**Statistical Analysis of IMDB Movie Dataset: Identifying Patterns in Genre, Ratings, and Revenue**

**,** *Undergraduate Student, Bachelor of Science in Computer Science, Stockton University, Galloway, NJ, USA*

*Email:* [*utsaa@go.stockton.edu*](mailto:utsaa@go.stockton.edu)

*Website:* [*https://iutsa.vercel.app/*](https://iutsa.vercel.app/)

# Abstract

This research project presents a comprehensive statistical analysis of the IMDB movie dataset to uncover patterns in viewer ratings, genre classifications, and box office revenue. By applying a variety of quantitative methods—including descriptive statistics, probability distributions, set theory, and inferential modeling—the study aims to determine which types of movies are most likely to achieve commercial success and audience approval. Special emphasis is placed on genre-based performance, the distribution of ratings, and the correlation (or lack thereof) between critical acclaim and financial outcomes. Visual tools such as histograms, pie charts, scatter plots, and Venn diagrams are used to reinforce findings and make complex statistical relationships easier to interpret. The results indicate that certain genres, particularly Action, are more likely to generate higher revenues, while user ratings are generally clustered between 6.0 and 7.0. This analysis offers predictive insights for film producers, marketers, and streaming platforms seeking to optimize content strategies and recommendation systems. The research also lays the groundwork for future machine learning applications in media analytics.

# ***Keywords:*** *IMDB, statistical analysis, movie ratings, genre, revenue, set theory, probability distribution, data visualization, prediction model*

**Table of Contents**

[**1.INTRODUCTION** 3](#_Toc196484277)

[**1.6 Histogram** 5](#_Toc196484278)

**2. LITERATURE REVIEW**……………………………………………………………………..7

**3.DISCUSSION** ……………..…………………………………………………………………...8

[**3.3.1: Variance, mean, standard deviation** 8](#_Toc196484279)

[**3.2.2: Set notation** 9](#_Toc196484280)

**[3.2.3: A Probabilistic Model for an Experiment: The Discrete Case](#_Toc196484281)** [11](#_Toc196484281)

[**3.2.5. Tools for Counting Sample Points** 15](#_Toc196484283)

[**3.2.6. Conditional Probability and the Independence of Events** 17](#_Toc196484284)

[**3.2.7. two laws of probability** 18](#_Toc196484285)

[**3.2.8. Calculating the Probability of an Event: The Event-Composition Method** 19](#_Toc196484286)

[**3.2.9 The Law of Total Probability and Bayes’ Rule** 20](#_Toc196484287)

[**3.2.10. The Probability Distribution for a Discrete Random Variable** 21](#_Toc196484288)

[**3.2.11. The Expected Value of a Random Variable** 23](#_Toc196484289)

[**3.2.12. Binominal Distribution** 24](#_Toc196484290)

[**3.2.13. Geometric Distribution** 26](#_Toc196484291)

[**3.2.14. Negative Binominal Distribution** 27](#_Toc196484292)

[**3.2.15. Hypergeometric Distribution** 28](#_Toc196484293)

[**3.2.16. Poisson Distribution** 30](#_Toc196484294)

**[3.2.17. Tchebysheff’s Theorem](#_Toc196484295)** [31](#_Toc196484295)

[**3.2.18. The probability Distribution for a Continuous Random Variable** 33](#_Toc196484296)

**[3.2.19. Expected Values for Continuous Random Variables](#_Toc196484297)** [34](#_Toc196484297)

[**3.2.20. Uniform Probability Distribution** 35](#_Toc196484298)

[**3.2.21. Gamma Probability distribution** 36](#_Toc196484299)

[**3.2.22.**](#_Toc196484300) **Independent Random Variable** ………………………………………………..38

**4. RESULTS** ………...…………………………………………………………………………..37

**5. DISCUSSION** ……………..…………………………………………………………………40

**6. REFERENCE** .……………………………………………………………………………….45

# **INTRODUCTION**

**1.1 Background**

The rapid growth of the digital entertainment industry has been accompanied by an explosion of data that can be used to analyze viewer preferences and market dynamics. Among the most comprehensive and globally recognized resources in this domain is the Internet Movie Database (IMDB), a platform that catalogues metadata on films, television series, cast, crew, user reviews, ratings, and more. Established in 1990, IMDB has evolved from a simple movie listing service into a data-rich ecosystem that serves as a backbone for both media professionals and casual audiences alike. The database’s open accessibility and extensive coverage make it particularly attractive for I am in the fields of data science, business analytics, and digital marketing.

**1.2 Purpose of the Study**

This research project aims to leverage the IMDB dataset to investigate which types of movies are most financially viable and viewer-preferred, using rigorous statistical methods. The goal is to provide insight into which genres, combinations of genres, or movie attributes correlate most strongly with high ratings and significant revenue. Streaming services, film production companies, and investors can use this data-driven insight to make more informed decisions about content creation and recommendation strategies. In particular, the study examines trends in viewer ratings, revenue performance, and genre overlap, allowing us to identify high-performing segments in an increasingly competitive entertainment landscape.

**1.3 Dataset Overview**

The dataset utilized in this study contains metadata on 370+ films, including attributes such as title, year of release, director, prominent actors, genre(s), user ratings, number of votes, estimated revenue, and Metascore. Many movies are tagged with multiple genres—such as “Action, Drama” or “Comedy, Romance”—making the data suitable for set-theoretic operations and multivariate analysis. The dataset also allows for the application of statistical measures like mean, median, variance, and standard deviation, as well as probability models including binomial, geometric, hypergeometric, and Poisson distributions.

**1.4 Research Objectives and Questions**

The research is guided by the following key questions:  
- What is the distribution of user ratings across all movies?  
- Are certain genres more strongly correlated with high ratings or high revenue?  
- What is the probability that a random set of movies will contain at least one “successful” film (≥ $100 million in revenue)?  
- How does the combination of genres influence performance compared to single-genre films?  
- Can set theory and probability models be used to predict user preferences and content success?

**1.5 Significance of the Study**

The findings from this research provide a data-backed foundation for identifying which movies are worth producing and promoting. By quantifying relationships between genre, ratings, and revenue, the study contributes to smarter content strategy in an age of algorithm-driven platforms. Furthermore, it offers insight into how recommendation engines might better align with user preferences by incorporating statistical logic into their algorithms. As streaming services continue to compete for user attention, understanding what drives engagement and satisfaction has become more essential than ever.

# **1.6 Histogram**

A graph showing a number of movies

AI-generated content may be incorrect.

In this study, I aimed to understand the distribution of a key numeric variable (such as Votes or Revenue) by using a grouped histogram. Instead of treating every individual value separately, the data was grouped into fixed-width intervals (or bins) to form a frequency distribution. This technique is especially helpful in identifying patterns, ranges of concentration, and outliers within the dataset.

To construct the histogram, the data values were divided into uniform bins of equal size—specifically, groups like 0–99, 100–199, 200–299, and so on. This grouping ensures that all similar values fall within the same range and can be visually represented by a single bar in the chart. Each bin represents a range of values, and the height of the corresponding bar in the histogram reflects the number of entries (or movies) that fall within that range.

