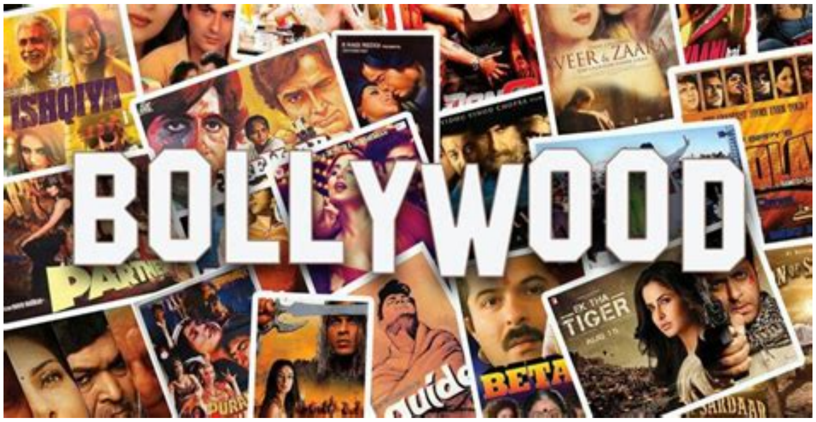
Indian Cinema through Data

By Vishwas Desai (visdesai), Swet Shah (shahswet)

D590 Data Visualization

Fall 2021



Abstract

The goal of this visualization is to analyze various aspects that contribute to the success of a film. A commercial picture can not only entertain the masses but also make a lot of money for the creators. A good director, excellent actors, production house, technicians such as editors/cinematographers, and the timing of the movie's release are all key factors in determining whether a film will make money or not. Indian cinema, one of the world's oldest cinemas is a broad term that refers to a variety of film industries in India, which are mostly split by languages and regions. The Hindi film industry, popularly known as Bollywood, will be our primary emphasis. We hope to visually explore what makes a Bollywood film successful as well as provide a brief overview of Indian cinema with this project. Furthermore, we will look at different aspects of movie making which will be linked to budget and box office returns.

Introduction

Why your project is important or interesting? why should we care?

India produces one out of every five films produced around the world. Indian film has tackled societal inequities such as caste, the subjugation of Indian women, religious intolerance, rural poverty, and the demands of living in expanding cities from its beginnings under colonial rule to the modern age. Indian Cinema with its political, economic, social affect has become a unique subculture and flag bearer of Indian pop culture. Though sometimes confused with Indian cinema, the word "Bollywood" refers only to the Hindi-language industry based in Mumbai. There are numerous regional cinema industries around the country, each producing films in a different language like Tamil, Telugu, Bengali, and Kannada. The regional cinemas share several similar motifs (music, dancing, fantastic costumes, high melodrama, action etc.). Considering the traction, the Bollywood movies are gaining all over the world, it motivated us to create this project for clearly visualizing the movies overall success factors based on different parameters. There are existing visualizations that emphasize essential elements of what makes a Bollywood film, but these renderings are generally monotonous and lack information, making the end user to seek for more accurate visualizations and infographics.

As for the current scope of this report being considered, we are focusing on the Bollywood movies since the data for other regional movies are not easily available. We will try to explore different attributes including actors, running time of a movie, genre, etc. to predict the performance of the movie along with the profits at box office. If we consider the profits only based on the only the total collection of money at box office than it would be a false, factor since we are not considering the amount it took to make this movie and hence, we are trying to consider all the other aspects as well apart from box office collection for visualizing the overall performance of the movie. As a result, developing visualizations that can show a movie's success before it is released and if it could be profitable.

Existing visualizations

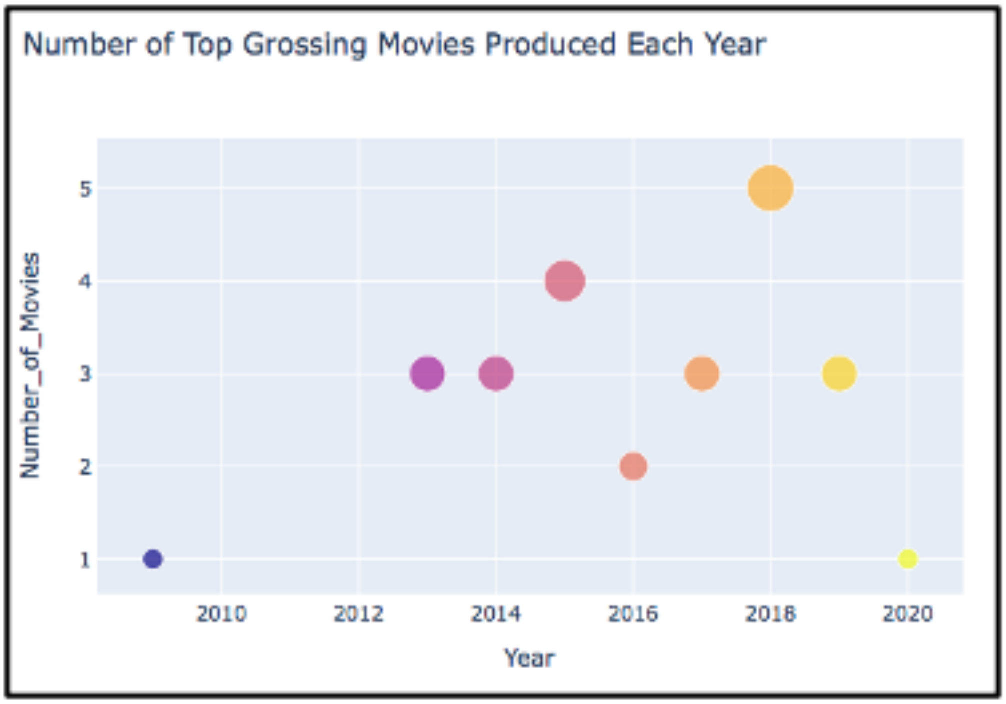


Figure 1. Number of Top Grossing Movies produced Each Year.

There are some of the Bollywood visualizations that exist, which are often restricted to genres, directors, or actors. There are no engaging visuals that convey a story or connect the key factors. As stated earlier, Indian film is a unique subculture, and Bollywood's contribution to that subculture should not just be visually limited to line graphs, bar graphs, and histograms. Our primary objective is to create visually enticing, aesthetically appealing visualizations with enough information to engage the user. Existing visualizations available only capture a few attributes for visualization and predict the rating but this alone doesn't define the performance of the movie. There are too many underrated movies with bad critiques but are a masterpiece based on the people's choice. Also, some movies which haven't done well in terms of box office collection but are some of the best critically acclaimed movies. Existing visualizations doesn't capture all this conditions or attributes for determining the performance and effectiveness of the movie on people, creator, actor, and critic writers.

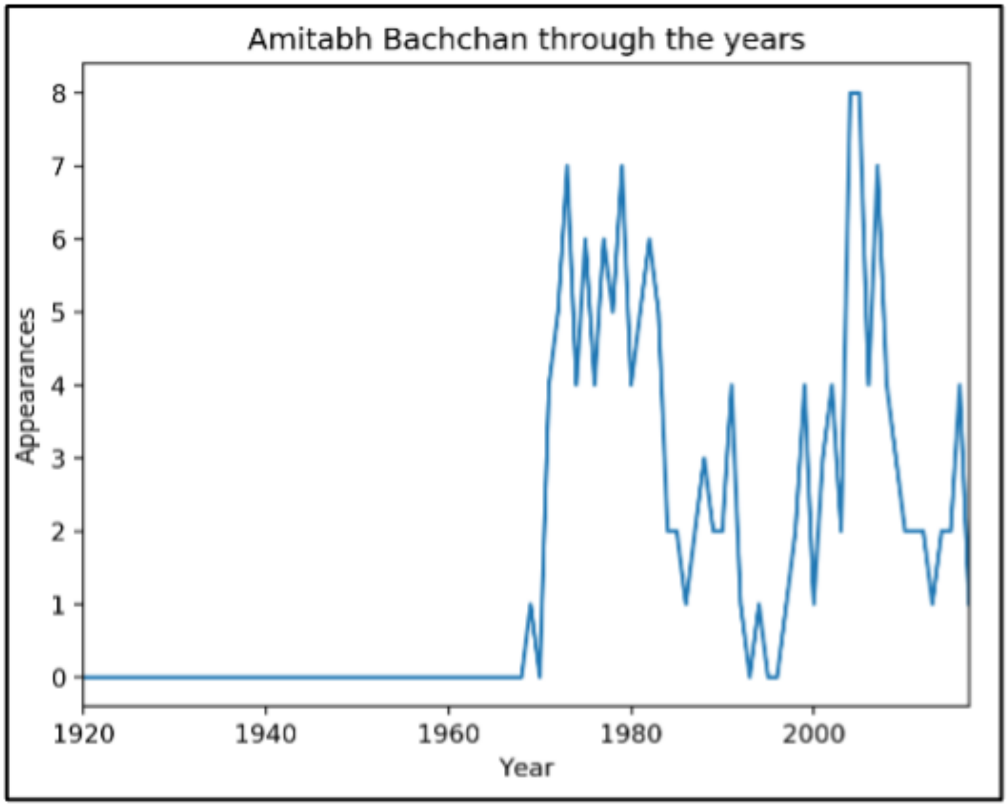


Figure 2. Number of movies casting Amitabh Bachchan over the years.

