Determine film's box office success using DSS and Regression model

By

SAJIB SAHA

171-35-191

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Software Engineering.

Department of Software Engineering

DAFFODIL INTERNATIONAL UNIVERSITY

Semester Summer – Year-2021

APPROVAL

This Thesis titled " **Determine film's box office success using DSS and Regression model**",
submitted by **Sajib Saha, ID:171-35-191** to the department of Software Engineering, Daffodil
International University has been accepted as satisfactory for the partial fulfillment of the
requirements for the degree of B.Sc. in Software Engineering and approved as to its style and
contents.

BOARD OF EXAMINERS

Marzia

Date: 11-07-21

Ms. Marzia Ahmed

Lecturer   **Supervisor**

 Department of Software Engineering

Faculty of Science and Information Technology   Daffodil
International University

Dr. Imran Mahmud

Associate Professor and Head   **Chairman**

Department of Software Engineering

Faculty of Science and Information Technology   Daffodil
International University

Md.Shariful Islam

Lecturer   **Internal Examiner**

Department of Software Engineering

Faculty of Science and Information Technology

Daffodil International University

Nayeem Hasan

Lecturer   (Senior Scale)   **Internal Examiner**

Department of Software Engineering

Faculty of Science and Information Technology

Daffodil International University

THESIS DECLARATION

I, Sajib Saha do hereby declare that this report has been done by me under the supervision of Ms. Marzia Ahmed, Lecturer, Dept. of Software Engineering, Daffodil International University. We also declare that this report nor any portion of this report has been submitted elsewhere for the award of any degree.

Supervised By,

Ms. Marzia Ahmed

Marzia

 Date: 11-07-21

Lecturer
Department of Software Engineering
Faculty of Science and Information Technology   Daffodil
International University

Submitted By,

Sajib Saha

ID: 171-35-191
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

ACKNOWLEDGEMENT

TABLE OF CONTENTS

Contents   Page

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Film is a surprisingly successful medium in passing on dramatization and particularly in the summoning of feeling. The specialty of movies is extremely mind-boggling, requiring commitments from practically the wide range of various expressions just as endless specialized abilities. As a business adventure, offering anecdotal stories to huge crowds in theaters, film was immediately perceived as maybe the principal genuinely mass type of diversion. By the last part of the 1930s, film had become a significant mass media outlet. Almost everybody in the Western world went to the cinema and numerous basically once per week. Film had made conceivable a gigantic development in efficiency in media outlets. In this study, we have tried to answer the question, "What determines a film's success at the box office?" We also mention a decision support system to support film investment decisions before going to movie productions. This study proposes four attributes and their significance in film's success .We used linear regression model and write down successful films and try to find out "why these film became so popular and successful."

Keywords: Decision support system, film's success, linear regression model

CHAPTER 1: INTRODUCTION

1.1 Background:

Like other significant advancements like the vehicle, power, synthetic substances, and the plane, film arose in most Western nations simultaneously. As the main type of industrialized mass-diversion, it was all-unavoidable. From the 1910s onwards, every year billions of film tickets were sold and buyers who didn't consistently visit the film turned into a minority. In Italy, today barely huge in worldwide amusement, the entertainment world was the fourthbiggest fare industry before the First World War. In the downturn struck U.S., film was the 10th most beneficial industry, and in 1930s France it was the quickest developing industry, trailed by paper and power, while in Britain the number of film tickets offered rose to right around one billion every year (Bakker 2001b). In spite of this financial importance, regardless of its fast development and development, notwithstanding its articulated impact on a regular day to day existence of shoppers, and notwithstanding it's anything but an early instance of the industrialization of administrations, the monetary history of the entertainment world has scarcely been inspected.

For about the initial ten years of its reality, a film in the United States and somewhere else was chiefly a stunt and a device. Before 1896 the coin-worked Cinematograph of Edison was available at fairs and in diversion settings. Observers needed to toss a coin in the machine and look through glasses to see the film. The primary projections, from 1896 onwards, pulled in huge crowds. Lumière had a gathering of administrators who went all throughout the planet with the cinematograph and showed the photos in theaters. Following a couple of years, films turned into a piece of the program in vaudeville and at times in the performance center also. Simultaneously voyaging film arose: films that went around with a tent or portable theater and set up for business for a brief time frame in towns and towns. These contrasted from the Lumière administrators and others in that they provided food for the general, famous crowds, while the previous were more upscale pieces of theater programs, or a unique program for the bourgeoisie (Musser 1990: 140, 299, 417-20).

