the model is robust and reliable. ***(Appendix: D, Tables 2,6,14,18).***

5.2 Non – Holiday Model

The Non – Holiday model investigates which factors influence box office revenue during non – holiday seasons. The movies that are released during the non – holiday season in the dataset are 437 movies. The model has **(R-squared = 0.446), 44.6 %** of the variation in box office revenue during non – holiday season is explained by the independent variables.

After the inclusion of the additional variables the model explains **(Adjusted R-squared = 0.403), 43.0% of** the variation in box office revenue which is close to R-squared. The model is also statistically significant with **F-statistic 28.46 ($p<0.001$).** The **AIC = 1868.92** and **BIC = 1921.96**, as was said to the holiday model chapter that lower values show a better model fit. ***(Appendix: D, Tables 3, 5, 13)***

For the non-holiday model with **ridge regression,** it seems that the **MSE = 3.97** which is smaller average prediction error compared to holiday – model which was mentioned that has MSE = 4.84. ***(Appendix D, Table18)***

Across all the four stages that were discussed the **log_budget** with coefficient 1.222 ($p<0.001$) shows that a 1% increase in budget, increase the revenue by 1.222% during non – holiday season. The **score (consumer reviews) across all the four stages has** coefficient 0.031 ($p<0.05$) which shows 1% increase in the volume of consumer reviews, increases box office revenue by 3.1%. While tomato **(professional reviews)** has a negative impact on box office revenue which means with 1% increase in the volume of professional reviews decreases the box office revenue by 1.28%. The **star_actor** seems to not significantly influence box office revenue during non-holiday period. The control variable **thriller (genre)** with coefficient 0.493 ($p<0.05$) has a better performance on box office revenue increasing it by 49.3% compared to other genres during the non-holiday period. The **sequel movies** seem also to perform better in the non-holiday season by increasing the revenue by 64% compared to non – sequels. ***(Appendix: D, tables 3, 5, 13).***

5.3 Baseline Model

The baseline model gives a more holistic view of what impacts box office revenue, including all the movies and on holiday and non – holiday season but excludes the interaction effect of Holiday and Non-Holiday.

The baseline model with **R-squared = 0.379**, means that it explains 37.9% of the variation in the revenue. The **Adjusted R – squared = 0.369** is lower than R-squared but close to it. The model is statistically significant with **F – statistic = 37.24 ($p<0.001$)** and **Baseline model 1** and **Baseline model 2** have the same value (because these models assume that the data has equal variance across all observations (homoscedasticity)) ***(Appendix: D, tables 1, 7).***

In **Baseline model 3**, robust standard errors are used to correct for this unequal variance (heteroscedasticity) so the **F-statistic** is declined to **19.85678 (p<0.001)** because it adjusts for the real variability in the data, making it more accurate because it recalculates the test by accounting for the unequal variances in the data. ***(Appendix: D, Table15)*** so it reflects a more accurate test after addressing the assumption of heteroscedasticity. ***(Appendix: D, Tables 1, 7, 15)***

For the baseline model the **ridge regression** is the **MSE = 4.48**, which seems to be merely better than the holiday model (MSE = 4.84), which means that indicates overall prediction accuracy when compared to holiday and non – holiday. ***(Appendix D, Table 18)***

The **log_budget** with coefficient = 1.102 ($p < 0.001$), means that a 1% increase in budget increases revenue by 1.102%. The **star_actor** with coefficient -0.510 ($p<0.01$), means that the inclusion of star actors in movies has a negative impact on box office revenue, which means that it reduces it by 51% compared to not including them. The **tomato (professional reviews)** with coefficient -0.007 ($p < 0.05$**)** means that the volume of professional reviews has a negative significant impact on revenue. The **score (volume of consumer reviews)** with coefficient 0.021 ($p<0.054$) is marginally significant which means that higher volume of consumer reviews has a marginally positive impact on revenue. The **seque**l movies with coefficient 0.506 ($p<0.05$) means that sequel movies increase revenue by 50.6 % compared to non-sequels. At the baseline model it seems that the **control variables genres** do not have a significant impact on box office revenue except for the action genre which has a small marginal significance, and it is negative***. (Appendix: D, Tables 1, 7, 15)***
Similarly, here at the baseline model the results are consistent at all the four stages which means that the model is robust and can give reliable findings***. (Appendix: D, Tables 1, 7, 15, 18)***

5.4 Combined Model

The combined model analyzes how different factors impact box office revenue across both holiday and non – holiday periods for all the 745 movies, while including interactions effects and accounting for how seasonal timing influences these relationships.

The **R-squared = 0.393** which explains the **39.3 %** of the variation in box office revenue for the **Combined model 1** and **Combined model 2** ***(Appendix: D (Tables: 4,8)*** also **Adjusted R – squared = 0.371** is same for the three model., While the **R-squared** = 0.38 for **Combined model 3** ***(Appendix: D Table: 16)*** slightly lower due to robust standard error.

The model is statistically significant because the **F-statistics = 27.70104 (p<0,001)** for **Combined model 1**, for **Combined model 2, 28.47605 (p<0.001**) and for **Combined model 3** is **15.72845 (p<0.001**) which is the lowest after robust standard errors. (***Appendix: D Tables: 4,8,16)***

In **Combined Model 1**, the **holiday interaction term** has a **VIF of 167.87 (*Appendix: C Table 8 VIF values Combined Model),*** indicating severe multicollinearity with other variables. To address this, the term was excluded in the *mean-centered (Model 2), robust standard error (Model 3),* ridge regression is chosen for the stages for more stable and reliable estimates.

The **AIC = 3249.89** and **BIC = 3332.93** are very close almost similar for all the three combined models. ***(Appendix: D, tables 4, 8, 16)***

For the **combined model** the ridge regression with **MSE = 4.49** it is very close to the MSE = 4.48 of the baseline model and slightly better than the holiday model which MSE = 4.84. This shows that the combined model does a good job of identifying trends during both holiday and non-holiday periods. ***(Appendix D, table 18)***

The coefficients of the **log_budget** for **combined models (2,3)** is 1.15 (p<0.01) while for **combined model 1** is 1.25 and significant also, it is merely higher, in all stages the log_budget is significant and positively impacts the revenue with almost 1.25 % increase. At **the combined model 1** the **holiday** coefficient is 6.39 and significant which means that a 1 % increase in movies released during the holiday season increases the revenue by 63.9% which is a significant increase. ***(Appendix D, Table 4).*** The **holiday_budget** at **combined model 1** with coefficient -0.31 seems to have negative impact on revenue during holidays also at **combined models 2,3** it is negative but not significant during holidays. ***(Appendix D, Tables 8, 16).***

The **holiday_star_actor** with coefficients of -0,69 at **combined model 1** and -0.78 at **combined model 2,3** seems to have a negative impact on the revenue during holidays (the presence of star actors in movies). The **tomato** coefficient is also negative and significant at the three stages which means that professional reviews have a consistent negative effect on revenue. While the **sequel** movies seem to have a positive and significant effect of increase 48% on revenue during holidays at **combined models 2,3*. (Appendix D, Tables 8, 16).***

Score (consumer reviews) seems to have a limited positive significant impact on revenue but it is not consistent across stages, and it appears at the **combines models 2,3** ***(Appendix D, Tables 8, 16).***

5.5 Combined Model OLS & Ridge

Ridge regression is applied to address high multicollinearity by shrinking coefficients close to zero, the model gains robustness, allowing for more reliable predictions and better handling of correlated predictors (Hoerl A. et al., 1970). Specifically, in the **Holiday** and **Combined Models**, where VIF values were excessively high ***(Appendix C, Tables 5,8).***

The **Combined Model** is chosen according to ***(Appendix D, table 17)*** because it offers a more complete view of box office performance. Ridge regression also improved

stability and reduced the **Mean Squared Error (MSE)**, with the **Combined Model** achieving a lower **MSE (4.49)** compared to the **Holiday Model (4.84)**, which means that it is closer to the actual values. This made the ***Combined Model*** more suitable for the current study.

The estimated **Mean Squared Error (MSE**) is the average squared difference between the observed values and the predicted values. A smaller MSE indicates better prediction accuracy. Moreover, according to (Chai, 2014), MSE is defined as one of the strongest metrics in regression analysis compared to RMSE and MAE in this case because it penalizes large deviations more heavily than small ones, it is used MSE.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Figure of MSE formula

Models	Significant Predictors	Non-Significant Predictors	R-squared	Adjusted R-squared	F-statistic	MSE (Ridge Regression)
Holiday Model	log_budget (+), holiday_star_actor (-), action (-, marginal), horror (+, marginal)	tomato (professional reviews), score (consumer reviews)	0.334	0.306	12.30 ($p < 0.001$)	4.84
Non-Holiday Model	log_budget (+), score (consumer reviews, +), tomato (professional reviews, -), thriller (+), sequel (+)	star_actor	0.446	0.403	28.46 ($p < 0.001$)	3.97
Baseline Model	log_budget (+), star_actor (-), tomato (professional reviews, -), score (consumer reviews, marginal +), sequel (+)	Genres (except action, marginal -)	0.379	0.369	37.24 ($p < 0.001$), 19.86 (robust)	4.48
Combined Model	log_budget (+), holiday (+), holiday_star_actor (-), tomato (-), sequel (+)	holiday_budget (-, not significant in 2,3), score (consumer reviews, limited significance)	0.393	0.371	27.70 (1), 28.48 (2), 15.73 (3) (all $p < 0.001$)	4.49

Table 10: Summary of Results of Models

6. Discussion

This chapter discusses how the results' findings align with the research question and the hypothesis, as well as the managerial implications and limitations for future research and the conclusion.

6.1 Hypothesis Testing & Findings

In this study, the primary focus was to estimate the dynamics of factors influencing box office revenue by analyzing multiple models and examining their significance. To delve deeper into these relationships, the isolated interaction effects of those factors with the holiday season, are discussed and the results are compared to the hypotheses and prior research.

Primarily, the interaction effect between **consumer reviews (score**) and **holidays** is not significant ($p > 0.05$), indicating that consumer reviews do not have a stronger effect during the **holiday season**. Also, the direct effect of **consumer reviews (score**) on **revenue** during **non – holiday** periods is not significant ($p > 0.05$) but could be defined as marginally significant ($p = 0.0998$) ***(Appendix D, Table 9)***. Same finding from **Baseline model** and **Non Holiday Model (Appendix D, Tables 5,15)** indicates that **consumer reviews** may have a limited but weaker impact on box office revenue during **non-holiday** periods. These findings **do not support *Hypothesis 1*** and align with prior research, such as (Liu, 2006), which highlights the role of consumer reviews in less saturated periods.

A possible explanation is that during non-holiday periods, audiences may rely on consumer reviews more to decide which movies to watch, as there are fewer major blockbusters dominating the market. In contrast, well-promoted films and franchises take center stage during holidays, making consumer reviews less influential in driving box office revenue.

Secondly, the interaction effect between **professional reviews (tomato)** and **holidays** is also non-significant ($p > 0.05$), ***not supporting Hypothesis 2*** which confirms the opposite ***(Appendix D, Table 10).*** Similarly, the direct effect of professional reviews is not significant during non-holiday periods, but from the results chapter it is concluded that professional reviews have a minimal but negative impact on revenue during non-holiday periods. These results diverge from prior studies, such as (Basuroy S. et al., 2003), highlighting professional reviews' predictive value.

This might happen because, during holiday periods, many blockbuster releases happen with marketing campaigns, likely to overshadow the role of professional reviews entirely during that peak period. As a result, marketers in film industries should use positive professional reviews in their promotional campaigns, like featuring them in trailers or ads, to match audience expectations and build excitement. This strategy can be beneficial during non-holiday periods when people might need more convincing about a movie's quality.

The interaction effect between **budget** and **holiday** is negative and significant (coefficient = -0.3, $p < 0.01$), which means that the budget has a negative impact on box office revenue during the holiday period. ***(Appendix D, Table 11).*** Also, it is confirmed from the results chapter from most of the estimated models. Those findings are supported by (Scott, 2019), highlighting that while high-budget films generally perform well, their impact can be lessened during holidays due to increased competition. While the direct effect of the budget on revenue during the non – holiday period is positive and significant, it is **not supported by *Hypothesis 3****, which* supports that budget has a positive significant impact on revenue during the holiday season.

High-budget films face tough competition from other big releases during the holidays, which eliminates the audience's attention and limits their individual success. A more effective strategy for studios could be to release high-budget films during quieter periods outside the holiday season.

The interaction effect between **star actors** and **holidays** is significant and negative **(coefficient = -0.957, p < 0.05**) which means that nearly 96% decrease in box office revenue during the holiday period which is highly significant ***(Appendix D, Table 12***). As a result, hypothesis ***4*** **is not supported** as it supports the opposite. Nevertheless, the direct effect of star actors on box office revenue is positive but not significant during non – holiday periods ***(Appendix D, Table 12***). One reason could be the strong competition during the holidays, when audiences often choose well-known series or popular movie types instead of films that rely on star actors. This aligns with (Gunter B, 2018), who found that the impact of star power depends on the context, such as how crowded the market is and what audiences prefer. Similarly, (Scott, 2019) highlighted that the holiday season adds extra challenges for star-focused films, as the intense competition makes their box office performance more unpredictable.

6.2 Managerial Implications

After analyzing the results and discussing them, significant managerial implications have been concluded for the entertainment industry. First, the findings indicate that seasonality analysis is important for marketers, as the hypothesis is not supported, supported they must follow more actionable planning, for example, focusing on predictive analytics, audience-centered campaigns, and strategic timing, studios can maximize box office revenue. To effectively conduct seasonality analysis, studios should study historical data to identify performance for various genres and marketing strategies during holiday and non-holiday periods.

During non-holiday periods, movie marketers can use positive reviews to build trust and attract audiences by featuring them in digital ads, social media posts, and trailers. Highlighting quotes like "A must-see!" or star ratings in marketing helps generate buzz and make the movie stand out when competition is lower. Regarding, professional reviews, movie marketers should strategically highlight positive reviews for niche or mid-budget films during non-holiday periods, incorporating them into pre-release

campaigns and using the feedback to improve technical aspects to align better with audience expectations.

High-budget movies should be scheduled during non-holiday periods to avoid intense competition with other blockbusters, with star actors' presence. Additionally, marketers should highlight unique aspects of these films, such as groundbreaking visuals or innovative storytelling over star actors' presence during non-holiday periods, and focus more on sequel movies during holiday periods when it comes to high-budget movies.

6.3 Limitations & Future Research

While this study provides valuable insights, some limitations should be acknowledged. First, multicollinearity among independent variables posed challenges during analysis. Despite using mean-centering, robust standard errors, and ridge regression, multicollinearity persisted to some extent, since it obscures the true relationship between variables and affects the reliability of findings. Advanced methods or alternative approaches should be explored to ensure more robust results. A possible reason for the multicollinearity could be the limited dataset, future research should focus on a more expansive global dataset.

Additionally, this study analyzed the holiday season as a single category because of limited time, without distinguishing the specific effects of individual holidays such as Christmas or Thanksgiving. This approach may overlook variations in audience behavior associated with distinct holiday periods. Future research should analyze specific holidays to capture nuanced audience behaviors and provide a more detailed understanding of seasonality effects.

Another limitation is the analysis of consumer and professional reviews, which are based on volume but not sentiment analysis. Future research could include sentiment analysis and analyzing how seasonality moderated the relationship of review variance

with box office revenue, providing a holistic understanding of how reviews influence box office performance.

Lastly, the growing role of streaming platforms has not been explored. Future studies should explore how streaming platforms impact box office performance, and how releasing movies on streaming services at the same time as theaters or after a delay impacts audience choices and overall revenue.

7. Conclusion

Previous studies have shown that box office success depends on various factors, including budget, star actors, audience reviews, and release timing. This study aimed to explore how these factors, along with seasonality influence box office performance. The research question was: "How do key factors such as budget, star actors, consumer and professional reviews interact with seasonality (holiday vs. non-holiday) to influence box office performance across specific genres?" The analysis rejected all null hypotheses and found that **budget** is the most important factor, especially during non-holiday periods, where it has the strongest effect due to less competition. The ridge combined model was identified as the best because it effectively captured the interactions between predictors and seasonality, providing the most reliable results by addressing multicollinearity issues more effectively.

This study highlights the importance of budgeting and strategic release timing while showing that star actors and reviews have a limited impact during holiday periods. These findings add to understanding box office dynamics and encourage further research into how new trends and platforms affect movie success. Future research is encouraged to expand on these findings, incorporating broader datasets and emerging market trends to refine predictive models and strategic approaches.

Bibliography

Ahmed et al. (2019, August 29). Pre-production box office success quotient forecasting. (p. o. Springer-Verlag GmbH Germany, Ed.) Soft Computing. doi:https://doi.org/10.1007/s00500-019-04303-w

Babic Rosario, A. et al. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. Journal of Marketing , 53(3), 297-318. doi:https://doi.org/10.1509/jmr.14.0380

Basuroy, S. et al. (2003). How critical are critical reviews? The box office effects of film critics, star power, and budgets. Journal of Marketing, 67(4), 103–117. doi:https://doi.org/10.1509/jmkg.67.4.103.18692

Berry et al. (1985). Multiple Regression in Practice. 96. doi:https://doi.org/10.4135/9781412985208

Bhattacharya, K. et al. (2009). Econophysics of Markets and Business Networks. Springer.

Breusch, T. S. (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica: Journal of the Econometric Society, 47(5), 1287–1294.

Casson et al. (2014). Understanding and checking the assumptions of linear regression: a primer for medical researchers. Clinical and Experimental Ophthalmology. 590-596. doi:https://doi.org/10.1111/ceo.12358

Chai, T. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geoscientific Model Development,. 1247–1250.

Charu C. Aggarwal. (2017). Outlier Analysis. Springer

.

Chatterjee, P. . (2001). Online Reviews: Do Consumers Use Them? . Advances in Consumer Research, 28, 129 - 133.

