Box office revenue Analysis and Prediction

Name: Wassem Messiha

Student #: 501113671

Supervisor Name: Dr. Bilgehan Erdem

Date: September 27, 2021

Table of Contents

Abstract

The outbreak of COVID-19 has hit the economy severely the past year, not only causing higher unemployment rates and lower income but also causing people to practice social distancing, and national and international travel restrictions. Reducing capacity in most businesses caused the complete shutdown of many industries. Speaking of the industries that were heavily impacted by COVID-19 pandemic, I would like to bring the light on the film industry, focusing on 2020 and the first two quarters of 2021. Due to the fear we have all developed during the pandemic and the new regulations implemented, people around the world were forced to stay home. This led to cinemas and movie theaters having to close, and big movies having to push back their release date for a year and others even more due to coronavirus uncertainty. The worldwide box office estimated revenue has reportedly reduced from 44.5 billion U.S. dollars to 16.3 billion in 2020 alone.

In this paper I will be using Regression analysis to focus on the revenue trends of the box office in previous years to get an insight about the top grossing months, top grossing movies based on actors and distributors, how are the trends for genres correlated with revenue, and the correlation between budgets and revenues of the movies.

The previous areas of focus along with other attributes in the dataset will also allow me to build a Predictive Analytics model to predict the box office revenue as the film industry starts to go back to normal with new releases announced and as audiences continue to have a strong appetite to enjoy the theatrical experience. I will be using the boxofficemojo dataset found on Kaggle, it has 26 attributes and 3243 observations for the last 30 years. First, I will start with Exploratory Data Analysis in order to understand the data description, the type of each attribute, and finding missing values and outliers. I will also do revenue classification based on genres, distributors,

mpaa and main actor/actress. Then I will perform analysis on the dataset to see which attributes are highly correlated with the revenue. And finally, I will use Regression analysis to predict the movies revenue.

Introduction

The growth of the film industry has increased and keeps increasing rapidly globally and locally. The global box office industry hit $42.3 billion in 2019, according to the Theme report (MPA, 2021) that was produced by the Motion Picture Association. Numerous previous studies have examined the prediction of box office revenue from a variety of viewpoints (Barry R Litman, 1989), (Mohanbir S. Sawhney, 1996). There are only three to four films out of ten that break even while they are produced and marketed, and only one out of 10 that becomes lucrative at the box office when they are released. According to an analysis by film data researcher Stephen Follows (Follows, 2016), only 51% of Hollywood movies make a profit. The pressure to perform at the box office is immense in a big entertainment industry with multimillion dollar productions, famous celebrities and strong film distributors. Many factors have an impact on the film's performance, such as quality of the story, the filming equipment used, marketing and the movie director. But what are some of the factors that affect the box office revenue? Previously, researcher made predictions based on factors such as budget, genre, star actors, distributor, runtime and release month which played a large role in predicting the box office revenue. In this paper, I will mainly focus on determining if these factors played a significant role in the prediction.

Literature review

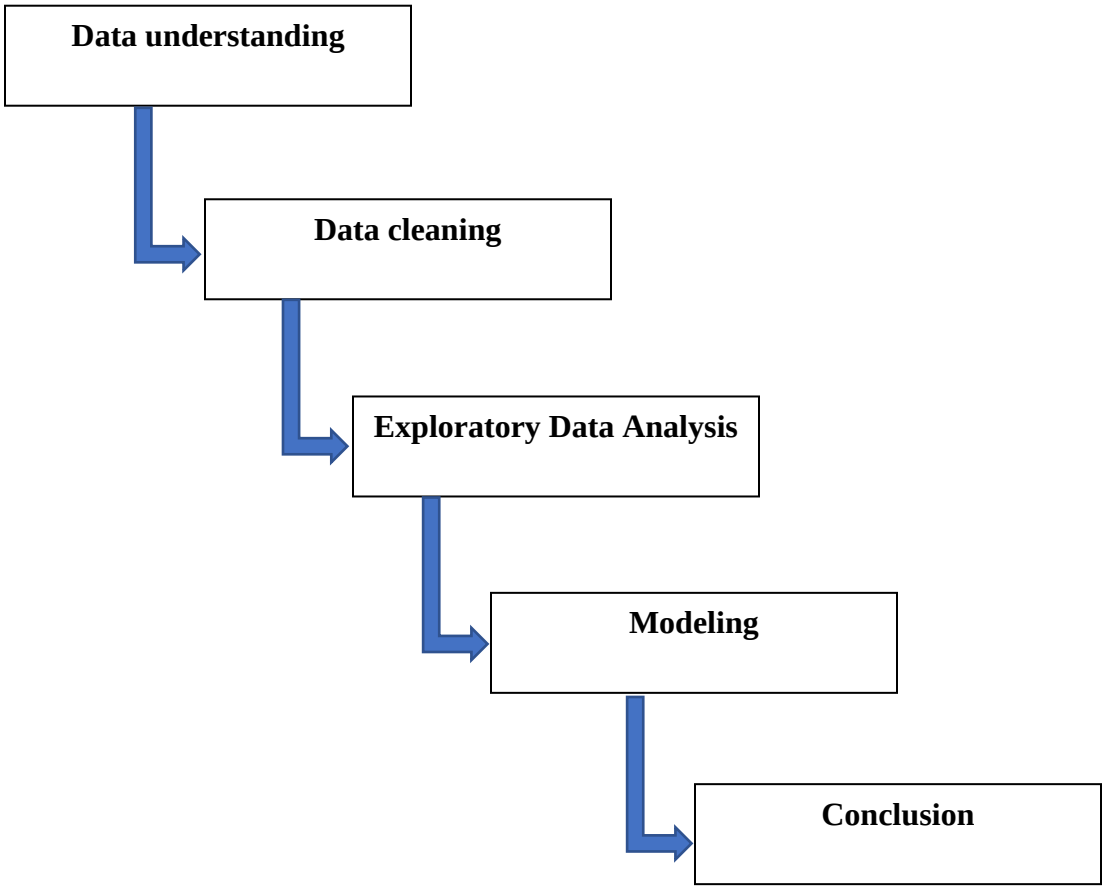
For the purpose of this study, I examined different publications that discussed the effect of multiple factors on the success of box office revenue. Predicting box office revenue is critical for film producers and directors because it is their primary source of income. And certainly, the movie budget would be the most important variable to predict the box office revenue. (Suman Basuroy, 2003) examined the effect of critical reviews on the success of a film, using star power and finances as moderators. The authors found that a movie with a star actor and big budgets positively impacted the box office success.