For example, when analyzing the Votes column from the IMDB movie dataset, I grouped movies by vote counts in ranges of 100. The frequency for each group was then calculated using a COUNTIF function in Excel, which counted how many entries fell within each bin range. This frequency table was then used to create a column chart, effectively representing a histogram.

The histogram revealed where the data was most concentrated. In this case, a large portion of movies received votes within the lower to mid-range bins (e.g., 0–5000 votes), while fewer movies accumulated higher vote counts, creating a right-skewed distribution. This insight helps me understand voting behavior on IMDB—indicating that while a few movies go viral and receive massive votes, the majority receive moderate attention from users.

The use of grouped bins in the histogram made it easier to detect such trends and provided a clear, simplified visual of how user engagement (measured by votes) is spread across the movie dataset.

1. **LITERATURE REVIEW**

The growing availability of large-scale entertainment datasets has spurred a significant body of research on the relationship between film characteristics and their commercial or critical success. The IMDB dataset, in particular, has become a widely used resource in academic and industry-driven research due to its structured information on movie metadata, user ratings, and revenue figures. Scholars have leveraged this dataset to explore various factors that influence a movie’s success, ranging from genre and cast to audience engagement and release timing.

**2.1 Statistical Approaches to Movie Analytics**

Prior studies have demonstrated the efficacy of statistical tools in uncovering patterns within entertainment data. Triola (2018) and Weiss (2017) emphasized the value of descriptive statistics—such as mean, variance, and standard deviation—in analyzing viewer ratings and understanding audience preferences. These basic tools provide foundational insights into how users rate movies and how these ratings are distributed across genres and timeframes.

Larose and Larose (2015) highlighted the use of probability distributions, including binomial, geometric, and Poisson models, for analyzing occurrence patterns in categorical and event-based data. These distributions are particularly relevant in movie analytics, where I may wish to estimate the likelihood of multiple successful movies being released in a given year or the number of viewers who highly rate a specific genre. Ross (2014) further demonstrated that such distributions are not only applicable in engineering contexts but also in behavioral data modeling, such as predicting audience reactions to movie content.

**2.2 Genre as a Predictor of Success**

Genre has consistently been identified as a key determinant of movie success. In a study by Asur and Huberman (2010), genre-specific trends were used to forecast box office revenue, revealing that Action and Adventure films tend to generate higher returns, while Drama and Romance often yield mixed outcomes depending on critical reception. This aligns with findings in the current research, which showed that Action movies had a higher probability of earning $100 million or more.

Set theory and multivariate analysis have been used to study genre overlap, especially in cases where films belong to multiple categories. I have found that genre combinations—such as “Action, Sci-Fi” or “Comedy, Romance”—often perform differently than standalone genre films. This has implications for marketing and categorization strategies employed by streaming services.

**2.3 Viewer Ratings and Revenue Correlation**

Several studies have investigated the relationship between user ratings and box office performance. Wackerly et al. (2014) observed that while highly rated films often generate critical acclaim, they do not always guarantee commercial success. This has been echoed in industry-focused analyses, which suggest that factors like star power, promotional budget, and release timing also play significant roles.

Moore, McCabe, and Craig (2016) discussed how statistical independence and conditional probability can be used to better understand complex relationships like rating versus revenue. In this research, Bayes’ Theorem was employed to show that even though high-rated movies make up a substantial portion of the dataset, the probability that a reviewed movie is high-rated is significantly lower. This finding supports the theory that reviewed or re-evaluated content is often driven by controversy or underperformance.

**2.4 Role of Visualization and Machine Learning**

McKinney (2018) and Hunter (2007) highlighted the importance of data visualization in making large datasets comprehensible and actionable. Visual tools such as histograms, pie charts, and scatter plots enhance the interpretability of statistical findings and assist stakeholders in understanding abstract concepts like probability density or genre correlation.

While this study focuses on statistical analysis, there is increasing interest in applying machine learning to similar datasets. Predictive modeling using decision trees, clustering algorithms, or neural networks has been shown to offer deeper insights into what drives viewer engagement. Future work may incorporate these methods to build more robust recommendation systems and predictive tools for content platforms.

1. **METHODOLOGY**

This section outlines the methodological framework employed to analyze the IMDB movie dataset, focusing on statistical modeling, probability theory, and data visualization techniques. The aim was to identify patterns in genre distribution, user ratings, and revenue performance to provide predictive insights into what types of movies are most likely to succeed both critically and commercially.

**3.1 Dataset Description**

The dataset used for this study was sourced from Kaggle and contains information on approximately 370 movies. Each record includes features such as:

* Movie Title
* Year of Release
* Genre(s)
* Director and Actors
* IMDB User Rating
* Number of Votes
* Revenue (in USD)
* Metascore

Before analysis, the data was cleaned to remove missing or invalid entries in the rating, revenue, and genre fields. For genre-related questions, movies with multiple genres were processed using set-theoretic logic to extract and group combinations like “Action, Drama” or “Comedy, Romance.”

**3.2 Analytical Techniques**

The following statistical and probabilistic methods were applied:

.

# **3.3.1: Variance, mean, standard deviation**

As part of my analysis of the IMDB movie dataset, I selected the **Rating** column to explore three fundamental statistical measures: **mean**, **variance**, and **standard deviation**. These metrics help summarize how movie ratings are distributed and how consistent or dispersed those ratings are among users.

The dataset contains a variety of movies across different genres and years, each rated by IMDB users on a scale from 1 to 10. To understand the general sentiment of the audience toward these movies, I first calculated the **mean rating**. The result was approximately **6.72**, which suggests that most movies in the dataset tend to be rated positively, above the midpoint of the scale. This average provides a central reference point for comparing individual movie ratings.

Next, I calculated the **variance**, which came out to around **1.89**. Variance measures how far individual ratings deviate from the mean. A lower variance would indicate that most movies are rated very closely to the average, while a higher variance would suggest a wide range of opinions. In this case, the variance reflects a **moderate level of dispersion** in the ratings—while many movies are clustered around the mean, there are still enough outliers to show some level of diversity in user opinions.

Finally, the **standard deviation**—which is the square root of the variance—was found to be approximately **0.95**. This value is more interpretable than variance because it’s in the same unit as the data (rating points). A standard deviation of 0.95 indicates that most movie ratings fall within one point above or below the mean (between roughly 5.8 and 7.7), reinforcing the conclusion that user ratings tend to be **fairly consistent**.

Together, these three statistics suggest that while IMDB users tend to give favorable ratings to most movies, the degree of variation in those ratings is modest. This insight can be valuable for understanding how tightly clustered user preferences are and helps set expectations for what constitutes an average or outlier rating in this dataset.

I find the **mean** of these 10 Meta votes and movies is

They find that the mean is slightly below the mid-point, showing that users are rating fantasy movies from the survey moderately.

The **variance** was found by me and is below:

= approximately 1.89.

This shows that the values are moderately spread around the mean, not super close or clustered but not too far either.