Cheng, L et al. (2022). The effect of online reviews on movie box office sales: An integration of aspect-based sentiment analysis and economic modeling. Journal of Global Information Management, 30(1), 1-15. doi:https://doi.org/10.4018/JGIM.298652

Cook, R. D. (1979). Influential observations in linear regression. Journal of the American Statistical Association, 169–174.

De Vany, A. (2004). Hollywood Economics: How Extreme Uncertainty Shapes the Film Industry. Routledge.

Dismuke, C. et al. (2006). Ordinary least squares regression and logistic regression: Better methods for analyzing dichotomous outcomes. 41(2), 782–805.

Duan et al. (2008). The dynamics of online word-of-mouth and product and product sales. An empirical investigation of the movie industry. Journal of Retailing, 84(2), 233-242.

Eliashberg et al. (2006). The Motion Picture Industry: Critical Issues in Practice, Current Research, and New Research Directions. Springer Science & Business Media. doi:https://doi.org/10.1007/978-3-319-71803-3

Eliashberg, J. et al. (1997). Film critics: Influencers or predictors? Journal of Marketing, 61(2), 68-78.

Eliashberg, J. et al.,. (2006). The Motion Picture Industry: Critical Issues in Practice, Current Research and New Research Directions. Marketing Science, 25(6),, 638-661.

Fitrianto, A et al. (2022). Comparisons between robust regression approaches in the presence of outliers and high leverage points. 243–252.

Franke, G. R. (2010). Multicollinearity Part 2. Marketing Research. doi:https://doi.org/10.1002/9781444316568.wiem02066

Gao et al., 2. (2019). How to make a successful movie: Factor analysis from both financial and critical perspectives. (Springer Nature Switzerland AG ed., Vol. Vol. LNCS 11420). (I. i. 2019, Ed.) Indiana University Bloomington, & Nanjing University of Science and Technology. Retrieved from https://doi.org/10.1007/978-3-030-15742-5_63

Gemser, G. et al. (2007). The impact of film reviews on the box office performance of art house versus mainstream motion pictures. Journal of Cultural Economics, 31(1), 43-63. doi:https://doi.org/10.1007/s10824-006-9025-4

Gunter B. (2018). How Significant Is Star Power? Springer International Publishing, 179 - 198. doi:https://doi.org/10.1007/978-3-319-71803-3

Gunter B. (2018). Some Genres More Profitable than Others? International Journal of Media Management, 20(3), 211-226.

Gunter. (2018). Predicting movie success at the box office. Palgrave Macmillan. doi:https://doi.org/10.1007/978-3-319-71803-3

Ha, L. (2011, October 1). Audience Evolution: New Technologies and the Transformation of Media Audiences. (P. Napoli, Ed.) Journal of Communication ISSN 0021-9916. Retrieved from

https://www.academia.edu/62234082/Audience_Evolution_New_Technologies_and_the_Transformation_of_Media_Audiences

Handke, C. (2021). The Box Office and the Long Tail: An Examination of the Effects of Streaming on the Distribution of Box office Revenue. Erasmus University Rotterdam.

Hayes, A et al. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. Behavior Research Methods, 41(3), 924-936. doi: https://doi.org/10.3758/BRM.41.3.924

Hoerl, A. et al. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 55–67.

Internet Movie Database. (n.d.). IMDb. Retrieved from IMDb: https://www.imdb.com/title/tt1073241/

Jehoshua Eliashberg et al. (2003). Demand and supply dynamics for sequentially released products in international markets: The case of motion pictures. Marketing Science. 22. doi:https://doi.org/10.1287/mksc.22.3.329.17740

John Paul et al. (2017). he Overlooked Importance of Constants Added in Log Transformation of Independent Variables with Zero Values: A Proposed Approach for Determining an Optimal Constant. 10(1), 26–29. doi:https://doi.org/10.1080/19466315.2017.1369900

Jorge I. et al. (2015). A modified Q-Q plot for large sample sizes. Comunicaciones en Estadística . doi:10.15332/s2027-3355.2015.0002.02

Kim et al. (2018). The effect of eWOM volume and valence on product sales - An empirical examination of the movie industry. International Journal of Advertising, 38(3), 471-488. doi:https://doi.org/10.1080/02650487.2018.1535225

King, T. (2007). Does film criticism affect box office earnings? Evidence from movies released in the U.S. in 2003. Journal of Cultural Economics, 31(3), 171–186. doi: https://doi.org/10.1007/s10824-007-9039-z

Lee et al. (2016). Predicting movie success with machine learning techniques: Ways to improve accuracy. Information Systems Frontiers, 20(3), 577-588. doi:10.1007/s10796-016-9689-z

Li et al. (2018). The effect of online user reviews on box-office revenue: The case of the movie industry. 36(2), 295–314. doi:https://doi.org/10.1108/EL-02-2018-0040

Litman et al. (1983). Predicting success of theatrical movies: An empirical study. Journal of Popular Culture, 16(4), 159-175. doi:https://doi.org/10.1111/j.0022-3840.1983.1604_159.x

Liu, Y. (2006). Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue. Journal of Marketing, 70(3), 74-89. doi:https://doi.org/10.1509/jmkg.70.3.74

Max Kuhn et al. (2013). Applied Predictive Modeling. Springer. Retrieved from https://link.springer.com/book/10.1007/978-1-4614-6849-3

Michalopoulos et al. (2024, March). National Bureau of Economic Research, Movies. (N. B. Research, Ed.) doi:https://doi.org/10.3386/w32220

Mirrlees, T. et al. (2013). Global entertainment media: Between cultural imperialism and cultural globalization.

Nelson et al. (2012). Movie stars and box office revenues: an empirical analysis. Journal of Cultural Economics, 36(2), 141–166. doi:http://www.jstor.org/stable/43549820

Noora Shrestha. (2020). Detecting Multicollinearity in Regression Analysis. American Journal of Applied Mathematics and Statistics., 39-42. doi:10.12691/ajams-8-2-1

Pangarker et al. (2013). The determinants of box office performance in the film industry revisited. South African Journal of Business Management, 44(3), 47-58. doi:10.4102/sajbm.v44i3.163

Parc, J. et al. (2021). Film Quality: The Golden Key for Sustainable Success. In: The Untold Story of the Korean Film Industry. Cultural Economics & the Creative Economy. doi: https://doi.org/10.1007/978-3-030-80342-1_6

Pek, J. et al. (2018). How to Address Non-normality: A Taxonomy of Approaches, Reviewed, and Illustrated. Frontiers in psychology,. doi:https://doi.org/10.3389/fpsyg.2018.02104

Reinstein et al. (2005). The influence of expert reviews on consumer demand for experience goods: A case study of movie critics. Journal of Industrial Economics, 53(1), 27-51. doi:https://doi.org/10.1111/j.0022-1821.2005.00249.x

Rosopa, P. J. et al. (2013). Managing heteroscedasticity in general linear models. Psychological methods, 18(3), 335–351. doi:https://doi.org/10.1037/a0032553

Russo, H. e. (2016). Eplaining Box Office Performance from the Bottom Up: Data, Theories and Models. George Mason University. doi:https://scholarworks.uni.edu/mtie

S. B. Kotsiantis et. (2006). Data Preprocessing for Supervised Learning. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE, 1(2), 111-117. Retrieved from https://www.researchgate.net/publication/228084519_Data_Preprocessing_for_Supervised_Learning

Scott, G. (2019). Determinants of Box Office Performance: Return of the Regressions. Major Themes in Economics, 21, 71–83. Retrieved from https://scholarworks.uni.edu/mtie/vol21/iss1/7

Squire. (2016). The Movie Business Book (4th ed.). Routledge. doi:https://doi.org/10.4324/9781315621968

Subramaniyaswamy et al. (2017). Predicting movie box office success using multiple regression and SVM. Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS 2017), 182-186. doi:https://doi.org/10.1109/ICISS.2017.8260040

Thomas, J. et al. (1971). Seasonal Variation in Regression Analysis. (S. A. (General), Ed.) Journal of the Royal Statistical Society, 134(1), 57–72. doi:https://doi.org/10.2307/2343974

Wilk, M. B. et al. (1965). An analysis of variance test for normality (complete samples). 52, 591–611. doi:10.1093/biomet/52.3-4.591

Wooldridge, J. M. (2013). Introductory Econometrics: A Modern Approach (5th ed.). (C. L. South Western, Ed.) Michigan State University. Retrieved from https://cbpbu.ac.in/userfiles/file/2020/STUDY_MAT/ECO/2.pdf

Zimbra et al. (2017). Movie aspects, tweet metrics, and movie revenues: The influence of iOS vs. Android. Decision Support Systems, 102, 98-109. doi:https://doi.org/10.1016/j.dss.2017.08.002

Appendix

A) Data Preparation

A1) Movie Titles

Appendix A: Movie Titles and Release Dates

Index	title	release_date
1	the super mario bros. movie	4/5/2023
2	supercell	3/17/2023
3	the devil conspiracy	1/13/2023
4	the passion of the christ	2/25/2004
5	consecration	2/10/2023
6	little dixie	2/3/2023
7	prey for the devil	10/28/2022
8	bandit	9/23/2022
9	breaking	8/26/2022
10	nocebo	11/4/2022
11	medieval	9/9/2022
12	detective knight: independence	1/20/2023
13	fast x	5/19/2023
14	air	4/5/2023
15	savage salvation	12/2/2022
16	kids vs. aliens	1/20/2023
17	glass onion: a knives out mystery	11/23/2022
18	the price we pay	1/13/2023
19	detective knight: rogue	10/21/2022
20	candy land	1/6/2023
21	the ledge	2/18/2022
22	harry potter and the half-blood prince	7/15/2009

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
23	the infernal machine	9/23/2022
24	abandoned	6/17/2022
25	the little mermaid	5/26/2023
26	pinball: the man who saved the game	3/17/2023
27	transformers: rise of the beasts	6/9/2023
28	spider-man: across the spider-verse	6/2/2023
29	harry potter and the order of the phoenix	7/11/2007
30	monstrous	5/13/2022
31	blood	1/27/2023
32	control	9/23/2022
33	the remaining	9/5/2014
34	crawlspace	3/31/2022
35	son of god	2/28/2014
36	luca	6/18/2021
37	the menu	11/18/2022
38	barbie	7/21/2023
39	maneater	8/26/2022
40	your place or mine	2/10/2023
41	puss in boots	10/28/2011
42	watcher	6/3/2022
43	wire room	9/2/2022
44	emancipation	12/2/2022
45	in the forest	1/28/2022
46	the amazing spider-man	7/3/2012
47	code name banshee	7/1/2022
48	god's not dead: a light in darkness	3/30/2018
49	the requin	1/28/2022
50	as they made us	4/8/2022
51	ice age: dawn of the dinosaurs	7/1/2009

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
52	samson	2/16/2018
53	the vatican tapes	7/24/2015
54	alita: battle angel	2/14/2019
55	last days in the desert	5/13/2016
56	terminator genisys	7/1/2015
57	spirit: stallion of the cimarron	5/24/2002
58	the friendship game	11/11/2022
59	black water: abyss	8/7/2020
60	the hangover part iii	5/23/2013
61	the case for christ	4/7/2017
62	unbroken: path to redemption	9/14/2018
63	the simpsons movie	7/27/2007
64	ready player one	3/29/2018
65	the flash	6/16/2023
66	the nun 2	9/8/2023
67	fortress: sniper's eye	4/29/2022
68	hunting ava bravo	4/1/2022
69	brian banks	8/9/2019
70	bullet proof	8/19/2022
71	aladdin	5/24/2019
72	sharper	2/10/2023
73	no manches frida 2	3/15/2019
74	heaven is for real	4/16/2014
75	a haunted house 2	4/18/2014
76	so cold the river	3/25/2022
77	american sicario	12/10/2021
78	fireproof	9/26/2008
79	a haunted house	1/11/2013
80	i'm not ashamed	10/21/2016

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
81	my father's dragon	11/4/2022
82	war room	8/28/2015
83	catch the fair one	2/11/2022
84	mindcage	12/16/2022
85	annabelle	10/3/2014
86	borrego	1/14/2022
87	you people	1/20/2023
88	assassin	3/31/2023
89	back to the outback	12/3/2021
90	god's not dead 2	4/1/2016
91	hidden	9/15/2015
92	the haunting of helena	6/21/2013
93	god's not dead	3/21/2014
94	facing the giants	9/29/2006
95	clean	1/28/2022
96	apex	11/12/2021
97	american murderer	10/21/2022
98	the young messiah	3/11/2016
99	kong: skull island	3/10/2017
100	scouts guide to the zombie apocalypse	10/30/2015
101	carriers	9/4/2009
102	troy	5/13/2004
103	blue beetle	8/18/2023
104	jeepers creepers 3	9/26/2017
105	strays	6/9/2023
106	warhunt	1/21/2022
107	the aviary	4/29/2022
108	mack & rita	8/12/2022
109	separation	4/30/2021

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
110	do you believe?	3/20/2015
111	bromates	10/7/2022
112	infinite storm	3/25/2022
113	amina	11/4/2021
114	indiana jones and the kingdom of the crystal skull	5/22/2008
115	the ultimate gift	3/9/2007
116	the cokeville miracle	6/5/2015
117	redeeming love	1/21/2022
118	dakota	4/1/2022
119	woodlawn	10/16/2015
120	indiana jones and the dial of destiny	6/30/2023
121	jumper	2/14/2008
122	spirit halloween: the movie	9/30/2022
123	season of the witch	1/7/2011
124	90 minutes in heaven	9/11/2015
125	gasoline alley	2/25/2022
126	the greatest beer run ever	9/30/2022
127	g.i. joe: retaliation	3/28/2013
128	jiu jitsu	11/20/2020
129	insidious: the last key	1/5/2018
130	elemental	6/16/2023
131	steven universe: the movie	9/2/2019
132	the outwaters	2/9/2023
133	dangerous game: the legacy murders	10/21/2022
134	dead silence	3/16/2007
135	wild card	1/30/2015
136	fifty shades of black	1/29/2016
137	american siege	1/7/2022
138	a day to die	3/4/2022

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
139	queen of spades	6/11/2021
140	as above, so below	8/29/2014
141	undisputed	8/23/2002
142	beverly hills chihuahua	10/3/2008
143	all saints	8/25/2017
144	all the old knives	4/8/2022
145	proximity	5/15/2020
146	the other side of heaven 2: fire of faith	6/28/2019
147	thirteen lives	7/29/2022
148	the matrix revolutions	11/5/2003
149	hot seat	7/1/2022
150	star trek	5/7/2009
151	seed of chucky	11/12/2004
152	the walk	6/10/2022
153	war of the worlds	6/29/2005
154	king kong	12/14/2005
155	overboard	5/4/2018
156	a walk among the tombstones	9/19/2014
157	his only son	3/31/2023
158	men in black ii	7/3/2002
159	the departed	10/6/2006
160	the adventures of sharkboy and lavagirl	6/10/2005
161	jigsaw	10/27/2017
162	the hunting	1/21/2022
163	i'm in love with a church girl	10/18/2013
164	a question of faith	9/29/2017
165	wifelike	8/12/2022
166	rogue one: a star wars story	12/16/2016
167	cirque du freak: the vampire's assistant	10/23/2009

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
168	ouija	10/24/2014
169	the protege	8/20/2021
170	paint	4/7/2023
171	asking for it	3/4/2022
172	no hard feelings	6/23/2023
173	the lone ranger	7/3/2013
174	offseason	3/11/2022
175	the seventh day	3/26/2021
176	left behind	10/3/2014
177	escape plan 2: hades	6/29/2018
178	my policeman	10/21/2022
179	showing up	4/7/2023
180	marriage story	11/6/2019
181	triple frontier	3/6/2019
182	kate	9/10/2021
183	the man from earth: holocene	10/13/2017
184	death wish	3/2/2018
185	son of the mask	2/18/2005
186	confess, fletch	9/16/2022
187	ouija: origin of evil	10/21/2016
188	unconditional	9/21/2012
189	exorcist: the beginning	8/20/2004
190	trigger point	4/23/2021
191	friday the 13th	2/13/2009
192	the letters	12/4/2015
193	the possession of hannah grace	11/30/2018
194	en brazos de un asesino	12/6/2019
195	i still believe	3/13/2020
196	how to blow up a pipeline	4/7/2023

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
197	clown	6/17/2016
198	the king's daughter	1/21/2022
199	accident man: hitman's holiday	10/14/2022
200	dominion: prequel to the exorcist	5/20/2005
201	2 hearts	10/16/2020
202	first love	6/17/2022
203	leatherface	10/20/2017
204	the unborn	1/9/2009
205	the unforgivable	11/24/2021
206	lamborghini: the man behind the legend	11/18/2022
207	spider-man 2	6/30/2004
208	collide	8/5/2022
209	what men want	2/8/2019
210	midnight sun	3/23/2018
211	red tails	1/20/2012
212	chips	3/24/2017
213	transformers: revenge of the fallen	6/24/2009
214	the last exorcism part ii	3/1/2013
215	guy ritchie's the covenant	4/21/2023
216	the cursed	2/18/2022
217	the mist	11/21/2007
218	batman beyond: return of the joker	12/12/2000
219	the babysitters	5/9/2008
220	secretary	9/20/2002
221	remember me	3/12/2010
222	el camino: a breaking bad movie	10/11/2019
223	dual	4/15/2022
224	the boy	1/22/2016
225	primeval	1/12/2007

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
226	jason x	4/26/2002
227	cinderella	9/3/2021
228	captain underpants: the first epic movie	6/2/2017
229	chasing mavericks	10/26/2012
230	slayers	10/21/2022
231	the ritual killer	3/10/2023
232	piranha 3dd	6/1/2012
233	call jane	10/28/2022
234	she's the man	3/17/2006
235	catherine called birdy	9/23/2022
236	breach	12/18/2020
237	the last house on the left	3/13/2009
238	jack and jill	11/11/2011
239	ash & dust	3/11/2022
240	the virtuoso	4/30/2021
241	the strangers	5/30/2008
242	teen titans go! to the movies	7/27/2018
243	how to be a latin lover	4/28/2017
244	bride wars	1/9/2009
245	aquaslash	6/23/2020
246	haunt	9/13/2019
247	encounter	12/3/2021
248	burn	8/23/2019
249	renegades	12/21/2018
250	the expendables 4	9/22/2023
251	siberia	7/13/2018
252	the fourth kind	11/6/2009
253	the ten	8/3/2007
254	leap year	1/8/2010