This entire time, which in the U.S. endured up to around 1905, was a period wherein film appeared to be only one of the numerous new designs, and it was not in any way sure that it would continue, or that it would be neglected or underestimated rapidly, for example, happened to the blast in skating arenas and bowling alleys at that point. This changed when Nickelodeons, fixed films with two or three hundred seats, arisen and immediately spread everywhere in the country somewhere in the range of 1905 and 1907. From this time onwards film changed into an industry by its own doing, which was unmistakable from different amusements, since it had its own structures and its own promoting. The rise of fixed films matched a gigantic development stage in the business all in all; film creation expanded incredibly, and film appropriation formed into a unique movement, regularly oversaw by huge filmmakers.

Nonetheless, until around 1914, other than the films, films likewise kept on being joined with a live diversion in vaudeville and different theaters (Musser 1990; Allen 1980).

In a good way, the film business may look pretty marvelous. Famous people and makers float down red floor coverings, grip their Oscars, and get-away in St. Barts—since they can. While there's a great deal of cash to be made in the entertainment world, the financial matters of making motion pictures are a long way from basic. Something you'll probably hear on the off chance that you stroll through the corridors of any film studio is "no one knows anything." And that is valid. General society can be whimsical, and the business is in motion. Pretty much any film is incredibly unsafe speculation, even a film featuring large-name entertainers and entertainers. As per the Motion Picture Association of America's (MPAA) Theatrical Market Statistics Report for 2019, the U.S. what's more, the Canadian film industry came in at $11.4 billion. All around the world, the movies for films hit $42.2 billion in 2019.It's anything but close to as direct as the beginning of film when a film would come out in theaters, make by far most of its incomes through ticket deals, and afterward vanish. Significant studios and independent movie producers the same currently go through a lot of their days searching for new wellsprings of income, since ticket deals are not, at this point the most important thing in the world for films. Lamentably, the end of most performance centers during mid-2020 makes different surges of pay more significant than ever.

Theater participation has been trying over ongoing years, making it considerably harder for studios and merchants to benefit from films. Normally, a segment of theater ticket deals goes to theater proprietors, with the studio and wholesaler getting the leftover cash.

Customarily, a bigger lump went to the studio during the initial few days of a film. As the weeks went on, the auditorium administrator's rate increased. A studio may make about 60% of a film's ticket deals in the United States, and around 20% to 40% of that on abroad ticket deals. The level of income an exhibitor gets relies upon the agreement for each film. Numerous agreements are expected to help auditorium support against films that lemon in the cinematic world. That is accomplished by giving auditoriums a bigger cut of ticket deals for such movies, so an arrangement may have the studio getting a more modest level of an inadequately performing film and a higher level of a hit film's take. You can see the protection filings for enormous auditorium chains to perceive the amount of their ticket income returns to the studios. Studios and wholesalers for the most part make more from homegrown income than from abroad deals since they get a bigger rate. In spite of this game plan, unfamiliar ticket deals turned out to be more significant in the mid-21st century.6 That is essential for the motivation behind why you see more science fiction, experience, dream, and superhuman films. Activity and enhancements require no interpretations. They're straightforward, regardless of whether you're in Malaysia or Montana. It is a lot harder to assemble an unfamiliar crowd for an independent satire.

1.2 Motivation of the Research:

The extension of the film business has been an overall marvel. As indicated by the yearly report from the Motion Picture Association of America, the worldwide film industry market came to $36.4 billion in 2014. In the United States, More than many movies are delivered with the expectation that they will get success. According to some websites, thirty-plus films are unsuccessful in every fifty films. For producers, it is not a good sign and it's quite a risky business too. MPAA ratings, budget, critic's review, promotion, rotten tomatoes score, Meta score, IMDb rating are important factors in the movie business. They are able to make an impact on movie success also at the same time they can hamper badly in a film business. In this study, we will try to figure out how much these factors make an impact on films' box office.

1.3 Research Objectives:

The purpose of the study is to achieve a general understanding in movie box office success.

- This study proposes four attributes and their significance in film's success.
- This study also explains the linear regression algorithm and its significance in our study.
- This study also shows top-rated movies and top-grossed movies.

1.4 Research Scope:

- Downloading a data set
- Using linear regression algorithm
- Using python and excel for the result
- Some important features of films and their significance.

1.5 Research Question:

- What are the important features of films?
- What are the significance of these features?
- What kind of algorithm needed to get the significant result?
- How to collect dataset for a selected topic and where**?**

1.6 Problem statement:

- ○ Work with important features
- ○ Able to achieve a good accuracy.
- ○ Able to show top-rated and top-grossed movies.
- ○ Linear regression algorithm and its significance in our study.

1.7 Thesis Organization:

In a certain section, the whole paper is organized. Where the relevant analysis is discussed in Section 2 to extract the conceptual framework. The Proposed methodology is presented in the section 3. I have clarified my results and final outcome in section 4 with the support of some graphs and network. In that section, all the outcomes have been described. A small summery of the whole research is presented as conclusion in section 5.