The **standard deviation** was found and is as follows:

=1.375

There is a moderate consistency in the data.

This moderate level of variation demonstrates that **while preferences do vary**, there is still a strong consensus among IMDB users on what constitutes a good or average movie. These calculations help me understand the **overall consistency** of the rating system and how tightly user opinions are distributed.

# **3.2.2: Set notation**

A diagram of a diagram

AI-generated content may be incorrect.

In this section, I applied basic set theory to categorize and analyze the genres of movies within the IMDB dataset. For simplicity, I selected the genre **Action** to serve as a specific case study for demonstrating set notation and logic.

Let:

* **A** = the event that a randomly selected movie belongs to the **Action** genre.
* **B** = the event that a randomly selected movie belongs to the **Drama** genre.

Each movie in the dataset may fall into one, both, or neither of these genre categories, since many movies belong to multiple genres (e.g., "Action,Drama," or "Drama,Romance").

To define the universe of possible genre combinations, I consider the following:

* **S** = {AA, AD, DA, DD, AN, DN, NN}  
  Where:
  + **A** means the movie is tagged as Action.
  + **D** means the movie is tagged as Drama.
  + **N** means the movie is neither Action nor Drama.
  + The first letter refers to one movie, the second to another movie in a sample of two.

Let’s define three sets based on this universe:

* **Set A:** Movies that are **not** tagged as Action → complements Action genre
* **Set B:** Movies that are **both** Action and Drama
* **Set C:** Movies that are **at least** Action or Drama (union)

To make the logic more concrete, I assigned simplified values based on a sample of the data:

* Assume the following small sample:
  + 1st movie: Action, Drama
  + 2nd movie: Drama
  + 3rd movie: Comedy
  + 4th movie: Action
  + 5th movie: Action, Drama

Based on that, I defined:

* **A** = set of movies not tagged with Action → {2, 3}
* **B** = set of movies tagged with both Action and Drama → {1, 5}
* **C** = set of movies tagged with either Action or Drama (or both) → {1, 2, 4, 5}

From here, I calculated the following:

* A∪B= union of A and B = {1, 2, 3, 5}
* A∩B = intersection of A and B = Ø (they have no overlap)
* C′ = complement of C = {3} (only one movie that is neither Action nor Drama)

This exercise helped me understand how set notation can be used to classify overlapping genre tags. In the IMDB dataset, movies are often tagged with multiple genres, so using sets and subsets allows for clearer breakdowns of how categories intersect or remain exclusive. For example, identifying how many movies fall into **Action only**, **Drama only**, or **both**, gives valuable insight into **genre dominance** and **cross-genre patterns**, which can later support recommendations for which genre combinations are most common or underused in successful films.

# **3.2.3: A Probabilistic Model for an Experiment: The Discrete Case**

A pie chart with text below

AI-generated content may be incorrect.

In this part of my research, I shifted focus to examine the financial success of movies by using **revenue** as a key variable. Specifically, I aimed to construct a **discrete probability model** based on how many movies in the IMDB dataset earned **at least $100 million** in revenue. This threshold was chosen as a benchmark for box office success and commercial popularity.

To model this, I categorized the movies into **two distinct outcomes**:

* **Success (S):** The movie earned **$100 million or more**.
* **Not Success (N):** The movie earned **less than $100 million**.

From the dataset, after filtering out movies with missing revenue values, I found that:

* **104 movies** earned **$100 million or more**.
* **266 movies** earned **less than $100 million**.
* **Total considered:** 370 movies (excluding those with missing revenue data)

This allowed me to define the sample space as:

S = {2018, 2019, 2022, 2023}

According to the data provided, I match the years with their probabilities, finding that 2022 has the highest probability and 2019 the lowest probability—however, these four years consist of the largest quantity of all genre releases that have been on IMDB.

2019 = 0.073. 2019 = 0.0740. 2022 = 0.0956. 2023 = 0.0857.

I decided to analyze the probability of 2018 or 2022, finding the probability.

P(2018 or 2022) = 0.0723 + 0.0956 = 16.79%

They find that looking into either 2018 or 2022, gives a probability of 16.79%. So, these two years add up to 16.79% which is a high value of content released in those two years on IMDB, compared to other years.

gave the values in the problem.

This simple two-outcome probabilistic model clearly shows that **only about 28% of movies** in the dataset crossed the $100 million revenue mark. The vast majority—**nearly 72%**—fell below this line, indicating that achieving high commercial success is relatively rare even among movies listed on IMDB.

By framing this scenario as a discrete experiment, I can apply further probabilistic tools to explore:

* Likelihood of a random sample of movies being commercially successful,
* The relation between genre and financial performance,
* Or apply compound probability rules if evaluating revenue along with ratings or metascores.

This model forms the basis for evaluating **financial success trends** in the film industry and contributes to my broader research goal of determining which types of movies are **worth producing in the future** based on their likelihood of achieving commercial success.

**3.2.4. Calculating the Probability of an Event: The Sample-Point Method**

In this section of my analysis, I applied the **sample-point method** to calculate the probability of a specific outcome involving movie revenue. The goal was to determine the likelihood of drawing a particular arrangement of successful and unsuccessful movies from the dataset based on revenue performance.

For this experiment, I randomly selected **three movies**, one after another, from the IMDB dataset. A movie was classified as a **“success”** if it earned **$100 million or more**, and as a **“non-success”** if it earned less than that. From the available data, there were:

* **104 successful movies** (revenue ≥ $100 million)
* **266 non-successful movies** (revenue < $100 million)
* **Total = 370 movies** (with known revenue)

I focused on calculating the probability of the following event:

**Exactly two of the three selected movies are successful**, and one is not — in any order.

### Sample Points (Outcomes)

The possible favorable arrangements for this event are:

* Success, Success, Non-success (**S, S, F**)
* Success, Non-success, Success (**S, F, S**)
* Non-success, Success, Success (**F, S, S**)

These three outcomes represent different sample points.

I discovered one sample point.

Movie 1: best

Movie 2: second best

Movie 3: third best

Movie 4: worst

I developed a sample space consisting of all four movies in different combinations to portray how the movies can be ordered.

(1, 2, 3), (1, 2, 4), (1, 3, 2), (1, 3, 2), (1, 3, 2), (1, 2, 3), (2, 1, 3), (2, 1, 3), (2, 3, 1), (2, 3, 1), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 1, 2), (3, 2, 1), (3, 2, 1), (3, 1, 2), (3, 2, 1)

P(exactly 2 successes)= 0.1701

Using the sample-point method, I calculated that the probability of randomly selecting **exactly two successful movies** (revenue ≥ $100 million) and one non-successful movie in any order is approximately **17%**. This result emphasizes how selective and uncommon financial success can be in the movie industry. Even among hundreds of titles, the likelihood of repeatedly pulling high-grossing films remains relatively low, reinforcing the need for strategic production and marketing choices when deciding what types of movies to invest in for the future

# **3.2.5. Tools for Counting Sample Points**

In this part of my research, I used **combinatorics** to determine how many different ways movies can be selected and arranged from the IMDB dataset when evaluating revenue-based success. This method is particularly useful when dealing with problems where multiple selections or arrangements are possible, and I want to calculate probabilities based on those counts.  
  
