Insights from IMDB Movie Data

Web Analytics

Uzair Ansar

1. Executive Summary

The goal of the project was to offer insights from movie data over the years and those that were relevant to the COVID-19 pandemic challenge and evolution. It was designed to pinpoint and quantify the variables that drove movie success, which included IMDb ratings and box-office sales. The report would help entertainment stakeholders, including film studios, production houses, and online streaming services, by providing insights that can be used to better manage their budgets, advertising, and content production.

We used data sourced from IMDb and Box Office Mojo; two reliable sites that provide movie-specific information. Titles, genres, ratings, production budgets, box office collections, and directors were a part of the dataset. We implemented this by scraping the data with Python libraries, parsing it, and presenting it with Tableau to find patterns and correlations between variables. It particularly highlighted the effects of COVID-19, which severely disrupted movie production and distribution and accelerated the move towards streaming.

We showed how specific elements, including budgets, genres, and directors, affected the movie's performance before and after the pandemic. Through the use of data analytics, this effort illustrates the value of good information in reducing risk and increasing profits in entertainment. Predictive modeling was not applied, but these findings are a basis for further research using forecasting methods.

2. Business Goal Analysis

The broader business purpose of this project was to discover how data insights could help improve decisions on movie production and distribution, particularly amid the uncertainties around the COVID-19 pandemic. Production and marketing decisions in the entertainment business have always been based heavily on instinct and subjective insight. But there are major downsides to such old-fashioned ways of doing things, especially when audiences change and outside forces intervene.

The original purpose of the project was to identify the key determinants of film success, in terms of box office and IMDb rating. Analyzing trends by production budget, genre, and other movie characteristics, the work sought to identify key elements that drive commercial and critical success. These insights allow studios to better focus resources, increasing spending on casting, marketing and post-production.

The second was adapting to the shift caused by the COVID-19 pandemic. With movie theaters shutting down and films being postponed, consumers began to gravitate toward streaming platforms. The project attempted to compare movie results during the pandemic with pre-

pandemic trends in order for studios to spot new revenue streams for content distribution. Audience insights during this awkward stage are particularly relevant to inform strategies in the post-pandemic landscape.

With the use of web analytics, the project solved the need for hard-headed, data-driven analysis that replaces guesswork with quantifiable data. These findings not only lower the potential for poor movies but also give studios the ability to capitalize on the trends that most appeal to consumers.

3. Dataset Description

For the project, the data we used were pulled from IMDb and Box Office Mojo, two of the most widely reputable movie data sources. The data incorporated metadata of movies from 1995 to the present, primarily those that had been made before, during, and after the COVID-19 pandemic. The dataset contained 6,000 records.