Movies are classified according to their genre, and it is much easier for audiences to determine what they enjoy and want to watch if the movies are categorized. Therefore, movie genre could be another important factor in predicting box office revenue. (Barry R Litman, 1989) conducted research in which they included sequels in addition to genre, MPAA ratings, star power, quality metrics, and release dates. The authors have used a dataset contains 697 films that covers the years 1981 through 1986. They used Regression model and one of the findings was sci-fi/Fantasy films do better at the box office while dramas do worse. In the meantime, a movie might be assigned to multiple genres at the same time. For example, the 2019 blockbuster 'Avengers: Endgame' is classified as Action, Adventure, Drama and Sci-Fi. Films in within the drama, adventure and action movie genre generated the most box office revenue in North America in the years between 1995 and 2021 (Navarro, 2021).

Academic studies have long examined the effect of celebrity power on the commercial success of motion movies. (Elberse, 2007) shed the light on the evidence that the presence of star power has an influence on the expected box office revenue. It was estimated that, on average, star actor/actress generate roughly $3 million in box office revenue per week (Elberse, 2007). In

other words, films featuring well-known actors and actresses appear to have much greater box office revenues, whereas films with unknown cast members appear to have substantially lower revenues.

Methodology



1. Data understanding:

Data Description

Attribute	Data Type	Description
movie_id	String Nominal data	IMDB ID of a movie
title	String Nominal data	Title of the movie
year	Integer Ordinal data	The release year
trivia	String Nominal data	Short description of the movie – Will be dropped as it is irrelevant to this analysis
mpaa	String Ordinal data	Motion Picture Association of America - Movie ratings
release_date	String Ordinal data	The release month and day
run_time	String Ordinal data	Total runtime of the movie
distributor	String Nominal data	Distributor company name
director	String Nominal data	Director name
writer	String Nominal data	Writer name - Will be dropped as it is irrelevant to this analysis
producer	String Nominal data	Producer name - Will be dropped as it is irrelevant to this analysis
composer	String Nominal data	Composer name - Will be dropped as it is irrelevant to this analysis
cinematographer	String Nominal data	Cinematographer name - Will be dropped as it is irrelevant to this analysis
main_actor_1	String Nominal data	Main actor / actress
main_actor_2	String Nominal data	Secondary actor / actress
main_actor_3	String Nominal data	Third actor / actress
main_actor_4	String Nominal data	Fourth actor / actress
budget	Float Continuous data	Production cost of the movie
domestic	Float Continuous data	Total domestic (US/Canada) revenue

international	**Float Continuous data**	**Total international revenue**
worldwide	**Float Continuous data**	**Total revenue worldwide**
genre_1	**String Nominal data**	**Main genre of the movie**
genre_2	**String Nominal data**	**Secondary genre of the movie**
genre_3	**String Nominal data**	**Third genre of the movie**
genre_4	**String Nominal data**	**Fourth genre of the movie**
html	**String Nominal data**	**Website that has information, crew member and actors of the movie - Will be dropped as it is irrelevant to this analysis**

Descriptive Statistics

	year	budget	Domestic_Revenue	International_Revenue	Worldwide_Revenue
count	3221.000000	3.221000e+03	3.221000e+03	3.221000e+03	3.221000e+03
mean	2006.656007	4.639630e+07	6.148491e+07	7.827260e+07	1.397575e+08
std	7.221364	4.714060e+07	8.041217e+07	1.434407e+08	2.165638e+08
min	1990.000000	2.200000e+02	0.000000e+00	0.000000e+00	3.000000e+01
25%	2001.000000	1.400000e+07	1.288293e+07	2.543849e+06	1.912640e+07
50%	2007.000000	3.000000e+07	3.537483e+07	2.509637e+07	6.267510e+07
75%	2012.000000	6.200000e+07	7.733913e+07	8.750000e+07	1.698528e+08
max	2020.000000	3.560000e+08	9.366622e+08	2.029931e+09	2.797801e+09

Some findings from the above table:

