Overview:

Here in this project we want to combine 3 data sets that are related to the Indian movies, and make some analysis to understand if there is any relation between the duration, number of languages a Indian movie is released , votes etc with the rating of the movie. We also want to provide some best Indian movies at the last of the EDA, the criteria of selecting some good Indian movies may depend on the rating and the number of languages it is available in. We also want to check out who are most famous actors among the Indian movies and what are most popular genres in the Indian movies.

India has the largest film industry in the world , in terms of films released with 1500 to 2000 movies a year across 20 languages. There can be n number of movies releasing each and every day.

So to do all the above mentioned stuff, we have been looking through the web to find some data and found 2 datasets on Kaggle, those are

50000+ Indian Movies Dataset

IMBD 10000 Indian Movies.

What is point of analyzing movie datasets, without knowing the language .The first data set did not contain the language of the movies, but the second dataset had the language, so we hereby merge them to know the movie also.

So we have merged these two data sets on the common column , title and we have received a dataframe of 52000 records, but unfortunately most of the records where having null values.The records having null values would not be helpful for analysis , so we have eliminated the records that have null values and then we had 6920 records.

We observed that some records have all the same information , but the same movie was released in different languages , so we basically made a single record for the single movie and made the language column , look like [‘hindi’,’telugu’], instead of having 2 records. So here we have optimized the records, and we also made sure that there were no movie records containing languages like [‘hindi’,’telugu’,’telugu’].We basically made sure that the movie languages are not repeated.

Then we found the third data set from data world, we had the movie names with the cast names, so we merged the third dataset with the above merged dataset and made the a unique data set .

So our final dataset is gone be of 2000 unique rows .

DATASET INFO:

Here in this project we wanted to know about the Indian movies, their ratings , number of languages they are available and etc. So in this process we have been searching for the data sets. Firstly on the Kaggle we have found a dataset with 50000 records but in this data set there was no language feature. There is no point in knowing about the movie if there is no information about language and the Indian movies are know to be released in many languges.

So we have been looking for some datasets that included the movie language, so we found one in the Kaggle with 10000 records in it .

Here is the information about both the datasets.

Dataset 1 – Features :Title,Year,Certificate,Runtime,Genre,Desc,Rating,Votes

Dataset 2- Features : Id,Title,Year,Timing(min),Rating,Votes,Genre,Language

Frankly speaking , in the data set 2 we only needed the language and id features , so we have sliced the data frames and merged the data frames and got the language added into the merged data frame. Then we that there were many records with the same title, and everything was same , only thing different was a language name, so we have clubbed the records into a single record and converted languages into a list and then converted the data type of the column into set , so as to eliminate the duplicate values, that is languages in the list to make sure that a language is not repeated while clubbing the records. So here we have optimized the language column of the data frame and we also created a new column that is the language count , by counting this list elements in the language column.

Then we have found out a third dataset from the data-world , which had movie titles along with the cast of the movie, and the director of the movie.We have combined this dataset with the above merged dataset and at last we got out final unique dataset.

The final dataset had these Features:
Title,Year,Language,Votes,Rating,Genre,Runtime,Certificate,Id,Director,Actor1,Actor2,Actor3,Language Count.

To create the above unique datasets we have done cleaning and wrangling stuff . There were roman numbers in between the values of year columns, we have made sure we only extracted the year numbers excluding the roman numbers, we replaced the commas in some of the columns and used astype to convert them into the integers and float datatypes to make the analysis easy.

Exploratory Data Analysis:

Here to perform the Exploratory Data Analysis, we have compared the different attributes of movies with the ratings, that we have checked if there are any factors that would be related with the rating of movies and we have also plotted some word clouds, which are nothing but a plot to depict the most frequent words of the given source. We have used the word clouds to see who most famous actors are, famous directors and what are most popular Genres.

1.Famous Actors In Indian Movies

2.Famous Directors In Indian Movies

3.Popular Genres

4.Relation Between The Votes and Rating of movies

5.Relation Between Language Count and Rating

6.Relation Between Runtime and Rating .

7.Relation Between Rating and Certificates

8.Suggesting Some Good Movies Based On Rating And Language Count.