CHAPTER 2: LITERATURE REVIEW

The majority of the past studies in regards to the film business have had the logical nature, researching factors that influence the movies exhibitions of motion pictures. The soonest works incorporate the exploration led by Litman (1983). He has researched how the creation cost, pundits' appraisals, classification, wholesaler, discharge season, and principle entertainer's honor history are identified with a's film industry execution. As the film business has continued developing since the Litman's investigation, the investigation of variables influencing film achievement has been an intriguing exploration region and consequently flourishing articles have been distributed inside the space. Vany and Walls (1999), Elberse (2007), and Nelson and Glotfelty (2012) have inspect the connection between a principle entertainer's star power and a film execution. Basuroy, Chatterjee, and Ravid (2003) have researched what basic surveys mean for a film achievement, setting star force and spending plans as arbitrators. Prag and Casavant (1994) have had a premium in recognizing the connection between elements like promoting costs, MPAA appraisals, and continuations and a film achievement.

As of late, in light of the information gathered from these studies, a couple of analysts have started to lead the examinations that have the prescient trademark. For instance, determining the motion pictures that are exceptionally conceivable to succeed is one of the sorts of such examination. Asur and Huberman (2010) have utilized Twitter information to foresee a film achievement and Mishine and Glance (2006) have anticipated film deals utilizing web blog information. Particularly, utilizing AI procedures, a few examinations have delivered the forecast models with the moderate degree of exactness (for example Sharda and Delen 2006; Eliashberg, Hui, and Zhang 2007; Zhang, Luo, and Yang 2009; Du, Xu, and Huang 2014). For example, Sharda and Delen (2006) have inspected the exhibition of the strategic relapse, discriminant examination, choice tree, and neural organizations to estimate film's prosperity.

They have utilized MPAA evaluations, rivalry level, principle entertainer's star esteem, type, embellishments, spin-off, and the quantity of screens at the underlying day of film discharge as highlights to foresee the film execution. Their best-performing model has anticipated the nine result factors with the 36.9% of exactness. Zhang, Luo, and Yang (2009) have proposed a multi-facet back spread neural organization that has further developed the neural organization model introduced by Sharda and Delen (2006). Their model effectively has ordered six result factors with 47.9% of precision. Eliashberg, Hui, and Zhang (2007) have anticipated a film's profit from venture dependent on its content data utilizing the choice tree calculation. Du, Xu, and Huang (2014) have assessed the exhibition of the straight relapse, support vector machine, and neural organizations on anticipating 5 the movies achievement, dissecting the opinions of the writings posted on Tencent Microblog.

While these studies have for the most part centered on methodological viewpoint to improve their model precision, we recommend more exhaustive strategy that upgrades the exhibition of the model. In this examination, we carry out DSS and Regression model both. Likewise in the wake of auditing the writing regarding the matter, it turned out to be evident that this undertaking adds to the writing by utilizing an old and ongoing informational index that better reflects current film inclinations from inside the most recent quite a while. Indeed, even the latest papers in video form income use information from the last part of the 1990s, with a couple wandering into the mid-2000's (Brewer, 2006; Collins, Hand, and Snell, 2002; Holbrook and Addis, 2008). Embellishments and PC innovation have made considerable progress over the most recent ten years, and may have added to an adjustment of customer tastes and inclinations for specific sorts of movies. By utilizing an informational index which analyzes the top grossing films, this undertaking will actually want to check if the factors that customarily foresee film industry achievement have changed after some time.

CHAPTER 3: METHODOLOGY

Box office achievement relies upon different things starting with MPAA ratings, critics review, release date, number of screens, advancements, star power, Production Company, distribution company, and so forth So in this paper, we will talk about these materials momentarily. We will utilize a linear regression model where Domestic box-office receipts served as the dependent variable and IMDB rating, critical reviews, no of votes, and release date acting as independent variables in the final regression.

3.1 DSS:

Decision support system is a sort of information system and its backing decision making activities. It supports decision makers use communications terminology, documents, information to fulfill decision process tasks. A Decision support system network accumulates and investigates information, combining it to deliver exhaustive data reports. Thusly, as an educational application, a DSS varies from a common tasks application, whose capacity is simply to gather information. The main role of utilizing a DSS is to introduce data to the client in a straightforward manner. A DSS framework is helpful on the grounds that it very well may be customized to produce numerous kinds of reports, all dependent on client particulars. For instance, the DSS can create data and yield its data graphically, as in a bar diagram that addresses projected income or as a composed report. .

The adaptability of the DSS is very gainful for clients who travel much of the time. This offers them the chance to be all around educated consistently, giving the capacity to settle on the best choices for their organization and clients in a hurry or even on the spot.