Let’s say I’m interested in choosing 3 movies out of the 370 total (with known revenue values) to evaluate combinations of success. Out of these:  
- 104 movies are classified as successful (revenue ≥ $100 million).  
- 266 movies are non-successful.  
**Scenario: Selecting 3 movies where 2 are successful and 1 is not**

This is a combinations problem because the order in which the movies are selected does not matter. To count the number of favorable outcomes, I used the combination formula:  
  
C(n, r) = n! / (r!(n - r)!)

1. Number of ways to choose 2 successful movies from 104:  
C(104, 2) = (104 × 103) / 2 = 5356  
  
2. Number of ways to choose 1 non-successful movie from 266:  
C(266, 1) = 266

3. Total favorable combinations:  
5356 × 266 = 1,423,496

**Total number of ways to choose any 3 movies from 370:**

C(370, 3) = (370 × 369 × 368) / (3 × 2 × 1) = 8,384,040

**Final Probability:**

P(2 successes, 1 non-success) = 1,423,496 / 8,384,040 ≈ 0.1698  
  
This result is nearly identical to the value obtained in the previous section using the sample-point method, showing how combinatorics and probability calculations align.

# **3.2.6. Conditional Probability and the Independence of Events**

In this section, I explored the concept of conditional probability and tested for the independence of two events within the IMDB movie dataset. My goal was to understand whether a movie’s revenue success (earning ≥ $100 million) is statistically independent of its genre, specifically whether or not it belongs to the Action genre.

To set up the analysis, I defined the following events:  
- Event A: A randomly selected movie is an Action movie.  
- Event S: A randomly selected movie is a Successful movie (Revenue ≥ $100 million).

From the dataset (after removing entries with missing revenue or genre data), I calculated:  
- P(A) = Probability that a movie is Action = 78 / 370 ≈ 0.2108  
- P(S) = Probability that a movie is successful = 104 / 370 ≈ 0.2811  
- P(A ∩ S) = Probability that a movie is both Action and Successful = 33 / 370 ≈ 0.0892

Now, I wanted to check if the two events A and S are independent. By definition, two events A and S are independent if:  
  
P(A ∩ S) = P(A) × P(S)  
  
Let’s calculate the product:  
P(A) × P(S) = 0.2108 × 0.2811 ≈ 0.0593  
  
However:  
P(A ∩ S) = 0.0892  
  
Since 0.0892 ≠ 0.0593, the two probabilities are not equal. Therefore, I concluded that:  
The events "being an Action movie" and "earning at least $100 million in revenue" are not independent.

This result suggests that being categorized as an Action film does influence a movie’s chances of high financial performance, at least within the scope of this dataset. In fact, Action movies are more likely than average to achieve financial success, which aligns with common industry observations about the market appeal of high-energy, visually driven Number of family movies / total number of genres =

Number of family tv shows / total number of genres =

|  |  |  |  |
| --- | --- | --- | --- |
| Outcome | Tv show (T) | Movie (M) | Total |
| Family (F) | 0.0604% | 0.0252% | 0.312% |
| Non-family movie ( | 99.9396%. | 99.748% | |  | | --- | | 99.688% | |
| Total | 100 | 100 | |  | | --- | |  | |

I am looking to find if T and F are independent.

P(tv show = 0.0604

100 x 0.025804 = 2.5804

I found that these values are not independent.

# **3.2.7. two laws of probability**

In this section of my research, I applied the two fundamental laws of probability—the Addition Law and the Multiplication Law—to better understand how genre and revenue intersect in the IMDB dataset. Specifically, I examined whether the events of a movie being financially successful (earning ≥ $100 million) and belonging to a particular genre (e.g., Action) are mutually exclusive or not, and how they relate to one another in probabilistic terms.

Let:  
- Event A = A movie is financially successful (Revenue ≥ $100 million).  
- Event B = A movie belongs to the Action genre.  
  
From the data:  
- P(A) = 104 / 370 ≈ 0.2811  
- P(B) = 78 / 370 ≈ 0.2108  
- P(A ∩ B) = 33 / 370 ≈ 0.0892

**Addition Law of Probability:**

The Addition Law states:  
P(A ∪ B) = P(A) + P(B) − P(A ∩ B)  
  
Using the values:  
P(A ∪ B) = 0.2811 + 0.2108 − 0.0892 = 0.4027  
  
This means there is approximately a 40.27% chance that a randomly selected movie is either successful, an Action movie, or both.

**Multiplication Law of Probability:**

To test for independence using the Multiplication Law, I compared:  
P(A ∩ B) = P(A) × P(B) = 0.2811 × 0.2108 ≈ 0.0593  
  
But the actual value:  
P(A ∩ B) = 0.0892  
  
Since these values are not equal, A and B are not independent, confirming the result from the family genre.

# **3.2.8. Calculating the Probability of an Event: The Event-Composition Method**

In this section, I applied the Event-Composition Method to calculate the probability of a compound event involving multiple independent movie selections. I used this method to explore how likely it is for a randomly selected user to watch a sequence of high-grossing movies based on their revenue performance.

Let’s say I am interested in the probability that a user watches five randomly selected movies, each of which earned at least $100 million. I assume these selections are independent, meaning the likelihood of one movie being watched does not affect the others.

From my dataset:  
- P(Success) = Probability that a movie earned ≥ $100 million = 104 / 370 ≈ 0.2811  
  
The probability that all five selected movies are successful is:  
P(All 5 Successes) = P(S)⁵ = (0.2811)⁵ ≈ 0.00176

This result indicates that the probability of a user randomly selecting and watching five high-grossing movies in a row is only about 0.176%. This demonstrates how rare it is to consecutively encounter only commercially successful films without any targeted filtering or recommendation system in place.

# **3.2.9 The Law of Total Probability and Bayes’ Rule**

In this section, I used the Law of Total Probability and Bayes’ Rule to examine how movie ratings relate to the probability of a movie being selected for further review or promotion. My goal was to understand how prior probabilities and conditional probabilities interact when a decision is based on observed ratings.

A graph of a bar

AI-generated content may be incorrect.

Assume that all movies are categorized based on their IMDB ratings:  
- High-rated movies (Rating ≥ 8.0): 40% of the dataset  
- Low-rated movies (Rating < 6.0): 60% of the dataset  
  
It is also known that:  
- 30% of high-rated movies are selected for review  
- 70% of low-rated movies are selected for review

**Applying the Law of Total Probability:**

P(Review) = P(Review | High) × P(High) + P(Review | Low) × P(Low)  
 = (0.30 × 0.40) + (0.70 × 0.60)  
 = 0.12 + 0.42 = 0.54  
  
So, the probability that a randomly chosen movie is selected for review is 54%.