Release	Gross	Theatres	Total Gross	Distributor	Genres	Rating	Budget	Director	Release Date
Inside Out 2	652980194	4440	652980194	Walt Disney Studios Motion Pictures	Drama	7.6	200000000	Kelsey Mann	2024-06-14
Deadpool & Wolverine	636745858	4330	636745858	Walt Disney Studios Motion Pictures	Comedy	7.7	200000000	Shawn Levy	2024-07-26
Despicable Me 4	361004205	4449	361004205	Universal Pictures	Comedy	6.2	100000000	Chris Renaud	2024-07-03
Beetlejuice Beetlejuice	294072781	4575	294072781	Warner Bros.	Comedy	6.8	100000000	Tim Burton	2024-09-06
Dune: Part Two	282144358	4074	282144358	Warner Bros.	Drama	8.5	190000000	Denis Villeneuve	2024-03-01
Twisters	267762265	4170	267762265	Universal Pictures	Thriller	6.5	155000000	Lee Isaac Chung	2024-07-19
Godzilla x Kong: The New Empire	196350016	3948	196350016	Warner Bros.	Thriller	6.1	135000000	Adam Wingard	2024-03-29
Kung Fu Panda 4	193590620	4067	193590620	Universal Pictures	Comedy	6.3	85000000	Mike Mitchell	2024-03-08
Bad Boys: Ride or Die	193573217	3885	193573217	Sony Pictures Releasing	Comedy	6.6	100000000	Adil El Arbi	2024-06-07
Kingdom of the Planet of the Apes	171130165	4075	171130165	20th Century Studios	Drama	6.9	160000000	Wes Ball	2024-05-10
It Ends with Us	148518266	3839	148518266	Sony Pictures Releasing	Drama	6.5	25000000	Justin Baldoni	2024-08-09
The Wild Robot	140727420	3997	138727420	Universal Pictures	Sci-Fi	8.3	78000000	Chris Sanders	2024-09-27
A Quiet Place: Day One	138930553	3708	138930553	Paramount Pictures	Drama	6.3	67000000	Michael Sarnoski	2024-06-28
Venom: The Last Dance	133825476	4131	129825476	Sony Pictures Releasing	Thriller	6.2	120000000	Kelly Marcel	2024-10-25
Wicked	114000000	3888	114000000	Universal Pictures	Romance	8.1	150000000	Jon M. Chu	2024-11-22
Ghostbusters: Frozen Empire	113376590	4345	113376590	Sony Pictures Releasing	Comedy	6.1	100000000	Gil Kenan	2024-03-22
IF	111149917	4068	111149917	Paramount Pictures	Drama	6.5	110000000	John Krasinski	2024-05-17
Alien: Romulus	105313091	3915	105313091	Walt Disney Studios Motion Pictures	Thriller	7.2	80000000	Fede Alvarez	2024-08-16
Bob Marley: One Love	96893170	3597	96893170	Paramount Pictures	Drama	6.2		Reinaldo Marcus Green	2024-02-14
The Fall Guy	92900355	4008	92900355	Universal Pictures	Drama	6.9	125000000	David Leitch	2024-05-03
The Garfield Movie	91956547	4108	91956547	Sony Pictures Releasing	Comedy	5.7	60000000	Mark Dindal	2024-05-24
Wonka	85272410	4213	218402312	Warner Bros.	Comedy	5.7		Timothe Chalamet	2024-12-15
Longlegs	74346140	2850	74346140	Neon	Thriller	6.7		Osgood Perkins	2024-07-12
Migration	73202330	3839	127306285	Universal Pictures	Comedy	5.5	70000	Joshua Philipp	2024-12-22
Mean Girls	72404248	3826	72404248	Paramount Pictures	Comedy	5.6	36000000	Samantha Jayne	2024-01-12

The dataset contained a few parameters that were fundamental to assessing movie performance: The movie title was used as the main identifier for each record, and this served as context for other features. Release year allowed the underlying trends in film success to be mapped out over time. The genre of each film, which was often the key factor in a movie's success, also appeared, making performance comparisons between films within drama, action, comedy, and thriller. The IMDb rating stood for audience satisfaction and engagement and provided a gauge of how well a movie was perceived by the public.

The production budget and the box office grosses were critical parameters in order to get a handle on the economics of filmmaking. The intersection of these two features offered a return on investment data for studios to use in their production and marketing budgets. More variables, including director and distributor, were incorporated to assess the impact of leadership and distribution techniques on film results.

It also contained pandemic-induced delays in release and box office losses. Data preprocessing involved missing values, eliminating inconsistent values, and merging datasets from both sides. It also assured the validity and integrity of the final dataset before it was imported into Tableau for processing and visualization.