DSS utilizes the rundown data, exemptions, examples, and patterns utilizing the logical models. It emotionally supportive network helps in dynamic however doesn't really give a choice itself. The leaders accumulate helpful data from crude information, archives, individual information, and additionally plans of action to recognize and take care of issues and decide.

©Daffodil International University

3.2 Linear regression model

Linear regression models are among the most normally utilized measurable techniques. They join boundless appropriateness with a profoundly evolved hypothetical premise Functionality for fitting linear regression models exists in every significant statistics software package (including Minitab, SAS, SPSS, and R) and also in some spreadsheet packages (e.g., Excel).

Linear regression is a machine learning algorithm. It builds a relationship between two variables within the linear equation. One variable is called to be an independent variable, and the other is called to be a dependent variable. In regression analysis, if we have more than one independent variable, then the model is not a simple linear model. This will be known as multiple regression

Linear regression has an equation and the equation is $Y = a + bX$, where the independent variable is X and the dependent variable is Y. The slope of the equation is b, and a is our intercept.

Linear regression models establish the main displaying approach for examining social and financial wonders. However, traditional regression examination doesn't consider issues related with conceivable cross-sectional relationships among observational units brought about by spatial reliance. Two types of spatial reliance among perceptions may refute relapse results: spatial mistake reliance and spatial slack reliance.

3.3 IMDb:

Internet Movie Database is an online informational collection of information related to films, TV programs, home accounts, PC games, and streaming substance on the web – including cast, creation gathering, and individual diaries, plot summations, arbitrary information, assessments, and fan, and essential reviews. Every one of the enlisted individuals from IMDb can project their votes/appraisals for any film. IMDb takes every one of the individual votes cast by the enlisted clients and utilizations them to ascertain a solitary rating.

Data on IMDb comes from an assortment of sources, like movie producers, film studios, onscreen credits, and other authoritative sources. Nonetheless, a significant part of the data comes from IMDb clients themselves, who can submit realities in a wiki-style design. Not at all like customary wiki destinations, IMDb consistently confirms data before it seems on the web - in spite of the fact that blunders do appear, and the site permits clients to report potential slip-ups so they can be fixed.

Users can likewise submit surveys of films and TV shows on one to ten scales, which are then used to make a weighted mean of all client audits to be shown on the film or TV show's page. Additionally, these appraisals are utilized to gather IMDb's Top Rated and Lowest Rated arrangements of motion pictures and TV shows.

IMDb was first distributed in 1990 by Col Needham, a software engineer, collectively of contents that permitted clients to look through a rundown of film acknowledges that were aggregated for the assistance of a Usenet bunch. The information base was fused in 1996. In 1998, Internet Movie Database Ltd. turned into an auxiliary of Amazon, which utilizes its anything but an approach to promote motion pictures and DVDs which it sells on its fundamental site.

In 2002, IMDb dispatched its paid membership administration for entertainment world experts, IMDbPro. IMDbPro permits entertainers, chefs, and different experts to deal with their IMDb page, transfer their list of references, and access more broad data not accessible on the free form of the site.

3.4 Meta score:

Metacritic is a site that totals audits of movies, TV shows, music collections, computer games, and in the past, books. For every item, the scores from each survey arrive at the midpoint of (a weighted normal). In 1999, Jason Dietz, Marc Doyle, and Julie Doyle Roberts had created Metacritic. The site gives a passage from each audit and hyperlinks to its source. A shade of green, yellow or red sums up the pundits' suggestions. It is viewed as the chief online audit conglomeration webpage for the computer game industry. [2][3]

Metacritic's scoring changes over each audit into a rate, either numerically from the imprint given, or what the site chooses abstractly from a subjective survey. The site won two Webby Awards for greatness as a conglomeration site. Analysis has zeroed in on the appraisal framework, the task of scores to surveys that do exclude evaluations, claimed outsider endeavors to impact the scores and the absence of staff oversight of client audits.

Metascore range from 0-100, with higher scores demonstrating better in general surveys, and lower scores showing less ideal audits from critics.

3.5 Dataset Explanation:

The data has been collected from kaggle.com. There are 98 successful films in this dataset. Data includes movie title, release year, IMDB rating, Meta score, and gross collection. Here, box office gross is the dependent variable and release date, IMDB rating, Meta score are the independent variable. How much US dollar a film earned? That is gross box office. How many public gave their votes in IMDB? That is basically no of votes.