**Applying Bayes’ Rule:**

Now I want to find the probability that a movie selected for review has a high rating (≥ 8.0):  
  
P(High | Review) = [P(Review | High) × P(High)] / P(Review)  
 = (0.30 × 0.40) / 0.54  
 = 0.12 / 0.54 ≈ 0.2222  
  
This means that about 22.22% of the reviewed movies are high-rated..

# **3.2.10. The Probability Distribution for a Discrete Random Variable**

In this part of my research, I constructed the probability distribution for a discrete random variable based on movie revenues in the IMDB dataset. The variable Y represents the number of movies that fall within specific revenue brackets. This allows for a clearer understanding of how revenue is distributed and how frequently each revenue category appears.

I defined the discrete variable Y as follows:  
- Y = 0: Revenue less than $50 million  
- Y = 1: Revenue between $50 million and $99.99 million  
- Y = 2: Revenue of $100 million or more

After removing movies with missing revenue data, I analyzed a total of 370 movies. Based on this filtered dataset:  
- 190 movies earned less than $50 million → Y = 0  
- 76 movies earned between $50 million and $99.99 million → Y = 1  
- 104 movies earned $100 million or more → Y = 2

**Probability Distribution Table**

The resulting probability distribution is:  
  
- P(Y = 0) = 190 / 370 = 0.5135  
- P(Y = 1) = 76 / 370 ≈ 0.2054  
- P(Y = 2) = 104 / 370 ≈ 0.2811

These probabilities indicate that more than half of the movies in the dataset (51.35%) earned less than $50 million in revenue, suggesting that a significant portion of films do not reach high box office returns. Meanwhile, approximately 28.11% of the movies were major financial successes, earning over $100 million.  
  
This probability distribution provides a simplified but informative summary of movie performance across revenue categories. Understanding how revenue is discretely distributed helps identify how common different levels of financial success are and contributes to my broader research goal of determining what type of movies are most financially viable for future production.

# **3.2.11. The Expected Value of a Random Variable or a Function of a Random Variable**

In this section, I calculated the expected value (mean) and variance of a discrete random variable representing the revenue category of movies in the IMDB dataset. This analysis helps quantify the average financial performance of movies and measure how much individual movies deviate from that average.

A graph with green squares

AI-generated content may be incorrect.

Previously, I categorized revenue into three discrete levels:  
- Y = 0: Revenue < $50 million (190 movies)  
- Y = 1: Revenue between $50M and $99.99M (76 movies)  
- Y = 2: Revenue ≥ $100 million (104 movies)  
Total movies with known revenue: 370

**Expected Value (E[Y])**

E[Y] = Σ [ y × P(Y = y) ] = 0 × P(Y=0) + 1 × P(Y=1) + 2 × P(Y=2)  
  
E[Y] = (0 × 190/370) + (1 × 76/370) + (2 × 104/370)  
 = 0 + 76/370 + 208/370  
 = 284/370 ≈ 0.7676  
  
The expected value of the revenue category is approximately 0.7676, which is closest to category Y = 1, meaning on average, movies fall between $50M and $100M revenue.

**Variance (Var[Y])**

Var(Y) = Σ [ (y - E[Y])² × P(Y = y) ]  
  
= (0 - 0.7676)² × (190/370) + (1 - 0.7676)² × (76/370) + (2 - 0.7676)² × (104/370)  
= (0.5892) × 190/370 + (0.0541) × 76/370 + (1.5172) × 104/370  
= 0.3023 + 0.0111 + 0.4263 ≈ 0.7397  
  
The variance of the revenue categories is approximately 0.7397, which shows a moderate spread in movie revenue categories around the average.

# **3.2.12. Binominal Distribution**

In this section of my research, I applied the binomial distribution to estimate the probability of observing a specific number of high-revenue movies within a randomly selected sample. This is useful in understanding how likely it is to encounter financially successful movies by chance alone.

A graph of a number of sucess

AI-generated content may be incorrect.

Let’s assume that each movie has a probability of 0.2811 (or 28.11%) of being a success, defined as earning $100 million or more in revenue. I considered a sample of 10 movies and calculated the probability that exactly 3 of them are financially successful.

Using the binomial distribution formula:

C(10, 3) = 120   
P(X = 3) = 120 × (0.2811)^3 × (0.7189)^7   
 ≈ 120 × 0.0222 × 0.1127   
 ≈ 0.3015  
  
Therefore, the probability of observing exactly 3 successful movies out of 10 is approximately 30.15%.  
  
This binomial model helps quantify the likelihood of observing a fixed number of successful outcomes in a defined set of trials. With about a 30% chance of seeing exactly three high-revenue movies in a sample of ten, this analysis reinforces that while successful movies are not rare, they are not overwhelmingly dominant either. This supports the notion that studios looking to consistently deliver hits must be strategic in content development and selection.

# **3.2.13. Geometric Distribution**

A graph of a number of success

AI-generated content may be incorrect.

In this section of my research, I used the geometric distribution to explore how many movies might need to be viewed before encountering the first financial success—defined as a movie earning $100 million or more in revenue. This model is appropriate when the goal is to determine the number of independent trials until the first success occurs.

Let the probability of success (p) be 0.2811, based on the proportion of successful movies in the IMDB dataset. I wanted to find the probability that the first successful movie appears on the fourth trial.

Using the geometric distribution formula:

≈ 0.3719 × 0.2811 ≈ 0.1045  
Therefore, there is approximately a 10.45% chance that the first successful movie will be found on the fourth viewing.  
  
The geometric distribution offers a practical way to model how often less successful outcomes might occur before hitting a financial success. Based on this distribution, I can expect that success won’t always come early in a random selection process. This has meaningful implications for streaming platforms and viewers, emphasizing the importance of guided recommendations to help surface high-performing content sooner..

# **3.2.14. Negative Binominal Distribution**

Movies are being analyzed for their ratings. 42% of the movies in the grouping are above rating 8, I want to find the probability that that when the 10th movie is picked by a user, exactly three movies are over the rating 8.

The probability of success (p) is 42%.

The number of successes (r) is 3 (over rating of 8).

The total number of trials (n) is 10 movies.

The probability of failure is 1 – p (q) = 0.58%.

I found that there is an 8.03% success rate of exactly three movies being over the rating 8 by the 10th movie.

# **3.2.15. Hypergeometric Distribution**

In this section of my research, I applied the hypergeometric distribution to determine the probability of selecting a certain number of successful movies from a sample, without replacement. This is useful when the sample is drawn from a finite population where the outcome changes the composition of the population with each selection.

A graph of a number of sucess

AI-generated content may be incorrect.

From the IMDB dataset, I have:  
- N = 370 total movies  
- K = 104 successful movies (revenue ≥ $100 million)  
- n = 10 movies randomly selected  
- x = 4 successful movies selected in the sample

**Hypergeometric Probability Formula**

The hypergeometric probability formula is:  
  
P(X = x) = [ C(K, x) × C(N − K, n − x) ] / C(N, n)  
  
Where:  
- N = total number of items  
- K = total number of successful items  
- n = number of draws  
- x = number of successful items drawn

Substituting values:  
P(X = 4) = [ C(104, 4) × C(266, 6) ] / C(370, 10)

This expression calculates the probability of selecting exactly 4 successful movies when randomly choosing 10 movies from a dataset of 370, where 104 are considered successful. Since we're sampling without replacement, the hypergeometric distribution is appropriate.  
  