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
255	grace unplugged	10/4/2013
256	the haunting in connecticut	3/27/2009
257	a babysitter's guide to monster hunting	10/14/2020
258	annie	12/19/2014
259	triple threat	3/19/2019
260	the night clerk	2/21/2020
261	a-x-l	8/24/2018
262	meg 2: the trench	8/4/2023
263	walking tall	4/2/2004
264	slender man	8/10/2018
265	diary of a wimpy kid: the long haul	5/19/2017
266	the tiger rising	1/21/2022
267	strange magic	1/23/2015
268	children of the corn	3/3/2023
269	the guilty	9/24/2021
270	the raven	4/27/2012
271	the kill team	10/25/2019
272	thirteen	8/20/2003
273	ken park	8/31/2002
274	p2	11/9/2007
275	legend of the guardians: the owls of ga'hoole	9/24/2010
276	sin nombre	3/20/2009
277	son	3/5/2021
278	12 hour shift	10/2/2020
279	time is up	9/9/2021
280	chevalier	4/21/2023
281	freaks	9/13/2019
282	winchester	2/2/2018
283	safer at home	2/26/2021

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
284	the world to come	2/12/2021
285	state of play	4/17/2009
286	brick mansions	4/25/2014
287	look away	10/12/2018
288	survive the game	10/8/2021
289	winnie the pooh	7/15/2011
290	the wild thornberrys movie	12/20/2002
291	caged	1/26/2021
292	shottas	2/27/2002
293	i love you, beth cooper	7/10/2009
294	the aeronauts	12/6/2019
295	the marine	10/13/2006
296	maybe i do	1/27/2023
297	gridiron gang	9/15/2006
298	moneyball	9/23/2011
299	taurus	11/18/2022
300	wish upon	7/14/2017
301	the strangers: prey at night	3/9/2018
302	into the deep	8/26/2022
303	punisher: war zone	12/5/2008
304	skiptrace	9/2/2016
305	the locksmith	2/3/2023
306	trial by fire	5/17/2019
307	get rich or die tryin'	11/9/2005
308	euphoria	6/28/2019
309	teenage mutant ninja turtles: mutant mayhem	8/4/2023
310	the hurricane heist	3/9/2018
311	marmaduke	6/4/2010
312	sinister 2	8/21/2015

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
313	the wicker man	9/1/2006
314	date movie	2/17/2006
315	unforgettable	4/21/2017
316	uncle drew	6/29/2018
317	jethica	1/13/2023
318	girls trip	7/21/2017
319	debt collectors	5/29/2020
320	resurrection	7/29/2022
321	the rental	7/24/2020
322	beowulf	11/16/2007
323	the loft	1/30/2015
324	the haunting in connecticut 2: ghosts of georgia	2/1/2013
325	frequency	4/28/2000
326	dreamcatcher	3/5/2021
327	the day	8/29/2012
328	the dark and the wicked	11/6/2020
329	here today	5/7/2021
330	papillon	8/24/2018
331	bright	12/22/2017
332	the system	10/28/2022
333	cha cha real smooth	6/17/2022
334	paranormal activity 4	10/18/2012
335	definitely, maybe	2/14/2008
336	s.w.a.t.: under siege	8/1/2017
337	guns akimbo	2/28/2020
338	13 minutes	10/29/2021
339	don't kill it	3/3/2017
340	vertical limit	12/8/2000
341	born a champion	1/22/2021

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
342	book of shadows: blair witch 2	10/27/2000
343	the collector	7/31/2009
344	the eye	2/1/2008
345	leprechaun: origins	8/22/2014
346	kidnap	8/4/2017
347	wicked little things	11/17/2006
348	the fog	10/14/2005
349	spell	10/30/2020
350	the butterfly effect 3: revelations	1/9/2009
351	captive state	3/15/2019
352	the invasion	8/17/2007
353	the condemned	4/27/2007
354	zombeavers	3/20/2015
355	7 guardians of the tomb	2/23/2018
356	when the bough breaks	9/9/2016
357	heist	11/13/2015
358	annapolis	1/27/2006
359	a score to settle	8/2/2019
360	the bad batch	6/23/2017
361	cold skin	9/7/2018
362	joy ride	10/5/2001
363	the omen	6/6/2006
364	easter sunday	8/5/2022
365	brothers by blood	1/22/2021
366	replicas	1/11/2019
367	oldboy	11/27/2013
368	aftermath	4/7/2017
369	dumb and dumberer: when harry met lloyd	6/13/2003
370	child 44	4/17/2015

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
371	kickboxer: retaliation	1/26/2018
372	universal soldier: day of reckoning	11/30/2012
373	day of the dead: bloodline	1/5/2018
374	inside	1/12/2018
375	halloween ii	8/28/2009
376	uptown girls	8/15/2003
377	the grudge 2	10/13/2006
378	v/h/s	10/5/2012
379	the green knight	7/30/2021
380	transformers: dark of the moon	6/29/2011
381	anger management	4/11/2003
382	the forest	1/8/2016
383	superfly	6/13/2018
384	book of dragons	11/15/2011
385	sister of the groom	12/18/2020
386	aliens in the attic	7/31/2009
387	oppenheimer	7/21/2023
388	the walking deceased	3/20/2015
389	max	6/26/2015
390	green zone	3/12/2010
391	falcon rising	9/5/2014
392	welcome home	11/16/2018
393	soul plane	5/28/2004
394	a jazzman's blues	9/16/2022
395	just like heaven	9/16/2005
396	escape from pretoria	3/6/2020
397	the crucifixion	10/6/2017
398	assassin 33 a.d.	1/24/2020
399	the ruins	4/4/2008

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
400	how high	12/21/2001
401	you're next	8/23/2013
402	the prince	8/22/2014
403	the bye bye man	1/13/2017
404	camp x-ray	10/17/2014
405	last holiday	1/13/2006
406	death sentence	8/31/2007
407	raymond & ray	10/14/2022
408	sgt. stubby: an american hero	4/13/2018
409	stay alive	3/24/2006
410	the forgotten	9/24/2004
411	osmosis jones	8/10/2001
412	wasabi	10/31/2001
413	siren	12/2/2016
414	primal	11/8/2019
415	three christs	1/10/2020
416	acrimony	3/30/2018
417	don't let go	8/30/2019
418	playing it cool	5/8/2015
419	dahmer	6/21/2002
420	2067	10/2/2020
421	the wolf of snow hollow	10/9/2020
422	swan song	12/17/2021
423	about fate	9/9/2022
424	see no evil	5/19/2006
425	stronger	9/22/2017
426	max 2: white house hero	5/5/2017
427	good mourning	5/20/2022
428	little manhattan	9/30/2005

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
429	addicted	10/10/2014
430	secretariat	10/8/2010
431	miss congeniality 2: armed and fabulous	3/24/2005
432	the grace card	2/25/2011
433	frozen	2/5/2010
434	wolves	11/14/2014
435	the reaping	4/5/2007
436	amulet	7/24/2020
437	detective knight: redemption	12/9/2022
438	rock dog	2/24/2017
439	dreamland	11/13/2020
440	6 below: miracle on the mountain	10/13/2017
441	beyond the reach	4/17/2015
442	midnight in the switchgrass	7/23/2021
443	turistas	12/1/2006
444	the whole truth	10/21/2016
445	american underdog	12/25/2021
446	traffik	4/20/2018
447	night train	1/13/2023
448	between worlds	12/21/2018
449	little children	10/6/2006
450	ida red	11/5/2021
451	dragonfly	2/22/2002
452	stolen	9/14/2012
453	dr. dolittle 2	6/22/2001
454	bloodthirsty	4/23/2021
455	the sky is everywhere	2/11/2022
456	stuck in love	7/5/2013
457	john and the hole	8/6/2021

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
458	hands of stone	8/26/2016
459	the poison rose	5/24/2019
460	the pyramid	12/5/2014
461	the flyboys	8/15/2008
462	backtrace	12/14/2018
463	the wedding date	2/4/2005
464	the double	10/28/2011
465	teen titans: trouble in tokyo	9/15/2006
466	kill switch	6/16/2017
467	one day as a lion	4/4/2023
468	the virginity hit	9/24/2010
469	rugrats go wild	6/13/2003
470	future world	5/25/2018
471	the bubble	4/1/2022
472	the novice	12/17/2021
473	beasts of no nation	10/16/2015
474	standoff	2/26/2016
475	once upon a time in venice	6/16/2017
476	the equalizer 3	9/1/2023
477	the last song	3/31/2010
478	nanny	11/23/2022
479	elephant	10/24/2003
480	shut in	11/11/2016
481	slugterra: return of the elementals	8/2/2014
482	bad company	6/7/2002
483	the abcs of death	3/8/2013
484	extraction	12/18/2015
485	the gift	1/19/2001
486	vengeance: a love story	9/15/2017

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
487	line of duty	11/15/2019
488	tom and jerry blast off to mars!	1/18/2005
489	211	6/8/2018
490	the con is on	5/4/2018
491	made of honor	5/2/2008
492	the choice	2/5/2016
493	the turning	1/24/2020
494	gone	2/24/2012
495	the devil inside	1/6/2012
496	bitch slap	1/8/2010
497	i.t.	9/23/2016
498	the first time	10/19/2012
499	hey arnold! the movie	6/28/2002
500	pay it forward	10/20/2000
501	jane got a gun	1/29/2016
502	every day	2/23/2018
503	nine lives	8/5/2016
504	rugrats in paris: the movie	11/17/2000
505	cooties	9/18/2015
506	the kings of summer	5/31/2013
507	nightride	3/4/2022
508	lullaby	12/16/2022
509	the last man	1/18/2019
510	southern gospel	3/10/2023
511	everyone's hero	9/15/2006
512	killing season	7/12/2013
513	the covenant	9/8/2006
514	miss bala	2/1/2019
515	nancy drew and the hidden staircase	3/15/2019

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
516	our house	7/27/2018
517	final score	9/14/2018
518	lakeview terrace	9/19/2008
519	acts of vengeance	10/27/2017
520	reprisal	8/31/2018
521	sweetwater	4/14/2023
522	honk for jesus. save your soul.	9/2/2022
523	think like a man too	6/20/2014
524	serenity	1/25/2019
525	the darkest hour	12/25/2011
526	fright night	8/19/2011
527	polaroid	10/11/2019
528	patients of a saint	1/3/2020
529	bone tomahawk	10/23/2015
530	in the blood	4/4/2014
531	the love witch	11/11/2016
532	life of crime	8/29/2014
533	iboy	1/27/2017
534	dope	6/19/2015
535	novitiate	10/27/2017
536	dance flick	5/22/2009
537	cantinflas	8/29/2014
538	alone in the dark	1/28/2005
539	white girl	9/2/2016
540	awake	11/30/2007
541	white boy rick	9/14/2018
542	radio	10/24/2003
543	the shaggy dog	3/10/2006
544	miss march	3/13/2009

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
545	thank you for your service	10/27/2017
546	mainstream	5/7/2021
547	savage dog	8/4/2017
548	the greatest game ever played	9/30/2005
549	the signal	6/13/2014
550	the hole in the ground	3/1/2019
551	confidence	4/25/2003
552	ted bundy: american boogeyman	8/16/2021
553	the messengers	2/2/2007
554	the trial of the chicago 7	10/16/2020
555	high strung	4/8/2016
556	don't knock twice	2/3/2017
557	what's your number?	9/30/2011
558	vicky cristina barcelona	8/15/2008
559	the vanished	8/21/2020
560	the flintstones in viva rock vegas	4/28/2000
561	24 hours to live	12/1/2017
562	street fighter: the legend of chun-li	2/27/2009
563	take me home tonight	3/4/2011
564	rust creek	1/4/2019
565	the promise	4/21/2017
566	vengeance	7/29/2022
567	just my luck	5/12/2006
568	life after beth	8/15/2014
569	2:22	6/30/2017
570	frontera	9/5/2014
571	bad samaritan	5/4/2018
572	martyrs	1/22/2016
573	lying and stealing	7/12/2019

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
574	black snake moan	3/2/2007
575	the silencing	8/14/2020
576	half brothers	12/4/2020
577	wanderlust	2/24/2012
578	the life of david gale	2/21/2003
579	pet	12/2/2016
580	life itself	9/21/2018
581	valentine	2/2/2001
582	saints and soldiers: airborne creed	8/17/2012
583	kate & leopold	12/25/2001
584	london fields	10/26/2018
585	navy seals vs. zombies	10/8/2015
586	all the devil's men	12/7/2018
587	rock star	9/7/2001
588	the stepfather	10/16/2009
589	old henry	10/1/2021
590	the ottoman lieutenant	3/10/2017
591	war of the worlds 2: the next wave	3/18/2008
592	ghost house	8/25/2017
593	enter the void	9/24/2010
594	confessions of a teenage drama queen	2/17/2004
595	slumber	12/1/2017
596	hart's war	2/15/2002
597	kickboxer: vengeance	9/2/2016
598	a dark place	4/12/2019
599	the black dahlia	9/15/2006
600	wrong place	7/15/2022
601	last looks	2/4/2022
602	trust	3/12/2021

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
603	saving silverman	2/9/2001
604	malum	3/31/2023
605	material girls	8/18/2006
606	boogeyman 2	10/20/2007
607	silent night	12/3/2021
608	clouds of sils maria	4/10/2015
609	the sisterhood of the traveling pants	6/1/2005
610	recess: school's out	2/16/2001
611	nazi overlord	11/13/2018
612	four good days	4/30/2021
613	sharp stick	7/29/2022
614	victor crowley	8/22/2017
615	a merry friggin' christmas	11/7/2014
616	black knight	11/21/2001
617	just wright	5/14/2010
618	careful what you wish for	6/10/2016
619	crossing over	2/27/2009
620	the possession of michael king	8/22/2014
621	the perfect game	4/16/2010
622	thunderstruck	8/24/2012
623	the sisterhood of the traveling pants 2	8/6/2008
624	mom and dad	1/19/2018
625	silk road	2/19/2021
626	the hitcher	1/19/2007
627	dream house	9/30/2011
628	duplex	9/26/2003
629	12 rounds 3: lockdown	9/11/2015
630	sorority row	9/11/2009
631	nick and norah's infinite playlist	10/3/2008

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
632	cold creek manor	9/19/2003
633	the hours	12/27/2002
634	the recall	6/2/2017
635	zoom	8/11/2006
636	the gallows	7/10/2015
637	skinwalkers	8/10/2007
638	proud mary	1/12/2018
639	the town that dreaded sundown	10/16/2014
640	awakening the zodiac	6/9/2017
641	viral	7/29/2016
642	my life in ruins	6/5/2009
643	don't tell a soul	1/15/2021
644	ramona and beezus	7/23/2010
645	crypto	4/12/2019
646	the roommate	2/4/2011
647	bent	3/9/2018
648	obsessed	4/24/2009
649	kill the messenger	10/10/2014
650	alfie	11/5/2004
651	ted k	2/18/2022
652	precious cargo	4/22/2016
653	the monster	11/11/2016
654	the human stain	10/31/2003
655	trauma center	12/6/2019
656	blood and chocolate	1/26/2007
657	the convent	5/3/2019
658	are we done yet?	4/4/2007
659	closed circuit	8/28/2013
660	waking life	10/19/2001

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
661	nefarious	4/14/2023
662	palo alto	5/9/2014
663	edge of darkness	1/29/2010
664	ultraviolet	3/3/2006
665	flypaper	8/19/2011
666	the trust	5/13/2016
667	hatchet ii	10/1/2010
668	reasonable doubt	1/17/2014
669	the painted veil	12/20/2006
670	reno 911!: miami	2/23/2007
671	the rocker	8/20/2008
672	elizabethtown	10/14/2005
673	keep watching	10/31/2017
674	the best of enemies	4/5/2019
675	the final cut	10/15/2004
676	first kill	7/21/2017
677	soul men	11/7/2008
678	machuca	2/24/2004
679	emergency	5/20/2022
680	the glass house	9/14/2001
681	legends of oz: dorothy's return	5/9/2014
682	last night	5/6/2011
683	welcome home roscoe jenkins	2/8/2008
684	pay the ghost	9/25/2015
685	jessabelle	11/7/2014
686	the jacket	3/4/2005
687	rules don't apply	11/23/2016
688	the art of getting by	6/17/2011
689	the girl who believes in miracles	4/2/2021

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
690	spinning gold	3/31/2023
691	criminal activities	11/20/2015
692	be cool	3/4/2005
693	the homesman	11/14/2014
694	autumn in new york	8/11/2000
695	survive the night	5/22/2020
696	the siege	3/10/2023
697	the stanford prison experiment	7/17/2015
698	the prodigy	2/8/2019
699	bloody hell	1/14/2021
700	the survivalist	10/1/2021
701	irrational man	7/17/2015
702	afternoon delight	8/30/2013
703	the siege of jadotville	10/7/2016
704	fighting	4/24/2009
705	in the bedroom	11/23/2001
706	out of the dark	2/27/2015
707	god's not dead: we the people	10/4/2021
708	birth	10/29/2004
709	the lords of salem	4/19/2013
710	the new guy	5/10/2002
711	mark felt: the man who brought down the white house	9/29/2017
712	kicks	9/9/2016
713	spree	8/14/2020
714	come and find me	11/11/2016
715	mary	10/11/2019
716	the starling	9/17/2021
717	atl	3/31/2006
718	trash	10/9/2015