As we can see from the graph using line graphs to show the number of movies Amitabh Bachchan has appeared in from 1960 to 2010 and ongoing. As we can see that the number of movies has decreased from the year of the late 2005's and this can be attributed to the age of an actor. This can conclude that as the actors age goes on increasing, the number of movies an actor appears in goes on decreasing but this alone can't be true since the time for which an actor is in the movie along with the roles an actor is playing should also be considered such that we can find the exact number of movie appearance with input attributes as actor's movie time appearance, age of an actor, actor's success, genre of movie an actor is comfortable performing, even the number of actors present at that time etc.

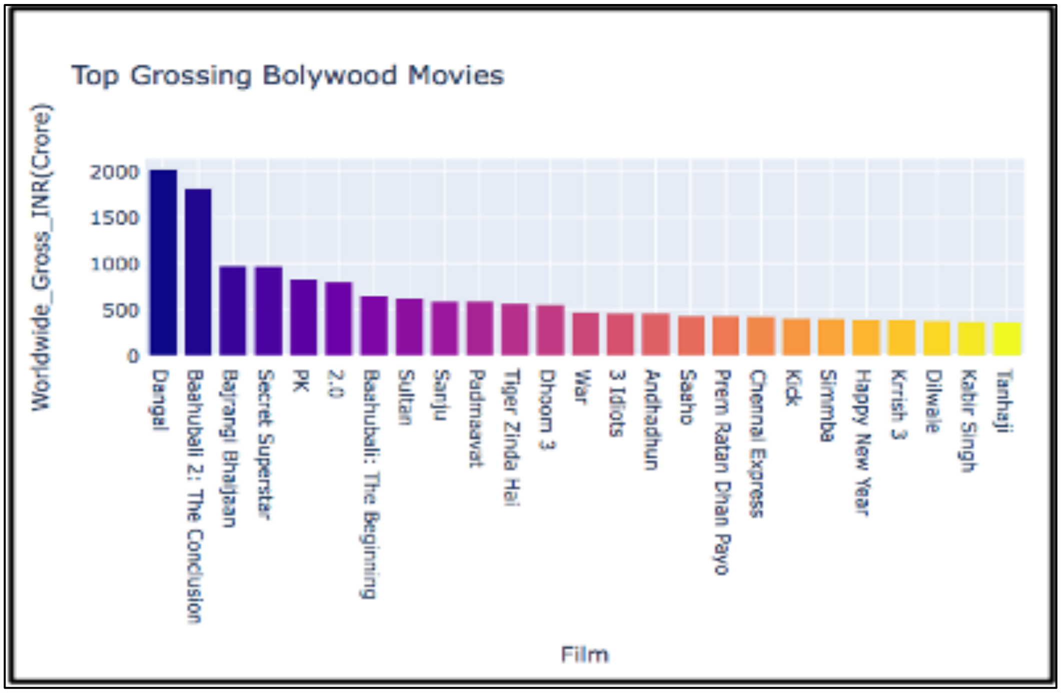


Figure 3. Top Grossing Bollywood movies over the last few years.

From the last graph, we can see the top grossing Indian movies in terms of Indian currency. It shows that “Dangal” made the most amount of money during that phase, but it is unclear of the attributes considered while plotting this graph. Over the period due to inflation the currency has made an effect, along with the price of tickets that have increased over this time and hence the ratio of budget of the movie to that of box office collection should be considered as a factor to calculate the top grossing movies. Also, the number of theatrical releases (number of screens it was released in), time of release (released on some occasion or festival) etc. are the parameters that needs to be considered for thoroughly visualizing and capturing the accurate results.

In our current project, along with visualizing different outcomes of movies based on various parameters, we are also trying to overcome some of the discrepancies that are present in the existing visualization related to budget, movies produced over the years, etc.

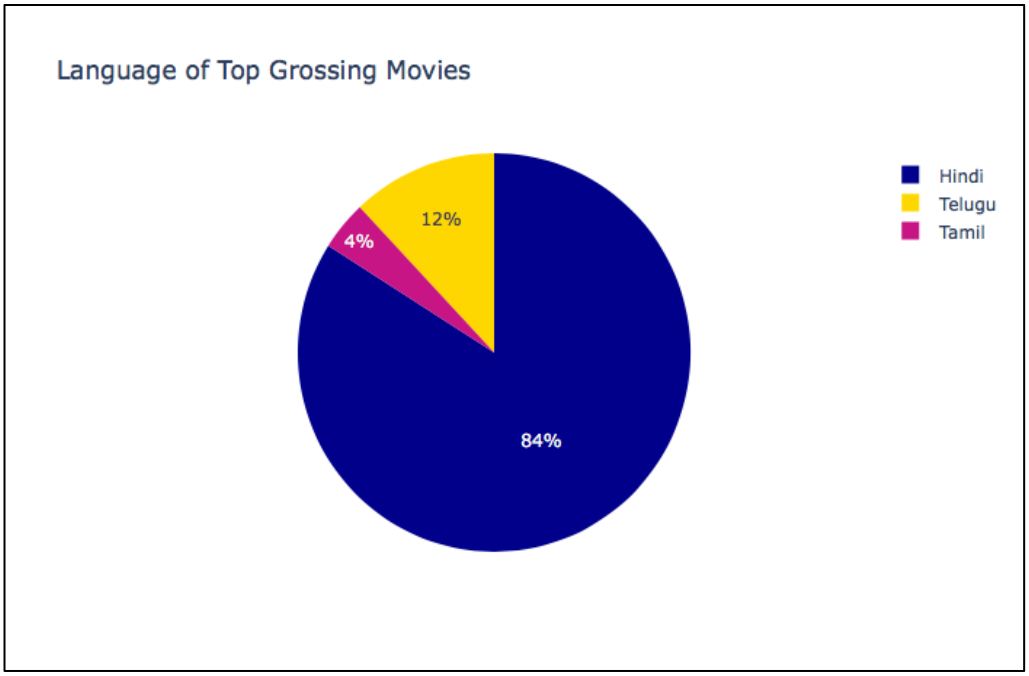


Figure 4. Top Grossing Indian Movies differentiated based on languages.

This pie diagram shows the top grossing movies with respect to Indian languages. As we can see the Hindi language has the most grossed movies in Indian cinema, which is clearly visualized using pie diagrams, but this does not describe the years these movies are produced or whether these movies were made in multiple languages with clear distinction. Also, this might be the data from the last decade since the number of movies released in different languages has seen an exponential growth in recent years.

One of the examples being "Bahubali" and "Bahubali 2" which was released in multiple languages and screened worldwide and the success of those kinds of movies cannot be attributed to just single language. Hence this can be further enhanced by making a clear distinction of the original language of the movie being produced and different graphs or visualizations showing the dubbed versions of movies earning overall.

Objective

The project's final objective is to look for patterns between key factors, identify outliers that have no bearing on our conclusion, and establish any relationships between the features in our dataset.

Finally, we would like to visualize answers for questions like:

- The number of movies produced each year with prediction of the number of movies that can be produced in coming years with different types of genres.
- The number of factors that affect the success or performance of the movie.
- When are movies most profitable?
- Which genres have been profitable over time?
- How has movie budget affected profit?
- Which directors make the most average profit?
- How many leading actors have worked with each other?

We will try answering each of these questions with different visualization to accurately derive some results for a movies successful performance based on different parameters.