- **Year**: The highest number of movies released was during 2006
- **Budget:** The highest budget was 356,000,000 and the average budget is about 46,148,491
- **Worldwide Revenue:** The highest was 2,797,801,000 and the average is 139,757,500

2. Data cleaning:

This phase is all about data cleaning. Initially, the dataset had 3243 movies. Then I recognize that there are many movies which do not have all data attributes available. MPAA has 161 records missing values, I decided to fill them with the most common MPAA which is PG-13. For Worldwide revenue, my target variable, there was only 7 movies missing the revenue, so I decided to delete these records. I also found movies with Worldwide revenue as low as $30 (figure 1), which didn't make sense, so I decided to delete records with Worldwide Revenue less than $100,000. I deleted other variables that was not relevant to this analysis.

movie_id	title	year	Worldwide_Revenue
tt0429277	Zyzzyx Rd	2006	30.0
tt1019449	The Rise and Fall of Miss Thang	2007	581.0
tt1235168	Redneck Carnage	2009	706.0
tt0431155	Issues	2005	783.0
tt0387057	Beat the Drum	2003	895.0

Figure 1

3. Exploratory Data Analysis:

MPAA (Motion Picture Association of America): MPAA rates the movies according to films' thematic and content suitability for certain audiences. The primary MPAA ratings are G (General Audiences), PG (Parental Guidance Suggested/Some material might not be suitable for children), PG-13 (Parents Strongly Cautioned/Some material may be inappropriate for children under the age of 13), R (Restricted/Under 17 not admitted without parent or adult guardian), and NC-17 (No One 17 and Under Admitted) (Wikipedia).

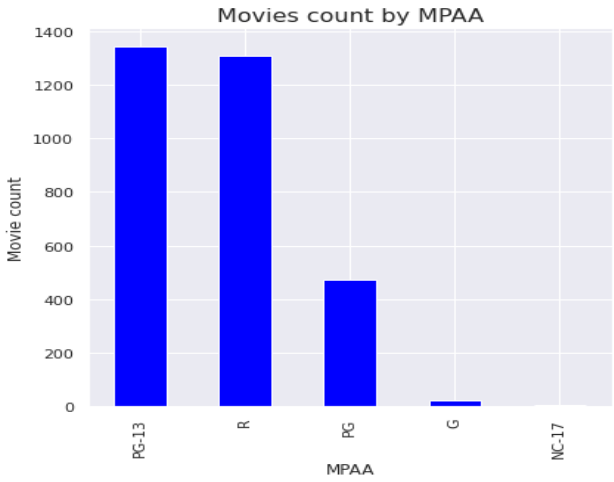


Figure 2

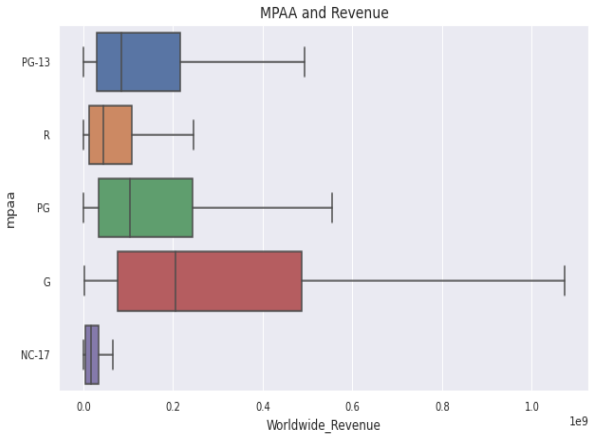


Figure 3

From figure1 we can see that PG-13 is the most common rating, followed by rating R. NC-17 is extremely rare because it restricts the movie to be seen by people under 17. But when we look at the revenue by MPAA (figure2), we see that G rating has the highest revenue grossing followed by the PG rating. Since this variable is Categorical, I converted it to a dummies to be able to use it in my model.

Genre: It normally describes a category of literature, music, or other forms of art and entertainment, whether written or spoken, audio or visual (Wikipedia).