Appendix A: Movie Titles and Release Dates		
	title	**release_date**
Index		
719	the meyerowitz stories (new and selected)	10/13/2017
720	black butterfly	5/26/2017
721	superman vs. the elite	6/12/2012
722	lords of dogtown	6/3/2005
723	parkland	10/4/2013
724	mara	9/7/2018
725	hell baby	9/6/2013
726	incarnate	12/2/2016
727	swallow	3/6/2020
728	a thousand and one	3/31/2023
729	driven	8/16/2019
730	agent game	4/8/2022
731	the final wish	1/24/2019
732	date and switch	2/14/2014
733	words on bathroom walls	8/21/2020
734	southbound	2/5/2016
735	maniac	6/21/2013
736	captain corelli's mandolin	8/17/2001
737	just getting started	12/8/2017
738	cabin fever	2/12/2016
739	norm of the north	1/15/2016
740	harsh times	11/10/2006
741	acts of violence	1/12/2018
742	hustle & flow	7/22/2005
743	hardball	9/14/2001
744	below	10/11/2002
745	payback	2/5/2021

A2) Sequel Movies

Title of sequels	release_date
detective knight: independence	1/20/2023
fast x	5/19/2023
glass onion: a knives out mystery	11/23/2022
detective knight: rogue	10/21/2022
harry potter and the half-blood prince	7/15/2009
transformers: rise of the beasts	6/9/2023
spider-man: across the spider-verse	6/2/2023
harry potter and the order of the phoenix	7/11/2007
puss in boots	10/28/2011
the amazing spider-man	7/3/2012
god's not dead: a light in darkness	3/30/2018
black water: abyss	8/7/2020
the hangover part iii	5/23/2013
unbroken: path to redemption	9/14/2018
the flash	6/16/2023
the nun 2	9/8/2023
no manches frida 2	3/15/2019
a haunted house 2	4/18/2014
a haunted house	1/11/2013
annabelle	10/3/2014
god's not dead 2	4/1/2016
god's not dead	3/21/2014
jeepers creepers 3	9/26/2017
indiana jones and the kingdom of the crystal skull	5/22/2008
indiana jones and the dial of destiny	6/30/2023
g.i. joe: retaliation	3/28/2013
insidious: the last key	1/5/2018
war of the worlds	6/29/2005
ouija	10/24/2014
escape plan 2: hades	6/29/2018
son of the mask	2/18/2005
ouija: origin of evil	10/21/2016
dominion: prequel to the exorcist	5/20/2005
leatherface	10/20/2017
spider-man 2	6/30/2004
transformers: revenge of the fallen	6/24/2009
the last exorcism part ii	3/1/2013
batman beyond: return of the joker	12/12/2000
jason x	4/26/2002
piranha 3dd	6/1/2012

the strangers	5/30/2008
teen titans go! to the movies	7/27/2018
the expendables 4	9/22/2023
the haunting in connecticut	3/27/2009
meg 2: the trench	8/4/2023
diary of a wimpy kid: the long haul	5/19/2017
12 hour shift	10/2/2020
the strangers: prey at night	3/9/2018
punisher: war zone	12/5/2008
teenage mutant ninja turtles: mutant mayhem	8/4/2023
sinister 2	8/21/2015
the haunting in connecticut 2: ghosts of georgia	2/1/2013
paranormal activity 4	10/18/2012
book of shadows: blair witch 2	10/27/2000
the fog	10/14/2005
the butterfly effect 3: revelations	1/9/2009
dumb and dumberer: when harry met lloyd	6/13/2003
kickboxer: retaliation	1/26/2018
universal soldier: day of reckoning	11/30/2012
day of the dead: bloodline	1/5/2018
halloween ii	8/28/2009
the grudge 2	10/13/2006
transformers: dark of the moon	6/29/2011
how high	12/21/2001
siren	12/2/2016
max 2: white house hero	5/5/2017
miss congeniality 2: armed and fabulous	3/24/2005
detective knight: redemption	12/9/2022
dr. dolittle 2	6/22/2001
teen titans: trouble in tokyo	9/15/2006
rugrats go wild	6/13/2003
the equalizer 3	9/1/2023
slugterra: return of the elementals	8/2/2014
tom and jerry blast off to mars!	1/18/2005
hey arnold! the movie	6/28/2002
rugrats in paris: the movie	11/17/2000
think like a man too	6/20/2014
the flintstones in viva rock vegas	4/28/2000
war of the worlds 2: the next wave	3/18/2008
kickboxer: vengeance	9/2/2016
boogeyman 2	10/20/2007
the sisterhood of the traveling pants	6/1/2005
the sisterhood of the traveling pants 2	8/6/2008
12 rounds 3: lockdown	9/11/2015
hatchet ii	10/1/2010
reno 911!: miami	2/23/2007
god's not dead: we the people	10/4/2021
superman vs. the elite	6/12/2012
cabin fever	2/12/2016

A3) Star Actors

Star Actors	Rank	Star Actors	Rank	Star Actors	Rank	Star Actors	Rank	Star Actors	Rank	Star Actors	Rank
Bruce Willis	1	Tom Holland	26	Grace Kelly	51	Rooney Mara	76	Rachel McAdams	101	Will Ferrell	126
Nicolas Cage	2	Michael B. Jordan	27	Sam Worthington	52	Lupita Nyong'o	77	John Krasinski	102	John C. Reilly	127
Johnny Depp	3	Zoe Saldana	28	Sam Rockwell	53	Naomi Watts	78	Emily Blunt	103	Owen Wilson	128
Tom Cruise	4	Salma Hayek	29	Paul Rudd	54	Shailene Woodley	79	Mark Ruffalo	104	Vince Vaughn	129
Leonardo DiCaprio	5	Kate Winslet	30	Paul Newman	55	Javier Bardem	80	Helen Mirren	105	Mel Gibson	130
Scarlett Johansson	6	Cate Blanchett	31	Eric Bana	56	Penélope Cruz	81	Kirsten Dunst	106	Kurt Russell	131
Will Smith	7	Samuel L. Jackson	32	Billy Bob Thornton	57	Charlize Theron	82	Jodie Foster	107	Patrick Stewart	132
Robert Downey Jr.	8	Hugh Jackman	33	Roy Scheider	58	Edward Norton	83	Diane Keaton	108	Ian McKellen	133
Chris Pratt	9	Ryan Gosling	34	Danai Gurira	59	Jeff Bridges	84	Glenn Close	109	Hugo Weaving	134
Angelina Jolie	10	Anne Hathaway	35	Jake Gyllenhaal	60	Liam Neeson	85	Marisa Tomei	110	Josh Brolin	135
Brad Pitt	11	Timothée Chalamet	36	Viola Davis	61	Robin Williams	86	Keira Knightley	111	Sean Penn	136
Matt Damon	12	Amy Adams	37	Harrison Ford	62	Rebel Wilson	87	Julianne Moore	112	Andy Serkis	137
Julia Roberts	13	Jessica Chastain	38	Mary Elizabeth Winstead	63	Anya Taylor-Joy	88	Saoirse Ronan	113	Eddie Redmayne	138
Ryan Reynolds	14	Daniel Craig	39	Sarah Jessica Parker	64	Oscar Isaac	89	Emma Thompson	114	Ralph Fiennes	139
Meryl Streep	15	Keanu Reeves	40	Sarah Paulson	65	Benedict Cumberbatch	90	Tilda Swinton	115	Matthew McConaughey	140
George Clooney	16	Sandra Bullock	41	Emma Watson	66	Taron Egerton	91	Dustin Hoffman	116	Halle Berry	141
Adam Sandler	17	Reese Witherspoon	42	Frank Sinatra	67	Viggo Mortensen	92	Al Pacino	117	Vivica A. Fox	142
Jennifer Lawrence	18	Channing Tatum	43	Bradley Cooper	68	James McAvoy	93	Robert De Niro	118	Anjelica Huston	143
Margot Robbie	19	Michael Fassbender	44	Daniel Radcliffe	69	Ezra Miller	94	Colin Farrell	119	Winona Ryder	144
Ben Affleck	20	Andrew Garfield	45	Henry Cavill	70	Dakota Fanning	95	Russell Crowe	120	Tim Roth	145
Chris Hemsworth	21	Florence Pugh	46	Idris Elba	71	Kristen Stewart	96	Jean Reno	121	Forest Whitaker	146
Dwayne Johnson (The Rock)	22	Elizabeth Olsen	47	Jason Momoa	72	Maggie Gyllenhaal	97	Morgan Freeman	122	Adrien Brody	147
Gal Gadot	23	Alexander Skarsgård	48	Millie Bobby Brown	73	Jeremy Renner	98	Denzel Washington	123	Jason Statham	148
Emma Stone	24	Peter Dinklage	49	Chris Evans	74	Sebastian Stan	99	Wesley Snipes	124	Catherine Zeta-Jones	149
Zendaya	25	Jack Nicholson	50	Jennifer Aniston	75	Donald Glover	100	Antonio Banderas	125	Kevin Costner	150

A4) Correlation Analysis (pre & post imputation)

```
Correlations with 'tomato_missing':
score                  -0.259818
tomato                       NaN
drama                  -0.060661
horror                  0.026291
action                  0.098987
comedy                 -0.086959
thriller                0.093809
sci-fi                  0.086448
adventure              -0.066544
is_sequel               0.040947
month                  -0.000887
summer                 -0.017803
christmas               0.008251
new_years_eve           0.035413
Thanksgiving_season    -0.038210
Easter                 -0.043168
Halloween               0.001392
Valentine's            -0.021941
Holiday                -0.031999
star_actor             -0.039792
tomato_missing          1.000000
Name: tomato_missing, dtype: float64
```