Regression Result:

Table 1: Coefficients of four attributes

Intercept	-1114371029.72723
Release Date	249484.050878238
IMDb rating	96798165.8236145
Meta score	-1429551.75092365
No. of votes	30.9098907180349

The one thing that is valuable to us is the coefficient, the coefficient is the number that goes into the equation of a line if we remember a linear equation on the line it has two items it has the y-intercept which is -111437 and has the slope of one variable that's the release date, so the slope of the release date is positive 249484 that means we have a positive relationship between release date and box office gross. IMDB rating is also positive and it's able to make a good impact on gross box office. Meta score is negative, so it's going downward and not able to make an impact on box office. To get this coefficient I ran the regression in Microsoft excel.

Table 2: Results of R square, Significance F and Observation

R Square	0.542934508654985
Observations	98
Significance F	0.262898362110876
Adjusted R squared	0.136179

One of the key items I like to look at is the R square, R square is a measure of how much of the change in ticket sales is driven by those four variables that we provided and this is indicating that 54% of the variation in ticket sales.. The other thing we need to take a look at is the perceptions, we see the quantity of perceptions is 98 which is right since I have taken 98 motion pictures. The importance f is to some degree significant and an overall objective is you need a number that is more modest than point zero five and we have 0.26 that is a very solid**.**
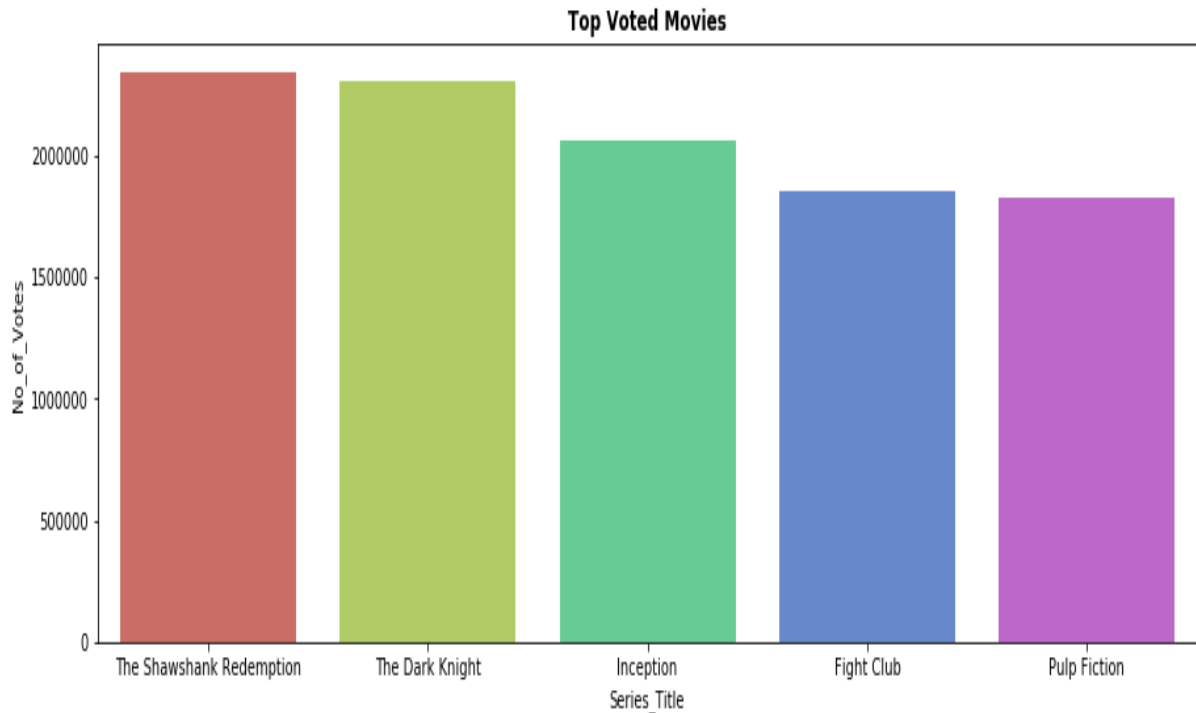
Movies with highest votes