Process

Data Preparation:

For our project, we will use the TIMDB (The Indian Movie Database) dataset. This is a "well-structured database for Indian Movies" that provides data on Bollywood films and other features from 1950 to 2019. The data frames are separated by decades and must be merged, to precisely visualize them. Then we must clean the collated datasets by replacing missing/null values, evaluate them by extrapolating features, and use our visualization techniques to create esthetically pleasing infographics. Also, we have used datasets from multiple sources. The datasets were in comma separated value (CSV) format. The dataset consisted of different csv files that were merged to form a single source of data. Additional attributes were added by writing API scripts to fetch data from IMDb databases. After the dataset was formed, pre- processing such as data cleaning, normalization and encoding was performed. Performing such processing on a dataset will help increase the quality of the dataset, which will improve the accuracy of visualizing the dataset and will give us a clear view of the insights.

Analysis of data, candidate visualization methods, failed experiments, why certain method did not work while others did, and so on

Analysis of data:

Reading bollywood_full.csv

bollywood_full.csv is the dataset we are using for this visualization project. It is small dataset of Bollywood movies form the years 1950 till 2019 having attributes like original_title , imdb_id, year_of_release, runtime, genres, imdb_rating, actors, wins_nominations.

Reading bollywood_box_clean.csv

The cleaned dataset has movies from the year 2017 to 2020 with important columns like box office that we intend to use. It also has other attributes like movie_name, movie_total, movie_total_worldwide, movie_genre, movie_director, runtime, etc.

Visualization Methods:

Line plot:

- We made use of a line plot to generate a time series graph for the number of movies produced per decade
- This enabled us to study the prevalent conditions of each decade with respect to number of movies.

WordCloud:

- WordCloud seemed a viable methos to visualize the most liked genre from our dataset
- Although WordCloud are not that effective, they provided us with most like genre from Bollywood

Barplots:

- One of the most prevalent styles of visual is a bar plot. It depicts the connection between a numeric and a categorical variable. Each categorical variable entity is represented by a bar. The numeric value is represented by the size of the bar.
- We used Barplots where we wanted to get he highest, the lengthiest, the topmost, the least, etc. of certain attributes.

Scatterplots and Bubble Charts:

- We made use of bubble charts to chart data and answer questions like most liked genre , find relations between runtime v box office, relation between runtime and rating etc.

Boxplots:

- Boxplots were used to find gross of the movies by month to find out which month made the most money and why so.?

Failed experiments

#1

- We made API calls to the IMDb database through IMDbPY to get important attributes like budget and box office in our bollywood_full.csv dataset

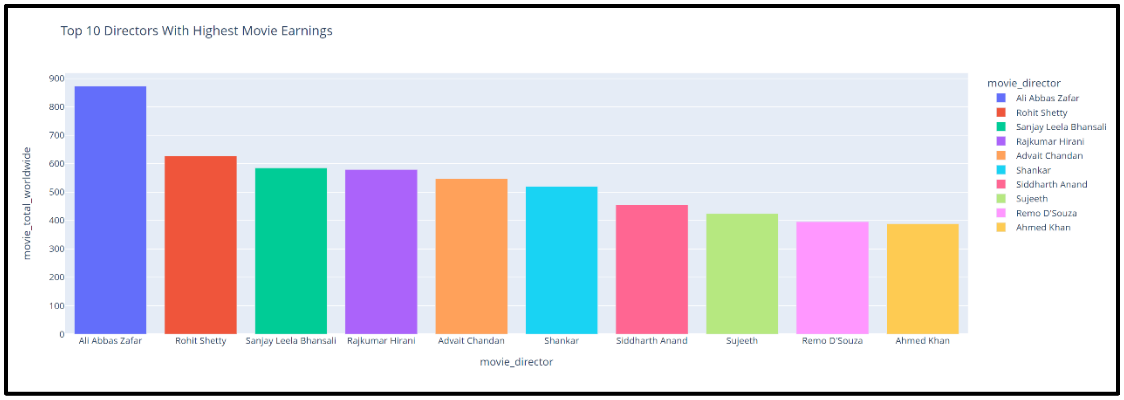
- Obtaining budget and box office data for the dataset would have revealed several relevant elements such as the link between runtimes, movie release dates, cast/crew relationships, and so on.

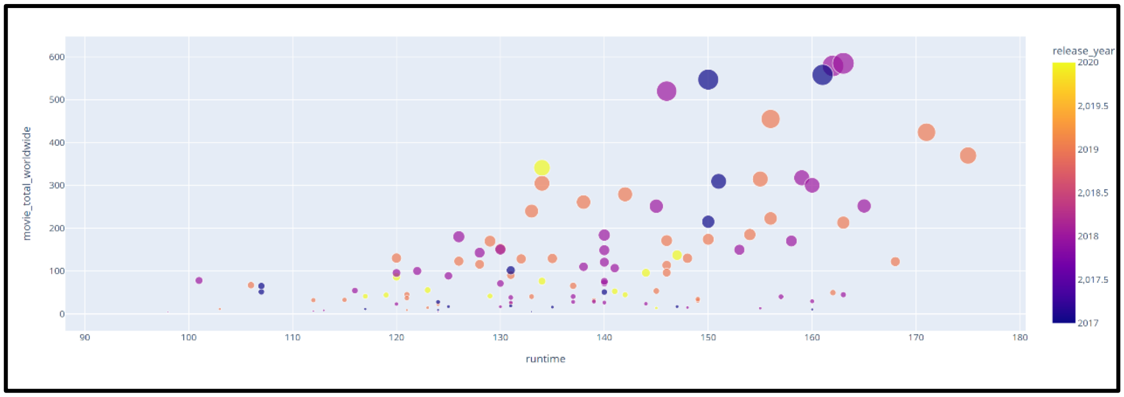
- Though we were able to collect budget and box office figures for the films, many of them were null values, and in certain cases, movies and box office figures were lacking budget figures.

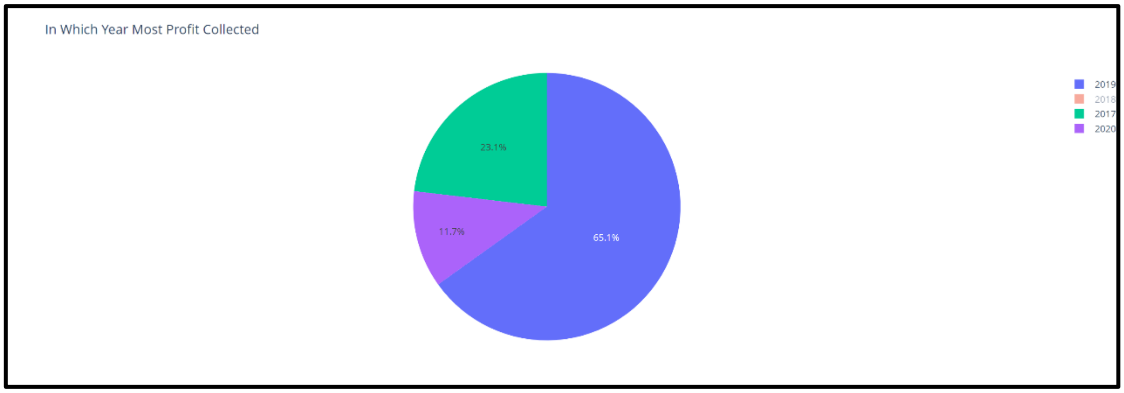
- This illustrates the main issue with data collection for Bollywood or Indian films in general: there is no proper data accounted for throughout the years, at least not in one source, and even if they do exist, most of the data from multiple sources contradict each other.