Table 1) Correlation Analysis

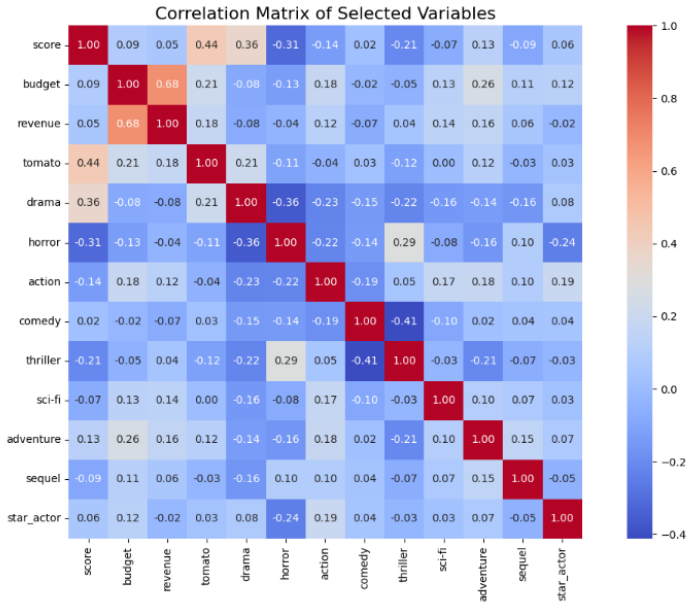


Figure 1) Correlation Matrix after Imputation

B) Data Description

Figure1

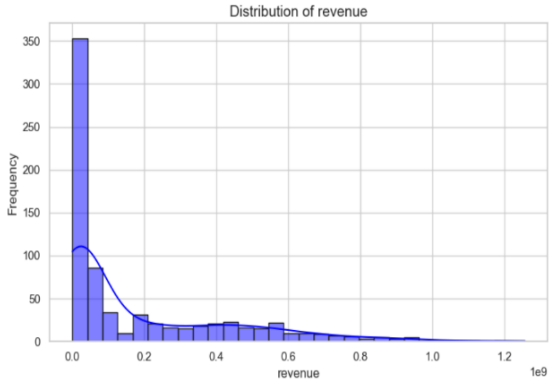


Figure 2

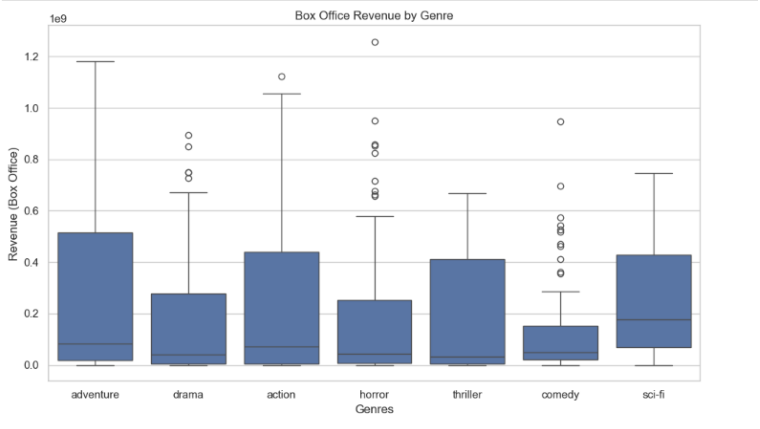


Figure 3

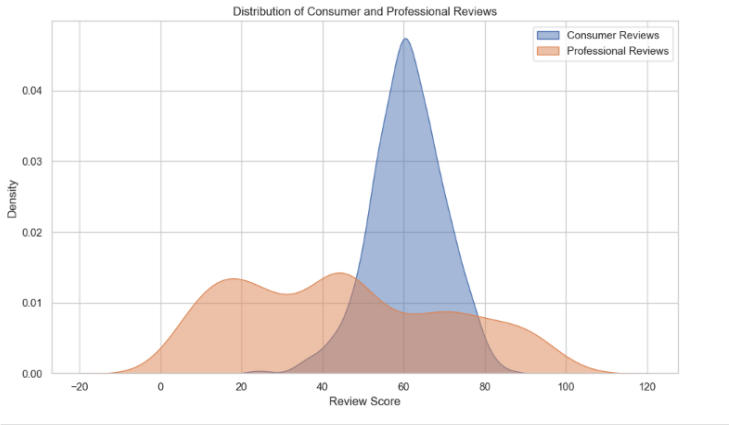


Figure 4

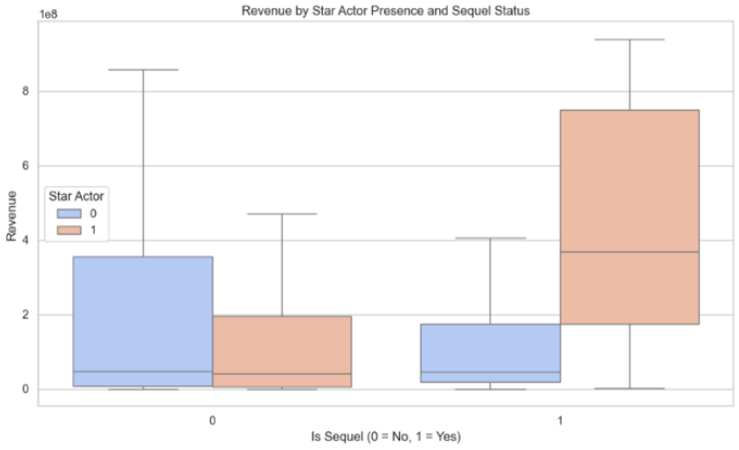


Figure 5

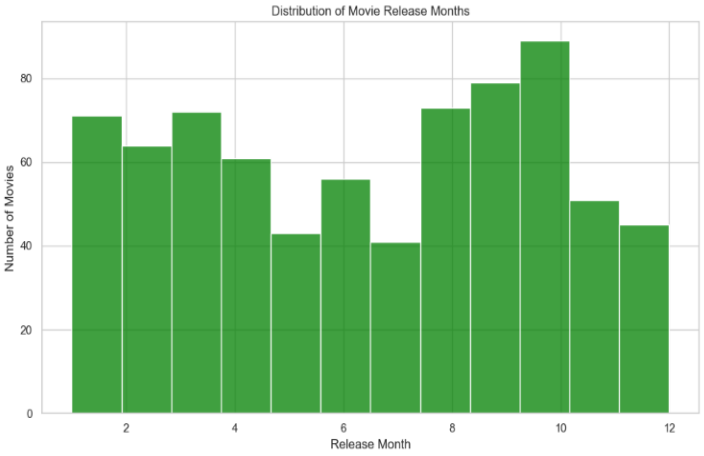


Figure 6

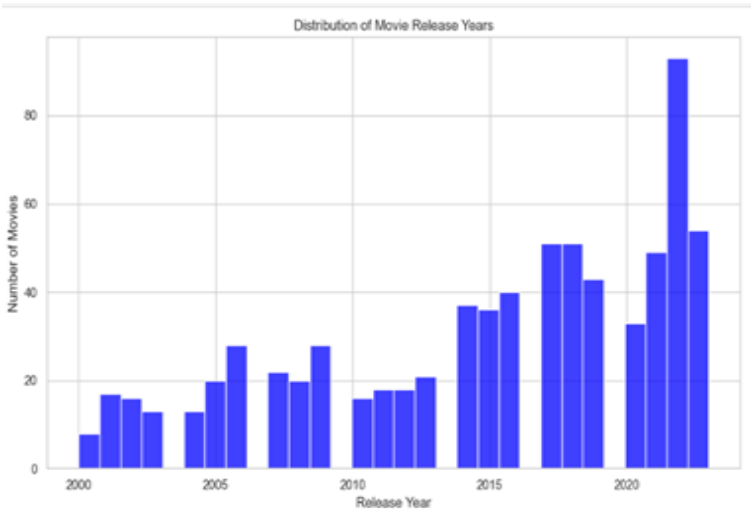


Figure 7

Figure 8

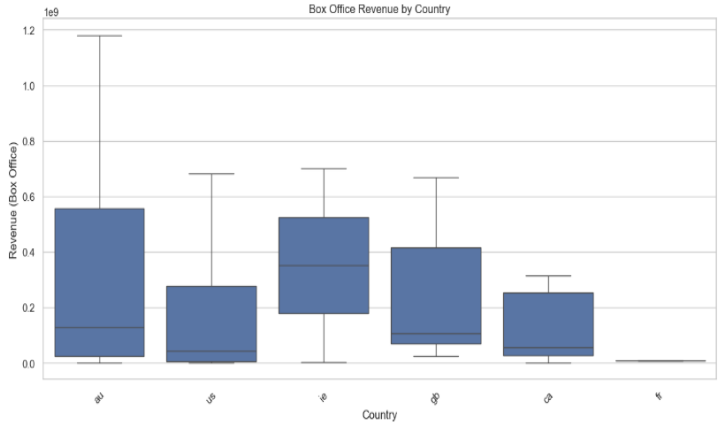


Figure 9

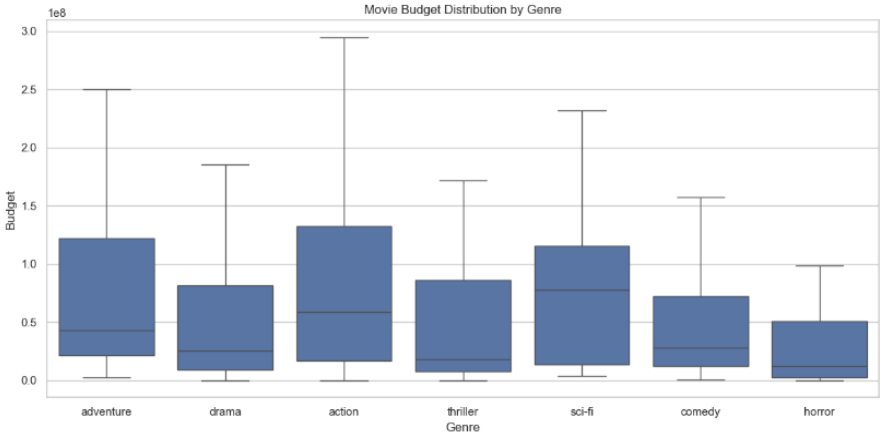


Figure 10

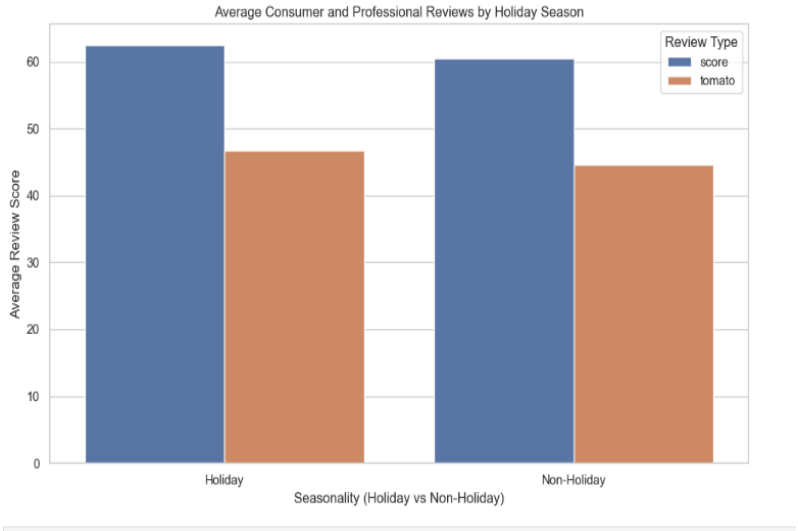


Figure 11

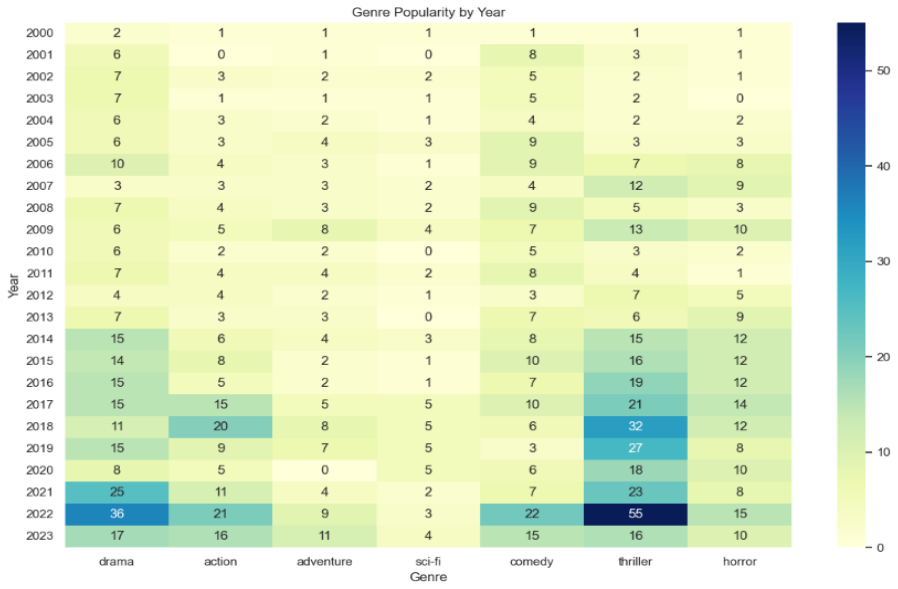


Figure 12

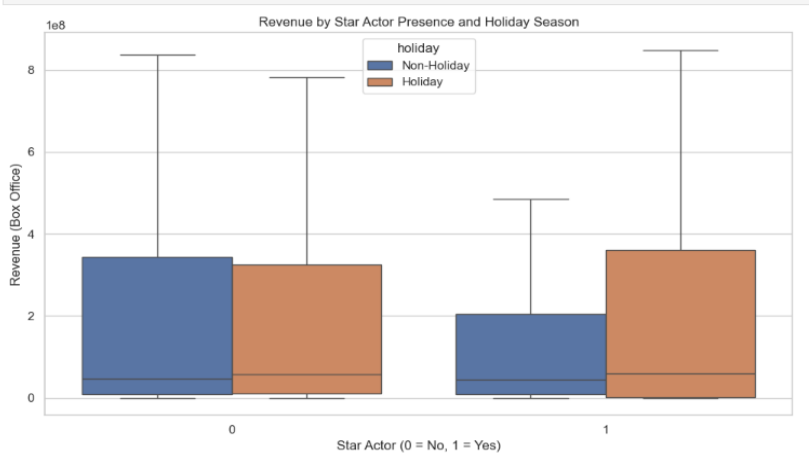


Figure 13

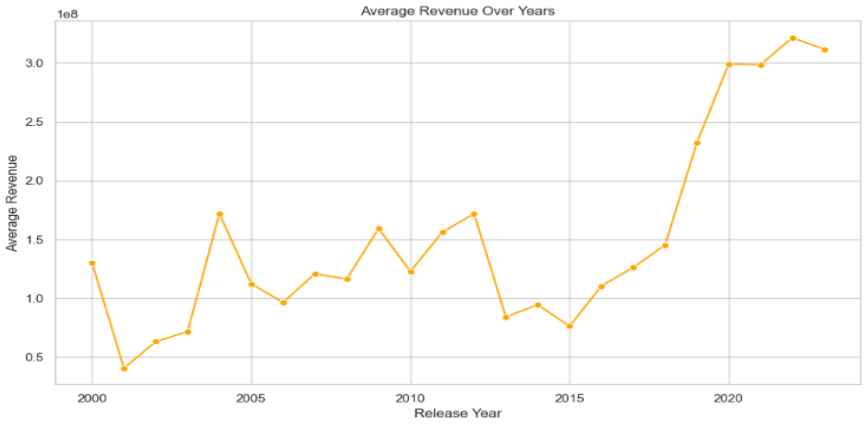


Figure 14

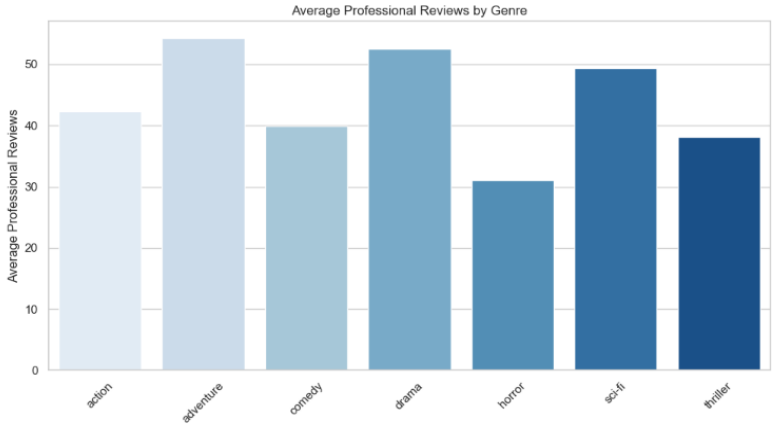


Figure 15

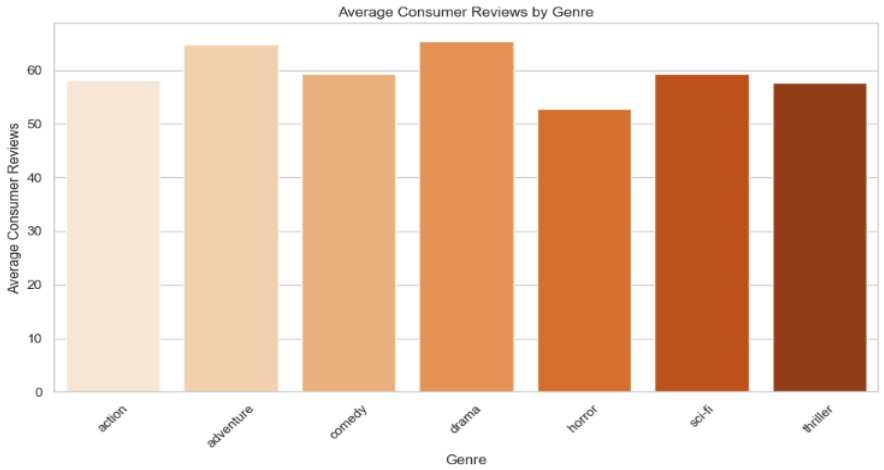


Figure 16

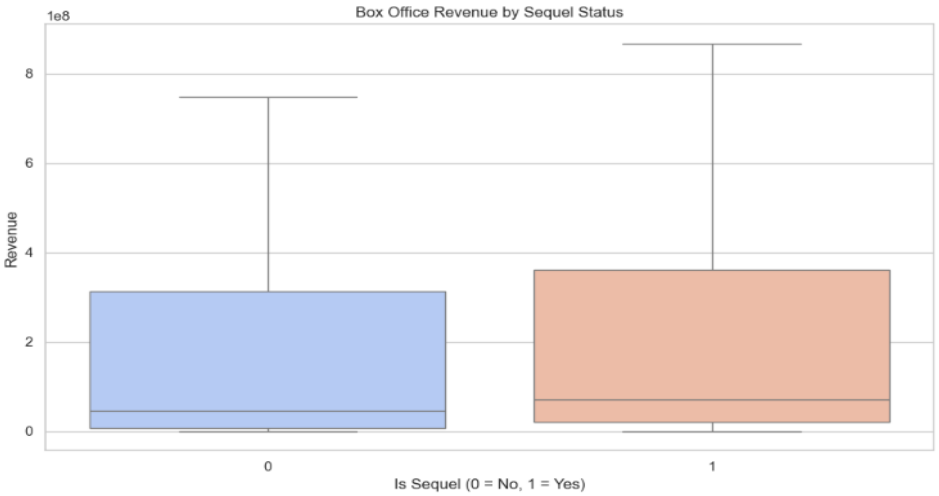


Figure 17

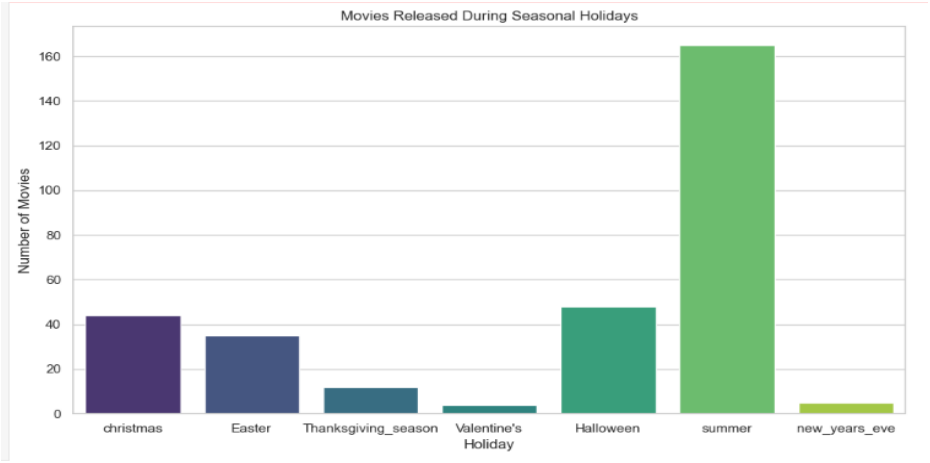


Figure 18

Descriptive Statistics:

	Variable	Mean	Std Dev	Min	Max
0	revenue	1.820426e+08	2.474920e+08	43.0	1.256888e+09
1	budget_x	5.625318e+07	5.911048e+07	85645.0	3.400000e+08
2	tomato	4.495570e+01	2.594342e+01	1.0	1.000000e+02
3	score	6.088591e+01	8.976171e+00	23.0	8.700000e+01
4	star_actor	3.073826e-01	NaN	0.0	1.000000e+00
5	sequel	1.194631e-01	NaN	0.0	1.000000e+00
6	Holiday	4.134228e-01	NaN	0.0	1.000000e+00

Number of unique genres (genre): 179

Table: Output from Python

C) Assumption Checks

Heteroscedasticity

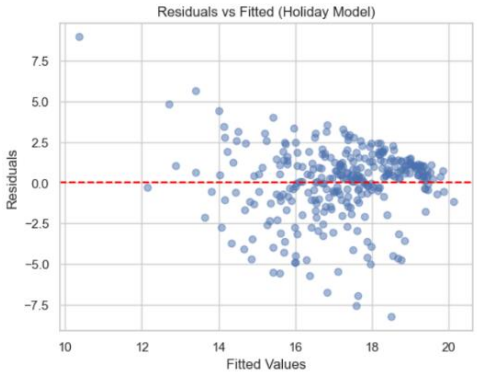


Table 1 Residual plot Holiday Model

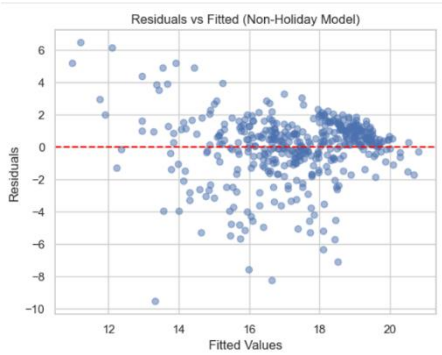


Table 2 Residual plot Non-Holiday Model

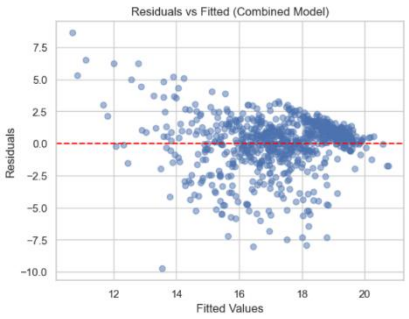


Table 3 Residual plot Combined Model

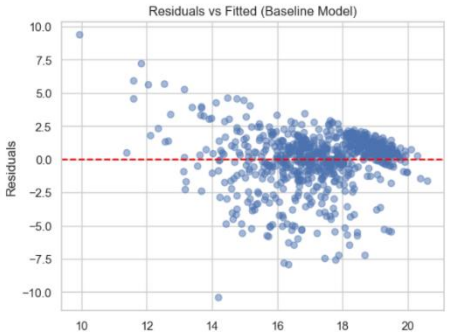


Table 4 Residual plot Baseline Model

Multicollinearity

VIF for Holiday Model:

	Variable	VIF
0	log_budget	inf
1	star_actor	inf
2	tomato	inf
3	score	inf
4	holiday_budget	inf
5	holiday_star_actor	inf
6	holiday_score	inf
7	holiday_tomato	inf
8	action	1.483335
9	adventure	1.337155
10	sci-fi	1.144403
11	drama	1.933447
12	comedy	1.605830
13	thriller	1.496159
14	horror	1.625684
15	sequel	1.100771

Table 5 VIF values Holiday Model

VIF for Non-Holiday Model:

	Variable	VIF
0	log_budget	1.184346
1	star_actor	1.144746
2	tomato	1.274533
3	score	1.488294
4	action	1.430132
5	adventure	1.217711
6	sci-fi	1.096593
7	drama	1.823636
8	comedy	1.576834
9	thriller	1.586395
10	horror	1.765190
11	sequel	1.077298

Table 6 VIF values Non-Holiday Model

VIF for Baseline Model:

	Variable	VIF
0	log_budget	1.120035
1	star_actor	1.105096
2	tomato	1.284312
3	score	1.510867
4	action	1.433609
5	adventure	1.234688
6	sci-fi	1.103473
7	drama	1.834887
8	comedy	1.575437
9	thriller	1.519038
10	horror	1.690159
11	sequel	1.074288

Table 7 VIF values Baseline Model

VIF for Combined Model:

	Variable	VIF
0	log_budget	2.098390
1	star_actor	1.861333
2	tomato	2.199800
3	score	2.502558
4	Holiday	167.877457
5	holiday_budget	131.290113
6	holiday_star_actor	2.212497
7	holiday_score	61.278680
8	holiday_tomato	6.129288
9	action	1.441649
10	adventure	1.248435
11	sci-fi	1.110156
12	drama	1.860032
13	comedy	1.589103
14	thriller	1.544939
15	horror	1.698915
16	sequel	1.081655