Top Voted Movies

Fig 1: Most-voted movies

This bar chart shows the top-voted movies. In our dataset, these films have got maximum votes on IMDb which means a huge number of people have rated these films. These films got huge votes because the maximum number of people have given their vote and made these films the most voted films. To find top-voted movies from the data set some python libraries like Numpy , Matplotlib were used.
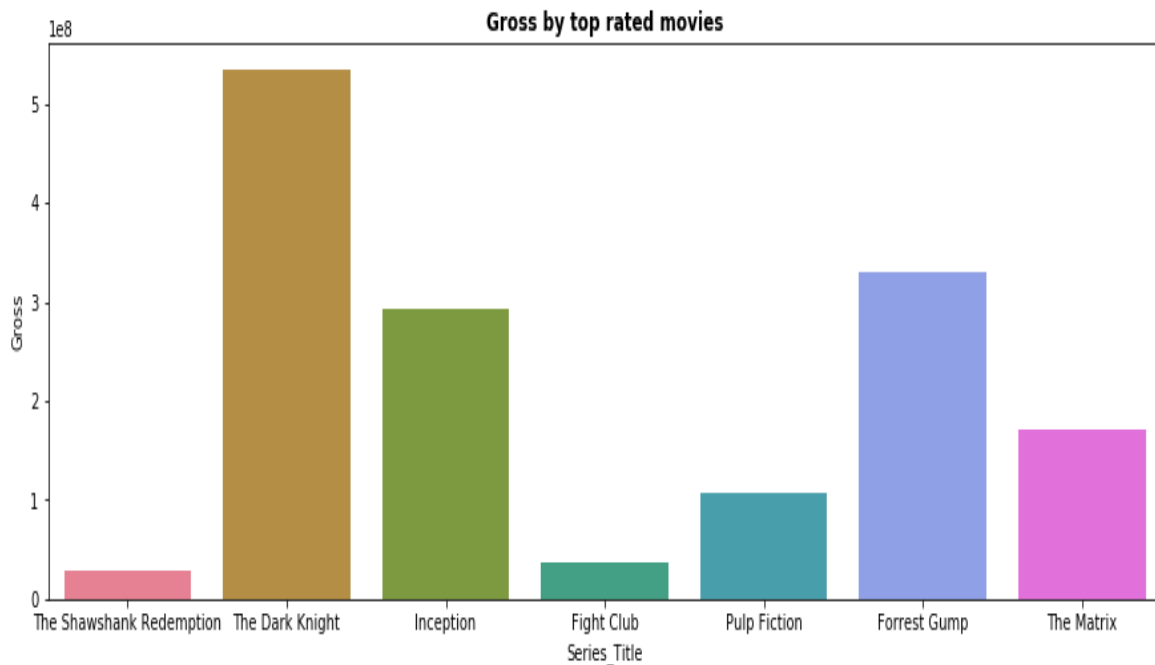
Top-rated movies with gross collection

Fig 2: Gross Collection of top-rated movies

This bar chart shows the Gross collection of top-rated movies. Gross collection means total revenue earned from cinema hall (But before TAX). It is nothing but the number of tickets sold multiplied by the ticket price. In our dataset, these films have got maximum ratings. This picture is showing, How many people have watched these top-rated movies in the cinema hall.