4. System Design
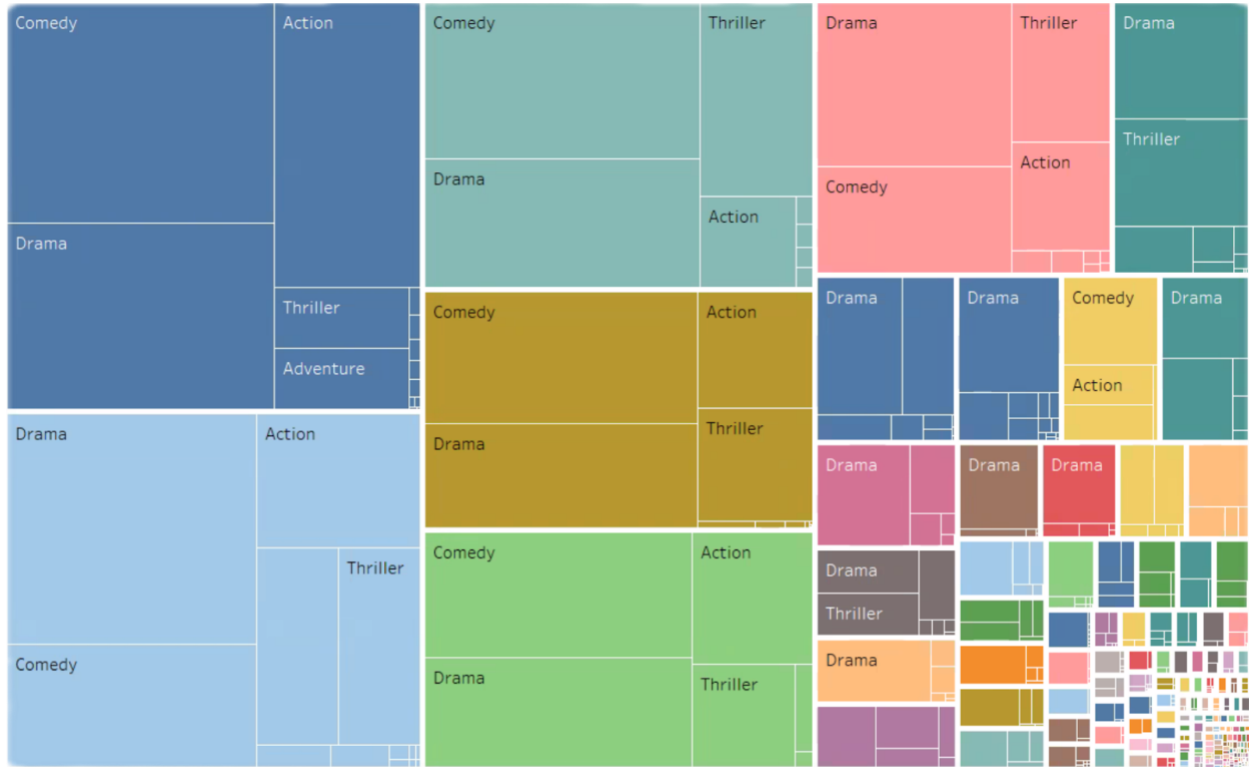
The Tableau analysis generated some helpful visualizations to visualize patterns in the data. The treemap graph, which categorised movies by genre, revealed that the most common genres were drama, action and thriller. Drama looked like the most common genre because it was always popular and often staged.
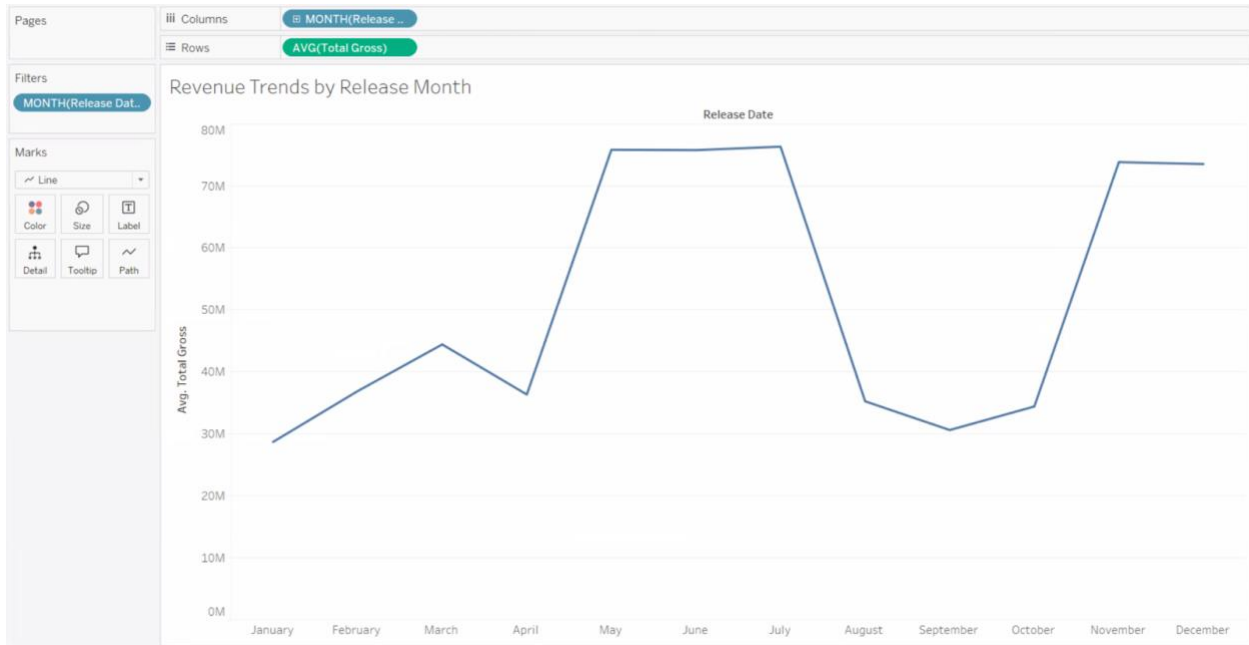
The line chart of box office earnings per release month showed that average box office