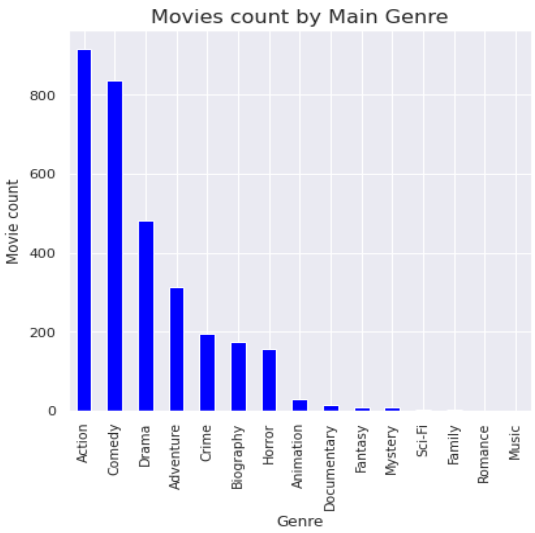


Figure 4

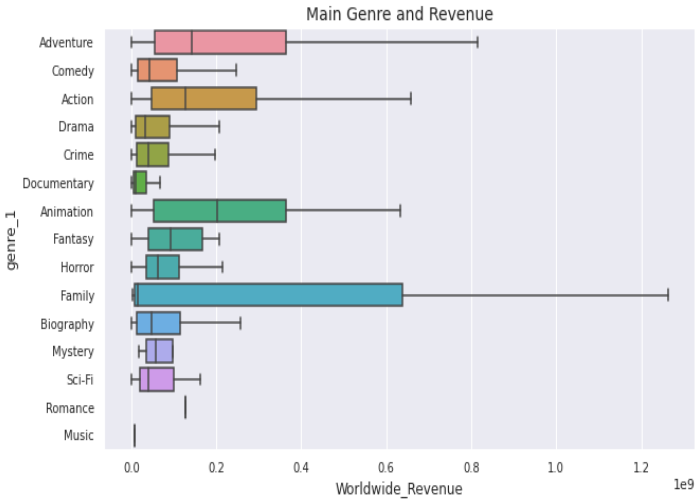


Figure 5

From the dataset, we can see that Action genre is the most popular followed by the Comedy and Drama movies (figure3). But in terms of revenue (figure4), Family genre recorded the highest revenue grossing followed by Adventure movies. There was a total of 15 unique genres in this dataset, and because it is a Categorical variable, I decided to create a mask of the top 5 genres and converted them to dummies to be used in the model.

Distributor: In other dataset they are called the producers or production companies. They work on the process of producing video content, and some of them are also responsible for the marketing of the movies domestically and internationally.

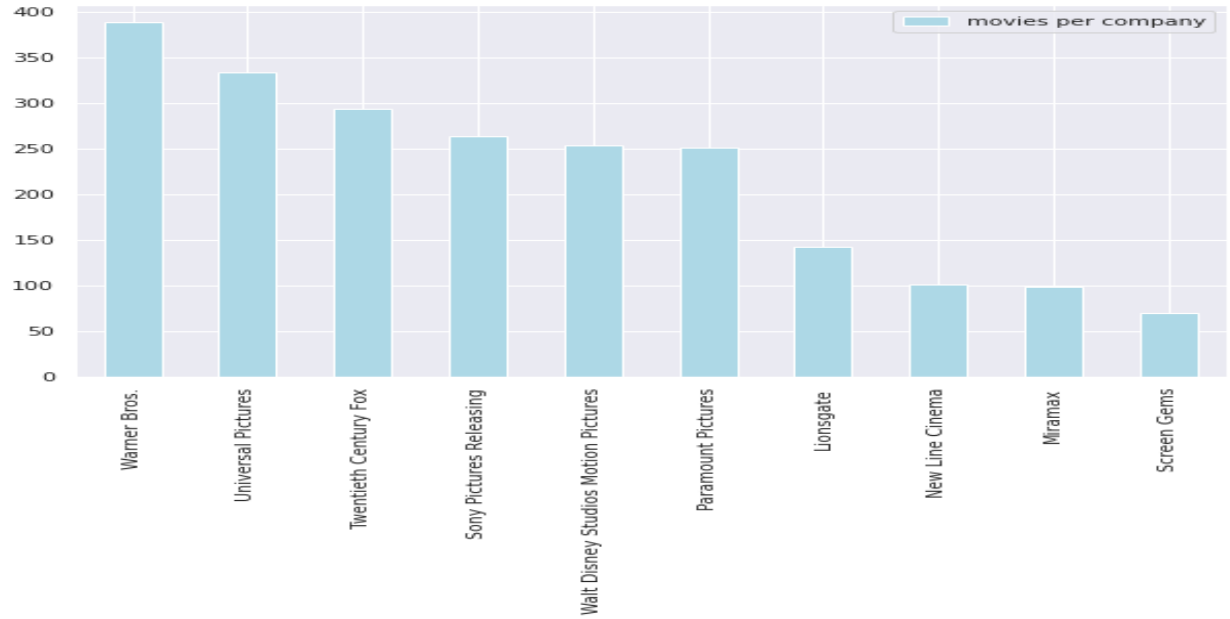


Figure 6

We can see in figure5 that Warner Bros has the highest number of movies produced during the last 30 years, followed by Universal Pictures. For revenue (figure6), Walt Disney Studios records the highest revenue grossing distributor with over $74 billion in the last 30 years, followed

	num_of_movies	total_revenue
Walt Disney Studios Motion Pictures	253	7.487085e+10
Warner Bros.	388	6.761542e+10
Universal Pictures	334	6.083843e+10
Twentieth Century Fox	294	5.987527e+10

Figure 7

by Warner Bros. There is 157 unique distributors in the dataset, I applied the dummy function on the top 10 distributors and used it in the model.

Director: A film director controls a film's artistic and dramatic aspects and visualizes the screenplay (or script) while guiding the film crew and actors in the fulfilment of that vision (Wikipedia). Director could have an impact on the revenue generated for a movie as he/she is the one who is choosing the cast members, production design and all the creative aspects of filmmaking. We can see in figure7 below that Clint Eastwood has the highest number of