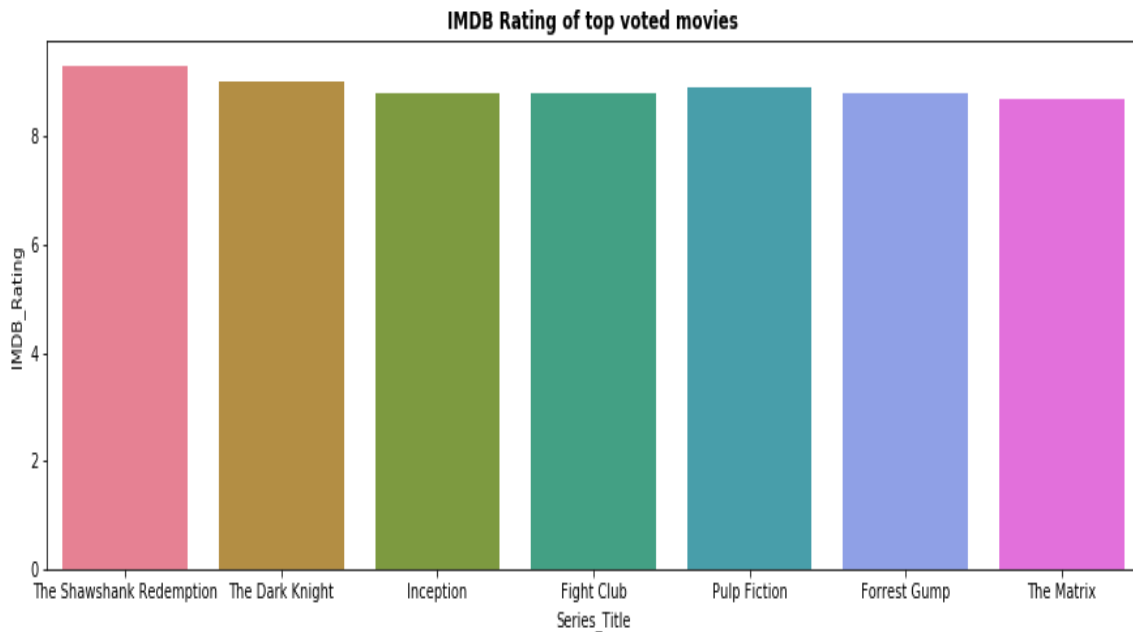**IMDB rating of top-voted movies**

Fig 3: Top IMDB rated movies

This bar chart shows the IMDB rating of top-voted movies. In our dataset, these films have got maximum ratings on IMDb which means a huge number of people have rated these films with the maximum number.
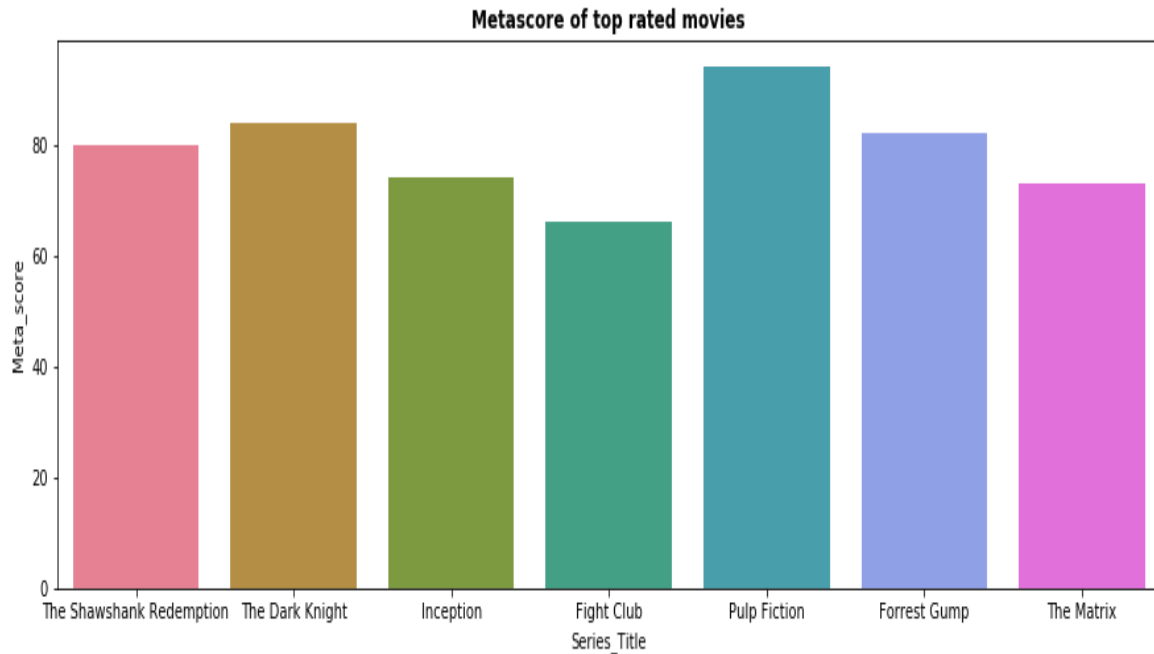
Meta Score rating of top-voted movies

Fig 4: Top Meta Score rated movies

This bar chart expresses the Meta Score of top-rated movies. In our dataset, these films have got maximum ratings on Meta Score which means a number of journalists, magazines, newspapers, critics have rated these films with the maximum number.

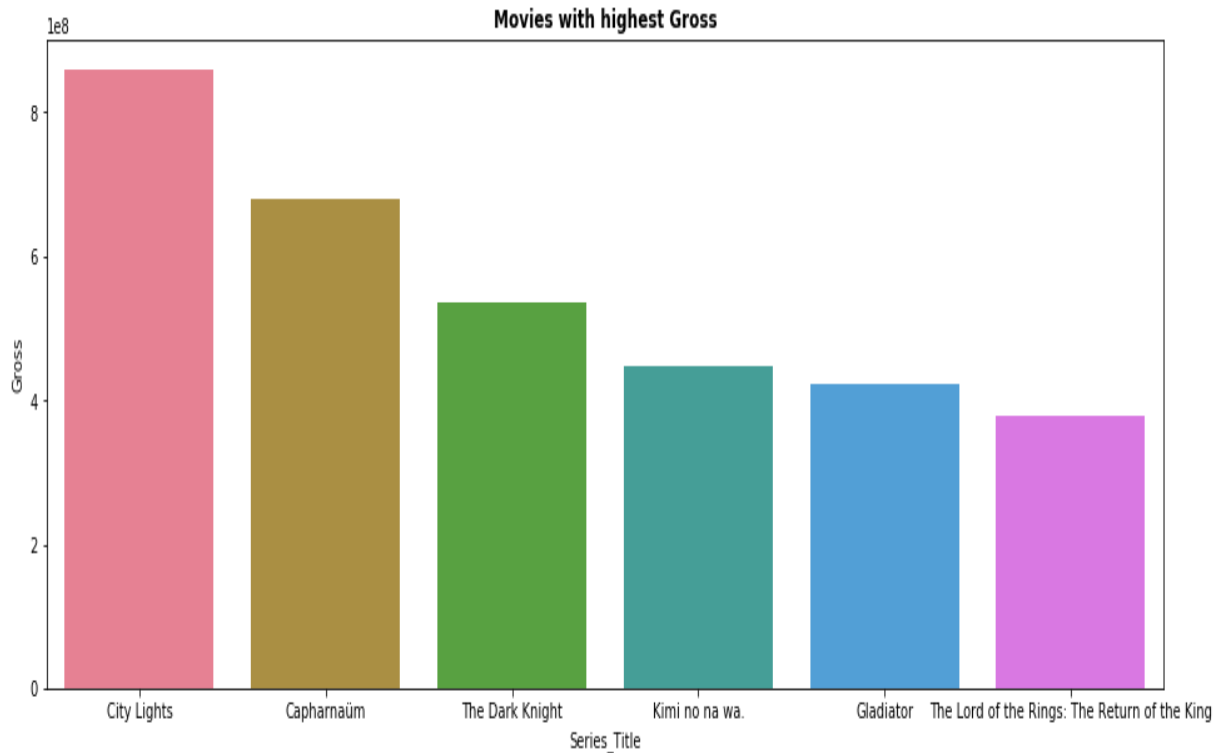**Top-grossing movies with the highest ticket sales**