revenues fluctuated year over year. Major revenue spikes came in the summer months of May, June, and July — the traditional blockbuster release window. Another peak during December also suggests that holiday movies also perform well due to heightened consumer attention around Christmas time. A significant drop in revenue in September and October refers to months where fewer high-grossing films are produced, perhaps because audiences are less demanding.
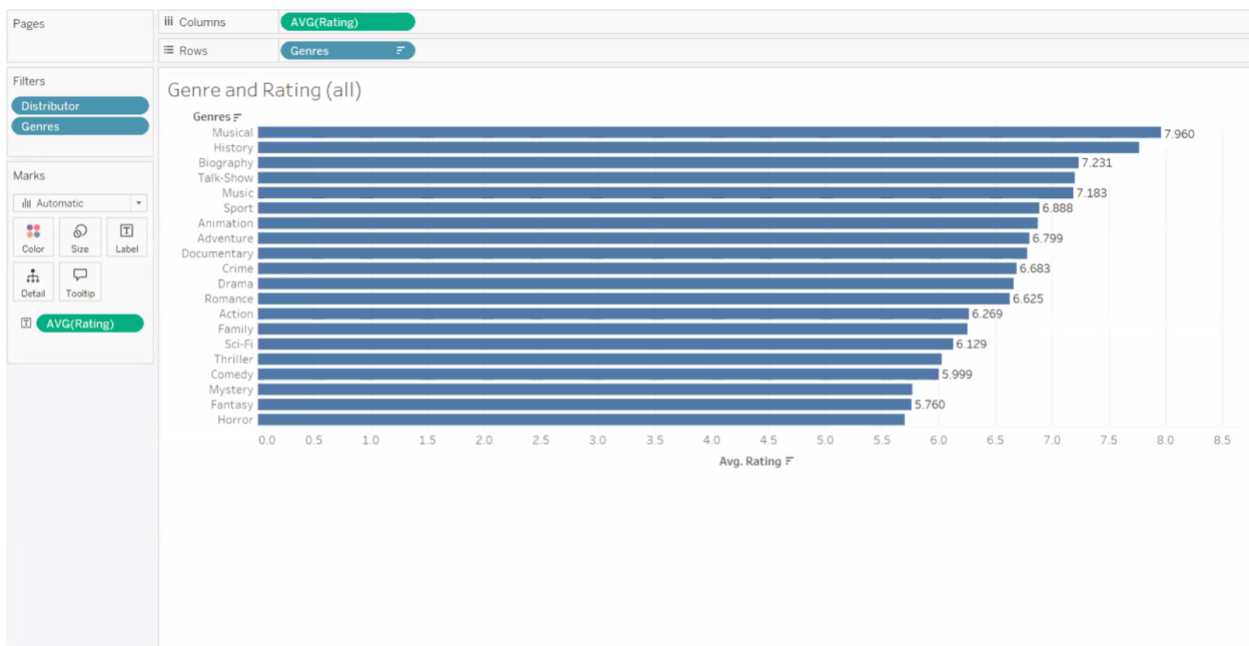