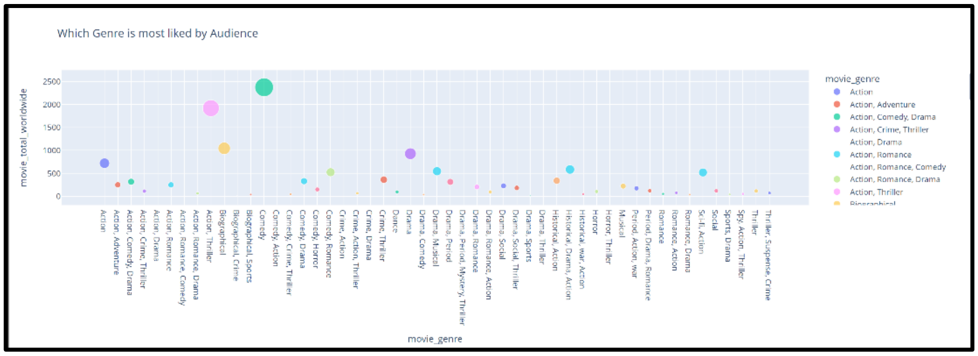
#2

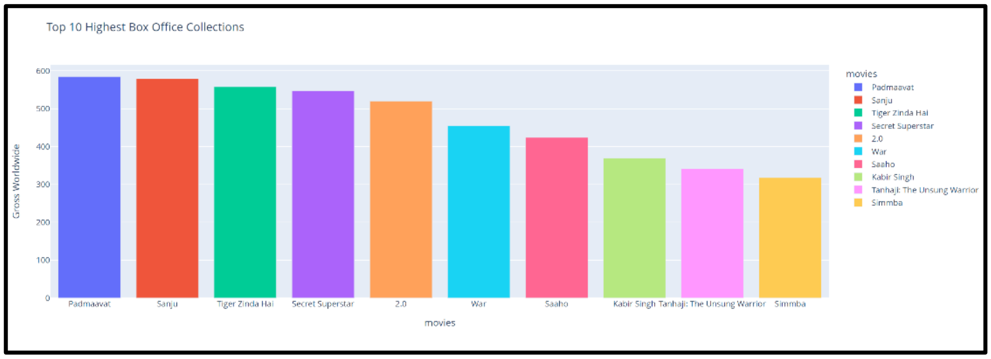
- We wanted to create interactive visuals for our datasets using plotly (specifically plotly express).
- •The plotly express renderings were visually dynamic, beautiful, and data from those representations was easy to interpret.
- The plotly visualizations are as below:

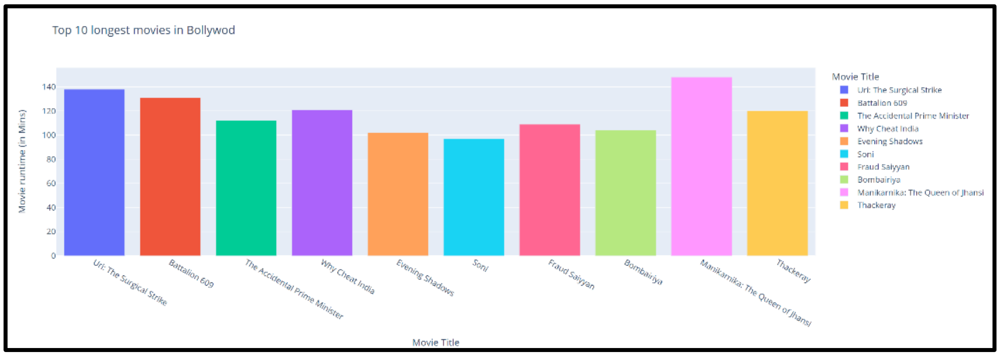












- Although we were able to generate visualizations with our data, the ix were not getting rendered on GitHub and when he ipynb was converted to html/pdf
- It was not rendering when the python notebook was run on another machine
- This was a peculiar problem since the libraries were installed on all the machines and the notebooks were converted to html via nbconverter
- Lastly, the notebook with plotly visualizations were not rendered even on nbviewer

Results:

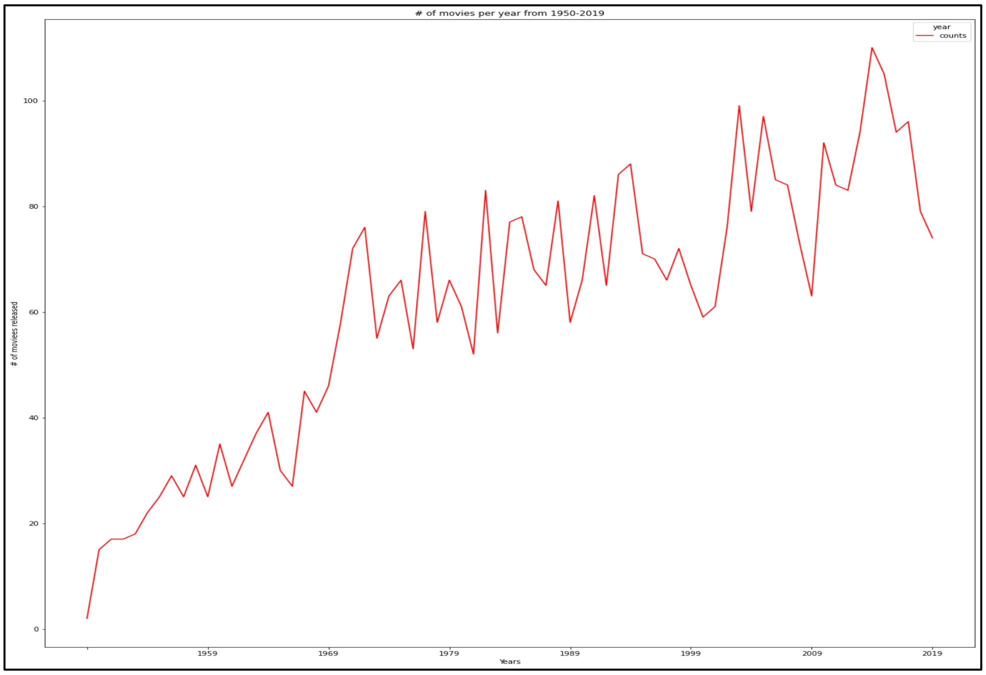


Figure 5: Number of movies per year from 1950 to 2019

The line plot offers us a quick overview of how Bollywood movies have changed over time.

- It is projected that just 10-12 movies were made and released for public consumption during the start of the 1950s. This is in keeping with the country's and industry's socio-political and economic realities. India was a young country, having gained independence in 1947, and the movie industry, which was only two decades old at the time, was never viewed as a viable source of revenue or employment.
- The number of releases increased as the country's economic conditions improved.
- The line plot shows a sharp spike between 1969 and 1977, which tells us two things:
    1. The movie industry had a higher per capita income, allowing it to produce more films.
    2. The country’s economic status has improved.

This was also the time when the Bollywood industry began to export its films to other "Third World Countries" that shared many political, economic, and cultural similarities with India, and the Bollywood films represented those themes.

A substantial increase in movie releases can be seen between 1990 and the early 2000s. This was the period of liberalization in India's economy, and the Bollywood business was one of the key beneficiaries of this period of liberalization. Several international companies, notably Disney and Sony, established operations in India to produce Bollywood films.

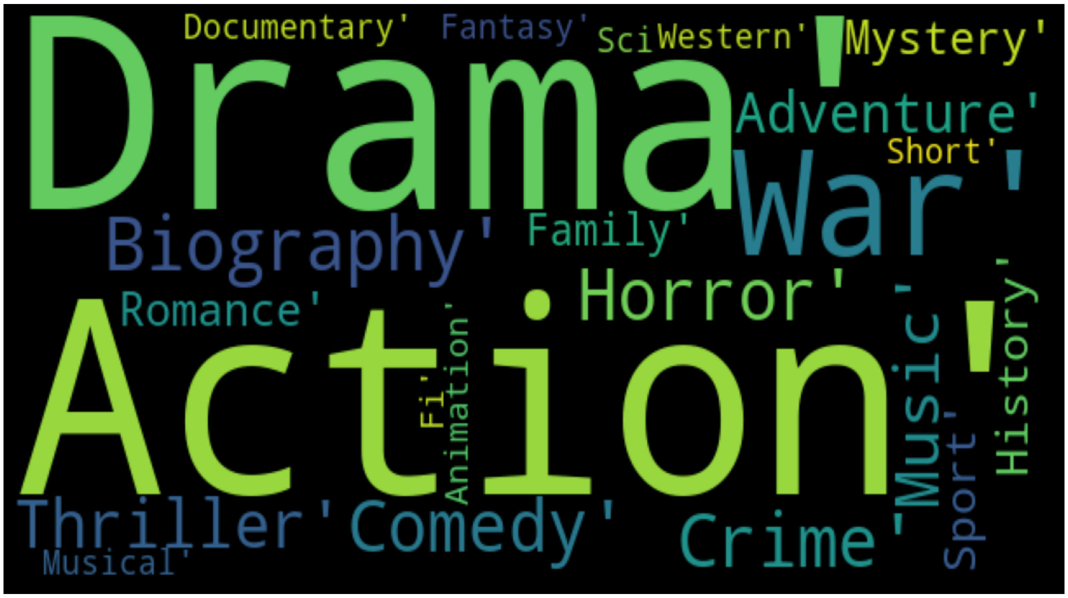


Figure 6: Word Cloud showing the most viewed movies by genre

Insights based on WordCloud

- The word cloud highlights that Drama and Action are the most popular genres among Bollywood moviegoers, as proven by the overwhelming number of films produced in the category.
- Inconsistencies and the difficulty of dealing with numerous genres
- It is common knowledge that many films cannot be classified into a single genre. Regardless of the language spoken in the films, this is true. The majority of these

genres are grouped, which might be problematic when analyzing genre data from a movie dataset.