Table 8 VIF values Combined Model

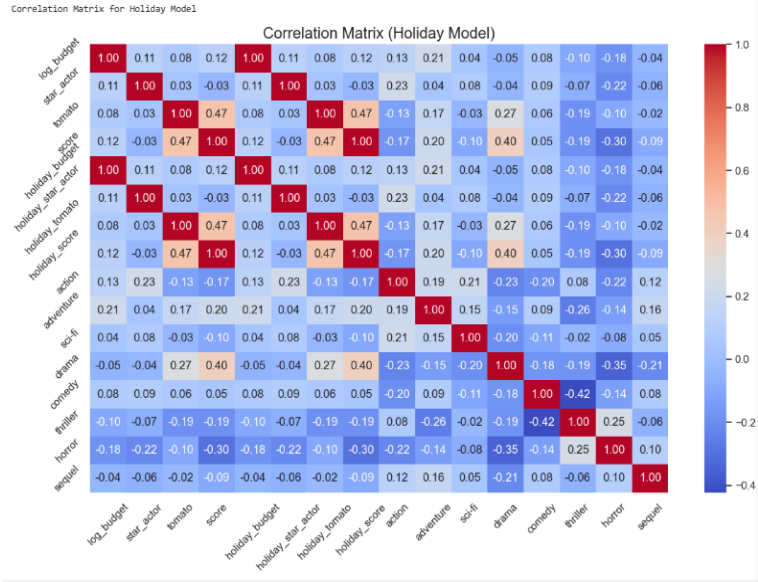


Table 10 Correlation Matrix Holiday Model

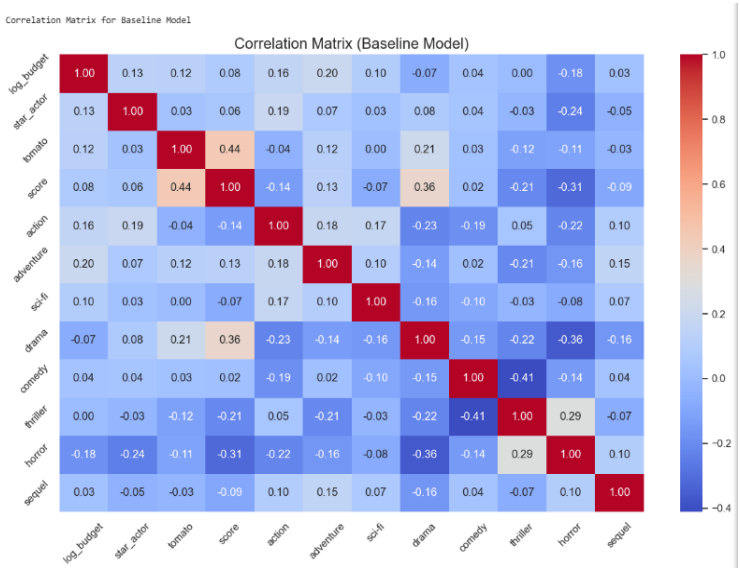


Table 11 Correlation Matrix Baseline Model

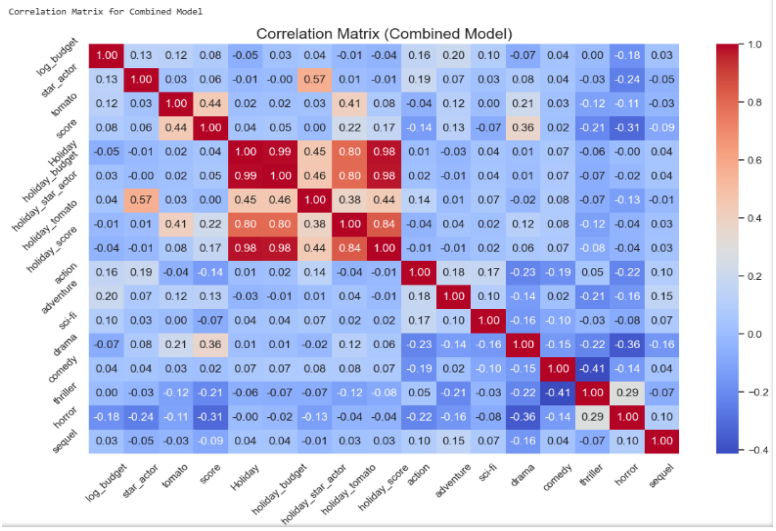


Table 12 Correlation Matrix Combined Model

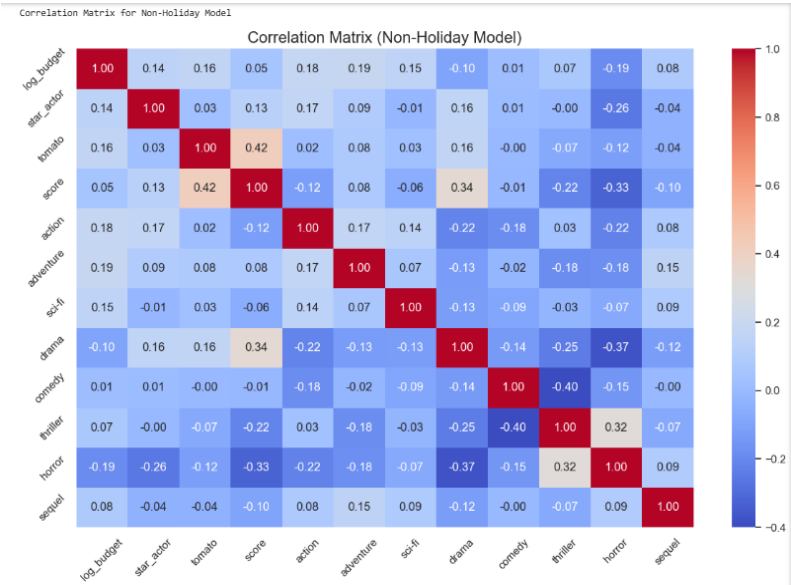


Table 9 Correlation Matrix Non – Holiday Model

Non- Normality

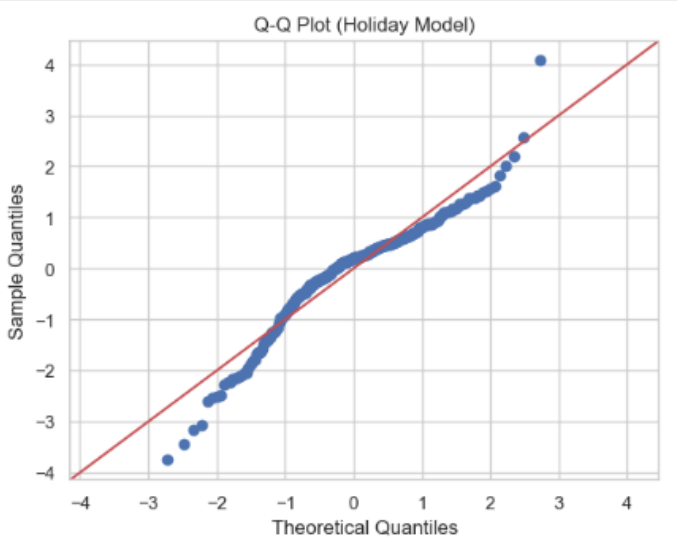


Table 13 QQ Plot Holiday Model

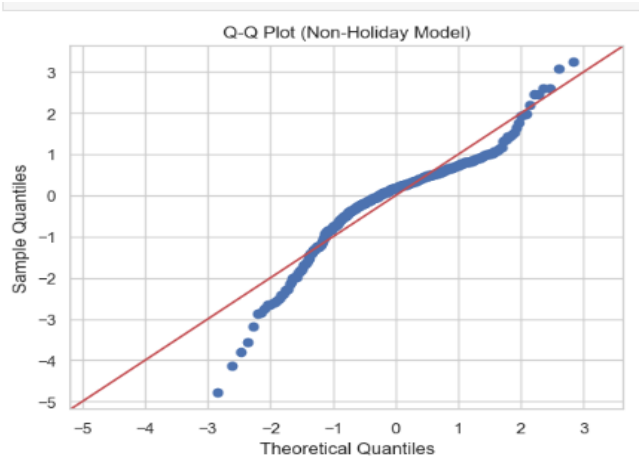


Table 14 QQ Plot Non-Holiday Model

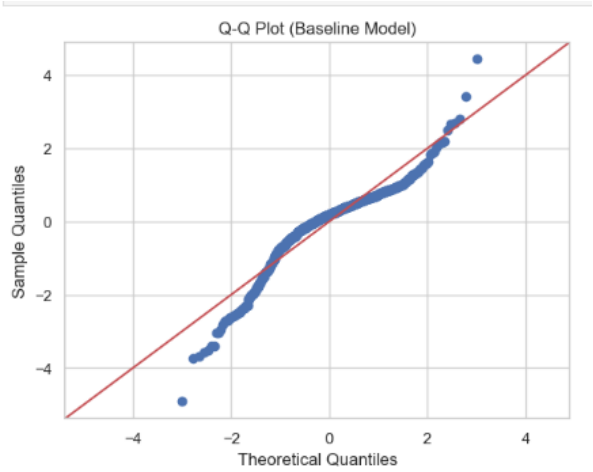


Table 15 QQ Plot Baseline Model

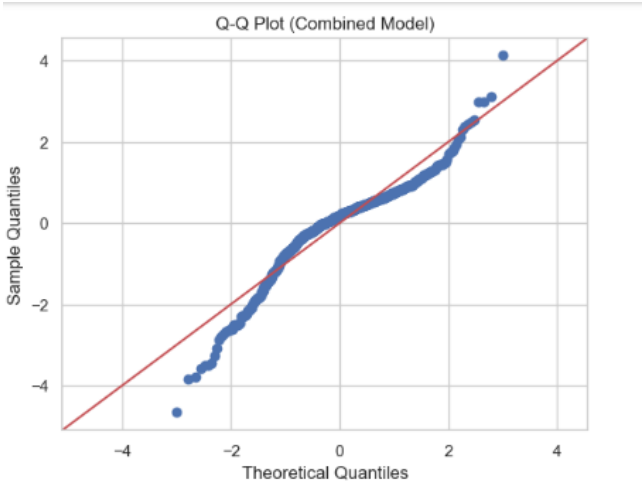


Table 16 QQ Plot Combined Model

Linearity

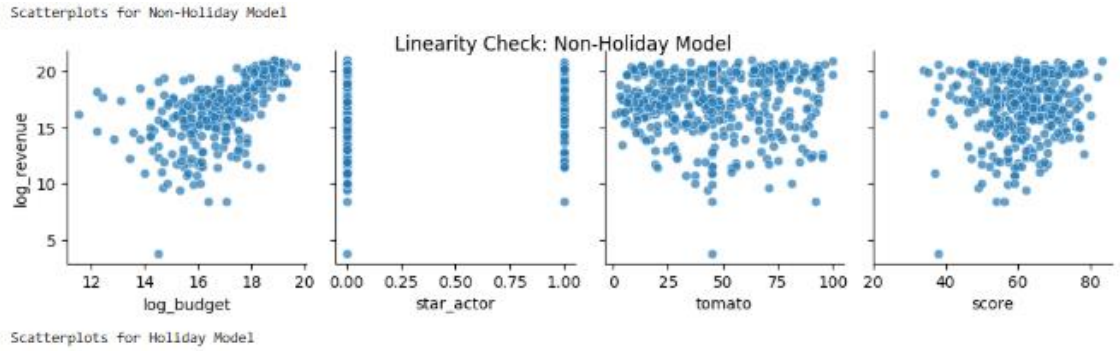


Table 17 Scatterplots Non-Holiday Model

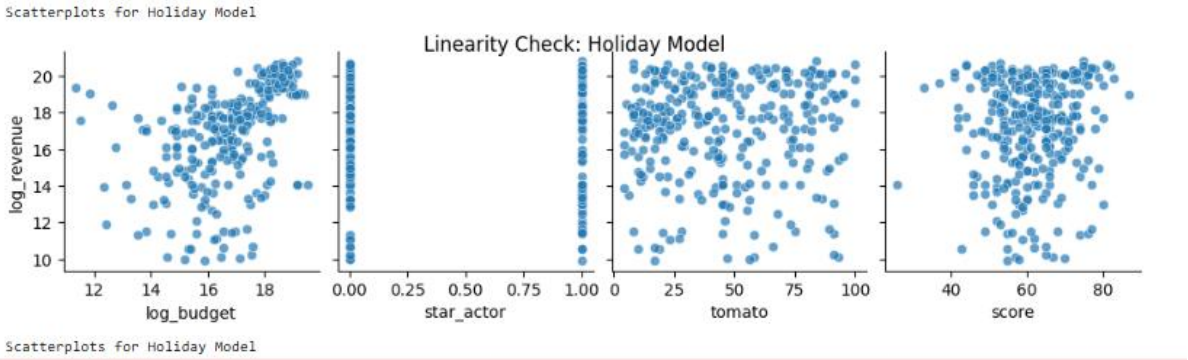
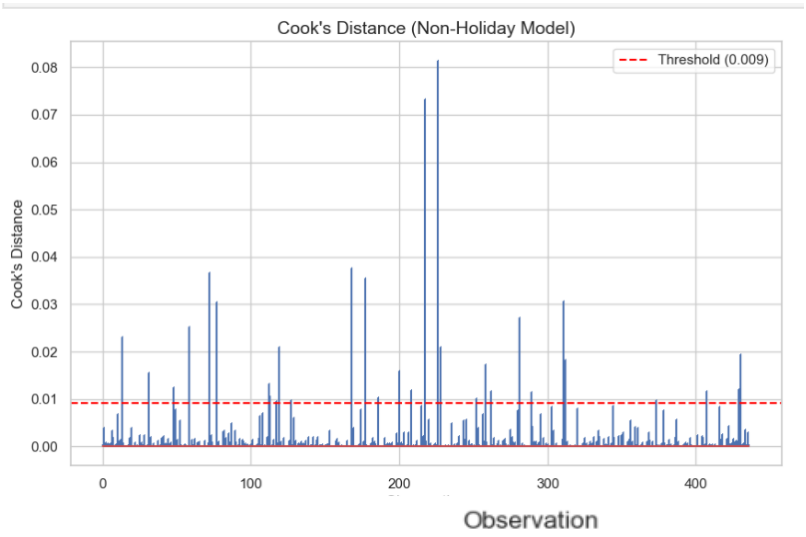


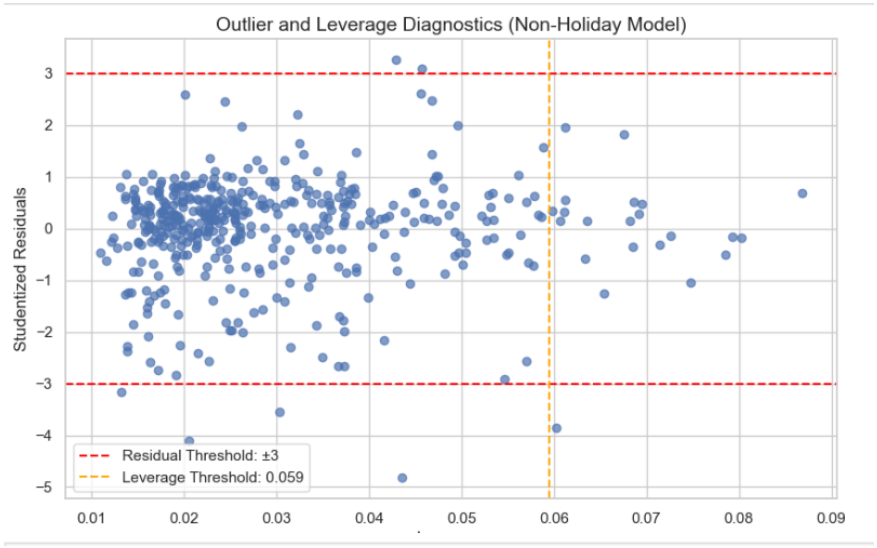
Table 18 Scatterplots Holiday Model

Influential outlier detection




```
High influence points: [ 13  31  48  58  72  77 112 113 117 119 127 168 177 186 200 208 217 226
 228 252 258 262 281 289 311 312 373 407 429 430]
```

Table 19 Cook's D chart -Non Holiday Model




```
High influence points: [  5   8  23  42  45  95 117 137 169 194 205 217 233 265 272 291 295 296]
```