Fig 5: Highest Grossing movies

This bar chart expresses the highest Grossing movies. This has nothing to do with ratings. In our dataset, these are the highest-grossing movies which mean the maximum number of people who have watched these films on theater.

CHAPTER 5: CONCLUSION

5.1 Findings:

So, the release date, IMDB rating, no of votes are showing a positive impact on gross ticket sales. In conclusion, the results of our study suggest, positive reviews, a large no of votes, and release date have positive and momentous reach on the domestic revenues of a film. We, along these lines, reason that despite the fact that the entertainment world is risky, certain highlights of a film can fundamentally raise its chance of homegrown film industry achievement.

5.2 Recommendation for Future works:

In our dataset, we have used 4 features in films, it is important to know the connection between these features and ticket sales in order to reach the destination. There are other variables like MPAA ratings, budgets, number of screens, sequels, stars that can also make an impact on film's box office, so researchers who willing to improve further need to find data and build another model with these attributes and find out the result.

References

[1]   Kim T, Hong J, Kang P. Box office forecasting considering competitive environment and word-of-mouth in social networks: a case study of Korean film market. Compute Intell   Neurosci. 2017.

[2]   Lash MT, Zhao K. Early predictions of movie success: the who, what, and when of profitability. J Manage Inf Syst. 2016;33(3):874–903

[3]   Nelson, R. A., and Glotfelty, R. 2012. "Movie Stars and Box Office Revenues: An Empirical Analysis," Journal of Cultural Economics, 36(2), 141-166.

[4]   Hur M, Kang P, Cho S. Box-ofce forecasting based on sentiments of movie reviews and independent subspace method. Inf Sci. 2016; 372:608–24

[5]   Duan J, Ding X, Liu T. "A Gaussian copula regression model for movie box-office revenue prediction with social media" edited by W. Z. H. X. Sun M. Zhang X. Commune Compute Inf Sci. 2015;568:28–37.

[6]   Duan J, Ding X, Liu T. A Gaussian copula regression model for movie box-office revenues prediction. Sci China Inf Sci. 2017;60(9):092103

[7]   Edwards, L. H. 2012. "Transmedia storytelling, corporate synergy, and audience expression," Global Media Journal, 12(20), 1-12.

[8]   Ghiassi, M., Lio, D., and Moon, B. 2015. "Pre-Production Forecasting of Movie Revenues with a Dynamic Artificial Neural Network," Expert Systems with Applications, 42(6), 3176-3193

[9]   Nelson, R. A., and Glotfelty, R. 2012. "Movie Stars and Box Office Revenues: An Empirical Analysis," Journal of Cultural Economics, 36(2), 141-166.

[10]   Lee K, Park J, Kim I, Choi Y. Predicting movie success with machine learning techniques: ways to improve accuracy. Inf Syst Front. 2018;20(3):577–88. 29.

[11]   Hur M, Kang P, Cho S. Box-office forecasting based on sentiments of movie reviews and independent subspace method. Info Sci. 2016; 372:608–24

[12]   Ru Y, Bo Li, Liu J, Chai J. An effective daily box office prediction model based on deep neural networks. Cognit. Syst Res. 2018; 52:182–91.

[13]   Sachdev S, Agrawal A, Bhendarkar S, Prasad BR, Agarwal S. Movie box-office gross revenue estimation. Adv Intell Syst Comput. 2018; 709:9–