- To circumvent this problem, we were able to create a list of distinct genres based on the number of IMDb votes each genre received. One hot encoding solved the problem of binarizing many genres.

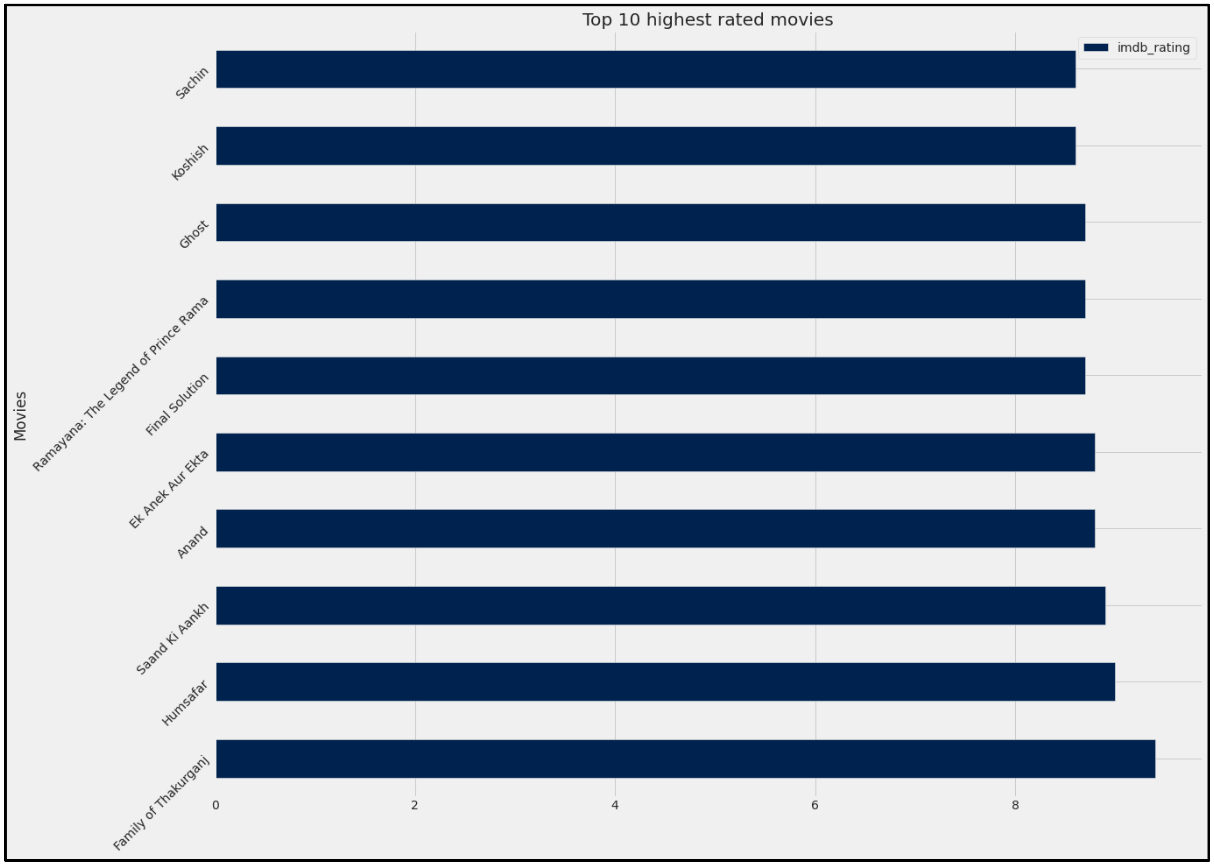


Figure 7: Barplot for Top 10 highest rated movies

Bar Plot based insights

The curious case of biases in the data

- From the bar plot, we get the top 10 highest rated movies in the dataset
- while it is not evident for those who do not follow Bollywood, all these movies are relatively unknown and most of them just barely qualify as a movie (The movie "Sachin" is a documentary that premiered in theatres and not on any streaming platforms)
- Now, we come to the faults in the data, IMDb calculates a movie's rating based on the number of votes it received i.e.

Avg Rating = summation of all ratings/summation of votes

- So, the data is skewed for movies that have fewer votes but a higher rating per vote. This is the case with our graph.
- Our highest-rated film, 'Family of Thakurgunj,' has an average IMDb rating of 9.4, but the total number of votes it has received is just 1600, proving that our presumption that it is the highest-rated film is inaccurate. The rectified graph is as below:

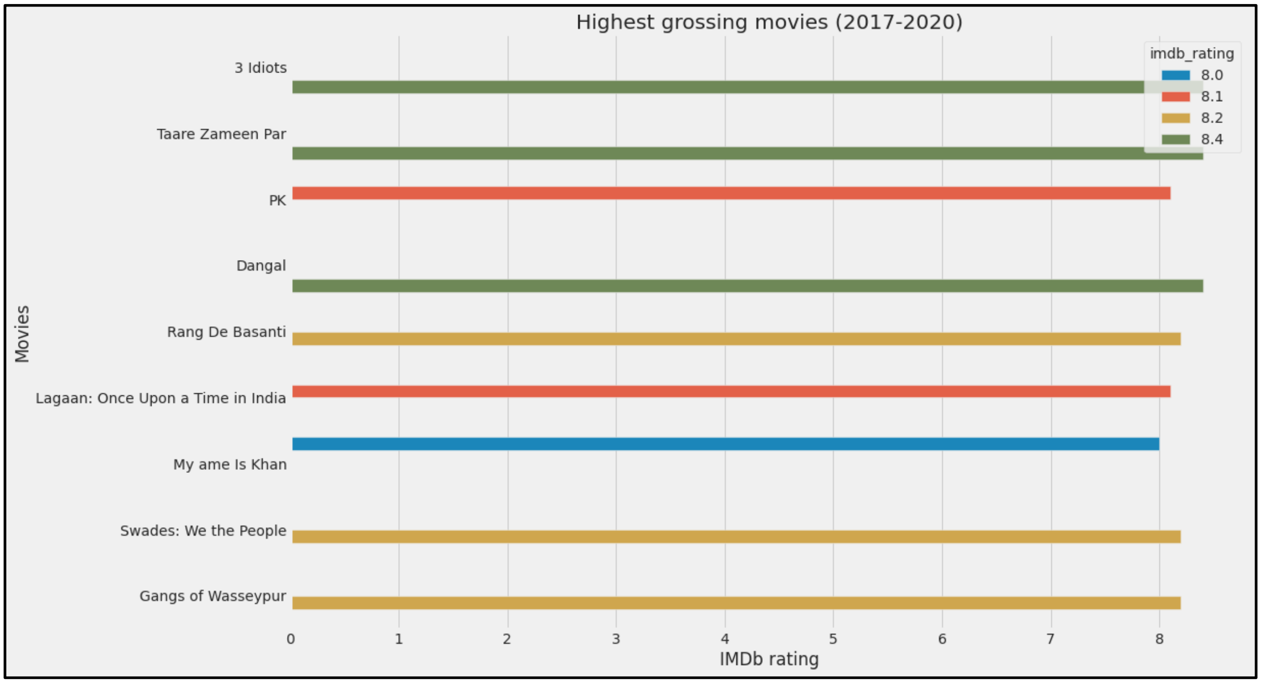


Figure 8: Rectified bar plot for Top 10 highest rated movies

Histograms

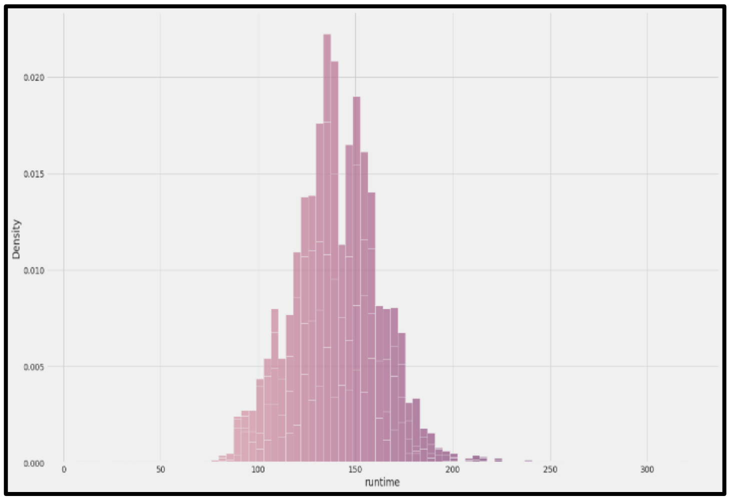


Figure 9: Movies by runtime on box office

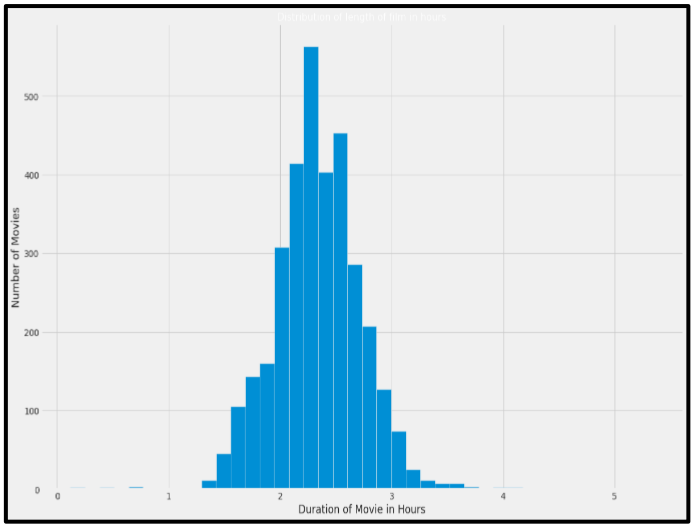


Figure 10: Movies by duration in hours

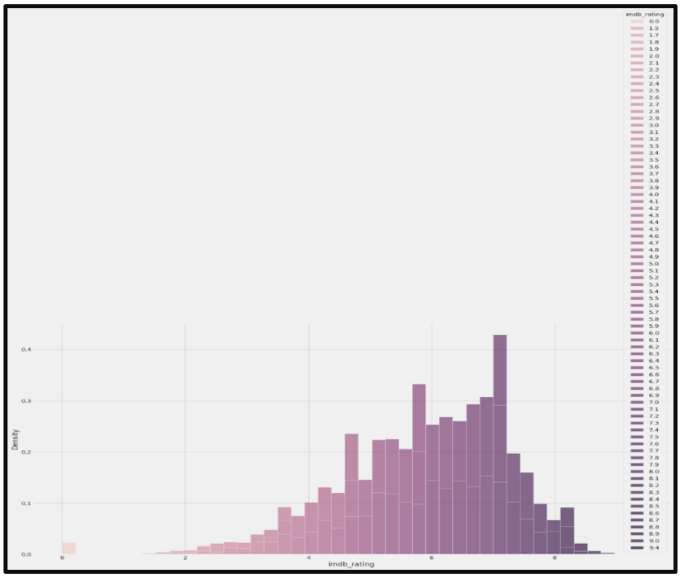


Figure 11: Movies plotted by imdb_ratings and runtime in minutes

We generated histogram plots for attributes like imdb_ratings, runtime, and duration of movie in hours etc. Through these plots we were able to get a general idea of Bollywood movies and its longer runtimes, higher ratings, and the average duration of a Bollywood movie through the years.

Histogram based insights

- From the above histogram, it is clear that there are a few outliers in the imdb_rating column
- Also, a lot of movies in the dataset have a rating in the range of 6 to 8
- There are approximately 1500 movies in our dataset having runtimes ranging between 2-3 hours
- We plot another hist based on runtime in minutes to get an average runtime as accurately as possible
- From the above hist plot, we gauge that Bollywood movies in our dataset have an average runtime of over 140+ minutes.
- This average runtime for Bollywood movies is approximately 40+ minutes more than the average runtime for movies made in Hollywood.

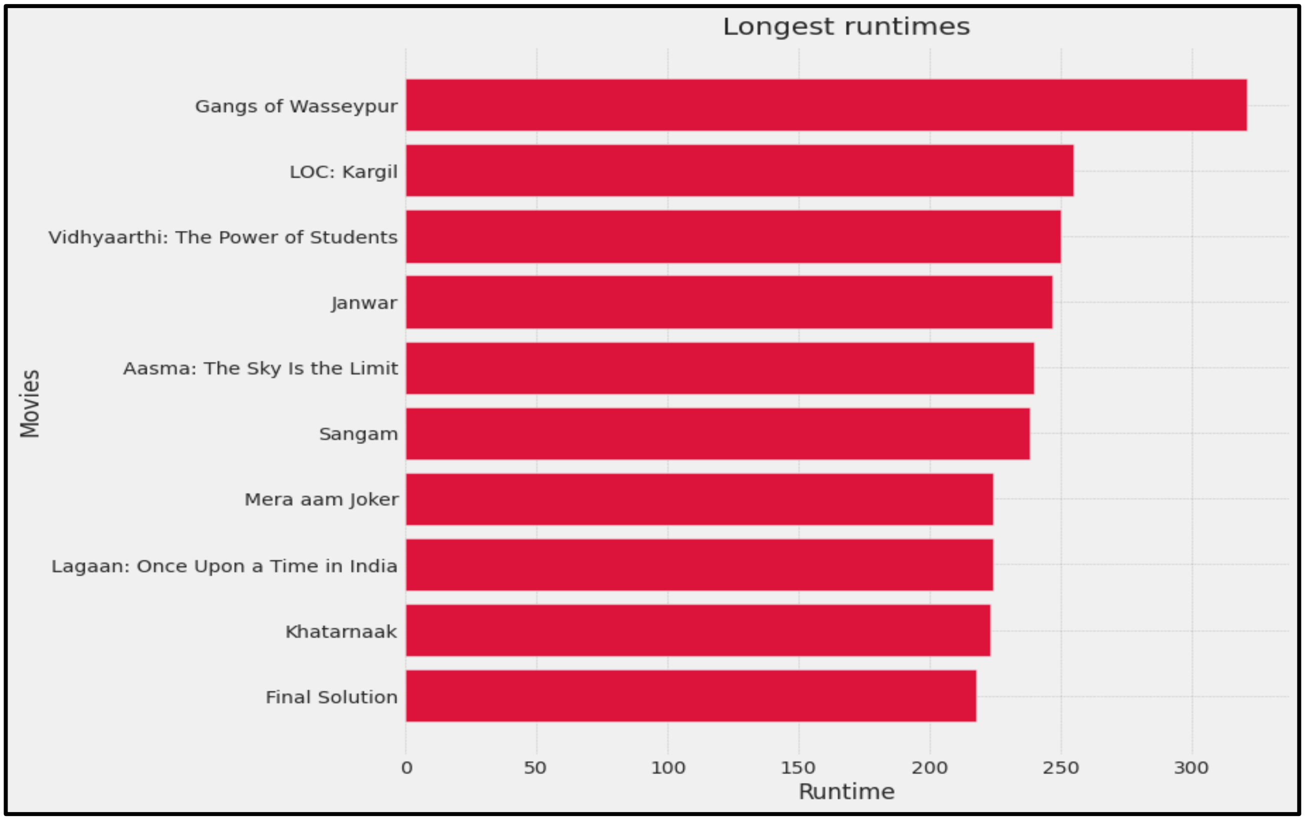


Figure 12: Top movies based on longest runtime

The bar plot above is straightforward, it gives the longest runtime for a Bollywood movie. But what is so interesting about a movie with a runtime of 321 minutes?

<u>Interesting fact</u>

The outlandish runtime of Gangs of Wasseypur

- As we see from the bar plot, the longest Bollywood movie is Gangs of Wasseypur having an approximately runtime of 321 minutes, i.e., just over 5 hours.
- Though the movie was released in 2 parts, IMDb considers the runtime when it was first screened at the Cannes film festival as its official runtime.
- According to the wiki of gangs of Wasseypur
- Both parts were originally shot as a single film measuring a total of 321 minutes and screened at the 2012 Cannes Directors' Fortnight, but, since no Indian theatre would volunteer to screen a more-than-five-hour film, it was split into two parts for that market.