Table 20 Outlier Leverage plot -Non Holiday Model

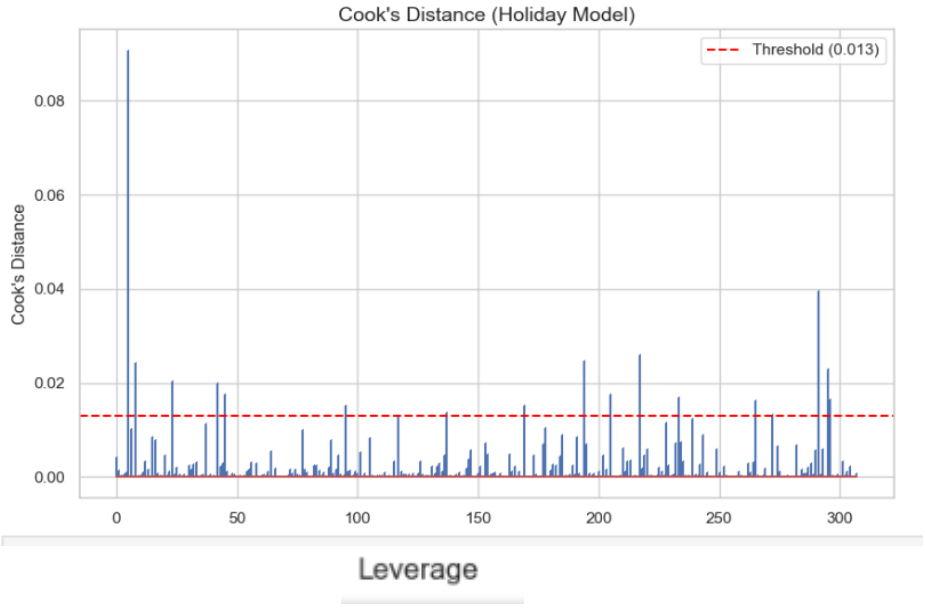


Table 21 Cook's D' chart - Holiday Model

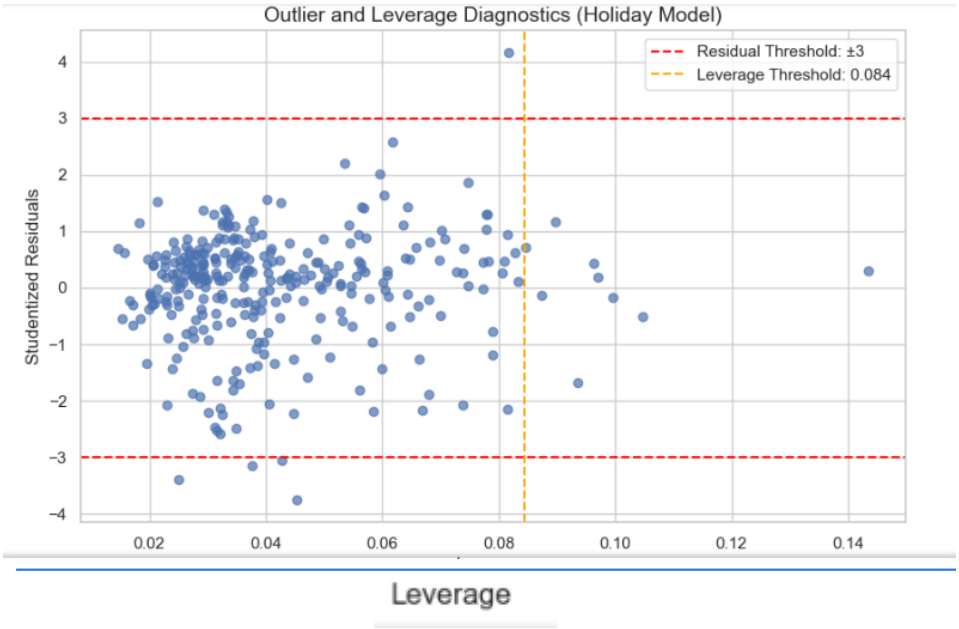
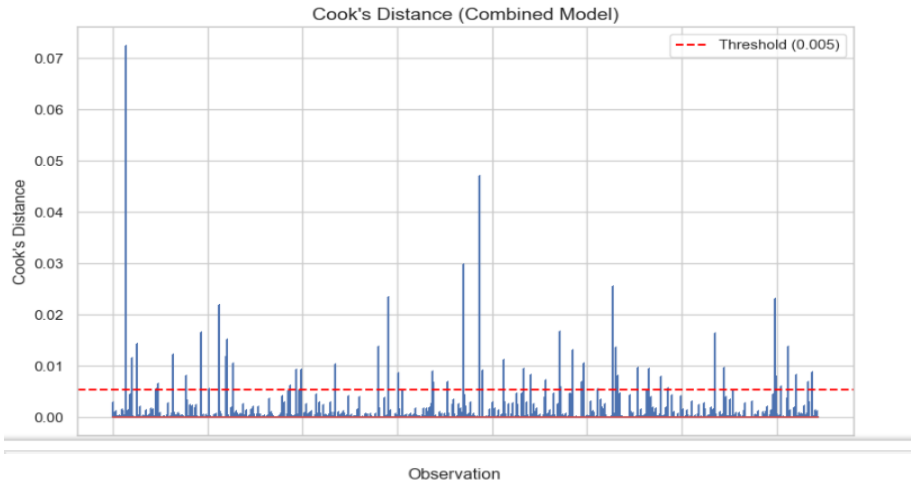


Table 22 Outlier Leverage Plot Holiday Model




```
High influence points: [ 14  20  26  46  48  64  77  93 102 112 119 121 127 187 193 197 199 235
 280 291 301 337 338 353 370 387 390 412 433 441 456 471 472 485 494 497
 511 527 530 532 553 565 578 585 635 644 698 699 704 712 720 733 737]
```

Table 23 Cook's D chart -Combined Model

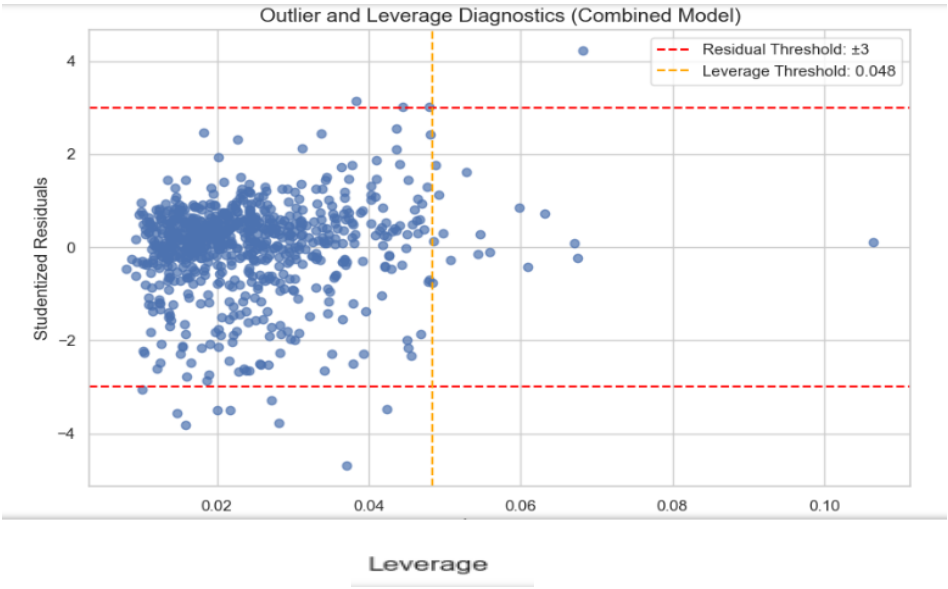


Table 24 Outlier Leverage Plott -Combined Model

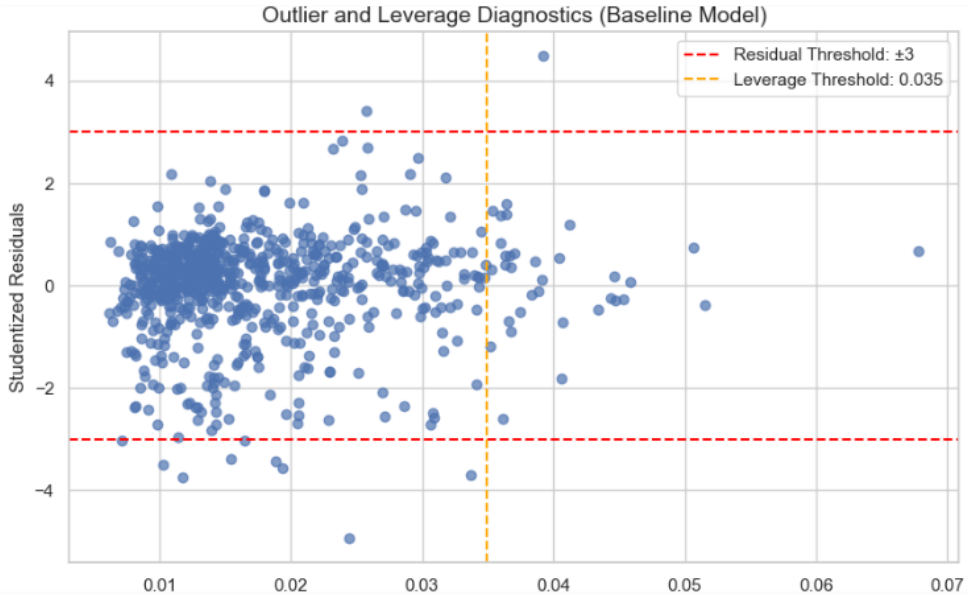


Table 25 Outlier Leverage Plott -Baseline Model

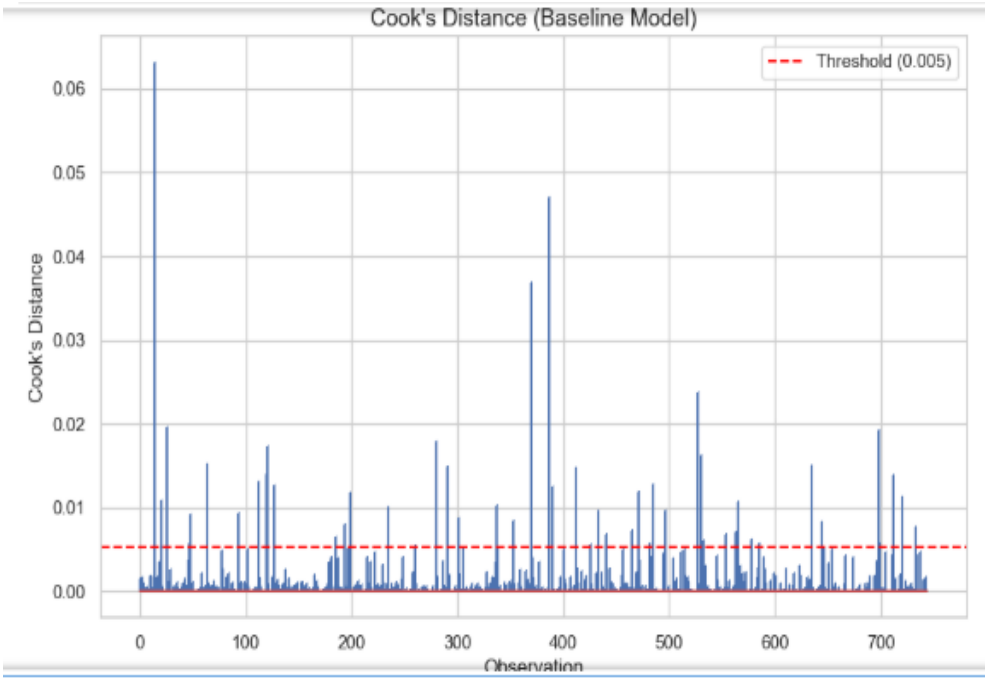


Table 26 Cook's ' D' chart-Baseline Model

D) Results

Pre-Assumption Stage

Baseline Model 1

log_revenue	Coef.	Std.Err	t-value	p-value	Sig.
const	-2.55925	1.114308	-2.29672	0.021916	**
log_budget	1.101575	0.054641	20.16014	3.13E-72	***
star_actor	-0.51046	0.17818	-2.86483	0.004292	***
tomato	-0.00738	0.003419	-2.15802	0.031251	**
score	0.021313	0.010717	1.988777	0.047098	**
action	-0.42371	0.230143	-1.84108	0.066014	*
adventure	0.024309	0.265382	0.091601	0.92704	
sci-fi	0.454036	0.316845	1.432991	0.152287	
drama	0.160176	0.223274	0.717395	0.47336	
comedy	-0.01323	0.230198	-0.05746	0.954197	
thriller	0.199289	0.195373	1.020044	0.308045	
horror	0.30598	0.243289	1.25768	0.208909	
sequel	0.506223	0.249928	2.025475	0.043181	**
Observations	745				
R-Squared	0.379042				
Adjusted R-Squared	0.368863				
AIC	3256.964				
BIC	3316.938				
Log-Likelihood	-1615.48				
F-Statistic	37.23536			8.31E-68	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 1: Baseline Model 1

Holiday Model 1

log_revenue	Coef.	Std.Err	t-value	p-value	Sig.
const	0.766738	1.726136	0.444193	0.657229	
log_budget	0.47695	0.041979	11.36166	4.69E-25	***
star_actor	-0.42955	0.147814	-2.90599	0.003938	***
tomato	-0.00088	0.002831	-0.30955	0.757122	
score	0.005932	0.008751	0.677828	0.498412	
holiday_budget	0.47695	0.041979	11.36166	4.69E-25	***
holiday_star_actor	-0.42955	0.147814	-2.90599	0.003938	***
holiday_score	0.005932	0.008751	0.677828	0.498412	
holiday_tomato	-0.00088	0.002831	-0.30955	0.757122	
action	-0.62188	0.380045	-1.63633	0.102836	
adventure	-0.01795	0.472467	-0.03799	0.969721	
sci-fi	0.289178	0.492689	0.586937	0.557695	
drama	-0.11809	0.373108	-0.31649	0.751852	
comedy	-0.27803	0.364318	-0.76316	0.445979	
thriller	-0.26317	0.322156	-0.81689	0.414649	
horror	0.656005	0.391552	1.675396	0.094916	*
sequel	0.225246	0.391452	0.575413	0.565451	
Observations	308				
R-Squared	0.333546				
Adjusted R-Squared	0.306436				
AIC	1385.519				
BIC	1434.01				
Log-Likelihood	-679.759				
F-Statistic	12.30343			2.88E-20	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 2: Holiday Model 1

Non – Holiday Model 1

Log_revenue	Coef.	Std.Err	t-value	p-value	Sig.
const	-5.2772	1.473835	-3.58059	0.000383	***
log_budget	1.222358	0.072799	16.79076	7.08E-49	***
star_actor	-0.35969	0.223675	-1.60811	0.108555	
tomato	-0.01277	0.004264	-2.99551	0.0029	***
score	0.031085	0.013566	2.291413	0.022429	**
action	-0.20762	0.286231	-0.72537	0.468625	
adventure	0.020553	0.317145	0.064805	0.94836	
sci-fi	0.547939	0.413902	1.323839	0.18627	
drama	0.305148	0.276647	1.103022	0.270643	
comedy	0.210371	0.295799	0.711194	0.477355	
thriller	0.493294	0.245367	2.010432	0.045018	**
horror	0.051842	0.30721	0.168751	0.866073	
sequel	0.6398	0.324266	1.973069	0.049137	**
Observations	437				
R-Squared	0.44609				
Adjusted R-Squared	0.430413				
AIC	1868.917				
BIC	1921.956				
Log-Likelihood	-921.459				
F-Statistic	28.45561			2.83E-47	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 3: Non - Holiday Model 1

Combined Model 1

log_revenue	Coef.	Std.Err	t-value	p-value	Sig.
const	-5.57982	1.491149	-3.74196	0.000197	***
log_budget	1.253051	0.074192	16.88923	3.10E-54	***
star_actor	-0.25744	0.229394	-1.12226	0.262123	
tomato	-0.01251	0.004438	-2.81875	0.004952	***
score	0.031109	0.013682	2.273706	0.023274	**
Holiday	6.38875	2.041234	3.129847	0.001819	***
holiday_budget	-0.30687	0.105638	-2.90489	0.003785	***
holiday_star_actor	-0.69468	0.349131	-1.98973	0.046995	**
holiday_score	-0.02332	0.019719	-1.18275	0.237295	
holiday_tomato	0.011463	0.006844	1.675046	0.094355	*
action	-0.4011	0.22894	-1.75197	0.0802	*
adventure	0.000469	0.26472	0.001771	0.998588	
sci-fi	0.429174	0.31526	1.361332	0.173831	
drama	0.12885	0.223	0.577805	0.563575	
comedy	-0.00799	0.229344	-0.03482	0.972233	
thriller	0.150967	0.195455	0.772387	0.440136	
horror	0.326292	0.241967	1.3485	0.177918	
sequel	0.435374	0.248777	1.750058	0.08053	*
Observations	745				
R-Squared	0.393114				
Adjusted R-Squared	0.378922				
AIC	3249.888				
BIC	3332.928				
Log-Likelihood	-1606.94				
F-Statistic	27.70104			1.69E-67	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 4: Combined Model 1

Mean-Centered Models

Non-Holiday Model 2

log_revenue	Coef.	Std.Err	t-value	p-value	Sig.
const	-5.2772	1.473835	-3.58059	0.000383	***
log_budget	1.222358	0.072799	16.79076	7.08E-49	***
star_actor	-0.35969	0.223675	-1.60811	0.108555	
tomato	-0.01277	0.004264	-2.99551	0.0029	***
score	0.031085	0.013566	2.291413	0.022429	**
action	-0.20762	0.286231	-0.72537	0.468625	
adventure	0.020553	0.317145	0.064805	0.94836	
sci-fi	0.547939	0.413902	1.323839	0.18627	
drama	0.305148	0.276647	1.103022	0.270643	
comedy	0.210371	0.295799	0.711194	0.477355	
thriller	0.493294	0.245367	2.010432	0.045018	**
horror	0.051842	0.30721	0.168751	0.866073	
sequel	0.6398	0.324266	1.973069	0.049137	**
Observations	437				
R-Squared	0.44609				
Adjusted R-Squared	0.430413				
AIC	1868.917				
BIC	1921.956				
Log-Likelihood	-921.459				
F-Statistic	28.45561			2.83E-47	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 5: Non-Holiday Model 2

Holiday Model 2

Log_revenue	Coef.	Std.Err	t-value	p-value	Sig.
const	0.766738	1.726136	0.444193	0.657229	
log_budget	0.47695	0.041979	11.36166	4.69E-25	***
star_actor	-0.42955	0.147814	-2.90599	0.003938	***
tomato	-0.00088	0.002831	-0.30955	0.757122	
score	0.005932	0.008751	0.677828	0.498412	
holiday_budget	0.47695	0.041979	11.36166	4.69E-25	***
holiday_star_actor	-0.42955	0.147814	-2.90599	0.003938	***
holiday_tomato	-0.00088	0.002831	-0.30955	0.757122	
holiday_score	0.005932	0.008751	0.677828	0.498412	
action	-0.62188	0.380045	-1.63633	0.102836	
adventure	-0.01795	0.472467	-0.03799	0.969721	
sci-fi	0.289178	0.492689	0.586937	0.557695	
drama	-0.11809	0.373108	-0.31649	0.751852	
comedy	-0.27803	0.364318	-0.76316	0.445979	
thriller	-0.26317	0.322156	-0.81689	0.414649	
horror	0.656005	0.391552	1.675396	0.094916	*
sequel	0.225246	0.391452	0.575413	0.565451	
Observations	308				
R-Squared	0.333546				
Adjusted R-Squared	0.306436				
AIC	1385.519				
BIC	1434.01				
Log-Likelihood	-679.759				
F-Statistic	12.30343			2.88E-20	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 6: Holiday Model 2

Baseline Model 2

log_revenue	Coef.	Std.Err	t-value	p-value	Sig.
const	-2.55925	1.114308	-2.29672	0.021916	**
log_budget	1.101575	0.054641	20.16014	3.13E-72	***
star_actor	-0.51046	0.17818	-2.86483	0.004292	***
tomato	-0.00738	0.003419	-2.15802	0.031251	**
score	0.021313	0.010717	1.988777	0.047098	**
action	-0.42371	0.230143	-1.84108	0.066014	*
adventure	0.024309	0.265382	0.091601	0.92704	
sci-fi	0.454036	0.316845	1.432991	0.152287	
drama	0.160176	0.223274	0.717395	0.47336	
comedy	-0.01323	0.230198	-0.05746	0.954197	
thriller	0.199289	0.195373	1.020044	0.308045	
horror	0.30598	0.243289	1.25768	0.208909	
sequel	0.506223	0.249928	2.025475	0.043181	**
Observations	745				
R-Squared	0.379042				
Adjusted R-Squared	0.368863				
AIC	3256.964				
BIC	3316.938				
Log-Likelihood	-1615.48				
F-Statistic	37.23536			8.31E-68	***