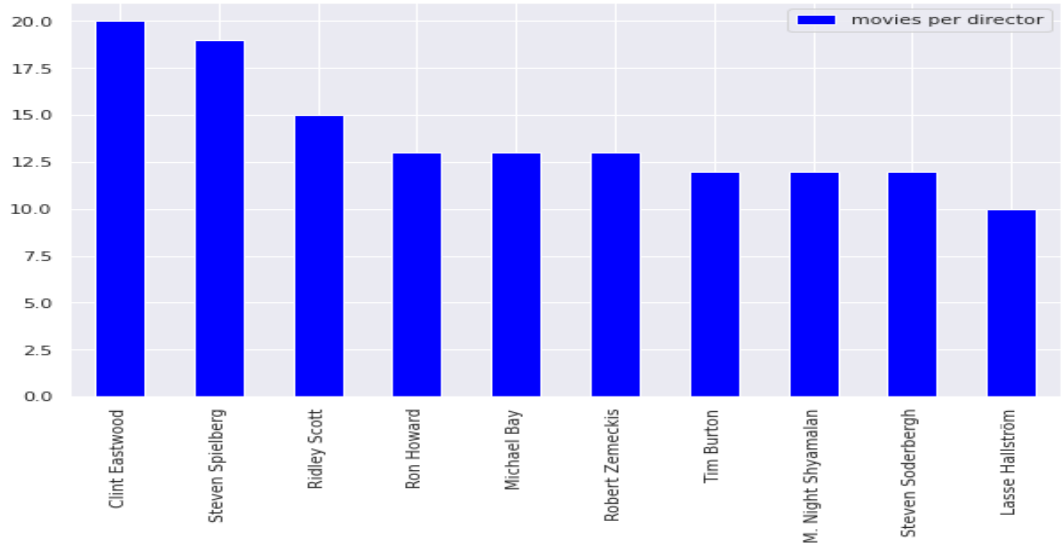


Figure 8

movies, followed by Steven Spielberg during the last 30 years that was covered by the dataset. On the contrast and in terms of revenue (figure8), Steven Spielberg tops the list of directors with most grossing movies and Clint Eastwood didn't make the top 10 list. That could be an indication that having a top director in a movie doesn't usually mean it can produce higher revenue. Since there is a different 1455 directors in the

	num_of_movies	total_revenue
Steven Spielberg	19	7.067664e+09
Michael Bay	13	6.451693e+09
James Cameron	4	5.884646e+09
Anthony Russo	4	4.796147e+09
Christopher Nolan	9	4.756854e+09
J.J. Abrams	6	4.653989e+09
Jon Favreau	7	4.294367e+09
Roland Emmerich	10	3.761203e+09
Gore Verbinski	10	3.753025e+09

Figure 9

dataset, I decided to convert the top 10 to dummies and use it in the model.

Actor: Star power evaluation has different opinion by researchers. Some stars are being evaluated by the number of awards or nominations received whether or not an actor was in top grossing movies before. In my study I used the total number of movies filmed as the evaluation of the star power. We can see in figure 9 that Adam Sandler followed by Tom Hanks and Johnny

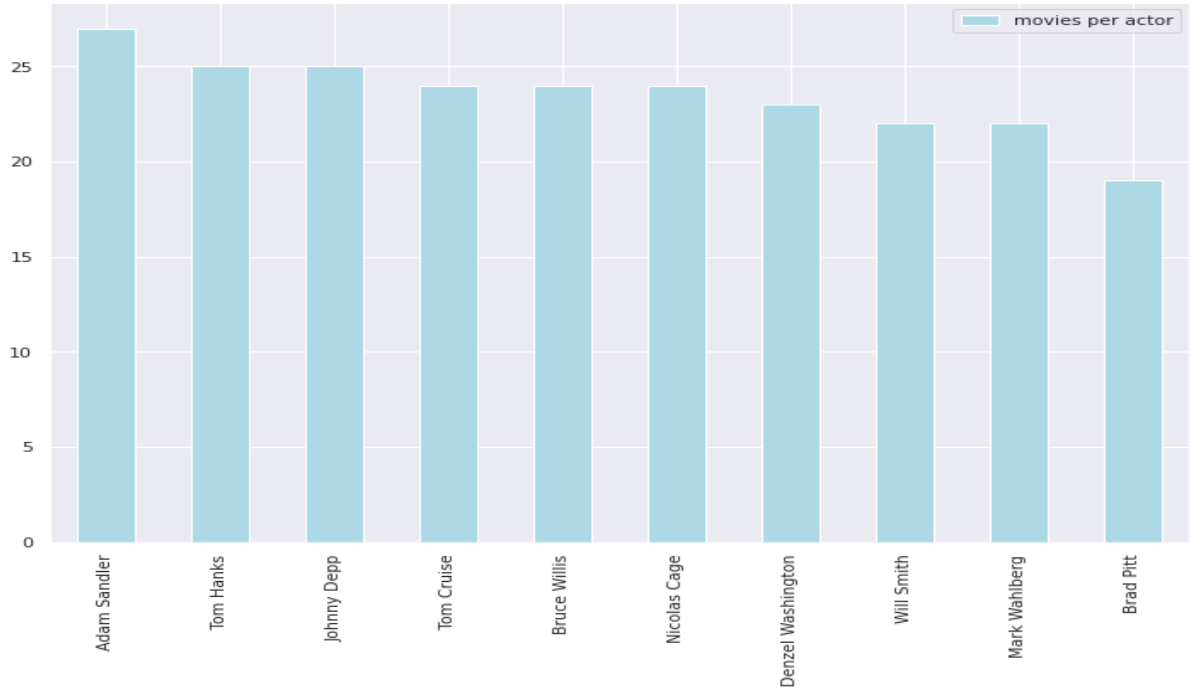


Figure 10

Depp has the highest number of movies filmed. And in the next table (figure 10), Robert Downey Jr. has highest revenue grossing followed by Tom Hanks and Tom Cruise. The dataset used in this analysis has 1238 unique main actor, so I converted the top 10 to dummies.