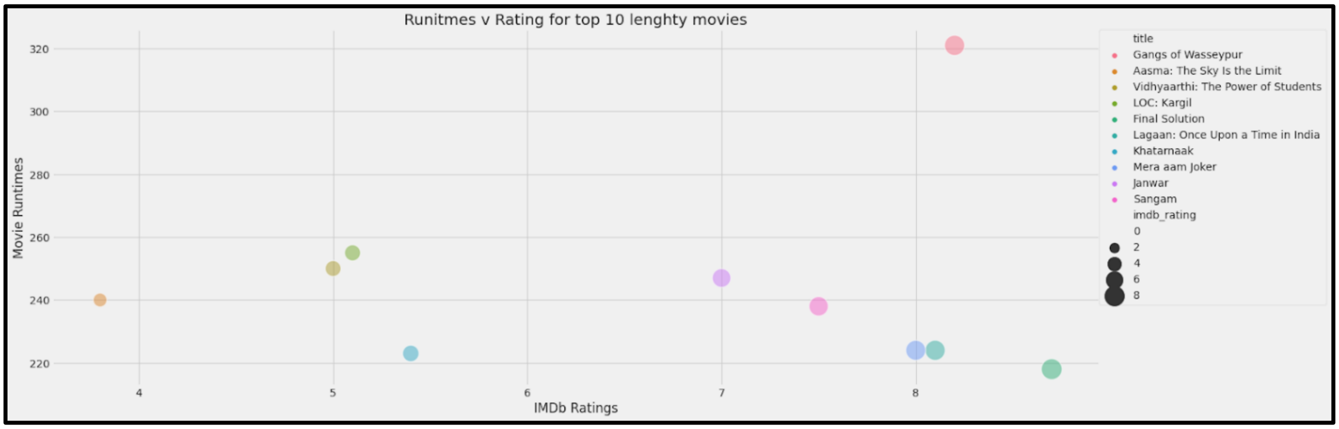


Figure 13: <u>Scatter plot for top 10 movies based on longest running vs its ratings</u>

A scatterplot of Runtime v Rating

- The majority of the top ten lengthy films have an average rating of 7.0 or above.
- The above graph is inconclusive as to whether there is a correlation between longer runtimes and higher ratings.
- No relation can be derived with our limited amount of data, but we do believe runtime does effect on how the audience views it and have an indirect effect on the rating

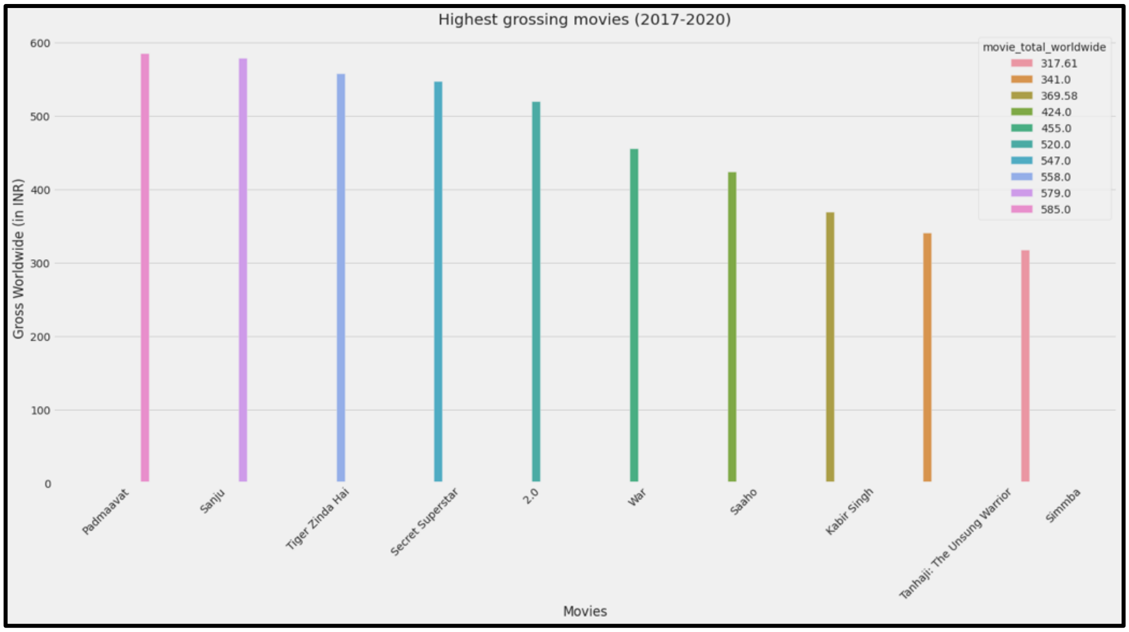


Figure 14: Highest grossing movie from 2017 to 2020

- Padmaavat is the highest grosser in the last few years.
- Data is limited and biased and the plot obtained is not accurate in a general scenario.

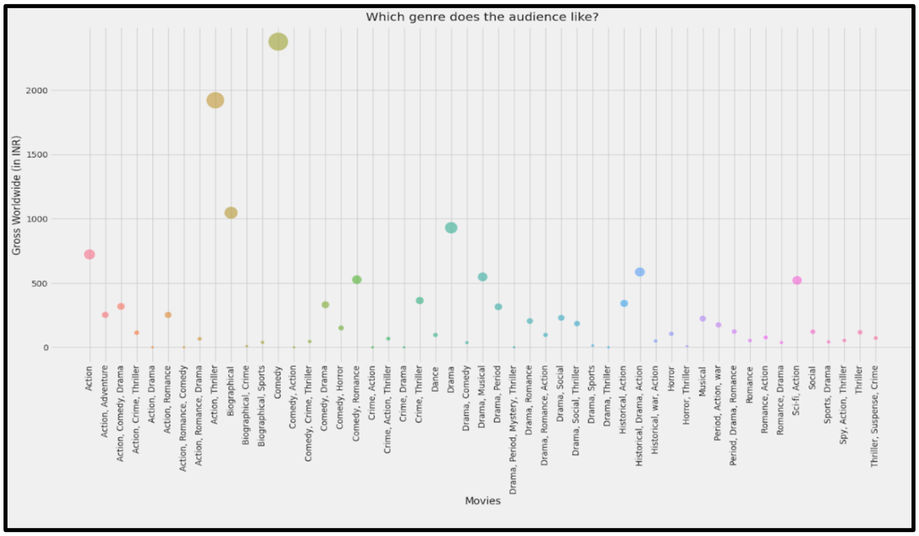
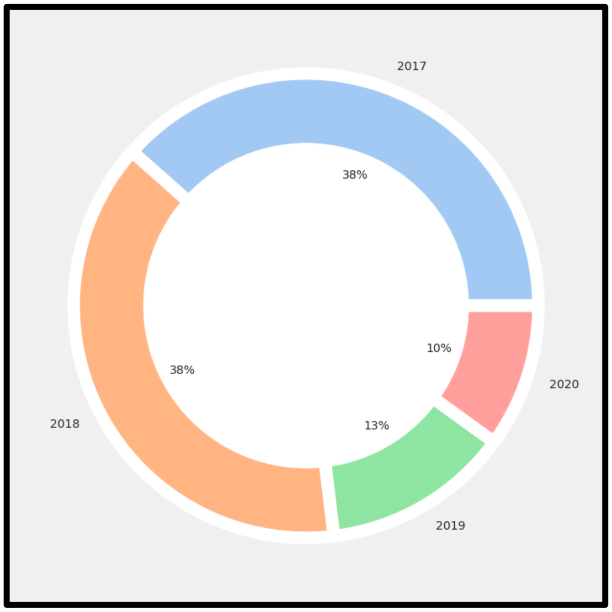