The bar graph of genres and their associated IMDb scores helped to understand the popularity of those titles. History, biography, and musical had the highest average ratings (with history taking the lead with 7.96). By comparison, horror and romance rated relatively low, showing

ambivalent reviews. It indicates that movies within specific or highly rated genres typically score higher on the quality scale, while mainstream genres such as action and thriller tend to attract more mass audiences even though they receive lower critical reviews.
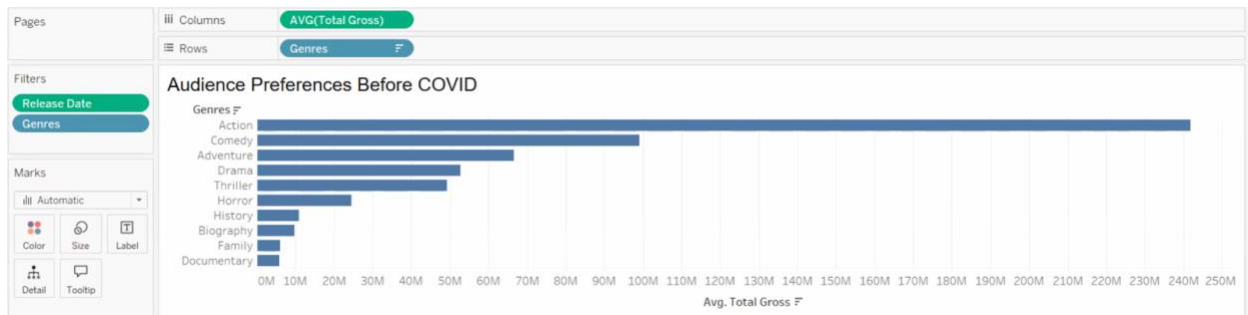

Comparing viewership levels before and after COVID-19 revealed a stark change in the level of viewership. Action and adventure films predominated overall gross revenue before the pandemic and mirrored audiences' appetite for massive theatrical spectacles. During the pandemic, however, comedy and family came into their own, perhaps because of the shift towards streaming services and the need for less heavily populated content during lockdowns.
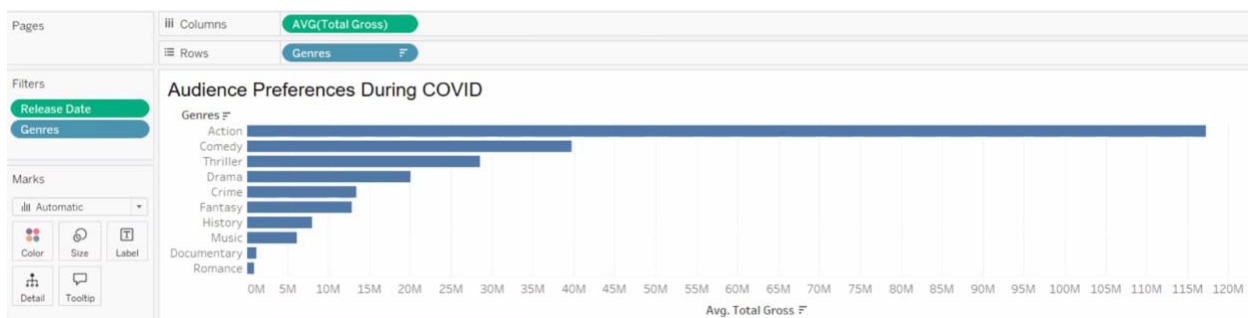
Comparing weekend to weekday performances, films that were released on weekends collected a significantly larger average gross than movies that were released on weekdays. This outcome is in line with moviegoing behaviour, which is that on weekends the majority will go to the movies.