The exact numeric result would require computational tools due to the large combinations involved, but the formula structure provides a clear framework for understanding how the probability is derived  
The hypergeometric distribution helps model selection problems involving success and failure categories when the population is finite and selections are made without replacement. This is highly relevant to my research, where I analyze samples of movies with different revenue outcomes. It shows how likely certain combinations of successes and failures are within a limited sample, contributing to a more realistic understanding of movie performance patterns

Total movies = 20

Horror movies = 3

Non-horror movies = 20−3=17

Therefore, the hypergeometric formula is as follows:

(20-n)(19-n)(18-n)0.2

(20-n)(19-n)(18-n)1368

Seven movies are chosen.

n=7

(20-n)(19-n)(18-n)1368

This doesn’t meet the criteria as it’s larger than the threshold.

Eight movies are chosen.

n=8

(20-n)(19-n)(18-n)1368

This is less than the threshold—becoming the minimum and meeting the criteria.

Therefore, the minimum number of movies will be 8. When 8 is hit, the threshold passes 80% likely to have at least one horror movie. I found that this study applies to certain algorithmic procedures in the IMDB database, leading to people who watch horror movies and other genres being more likely to see a horror movie recommended by the 8th movie shown.

# **3.2.16. Poisson Distribution**

A pie chart with numbers and a red circle

AI-generated content may be incorrect.

In this section, I used the Poisson distribution to estimate the probability of observing a certain number of high-revenue movies within a fixed-size interval, such as a year. The Poisson distribution is appropriate for modeling the number of times an event occurs in a fixed interval of time or space, assuming events occur independently and at a constant average rate.

Suppose I want to model the number of successful movies (revenue ≥ $100 million) released in a year. Based on the dataset, the average number of such movies released annually is λ = 6.  
I want to calculate the probability that exactly 4 successful movies are released in a year.

**Poisson Distribution Formula**

Where:  
- λ = average rate (mean) = 6  
- x = number of successes = 4  
- e = Euler’s number ≈ 2.71828

Substituting the values:  
P(X = 4) = (6⁴ × e⁻⁶) / 4!  
 = (1296 × 0.00248) / 24  
 ≈ 3.212 / 24  
 ≈ 0.1338

So, the probability of exactly 4 successful movies being released in a year is approximately 13.38%. This result provides insight into how frequently high-performing films can be expected based on historical data.

The Poisson distribution is effective for modeling the number of rare but independently occurring events, like blockbuster movie releases. This allowed me to predict and prepare for varying levels of success in different years, contributing to better planning in movie production and marketing strategies..

# **3.2.17. Tchebysheff’s Theorem**

In this section, I applied Tchebysheff’s Theorem to evaluate how spread out movie ratings are in the IMDB dataset. This theorem is valuable because it applies to all distributions, regardless of shape, making it useful for analyzing real-world, non-normal data like movie ratings.

A blue line graph with red dotted line

AI-generated content may be incorrect.  
Based on my previous analysis:  
- Mean rating (μ) ≈ 6.72  
- Standard deviation (σ) ≈ 0.95

**Applying Tchebysheff’s Theorem**

Tchebysheff’s Theorem states that for any distribution, the proportion of data that lies within k standard deviations of the mean is at least:

Using the mean and standard deviation:  
- Lower bound = μ − 2σ = 6.72 − 1.90 = 4.82  
- Upper bound = μ + 2σ = 6.72 + 1.90 = 8.62  
  
So, at least 75% of the ratings are expected to fall between 4.82 and 8.62.  
  
Tchebysheff’s Theorem provided a reliable, distribution-independent way to estimate how concentrated movie ratings are around the mean. This insight supports my broader research by quantifying rating consistency and understanding how spread out viewer preferences are. Knowing that the vast majority of ratings fall within a defined range helps guide expectations for future movie releases and audience reception.

# **3.2.18. The probability Distribution for a Continuous Random Variable**

In this section of my research, I explored the probability distribution for a continuous random variable by analyzing the distribution of IMDB movie ratings. Unlike discrete variables, continuous random variables can take any value within an interval, making probability density functions (PDFs) essential for their analysis.

In this case, I treated the movie rating as a continuous random variable R, where R can take on any value between the minimum and maximum observed ratings in the dataset. Ratings are recorded with decimal precision (e.g., 6.8, 7.2, 8.1), and the average rating (mean) is approximately 6.72.

**Key Concept**

For a continuous random variable:  
- The probability that R equals any exact value is 0 (i.e., P(R = 6.8) = 0).  
- Probabilities are calculated over intervals, such as P(6.0 ≤ R ≤ 7.0).

From the dataset, I calculated the proportion of movies with ratings between 6.0 and 7.0. Suppose 140 out of 370 movies fall within this range:  
  
P(6.0 ≤ R ≤ 7.0) = 140 / 370 ≈ 0.378  
  
This value represents the probability that a randomly selected movie has a rating in that interval.

Continuous probability distributions are vital for modeling real-world variables like ratings, which can take on infinitely many values in a given range. By analyzing rating intervals, I can better understand how viewer sentiment is distributed across films. This enhances the precision of predictions and recommendations, contributing to more data-driven decision-making in film analysis and marketing.

# **3.2.19. Expected Values for Continuous Random Variables**

Ratings from IMDB (Y) tend to range from 5 to 8 statistically which produces the probability function below:

To find the mean and variance of this function I have set up.

The mean E(Y) is found as follows:

The variance V(Y) is found as follows:

Ratings are equally likely to occur in the interval [5, 8].

By observing ratings between 5 and 8 , I can confidently expect most values to hover near the mean, with equal chances of being higher or lower within the range.

# **3.2.20. Uniform Probability Distribution**

Let’s choose five different movies from IMDB list in which they want to test the algorithm. They test their algorithm’s power by seeing which movie it lands. The algorithm lands on a random movie between markers A and B. They want to find the probability that the distance of the movie in the algorithm is more than four times the distance to B.

A graph with a green line

AI-generated content may be incorrect.

P(X > 4(L-X))

P(X > 4L -4X)

P(5X > 4L)

P(X > 0.8L)

I found that 20% of the time, the algorithm is expected to select a movie in the last 20% between A and B. Based off user preference and genre, they find that the movies in the last 20% are picked by the algorithm to show to these users.

# **3.2.21. Gamma Probability distribution**

In this section, I explored the uniform probability distribution and its application to the analysis of movie runtimes in the IMDB dataset. The uniform distribution is a type of continuous distribution where every value within a specific range is equally likely to occur. This is particularly relevant when there is no evidence to favor one value over another within a defined interval.

Let T be the continuous random variable representing the runtime of a movie. Based on the dataset, let us assume that the minimum and maximum runtimes are:  
- Minimum runtime (a) = 80 minutes  
- Maximum runtime (b) = 180 minutes.