	num_of_movies	total_revenue
Robert Downey Jr.	12	9.206894e+09
Tom Hanks	25	8.491896e+09
Tom Cruise	24	8.125249e+09
Will Smith	22	7.933443e+09
Johnny Depp	25	7.268864e+09

Figure 11

Release Date: I did some EDA about the release day, week, and month but didn't find high correlation between these variables and revenue. Here are some findings, figure 12 is very crowded but it is clear to notice that Christmas time has the highest number of movies released. And in figure 13, it is also showing that Christmas week has the highest number of releases and that is because of the popularity of going to the movies during the holiday season. The second largest spike is during the first week of October. Also in figure 14, it shows that the month of October and December has the highest number of released movies.

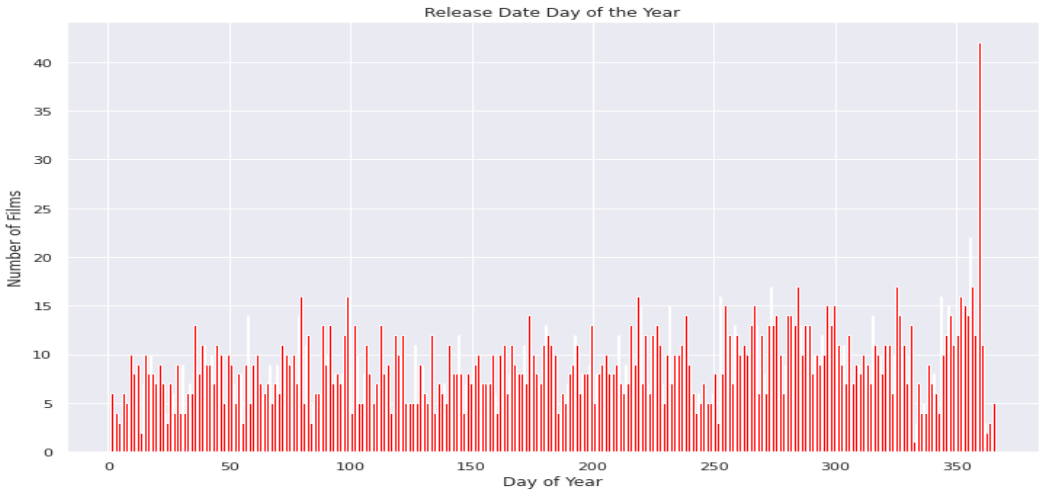


Figure 12

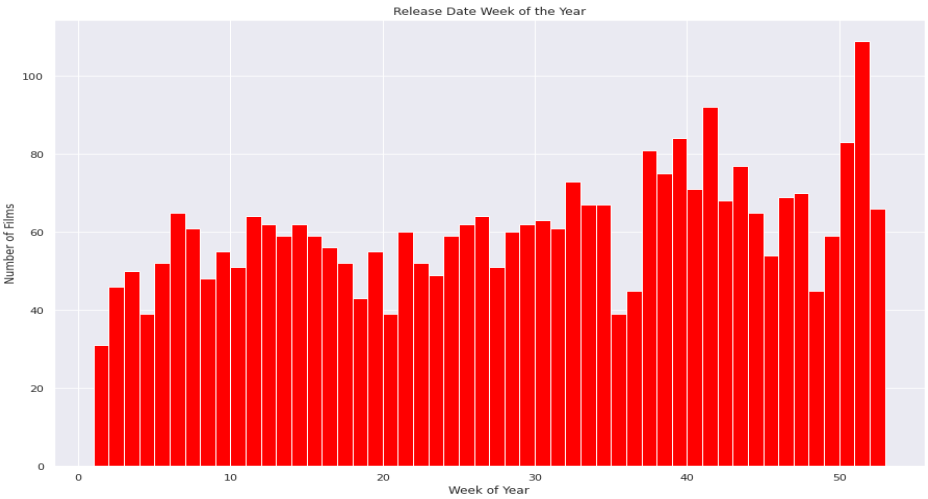


Figure 13

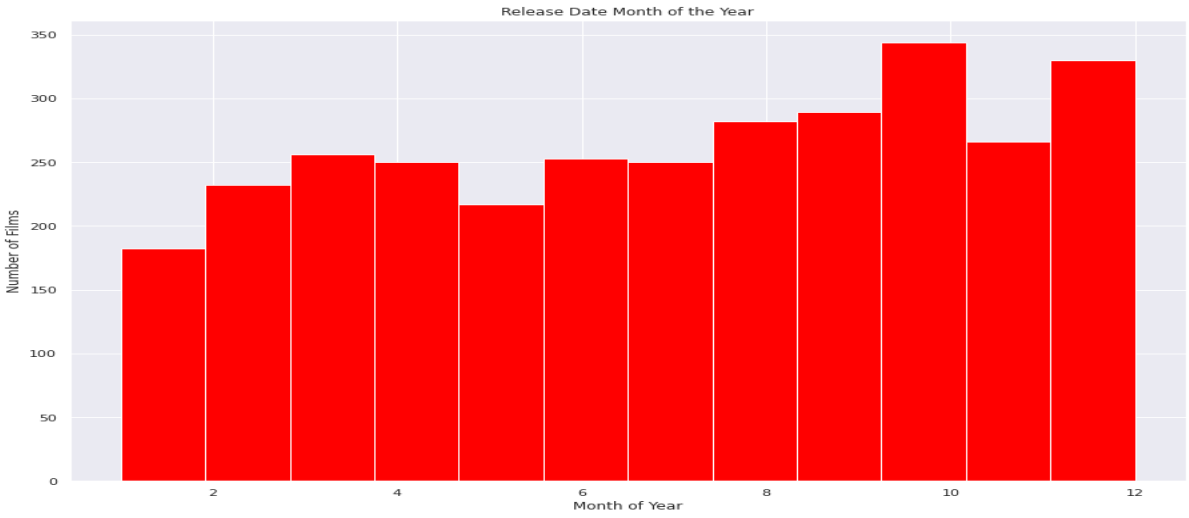


Figure 14