The line graph showing production budgets over time revealed a constant growth in budgets from the late 1990s through 2019 as film productions grew in scale and expense. But in 2020, it went into decline as the COVID-19 pandemic hit. This decline reflects the studios' struggles financially at this time, and the rebound they experienced over the next few years.

Revenue over the Years
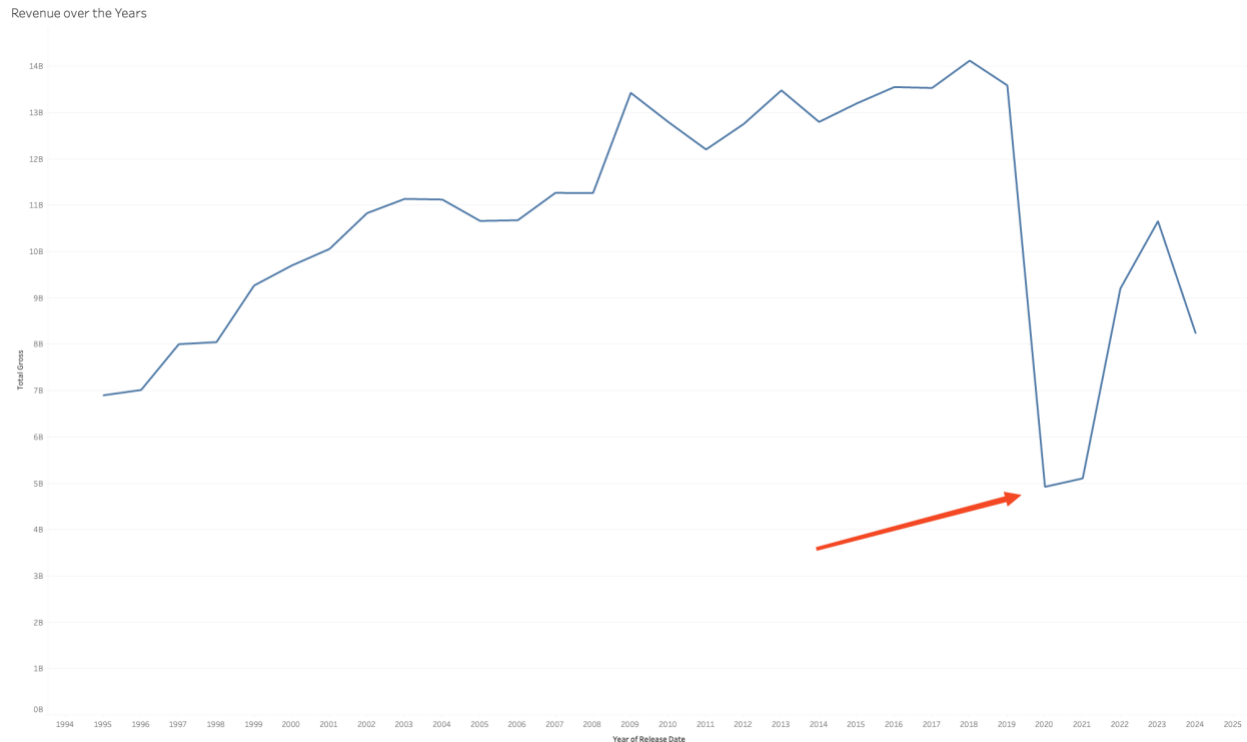
This line graph shows total revenue for each year from 1995 to 2024. Yearly revenue has been on a steady incline, however we notice a sudden downward trend in 2020 and 2021. This is most likely due to lockdowns during the Covid 19 pandemic. With theatres closed and audiences staying at home, box office revenues took a massive hit. It is interesting to note that revenues have not climbed back to pre-Covid levels. This may be due to a number of reasons. Most likely, audiences got accustomed to staying at home and streaming movies on providers like Netflix, Hulu, Amazon and HBO. Therefore it is safe to assume that these online streaming providers have captured a significant portion of the market and it may be a good idea for distributors to heavily consider these platforms as a part of their revenue strategy.