Notes: * p < 0,1, ** p <0,5, *** p < 0,001

Table 7: Baseline Model 2

Combined Model 2

log_revenue	Coef.	Std.Err	t-value	p-value	Sig.
const	-2.45588	1.114504	-2.20356	0.027868	**
log_budget	1.115369	0.060104	18.55727	3.06E-63	***
star_actor	-0.187	0.229662	-0.81423	0.415781	
tomato	-0.01025	0.004405	-2.32647	0.020267	**
score	0.016898	0.012984	1.301421	0.193526	
holiday_budget	-0.02954	0.057864	-0.51055	0.609824	
holiday_star_actor	-0.7796	0.350172	-2.22634	0.026297	**
holiday_tomato	0.007647	0.006775	1.12877	0.259367	
holiday_score	0.006554	0.017358	0.377574	0.705857	
action	-0.41157	0.230295	-1.78716	0.074328	*
adventure	-0.03111	0.266121	-0.11691	0.906966	
sci-fi	0.47476	0.31682	1.498514	0.134433	
drama	0.100501	0.224158	0.448348	0.654035	
comedy	-0.02598	0.230653	-0.11265	0.910335	
thriller	0.169895	0.196538	0.864437	0.387633	
horror	0.286424	0.243087	1.178279	0.23907	
sequel	0.483462	0.249798	1.935414	0.053326	*
Observations	745				
R-Squared	0.384936				
Adjusted R-Squared	0.371418				
AIC	3257.859				
BIC	3336.287				
Log-Likelihood	-1611.93				
F-Statistic	28.47605			3.44E-66	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 8: Combined Model 2

Hypothesis:(isolated interactions effects)

Interaction Effect of Score and Holiday

Variable	Coef.	Std.Err	t-value	p-value	Sig.
const	15.4767378	1.016964	15.21857	1.19E-45	***
score	0.02590883	0.015676	1.652734	0.098813	*
Holiday	-0.22868539	1.357573	-0.16845	0.866274	
score_holiday	-0.00155859	0.022009	0.070818	0.943562	
action	-0.21866966	0.282233	-0.77478	0.438717	
adventure	0.68586602	0.328054	2.090713	0.036898	**
sci-fi	0.78809425	0.394354	1.998444	0.046037	**
drama	-0.22762248	0.275256	-0.82695	0.408535	
comedy	0.13392934	0.286069	0.468171	0.639801	
thriller	0.4882012	0.242722	2.011355	0.044654	**
horror	-0.30528167	0.297064	-1.02766	0.304448	
sequel	0.57371001	0.311668	1.840773	0.066059	*
Observations	308				
R-Squared	0.03320761				
Adjusted R-Squared	0.01869913				
AIC	3584.79095				
BIC	3640.15156				
Log-Likelihood	-1780.39547				
F-Statistic	2.28884125			0.00941	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 9: Score, Hypothesis 1

Interaction Effect of Tomato and Holiday

Variable	Coef.	Std.Err	t-value	p-value	Sig.
const	16.98672	0.367311	46.24609	1.53E-219	***
tomato	0.002305	0.005052	0.456183	0.6483935	
Holiday	-0.18919	0.399371	-0.47372	0.6358427	
tomato_holiday	0.00171	0.007674	0.222773	0.8237743	
action	-0.30922	0.279937	-1.10461	0.2696925	
adventure	0.757969	0.327648	2.313366	0.0209784	**
sci-fi	0.733602	0.395167	1.856435	0.0637927	*
drama	-0.15485	0.2756	-0.56187	0.5743797	
comedy	0.095485	0.286518	0.333261	0.7390323	
thriller	0.457329	0.242549	1.885516	0.0597554	*
horror	-0.4393	0.29167	-1.50616	0.1324575	
sequel	0.544531	0.312208	1.744128	0.0815559	*
Observations	308				
R-Squared	0.0279				
Adjusted R-Squared	0.013312				
AIC	3588.87				
BIC	3644.23				
Log-Likelihood	-1782.43				
F-Statistic	1.912527			0.0347604	**

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 10: professional reviews (as tomato), hypothesis 2

Interaction Effect of Budget and Holiday

Variable	Coef.	Std.Err	t-value	p-value	Sig.
const	-3.75646	1.271294	-2.95483	0.003229	***
log_budget	1.225693	0.073523	16.67096	3.80E-53	***
Holiday	5.170829	1.794567	2.88138	0.004075	***
log_budget_holiday	-0.3	0.104908	-2.85964	0.004362	***
action	-0.60838	0.224872	-2.70543	0.00698	***
adventure	0.005972	0.262922	0.022714	0.981885	
sci-fi	0.342792	0.317756	1.078792	0.281035	
drama	0.100367	0.21719	0.462117	0.644135	
comedy	-0.12476	0.229879	-0.5427	0.587499	
thriller	0.096676	0.195931	0.493418	0.621865	
horror	0.297173	0.236746	1.255241	0.209791	
sequel	0.471374	0.251011	1.877904	0.060791	*
Observations	308				
R-Squared	0.374002				
Adjusted R-Squared	0.364608				
AIC	3260.987				
BIC	3316.348				
Log-Likelihood	-1618.49				
F-Statistic	39.8118			2.32E-67	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 11: budget, hypothesis 3

Interaction Effect of Budget and Holiday

Variable	Coef.	Std.Err	t-value	p-value	Sig.
const	17.03477	0.307577	55.38376	3.72E-264	***
star_actor	0.148664	0.284456	0.522624	0.601393651	
Holiday	0.176957	0.238706	0.741318	0.458737981	
star_actor_holiday	-0.95661	0.433281	-2.20783	0.027564936	**
action	-0.25337	0.282649	-0.8964	0.370334175	
adventure	0.771554	0.323533	2.384777	0.017342731	**
sci-fi	0.775877	0.393655	1.970959	0.049104399	**
drama	-0.14312	0.271244	-0.52764	0.597910238	
comedy	0.13691	0.286138	0.478474	0.632455205	
thriller	0.440532	0.242483	1.816755	0.069662951	*
horror	-0.47916	0.293037	-1.63516	0.102444919	
sequel	0.508545	0.31134	1.633406	0.102812972	
Observations	308				
R-Squared	0.035085				
Adjusted R-Squared	0.020605				
AIC	3583.343				
BIC	3638.703				
Log-Likelihood	-1779.67				
F-Statistic	2.422943			0.005776001	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 12: hypothesis 4, star actor model

Robust Standard Error Models

Non- Holiday Model 3

log_revenue	Coef.	Std.Err	t-value	p-value	Sig.
const	-5.2772	2.022518	-2.60922	0.009396	***
log_budget	1.222358	0.096345	12.68728	1.72E-31	***
star_actor	-0.35969	0.198227	-1.81456	0.070299	*
tomato	-0.01277	0.00467	-2.73512	0.006497	***
score	0.031085	0.01537	2.02241	0.043761	**
action	-0.20762	0.28377	-0.73166	0.464779	
adventure	0.020553	0.307308	0.06688	0.946709	
sci-fi	0.547939	0.350547	1.563096	0.118776	
drama	0.305148	0.263856	1.156496	0.248129	
comedy	0.210371	0.305855	0.687812	0.491947	
thriller	0.493294	0.248973	1.981313	0.048202	**
horror	0.051842	0.336001	0.154291	0.877454	
sequel	0.6398	0.306198	2.089497	0.037259	**
Observations	437				
R-Squared	0.44609				
Adjusted R-Squared	0.430413				
AIC	1868.917				
BIC	1921.956				
Log-Likelihood	-921.459				
F-Statistic	17.74589			5.70E-31	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 13: Non-Holiday Model 3

Holiday Model 3

log_revenue	Coef.	Std.Err	t-value	p-value	Sig.
const	0.766738	2.255072	0.340006	0.734094	
log_budget	0.47695	0.05721	8.336896	2.94E-15	***
star_actor	-0.42955	0.159595	-2.69149	0.007519	***
tomato	-0.00088	0.002893	-0.30283	0.762231	
score	0.005932	0.007711	0.769216	0.44238	
holiday_budget	0.47695	0.05721	8.336896	2.94E-15	***
holiday_star_actor	-0.42955	0.159595	-2.69149	0.007519	***
holiday_tomato	-0.00088	0.002893	-0.30283	0.762231	
holiday_score	0.005932	0.007711	0.769216	0.44238	
action	-0.62188	0.401761	-1.54788	0.122722	
adventure	-0.01795	0.3931	-0.04566	0.963612	
sci-fi	0.289178	0.484093	0.59736	0.550725	
drama	-0.11809	0.390172	-0.30265	0.762368	
comedy	-0.27803	0.375469	-0.74049	0.459589	
thriller	-0.26317	0.349575	-0.75282	0.452159	
horror	0.656005	0.404893	1.620194	0.106259	
sequel	0.225246	0.380474	0.592015	0.554294	
Observations	308				
R-Squared	0.333546				
Adjusted R-Squared	0.306436				
AIC	1385.519				
BIC	1434.01				
Log-Likelihood	-679.759				
F-Statistic	7.916518			8.04E-13	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 14: Holiday Model 3

Baseline Model 3

log_revenue	Coef.	Std.Err	t-value	p-value	Sig.
const	-2.55925	1.559359	-1.64122	0.101181	
log_budget	1.101575	0.076256	14.44579	8.43E-42	***
star_actor	-0.51046	0.173462	-2.94276	0.003356	***
tomato	-0.00738	0.003643	-2.02484	0.043247	**
score	0.021313	0.011041	1.93043	0.05394	*
action	-0.42371	0.240227	-1.76379	0.078184	*
adventure	0.024309	0.241608	0.100614	0.919884	
sci-fi	0.454036	0.287711	1.578095	0.114976	
drama	0.160176	0.224496	0.713488	0.475771	
comedy	-0.01323	0.240875	-0.05491	0.956225	
thriller	0.199289	0.208166	0.957358	0.338703	
horror	0.30598	0.261687	1.16926	0.242679	
sequel	0.506223	0.24264	2.08631	0.037296	**
Observations	745				
R-Squared	0.379042				
Adjusted R-Squared	0.368863				
AIC	3256.964				
BIC	3316.938				
Log-Likelihood	-1615.48				
F-Statistic	19.85678			8.55E-38	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 15: Baseline Model 3

Combined Model 3

Variable	Coef.	Std.Err	t-value	p-value	Sig.
const	-2.45588	1.559245	-1.57504	0.115681	
log_budget	1.115369	0.076573	14.56604	2.29E-42	***
star_actor	-0.187	0.200759	-0.93145	0.351927	
tomato	-0.01025	0.004542	-2.25664	0.024326	**
score	0.016898	0.014043	1.203329	0.22924	
holiday_budget	-0.02954	0.057403	-0.51465	0.606952	
holiday_star_actor	-0.7796	0.365622	-2.13226	0.033319	**
holiday_tomato	0.007647	0.006987	1.094494	0.2741	
holiday_score	0.006554	0.018036	0.363399	0.716413	
action	-0.41157	0.238097	-1.72859	0.084306	*
adventure	-0.03111	0.243513	-0.12776	0.898374	
sci-fi	0.47476	0.28122	1.688214	0.091798	*
drama	0.100501	0.228689	0.439465	0.660455	
comedy	-0.02598	0.241002	-0.10782	0.914171	
thriller	0.169895	0.210451	0.80729	0.419763	
horror	0.286424	0.263704	1.086158	0.277768	
sequel	0.483462	0.239018	2.022699	0.043469	**
Observations	745				
R-Squared	0.384936				
Adjusted R-Squared	0.371418				
AIC	3257.859				
BIC	3336.287				
Log-Likelihood	-1611.93				
F-Statistic	15.72845			1.65E-37	***

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 16: Combined Model 3

Ridge Regression Models

Comparison of Ridge & OLS Regression Holiday & Combined Models

log_revenue	OLS Coef.	Std.Err	t-value	p-value	Ridge Coef.
const	-2.45588	1.114504	-2.20356	0.027868	0
log_budget	1.115369	0.060104	18.55727	3.06E-63	1.0327402
star_actor	-0.187	0.229662	-0.81423	0.415781	-0.253621
tomato	-0.01025	0.004405	-2.32647	0.020267	-0.009641
score	0.016898	0.012984	1.301421	0.193526	0.0193014
holiday_budget	-0.02954	0.057864	-0.51055	0.609824	-0.013557
holiday_star_actor	-0.7796	0.350172	-2.22634	0.026297	-0.30023
holiday_tomato	0.007647	0.006775	1.12877	0.259367	0.0075863
holiday_score	0.006554	0.017358	0.377574	0.705857	-8.65E-05
action	-0.41157	0.230295	-1.78716	0.074328	-0.221136
adventure	-0.03111	0.266121	-0.11691	0.906966	0.0002947
sci-fi	0.47476	0.31682	1.498514	0.134433	0.1373077
drama	0.100501	0.224158	0.448348	0.654035	0.0127104
comedy	-0.02598	0.230653	-0.11265	0.910335	-0.041562
thriller	0.169895	0.196538	0.864437	0.387633	0.1093563
horror	0.286424	0.243087	1.178279	0.23907	0.1693673
sequel	0.483462	0.249798	1.935414	0.053326	0.2136373
OLS R-Squared	0.384936				
Ridge R-Squared					0.3846236
OLS MSE (Holiday)					4.8362716
OLS MSE (Combined)					
Ridge MSE					4.4864072

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 17: Comparison of Ridge & OLS Regression Holiday & Combined Models

Models 'comparisons OLS & Ridge

Model	Variable	Coefficient	R^2	MSE
Holiday Model	const	0.7667	0.333546	4.836272
Holiday Model	log_budget	0.477		
Holiday Model	star_actor	-0.4295		
Holiday Model	tomato	-0.0009		
Holiday Model	score	0.0059		
Holiday Model	holiday_budget	0.477		
Holiday Model	holiday_star_actor	-0.4295		
Holiday Model	holiday_tomato	-0.0009		
Holiday Model	holiday_score	0.0059		
Holiday Model	action	-0.6219		
Holiday Model	adventure	-0.0179		
Holiday Model	sci-fi	0.2892		
Holiday Model	drama	-0.1181		
Holiday Model	comedy	-0.278		
Holiday Model	thriller	-0.2632		
Holiday Model	horror	0.656		
Holiday Model	sequel	0.2252		
Non-Holiday Model	const	-5.2772	0.44609	3.972218
Non-Holiday Model	log_budget	1.2224		
Non-Holiday Model	star_actor	-0.3597		
Non-Holiday Model	tomato	-0.0128		
Non-Holiday Model	score	0.0311		
Non-Holiday Model	action	-0.2076		
Non-Holiday Model	adventure	0.0206		
Non-Holiday Model	sci-fi	0.5479		
Non-Holiday Model	drama	0.3051		
Non-Holiday Model	comedy	0.2104		
Non-Holiday Model	thriller	0.4933		
Non-Holiday Model	horror	0.0518		
Non-Holiday Model	sequel	0.6398		
Baseline Model	const	-2.5593	0.379042	4.477152
Baseline Model	log_budget	1.1016		
Baseline Model	star_actor	-0.5105		
Baseline Model	tomato	-0.0074		
Baseline Model	score	0.0213		
Baseline Model	action	-0.4237		
Baseline Model	adventure	0.0243		
Baseline Model	sci-fi	0.454		
Baseline Model	drama	0.1602		
Baseline Model	comedy	-0.0132		
Baseline Model	thriller	0.1993		
Baseline Model	horror	0.306		
Baseline Model	sequel	0.5062		
Combined Model with Ridge	const	0	0.377759	4.486407
Combined Model with Ridge	log_budget	1.0327		
Combined Model with Ridge	star_actor	-0.2536		
Combined Model with Ridge	tomato	-0.0096		
Combined Model with Ridge	score	0.0193		
Combined Model with Ridge	holiday_budget	-0.0136		
Combined Model with Ridge	holiday_star_actor	-0.3002		
Combined Model with Ridge	holiday_tomato	0.0076		
Combined Model with Ridge	holiday_score	-0.0001		
Combined Model with Ridge	action	-0.2211		
Combined Model with Ridge	adventure	0.0003		
Combined Model with Ridge	sci-fi	0.1373		
Combined Model with Ridge	drama	0.0127		
Combined Model with Ridge	comedy	-0.0416		
Combined Model with Ridge	thriller	0.1094		
Combined Model with Ridge	horror	0.1694		
Combined Model with Ridge	sequel	0.2136		

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 18: Models' Comparison OLS & Ridge

Variance Inflation Factor (VIF) Diagnostics for Holiday and Combined Model

VIF for Holiday Model:

	Variable	VIF
0	const	181.744501
1	log_budget	inf
2	star_actor	inf
3	tomato	inf
4	score	inf
5	holiday_budget	inf
6	holiday_star_actor	inf
7	holiday_tomato	inf
8	holiday_score	inf
9	action	1.483335
10	adventure	1.337155
11	sci-fi	1.144403
12	drama	1.933447
13	comedy	1.605830
14	thriller	1.496159
15	horror	1.625684
16	sequel	1.100771

VIF for Combined Model:

	Variable	VIF
0	const	203.908130
1	log_budget	1.360696
2	star_actor	1.843416
3	tomato	2.141510
4	score	2.226966
5	holiday_budget	38.921884
6	holiday_star_actor	2.199133
7	holiday_tomato	5.934715
8	holiday_score	46.918520
9	action	1.441341
10	adventure	1.246621
11	sci-fi	1.107787
12	drama	1.856963
13	comedy	1.588104
14	thriller	1.543460
15	horror	1.694207
16	sequel	1.077530

*Notes: * p < 0,1, ** p <0,5, *** p < 0,001*

Table 19: VIF for Holiday & Combined Model