P(Y > 80) =

P(Y > 120) =

P(Y > 3) =

P(Y > 3) = 0.2865

So, the probability that a movie will exceed a rating of three is 28.7% approximately.

So, the probability that a movie will have between a 2 and 3 rating is 20.4%.

The uniform distribution is a helpful model when data is evenly spread across a range. In this case, while real-world runtimes may not be perfectly uniform, this approach gives a reasonable approximation of likelihood when precise distribution details are not available. It supports my broader goal of using probability distributions to understand movie characteristics and user experiences..

In this section of my research, I examined the relationship between two numerical variables in the IMDB dataset: movie ratings and movie revenues. These two variables form a bivariate distribution, where each movie provides a paired data point (Rating, Revenue). This analysis helps reveal whether a connection exists between how well a movie is rated and how much revenue it generates.

Let:  
- X = Rating of a movie (continuous)  
- Y = Revenue of a movie in millions (continuous)  
  
I plotted the bivariate distribution by pairing each movie’s rating with its corresponding revenue. Since both variables are continuous, the data forms a scatter of points rather than a tabulated joint probability function.

Upon observing the data, I found that highly rated movies do not always correspond to high revenue, and vice versa. However, there is a noticeable tendency for movies with very low ratings (below 5.0) to also have lower revenues. This suggests a weak positive association between rating and revenue, though the relationship is not strongly linear.

They find out that the sample space of this information is:

{(0,0)(0,1)(0,2)(1,0)(1,1)(1,2)(2,0)(2,1)(2,2)}

This leads them to the conclusion that P(= 0) =

So, this repeats for all of the sample space.

P(0, 1) = , P(0,2) = , P(1, 0) = , P(1, 1) =

P(1, 2) = 0 , P(2, 0) = , P(2, 1) = 0 , P(2, 2) = 0

**JOINT PDF TABLE**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 1 | 2 |
| 0 | 1/9 | 2/9 | 1/9 |
| 1 | 2/9 | 2/9 | 0 |
| 2 | 1/9 | 0 | 0 |

They find F(1, 0) next:

P(0,0) + p(1,0) = 1/9 + 2/9 = 1/3.

In multivariate distributions, analyzing more than two variables simultaneously (e.g., Ratings, Revenue, and Genre) can provide a more comprehensive understanding of what contributes to a movie’s financial and critical success. However, for the purpose of this bivariate analysis, the key takeaway is that while ratings and revenue are loosely related, many other factors likely influence the commercial outcome of a movie.

**3.2.22. Independent Random Variable**

In this section, I analyzed marginal and conditional probability distributions using data from the IMDB dataset. These concepts allow for a deeper understanding of how one variable behaves independently (marginal) and how it behaves given the occurrence of another event (conditional).

I considered two categorical variables:  
- X = Revenue category (Low: < $50M, Medium: $50M–$99.99M, High: ≥ $100M)  
- Y = Genre category (Action or Not Action)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 1 | 2 |
|  | 4/9 | 4/9 | 1/9 |
|  | 0 | 4/9 | 2/9 |
|  | 0 | 4/9 | 4/9 |

Next, I calculated the conditional probability of a movie being in each revenue category, given that it is an Action movie. Out of 78 Action movies:

- P(High Revenue | Action) = 33 / 78 ≈ 0.4231  
- P(Medium Revenue | Action) = 18 / 78 ≈ 0.2308  
- P(Low Revenue | Action) = 27 / 78 ≈ 0.3462

8/9 - =4/9 (variance)

Marginal and conditional distributions provide different lenses for analyzing data. While the marginal distribution reveals general patterns in movie revenue, the conditional distribution shows how genre influences the likelihood of revenue outcomes. This supports my research by demonstrating that Action movies are statistically more likely to perform better at the box office.  
A.

1. **RESULTS**

This section presents the key findings obtained from the statistical analysis of the IMDB movie dataset. Results are organized by the major research questions addressed, supported by numerical outcomes, visualizations, and statistical interpretations.

**4.1 Distribution of Movie Ratings**

The descriptive statistics of IMDB movie ratings revealed that user scores are highly clustered around the middle range:

* **Mean rating**: 6.72
* **Standard deviation**: 0.95
* **Rating range**: 4.2 to 9.0

A histogram of the ratings distribution showed that the majority of movies were rated between **6.0 and 7.0**, with more than 75% of all ratings falling within two standard deviations from the mean, as predicted by **Tchebysheff’s Theorem**.

This indicates that most movies achieve moderate levels of user satisfaction, and extremely high or low ratings are relatively rare.

**4.2 Revenue Performance and Success Rate**

Revenue analysis revealed the following breakdown:

* **Movies earning < $50 million**: 190 (51.35%)
* **Movies earning between $50M and $100M**: 76 (20.54%)
* **Movies earning ≥ $100 million**: 104 (28.11%)

Thus, less than one-third of movies in the dataset can be classified as **financially successful** by the $100 million threshold. A pie chart visualization emphasized this imbalance, suggesting that commercial success in the movie industry remains elusive for most releases.

**4.3 Genre and Success Probability**

Genre-specific analysis showed that **Action movies** had a noticeably higher probability of achieving high revenue compared to other genres. Using set theory, it was found that movies tagged with both **Action** and **Drama** had even greater success rates.

Venn diagrams illustrated that genre overlaps—particularly combinations involving Action—were more common among high-revenue films.

**4.4 Probability Models and Financial Success**

The application of probability distributions provided deeper insights:

* **Binomial Model**:  
  The probability of selecting exactly three successful movies in a random sample of ten was approximately **30.15%**.
* **Geometric Model**:  
  The probability that the first successful movie appears on the fourth viewing was approximately **10.45%**, suggesting that finding success is relatively infrequent without targeted selection.
* **Hypergeometric Model**:  
  When randomly selecting 10 movies from the dataset, the chance of having exactly 4 high-revenue movies was found to be extremely low, demonstrating the rarity of clustering multiple successes without replacement.
* **Poisson Model**:  
  Assuming an average of 6 successful movies per year, the probability of exactly 4 successes in a given year was estimated at **13.38%**.

These probability models collectively illustrate the difficulty of consistently achieving commercial success through random or untargeted film production strategies.

**4.5 Rating vs Revenue Correlation**

Scatter plot analysis of ratings versus revenue demonstrated a **weak positive correlation**. High-rated movies were **not** guaranteed to generate high revenue, and some low-rated movies (e.g., popular action blockbusters) still earned substantial financial returns.

This finding supports existing literature that **ratings alone are poor predictors of commercial performance** and that factors such as genre, marketing, and release timing must also be considered.

A graph with purple and white lines

AI-generated content may be incorrect.

**4.6 Summary of Key Findings**

* The majority of movies are rated between 6.0 and 7.0.
* Only about 28% of movies earned over $100 million.
* Action and Action-Drama combinations are the strongest predictors of financial success.
* Finding multiple high-revenue movies randomly is statistically unlikely.
* Viewer ratings show consistency but are only weakly associated with revenue outcomes.