Figure 15: Most liked genres based on audience's ratings

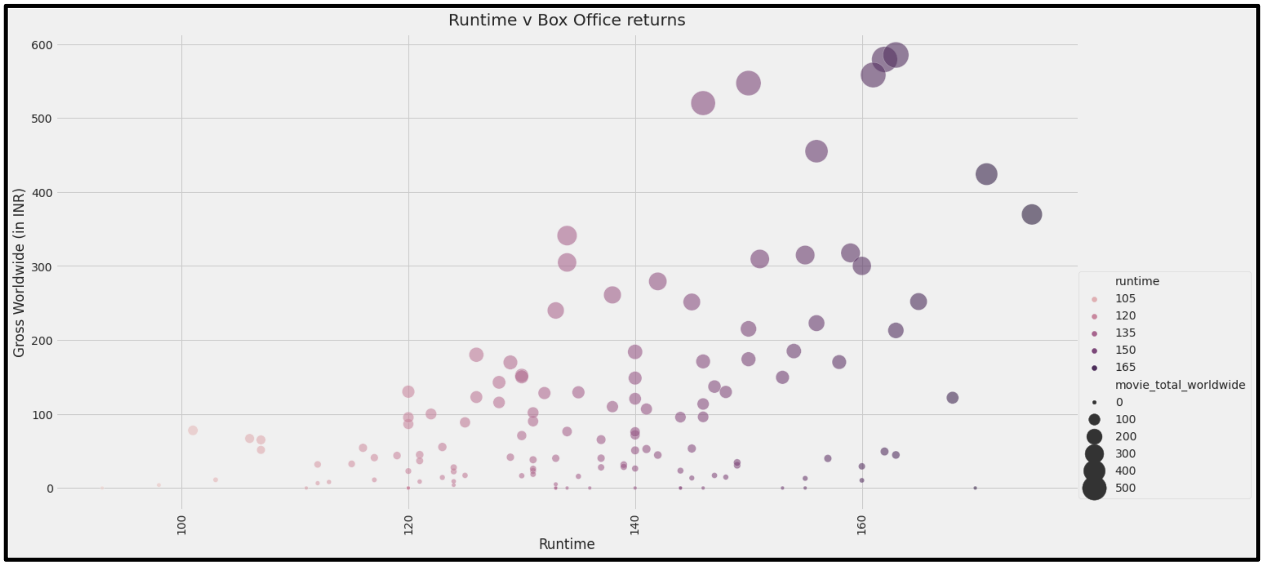
- From the previous WordCloud about genres, we inferred that Action, Drama and its related genres were the most liked by the movie going audience.
- This trend has changes in the last few years since " Comedy" is now the most like genre along with Action and Thriller
- This shows a change in the liking of the audience where they ask for more diversity in movie genres



This is a donut chart about the percentage of movies released in theatres per year since 2017, Although this visualization does not reveal a whole lot of things, an important thing to see is that the number of releases has decreased considerably from 2018 to 2019. This change in tides can be directly attributed to the advent of streaming platforms on Indian cinema. As the popularity of platforms such as Netflix increase, the way we analyze movie releases and audience

reactions also needs to be reconsidered

Figure 16: Percentage of movies released each year since 2017



.

Figure 17: Scatter plot showing the top grossing movies based on its runtime

- The scatter plot tells us that the highest grossing movies have runtimes over 160+
- This also validates our earlier plot where we had an average runtime of 140+ minutes for movies between 1950-2019.

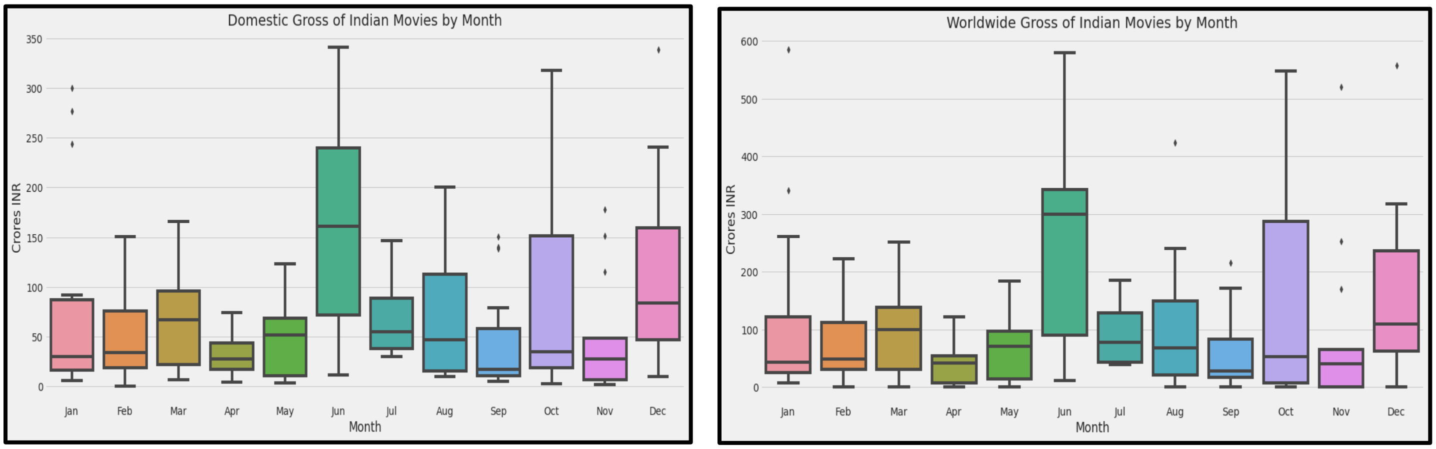


Figure 18, 19: Monthly domestic and worldwide gross of Indian movies

Boxplot based Insights:

- The boxplots give us a range of earnings per month.
- The boxplots are mostly similar thought the year except for months like Oct and Dec.
- These months have extended holidays owing to festivals etc. which makes them viable for a movie to be released during this time.
- The earnings for June are an outlier since it is directly influence by one movie in the dataset.

Challenges:

The problem with the dataset is that it is confined to films released between 2017 and 2020 and lacks a budget column, making our goal of establishing relationships between profit and other variables difficult. Due to lack of accurate data available, we were not able to accurately find the outcome of the movies success in terms of running time or highest grossing if compared with results available online. Also, since both of us were new to the Data Analytics/Visualization field, it was an exciting and challenging project considering the knowledge of visualizations we had before starting this project. We had to understand the data preprocessing, data preparation, implementing different algorithms and visualizations and finally coming up with the outcome based on different combinations of the attributes. This course has helped us a lot in that aspect considering the number of visualizations we have learnt along with the use case of a particular visualization to use. The most important challenge was to visualize the profits or success that a movie will have based on different permutations of attributes available since even a slight change in input parameters lead to changing the outcome.

Discussion and Future Scope:

There are a lot of derivations that can be made and insights that can be generated using the dataset but considering the constraints, we are currently focusing on few of those, and this can be scaled to include so many other visualizations for comparison few are which are discussed. We can expand this project to include all the other cinemas apart from just Bollywood and make it truly the "Indian Cinema through Data". We can also combine the visualizations from different cinemas of various countries and find interesting differences between each of them. This could form the basis of what each country could adapt from other countries' cinema to make the movie successful. Also, current scope only considers movies that are released in theatres or cinema halls and doesn't consider the movies released on OTT platforms and hence this can be expanded to include those data as well and the series that are released. Furthermore, different interesting comparisons can be visualized between Movies released in cinema halls, movies released on OTT platforms, series released etc.

Conclusion:

As stated earlier, because we were both new to data visualization when we started this project, we had to confront a lot of hurdles and learn a lot of new things. When it comes to the data part of data visualization, our first issue was with the data. The information was insufficient. It was confined to only a few elements and left out several important characteristics that would have helped us accomplish our goals for this project. Secondly, our efforts to call the IMDB API to obtain these attributes were unsuccessful. After resolving these problems with the data, we were able to generate line plots, box plots, histograms, bar graphs, bubble charts and pie charts for our datasets. We were able to generate intriguing visualizations and relationships between box office collection, month of release, and number of releases every decade with connection to the nation and the industry's socio-economic, economy, and cultural milieu, as demonstrated in the results and insights. We learnt several aspects of visualization because of this project, and we were able to analyze and visualize Bollywood films and its idiosyncrasies.

References

[1] https://www.indiewire.com/2013/07/10-things-you-should-know-aboutindian-cinema-37021/

[2]]https://www.veryshortintroductions.com/view/10.1093/actrade/9780198723097.001.0001/actrade-9780198723097

[3] [https://github.com/pncnmnp/TIMDB

[4] https://github.com/fasouto/awesome-dataviz

[5] https://datavizcatalogue.com/

[6] https://towardsdatascience.com/visualisation-of-bollywood-2c973328ddc9

[7] [https://www.indiewire.com/2013/07/10-things-you-should-know-aboutindian-cinema-37021/

[8] https://blog.gramener.com/movie-data-visualizations-data-storytelling-hackathon/

[9] https://towardsdatascience.com/visualisation-of-bollywood-2c973328ddc9

[10] https://sudevsheth.medium.com/guftagoo-part-2-visualizations-85d78482bf2d

[11] https://www.varsity.co.uk/film-and-tv/17664

[12] https://www.livemint.com/Consumer/BVVx6EV79uZcrkZBgAky3H/The-liberalization-of-Bollywood.html