When it comes to release date and revenue, there is a high peak in revenue during the last week of May and through out the weeks of summer followed by the Christmas break (Figure 15). In summary, the number of movies released in a given month or day during the year is not as important as some months that has special events happening, for example, January and February could be generating more revenue with less movies released because of the Oscar season.

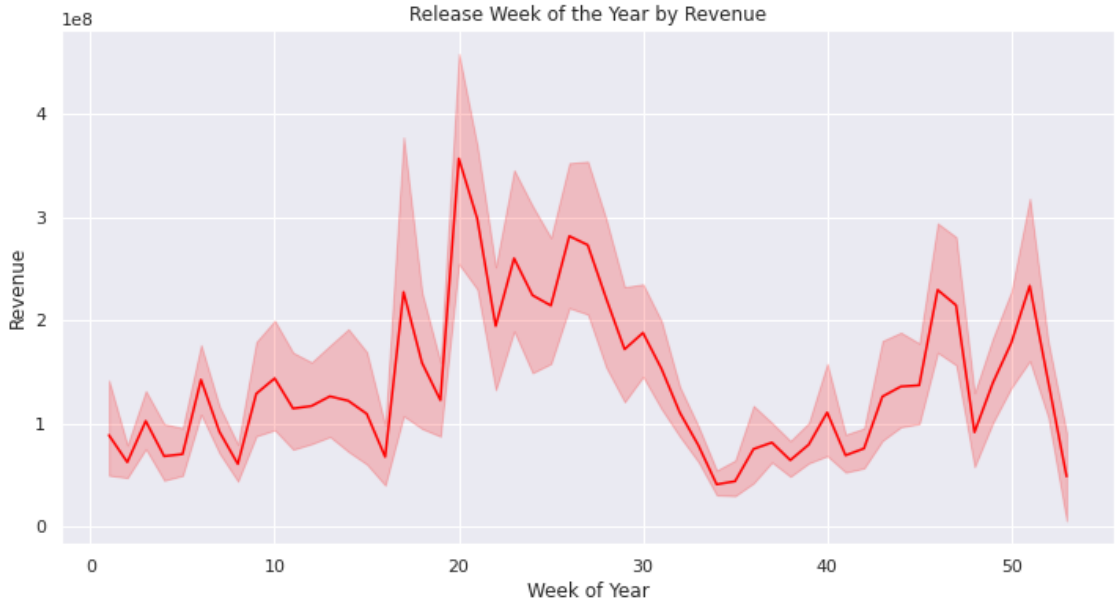


Figure 15

Budget: Big budget movies with top actors, and large advertising budgets could have an advantage on drawing the crowd to the box office. Litman argues in his study that big budgets reflect higher quality and greater box office popularity (Litman, 1983). In general, there is a positive agreement on the effect of budget on box office revenue and in this analysis budget had the highest correlation with box office revenue. In figure 16, showing the top 10 profitable movies in the last 30 years, it shows that higher budget produces higher revenue and profit.