5. System Implementation

Project delivery was organized in a gradual fashion, beginning with data collection and ending with interactive visualizations. Every step involved a mix of tools and methods that ensured the data was robust, clean, and ready for analysis.

The data collection phase started with extracting movie information from IMDb and Box Office Mojo. We scraped the relevant data — movies, release dates, genres, production costs, and box office revenue using the Python package BeautifulSoup. Sleep timers were added to the scraping operation to stay within the bounds of the site's access controls and not overload the servers. Scraping was not that simple; it took about six weeks to extract the data because of its size and precision.

6000 rows of data were first scraped from BoxOffice Mojo and then 6000 rows of data were scraped from Imdb. These data sets were then combined into a single dataframe for analysis.

The data was cleaned and preprocessed using Python's pandas library once the raw data had been downloaded. This step corrected missing values and data inconsistencies. Duplicates were removed, and null values were imputed or excluded when needed. Data formats were standardized for consistency. We also added calculated fields (revenue/production budget, etc) to get further insights.

The cleaned data was then imported into Tableau, a data visualization tool. Tableau was selected for its ability to build interactive dashboards and provide intuitive, intuitive views of large amounts of data. Tableau generated multiple visualizations to compare relationships between key variables. For instance, bar charts were used to show what genres were performing best in terms of average IMDb ratings, and scatter plots were used to plot production budgets against box office receipts. We designed time series tables to compare movie releases and revenues over the years to compare before, during, and after the pandemic.

Additionally, Tableau made it easy to create dashboards where stakeholders could engage with the data. Users might sort the output by year, genre or any other feature to detect patterns and find data for their needs. The graphs showed the pandemic effect on box office collections, showing how tastes evolved over time.

6. Evaluation

This project was judged on how well the system implementation went and how easily the results of the analysis were understood. Data preprocessing, analysis, and visualization helped to fulfill the project's mission of identifying what makes a movie successful, and showing trends during the COVID-19 outbreak. Tableau allowed us to build dashboards interactively, which was useful when presenting complicated interdependencies between variables in a simple way.

The analysis did a good job at breaking down observable patterns, like genre dominance, box-office trends, and the changing tastes of audiences in the context of the pandemic. These observations supported the project's commercial objectives by delivering data to stakeholders that could be used for production, marketing, and distribution decisions. For instance, knowing

which months generate the most revenue and what genres get a high share during these time periods enables studios to schedule release dates and target specific audiences in a more effective way.

Although the project achieved its objectives, there were some limitations. The research was retrospective and not predictive modelling which could further boost its utility by making predictions about how movies would perform going forward. In addition, the data did not include external factors, like streaming platform performance or world events, that might affect viewers' choices. Other versions of the project could address these shortcomings by using more sophisticated methods, like machine learning predictive analytics, and expanding the data set to include newer streaming data.

In all, the project helped to validate the importance of using data in the entertainment sector, setting the groundwork for work in the future that will help drive decisions and increase profits.

7. Conclusion and Future Direction

The Web Analytics Project was able to show just how much movie data can be used to inform entertainment decisions. Using a rich data set from IMDb and Box Office Mojo, we identified production budgets, genres, and viewership as the main drivers of movie success. With Tableau, we were able to create interactive visualizations that conveyed these insights in a simple and easy-to-understand way.

The project shows that COVID-19 destabilized the metrics used to measure movie quality and sped up the transition to streaming. Such results offer an initial template for future research, which may use predictive modeling to predict box office and rating revenues. Other parameters, like audience demographics and streaming platform performance, might add value to the analysis and give a better understanding of audience behavior.

In conclusion, this project illustrates how critical it is to use data-driven solutions to make better production and marketing decisions. When entertainment stakeholders replace instinct with quantifiable knowledge, they reduce risk, increase profitability, and respond to their audiences' changing tastes.

8. References

https://www.boxofficemojo.com/
https://www.imdb.com/