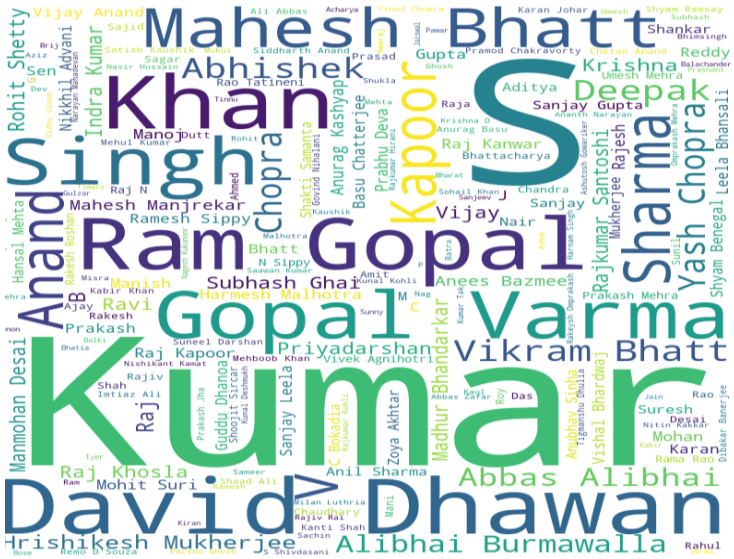
The above points were analyzed in the EDA process and the for the first three point we have used the word cloud approach to visualize.

1.Famous Actors In Indian Movies

Here we can see that there are many famous actors in Indian Movies that can be noticed from the below word cloud they are Amitabh Bachchan, Shah Ruhk, Salman Khan, Sanjay Dutt and many more.

2.Famous Directors In Indian Movies

Here we can see that there are many famous directors in Indian Movies that can be noticed from the below word cloud they are Kumar, Gopal Varma,Sharma, David Dhawan and many more.

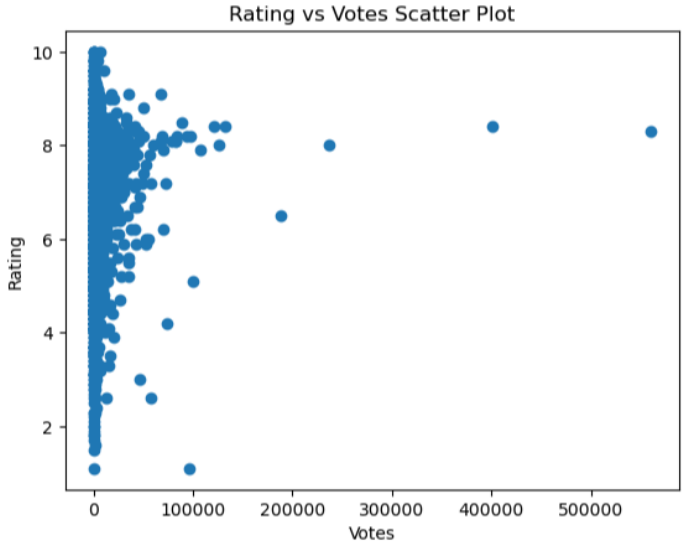


3.Popular Genres

Here we can see popular genres in Indian Movies that can be noticed from the below word cloud. Most of the popular genres are a combination of comedy and drama, drama and romance, action and crime and thriller and action and many more.

4. Relation Between The Votes and Rating of movies

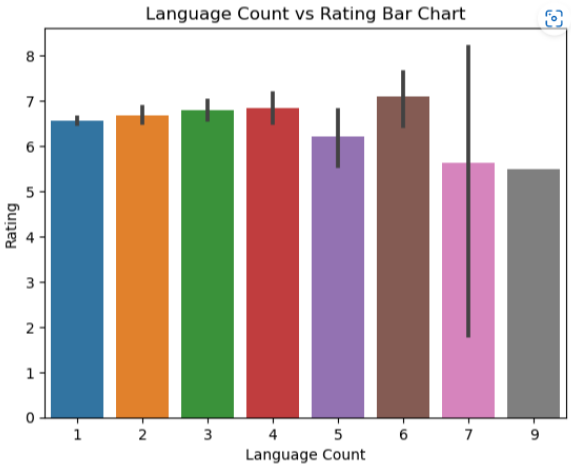
Here we can notice that votes of the movies that lie under 20000 does not have a certain pattern for the rating , many of the movies that are voted between 0 to 10000, have a scattered rating from 0 to 10,there are no noticeable amount of movies from 10000 votes to 20000 votes, there are only some movies with votes more than 20000.We can also notice that the movies which have the votes above 20000 for sure have rating above 6.



5.Relation Between Language Count and Rating

We can see that there is no noticeable relation between the language count and the rating.