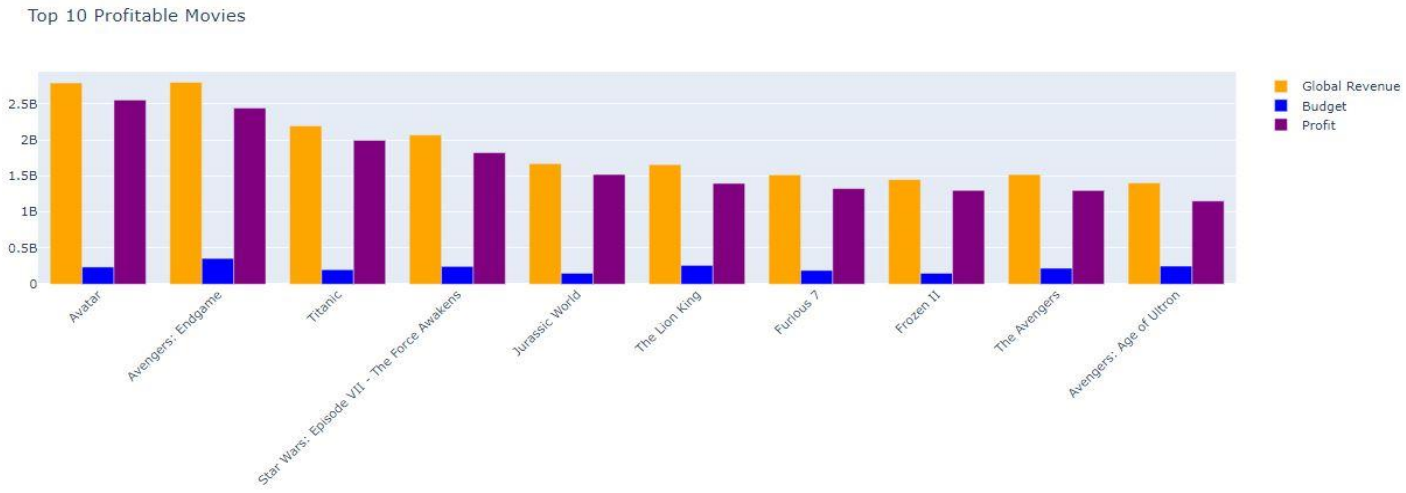


Figure 16

4. Modeling:

After looking at the correlation between the discussed variables and the Worldwide Revenue, there is a very little correlation between all of them except the budget. I started building the model by combining the budget with other features and from figure 17, we can see that the best model ended up being the Linear Regression model, with Budget and Genre as the 2 main features. The model scores RMSE of 1.2554 and run time of 0.11 and the Random Forest model was a second close with RMSE value of 1.2680 with run time of 0.16

	Model	Dependent Var	R-Square	MSE	RMSE	Run Time
3	Linear Regression	Budget & Genre	0.509896	1.576106	1.255431	0.11
5	RandomForest	Budget & Genre	0.502681	1.599311	1.264639	0.16
14	RandomForest	Budget & Actor	0.500687	0.597399	0.772916	0.51
15	Linear Regression	Budget & MPAA	0.487490	1.509833	1.228753	0.55
2	RandomForest	Budget	0.483526	1.521510	1.233495	0.06

Figure 17

The regression equation:

Revenue = 1.3556 + (0.9602 * Log_budget) + (0.2145 * Adventure) + (-0.1376 * Comedy) + (-0.2899 * Crime) + (-0.2591 * Drama)

From the above Multiple Linear Regression equation, I can conclude that for an increase in Budget the Revenue increases by 0.9602 and movies with Genre Adventure increases Revenue by 0.2145

	title	budget	releaseYear	Worldwide_Revenue	Predicted_Revenue	Revenue_diff	Diff_percentage
0	Back to the Future Part III	40000000.0	1990.0	246144250.0	4.164916e+06	-2.419793e+08	-0.983079
1	The Bonfire of the Vanities	47000000.0	1990.0	15691192.0	2.327033e+07	7.579139e+06	0.483019
2	Dances with Wolves	22000000.0	1990.0	424208848.0	5.753357e+06	-4.184555e+08	-0.986437
3	Dick Tracy	47000000.0	1990.0	162738726.0	9.628572e+07	-6.645301e+07	-0.408342
4	Die Hard 2	70000000.0	1990.0	240247433.0	2.199198e+08	-2.032768e+07	-0.084611

5. Conclusion:

Even for the best model in this analysis that had a prediction accuracy of only 50%, we can see that Revenue is still off by an average of $76 million on each prediction. It is a very significant amount, but if we look at the data on hand, for a blockbuster movies that make over $800 million in revenue, being off by $76 million is very close and a good start for more accurate prediction model.

Limitations and Potential Improvements:

With this being my first full machine learning project that I have worked on, there were some limitations as well as some improvements.

Starting with the dataset, that I found on Kaggle, it was very limited in terms of features that could be used in training the model. For example, there were other datasets that had movie ratings and audience reviews which could be used as features in determining the potential box office revenue after the first week of release. Also, information about the first weekend box office revenue could be a good feature in predicting the total global revenue of the movie. There were some columns that I didn't fully utilize, for example the title of the movie, which can be used to find the most popular keywords in titles and the length of the title as well. I also could have added more features together, like budget, genre, and actor which could return an overall better model accuracy.

Lastly, with my lack of experience in building a machine learning prediction model, there may be a better feature selection technique, and better models and algorithms to train the dataset. I did extensive research and I applied what I thought would be a good fit for the dataset I have.

References

Barry R Litman, L. S. (1989). Predicting financial success of motion pictures: The'80s experience. *Media economics*, 2(2): 35–50.

Elberse, A. (2007). The Power of Stars: Do Star Actors Drive the Success of Movies? *Journal of Marketing*, Vol. 71, No. 4 (Oct., 2007), pp.102-120.

Follows, S. (2016, July). Retrieved from https://stephenfollows.com/hollywood-movies-make-a-profit/

Litman, B. R. (1983). Predicting Success of Theatrical Movies: An Empirical Study. *Journal of popular culture*, Vol 16, Issue 4.

Mohanbir S. Sawhney, J. E. (1996). A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. *Marketing Science*, 15(2): 113–131.

MPA. (2021, March). Retrieved from Motion Picture Association: https://www.motionpictures.org/research-docs/2020-theme-report/

Navarro, J. G. (2021, August 12). *Statista*. Retrieved from https://www.statista.com/statistics/188658/movie-genres-in-north-america-by-box-office-revenue-since-1995/

Suman Basuroy, S. C. (2003). How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. *Journal of Marketing*, 43(2), 287-295.