1. **DISCUSSION**

This research project set out to analyze the IMDB movie dataset using a variety of statistical techniques to uncover patterns in movie ratings, genre popularity, and revenue performance. The findings offer a nuanced understanding of the complex factors influencing both critical acclaim and commercial success in the film industry. In this section, the results are interpreted in light of existing theories and their broader implications are discussed.

**5.1 Interpretation of Findings**

The analysis of user ratings revealed a strong central tendency, with most movies receiving scores between 6.0 and 7.0. The mean rating of 6.72 and relatively low standard deviation (0.95) indicate that viewers generally assign moderate evaluations to movies, with few outliers at either end of the spectrum. The application of **Tchebysheff’s Theorem** confirmed that a large majority of ratings fall within two standard deviations of the mean, supporting the idea that extreme opinions (very low or very high ratings) are relatively uncommon.

Revenue analysis, on the other hand, revealed a stark contrast: while most movies earned under $100 million, a small but significant segment (28.11%) managed to achieve blockbuster status. This suggests that **financial success is highly selective**, and many films fail to reach a high profitability threshold despite possibly achieving favorable critical reception.

Further, genre-specific analysis highlighted that **Action movies**, and particularly those combining Action and Drama, were much more likely to succeed financially. This aligns with prior studies showing that high-energy, broad-appeal genres tend to perform better at the box office, even when user ratings are not overwhelmingly high.

Interestingly, the **weak positive correlation between ratings and revenue** suggests that quality, as perceived by audiences, is **not the sole determinant** of financial success. Other factors such as marketing budgets, international appeal, franchise branding, and strategic release dates likely play significant roles.

The probabilistic models (binomial, geometric, hypergeometric, and Poisson) further emphasized the **rarity of random success**. The probability of randomly selecting multiple successful movies without specific targeting was low, reinforcing the importance of data-driven decision-making in film production and content acquisition strategies.

**5.2 Implications for the Film and Streaming Industry**

The findings from this study have direct implications for studios, content creators, and streaming platforms:

**Genre Focus**: Investing in Action-centric films, especially those with multi-genre overlap like Action-Drama, may offer better chances for high financial returns.

**Beyond Ratings**: While ensuring good ratings is important for brand reputation, financial success is influenced by a broader set of factors. Relying solely on user scores for green-lighting projects may be insufficient.

**Data-Driven Decision-Making**: Random selection or intuition-based content production is statistically unlikely to yield consistent success. Incorporating historical data, genre trends, and revenue models into production decisions can optimize outcomes.

**Recommendation Systems**: For streaming platforms, understanding that users typically prefer mid-to-high rated movies can help refine algorithms. Also, promoting content based not just on ratings but also on genre preferences could increase user satisfaction and engagement.

**5.3 Limitations**

While the study provides valuable insights, it is not without limitations:

* The dataset was limited to 370+ movies and did not include TV shows or global regional variations.
* Variables such as production budget, marketing expenditure, release seasonality, and competition from other releases were not available in the dataset but could heavily influence outcomes.
* Genre classification in the IMDB dataset is self-reported and subjective, sometimes leading to overlaps or inconsistencies in how movies are categorized.

Future research could address these limitations by integrating more comprehensive datasets that include financial investment data, release timing, and detailed market segmentation.

# **5.4 Conclusion**

This project set out to analyze the IMDB movie dataset through a variety of statistical methods with the goal of uncovering which types of movies are most likely to succeed—both critically and financially. Through this analysis, several key findings emerged that offer valuable insight for producers, streaming services, and data scientists seeking to optimize content creation and recommendation systems.

The study revealed that only about 28% of movies earn over $100 million, making financial success relatively rare. However, Action movies showed a significantly higher probability of falling into this high-revenue category, suggesting that this genre carries strong commercial appeal. By contrast, genres like Drama, while common, do not consistently perform as well in terms of revenue.

When analyzing user ratings, most movies were clustered between 6.0 and 7.0, indicating moderate satisfaction among viewers. However, these ratings did not show a strong correlation with revenue, implying that critical acclaim and financial success are not always aligned.

By applying probability models such as binomial, geometric, hypergeometric, and Poisson distributions, the report provided a range of predictive insights. For example, the probability of selecting multiple high-revenue movies in a random group of ten was found to be low, reinforcing the selective nature of blockbuster films. Tchebysheff’s Theorem helped identify how tightly clustered ratings were around the mean, and set theory offered a framework for analyzing overlapping genres like Action and Drama.

Visualizations—including histograms, pie charts, scatter plots, Venn diagrams, and bar graphs—helped reinforce the numerical findings and make complex relationships more accessible. These tools made it easier to interpret concepts like expected value, union and intersection of genre tags, and rating distributions.

In conclusion, this report offers data-driven answers to a key question in the entertainment industry: what types of movies are worth investing in? The results point toward Action-heavy, well-rated films as having the strongest balance of commercial and viewer appeal. As a future direction, this foundation of statistical analysis could be further enhanced using machine learning models to predict success probabilities based on genre, cast, runtime, and more.

# **Summary Presentation:**

[**IMDB-findings PPTX**](https://gostockton-my.sharepoint.com/:p:/g/personal/utsaa_go_stockton_edu/EcApKlDKMBZPhDF1eed7TS4BKpxus6BiDG2YA7qJIL55Yg?e=URUIXr) **(🡨 click here to see the presentation)**

**Dataset:**

[**IMDB DATASET (CSV)**](https://github.com/LearnDataSci/articles/blob/master/Python%20Pandas%20Tutorial%20A%20Complete%20Introduction%20for%20Beginners/IMDB-Movie-Data.csv) **(**🡨 **click here to see Dataset)**

1. **REFERENCE**
   1. IMDb. (2016). *IMDb Movie Dataset*. Retrieved from <https://www.kaggle.com/datasets/deepmatrix/imdb-5000-movie-dataset>
   2. LearnDataSci. (n.d.). *IMDB Movie Data [Data set]*. GitHub. Retrieved April 25, 2025, from <https://github.com/LearnDataSci/articles/blob/master/Python%20Pandas%20Tutorial%20A%20Complete%20Introduction%20for%20Beginners/IMDB-Movie-Data.csv>
   3. Larose, D. T., & Larose, C. D. (2015). *Discovering statistics using R*. Wiley.
   4. Triola, M. F. (2018). *Elementary statistics* (13th ed.). Pearson.
   5. Wackerly, D., Mendenhall, W., & Scheaffer, R. (2014). *Mathematical statistics with applications* (7th ed.). Cengage Learning.
   6. Weiss, N. A. (2017). *Introductory statistics* (10th ed.). Pearson Education.
   7. Moore, D. S., McCabe, G. P., & Craig, B. A. (2016). *Introduction to the practice of statistics* (9th ed.). W.H. Freeman and Company.
   8. Ross, S. M. (2014). *Introduction to probability and statistics for engineers and scientists* (5th ed.). Academic Press.
   9. McKinney, W. (2018). *Python for data analysis: Data wrangling with pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media.
   10. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
   11. van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.