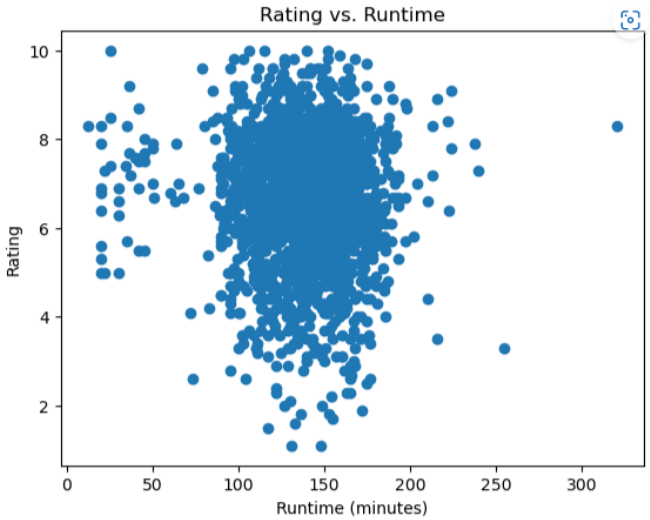
We can here notice that some Indian Movies are also released in 9 languages.



6.Relation Between Runtime and Rating .

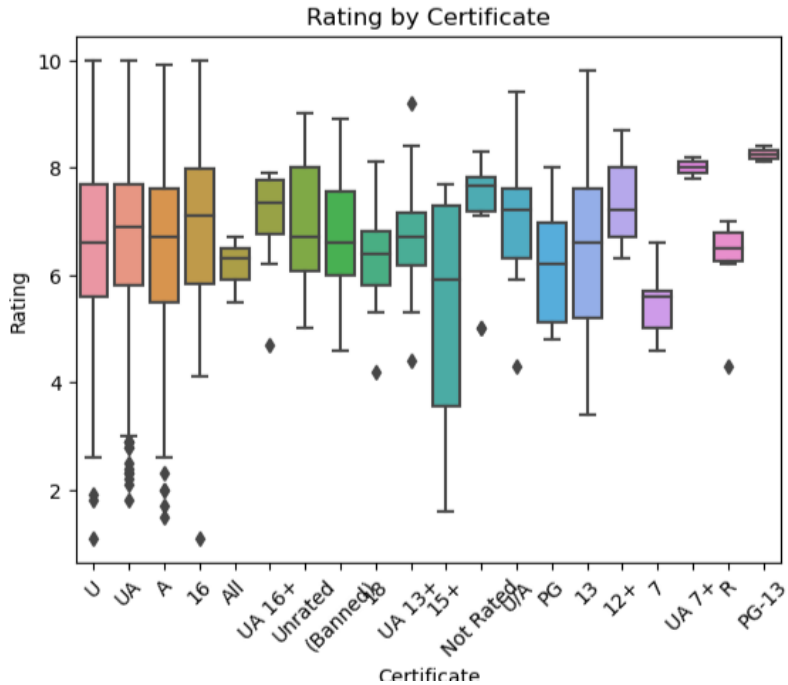
Here we can see the relation between the duration (runtime) and the ratings of the Indian movies.We can notice that there are many movies that lie in the duration between 100 mins to 200 mintues.Most of the movies that are in this interval,are rated between 4 to 10.

We can notice that the movies with low duration that is below 100 minutes have only ratings above 5 and majority of them are above 7.There are very few movies which are above time 200 minutes.



7.Relation Between Rating and Certificates

This is a box plot that shows the relation between the certification of the movie and rating. Observations are UA conists of highest number of outliers followed by 16 and there are no outilers in ALL,UA 16+,BANNED,15,+PG,13,7 UA7+, PG-13



8.Suggesting Some Good Movies Based On Rating And Language Count.

Here I have suggested some movies that are rated above 8 and are released in more than 3 languages.

	title	year
25	777 Charlie	2022
52	Aaina	1993
112	Adhurs	2010
417	Bobby	1973
419	Bodyguard	2011
541	Coolie	1983
621	Devi	1960
688	Don	2006
699	Double Dhamaal	2011
706	Dulha Mil Gaya	2010
1101	Jackpot	2019
1132	Jeeva	2014
1155	Johnny	2003
1195	K.G.F: Chapter 2	2022
1369	Kurup	2021
1447	Love Story	1981
1990	Rocketry: The Nambi Effect	2022
2199	Singham	2011
2316	Thank You	2011
2406	Tumse Achha Kaun Hai	1969

CONCLUSION:

In this project we have saw if there is any relation between the rating and the other attributes of the dataset like duration, votes and many more with the ratings. We can also known the famous Indian movie actors and directors. We also came to know the about the popular genres in the Indian Movies. At last we also have suggested some good Indian movies on basis of the language's count it is available in and also based on a good rating.