Exploratory Visual Study of Movies and its related factors

Sanchez Innocencia D[1], Kannammal. A[2]
[1]M. Sc Decision and Computing Sciences, Coimbatore Institute of Technology
[2]Dept. of Decision and Computing Sciences, Coimbatore Institute of Technology
sanchezinno1@gmail.com[1], kannammal@cit.edu.in[2]

Abstract

The film Industry is one of the most evergreen industries in the field of entertainment world-wide. Film Industry plays a vital role in the global economy. Though there are a huge number of movies released on a daily basis, only a countable number of movies hit an enormous audience. This paper deals with the analysis of the movies taking into consideration the features like the Genre, directors, the year of release, the budget and the revenue, the success rate of the movies etc. The analysis is being displayed in the form of a dashboard. This dashboard uses Kibana to access and import the dataset of the movies for parsing and processing within Logstash. From the Logstash, the data discovers the top genres, the movie with highest revenue, the top actors and directors as per the number of movies and the countries which make more movies and also the highest ratings of the movies based on the language. Several analyses have been carried out for visualizing the movies and its related factors. This paper helps the film industry understand the success of the movies. The idea of this paper is to build a self-explorable dashboard in favor of the users so that they take control of the dashboard and select the necessary dimensions and measures. And also, it helps the people in the film industry to have an analysis on the revenue.

Keywords: *Logstash, Kibana, Dashboard, Film Industry*

1. Introduction

A major problem in watching a movie is selecting the good movie. Most of the time we spend our time checking for the good movies ending up wasting anywhere between an hour to two just for choosing the movie. In order to minimize it, a highly interactive user-friendly dashboard is created to help the user with a list of movies based on their preferences. Based on the preferences of the users, the interactive views in a single dashboard would allow the user to explore the preferences and further arrive at a shortlisted movie set wherein they do not need to scroll up to multiple pages for getting a particular information rather everything is viewed in the same page which is useful when compared to the traditional rating sites. Kibana dashboard is one where the insights of the data are taken together to provide decisions. The aim of Kibana has always been to enable the users and their organization to go even further with Kibana dashboards by not just supporting data-driven insights, but also enabling data-driven action. This model is a best fit for the producers or the movie enthusiasts where they understand the ratings of a movie even before it is being released based on the factors that can all be determined prior to the release of the movie.

2. Literature Review

Andrei Oghina, et. al.,[1] have predicted the IMDb movie ratings and have considered-two sets of features: surface and textual features. They initially extract textual features from each channel and then use the predicted model and then they explore the data from the channels which helps them get a better set of textual features used for prediction. The best performing model in their analysis helps to rate the movies close to the observed values.

Bharadwaj Naidu Muthuluru, et.al.,[2] have created an interactive dashboard for IMDB movies dataset in order to allow users to explore the same visually. This has been implemented using packages offered in R. This dashboard attempts to demonstrate some aspects of data exploration which are uniquely suitable for visual interactions with movie data. Their dashboard translates these interactions into filters at data level and updates the visualizations accordingly

Abba Suganda Girsang, et.al.,[3] have developed data warehouse architecture for IMDb using kimball methodologies specifically for making a reporting system to analyze actor data like popularity and movie statistics. The dashboard results are expected to be valuable assets to be used by movie stakeholders and future data analysis. ETL is also proven as a best method for converting databases into a data warehouse easily without the need to interfere with the production database.

Nithin VR, et.al.,[4] provides a detailed study of Logistic Regression, SVM Regression and Linear Regression on IMDB data to predict movie box office. The budget of the movies has been of the order of hundreds of millions of dollars, making their box office success absolutely essential for the survival of the industry. Knowing which kind of movies are more likely to be successful and which kind are more likely to fail before the release could benefit the production house greatly as it enables them to focus on their advertising campaigns which itself cost millions of dollars, accordingly.

Klaus Dodds[5], in his paper is concerned with the further theoretical development of popular geopolitics with explicit reference to audience dispositions and reception more generally. Using the film series of James Bond particularly Die Another Day (2002) and the Internet Movie Database (IMDb), it is contended that political geographers need to better understand how the audiences and the fans of the film in particular can interpret the popular geopolitics of film.

3. Problem Definition

Based on the massive movie information, the interesting fact in it is that one can understand the important factors that have been in the backend for making a movie more successful than the others. So, an analysis is made determining what kind of movies are more successful, in other words, get higher IMDB scores. We also want to show the results of this analysis in an intuitive way by visualizing the outcome. Here, we take into consideration the IMDB scores as response variables and focus on operating predictions by analyzing the rest of variables in the IMDB 5000 movie data. A global filter is being created to show a quick filter with appropriate style for every measure and dimension except the URL of the movie. Note that we have Movie Title configured as a wildcard match filter and Release Year as multiple values. We will create a simple dashboard with Filters and use dashboard actions and little bit of formatting to make it interactive.

The analyzed data will be displayed as a dashboard making the viewers have a knowledge of which movie they want to watch and also it will be useful for the persons in the film industry letting them know the Movie statistics along with their revenues and success rate of the movie.

4. Dataset

The dataset of the movies is available on IMDb and many other websites with average votes, vote numbers, reviews and descriptions. The dataset has been taken from the Kaggle website. The dataset contains 28 variables for 5043 movies, spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors/actresses. “imdb_score” is the response variable while the other 27 variables are possible predictors.[6]

The Attributes of the dataset are movie_title - Title of the Movie, duration - Duration in minutes, director_name - Name of the Director of the Movie, director_facebook_likes - Number of likes of the Director on his Facebook Page, actor_1_name - Primary actor starring in the movie, actor_1_facebook_likes - Number of likes of the Actor_1 on his/her Facebook Page, num_voted_users - Number of people who voted for the movie, cast_total_facebook_likes - Total number of facebook likes of the entire cast of the movie, movie_facebook_likes - Number of Facebook likes in the movie page, plot_keywords - Keywords describing the movie plot, facenumber_in_poster - Number of the actor who featured in the movie poster, color - Film colorization. ‘Black and White’ or ‘Color’, genres - Film categorization like ‘Animation’, ‘Comedy’, ‘Romance’, ‘Horror’, ‘Sci-Fi’, ‘Action’, ‘Family’, title_year - The year in which the movie is released (1916:2016), language - English, Arabic, Chinese, French, German, Danish, Italian, Japanese etc, country - Country where the movie is produced, content_rating - Content rating of the movie, aspect_ratio - Aspect ratio the movie was made in, movie_imdb_link - IMDB link of the movie, gross - Gross earnings of the movie in Dollars, budget - Budget of the movie in Dollars, imdb_score - IMDB Score of the movie on IMDB.

4.1. Data Preprocessing

The Data Preprocessing has been done using Python programming. The Preprocessing techniques which would be carried out are Removing the Duplicates, Data Cleaning, wher you remove the missing values and add and remove the needed and unwanted columns respectively. Since gross and budget have too many missing values, we want to keep these two variables for the following analysis. The data contains 1660 directors, and 3621 actors. Since the names are different for the whole dataset, we needn’t have to use the names to predict the score. Two variables have been added based on the existing variables for the purpose of data exploration, we added two variables based on existing variables. We plot the correlation heatmap for our data based on which we see higher and lower correlations between predictors. We want to reorder the columns to make the dataset easier to understand. And has renamed the columns to make the names shorter.

The Number of missing values in the dataset has been calculated and displayed in the figure 2 and further removed from the dataset. From which, 6% of the data has been lost which is acceptable.

By the correlation matrix in figure 1, it was found that the cast_total_facebook_likes and actor_1_facebook_like are highly correlated to each other. Both actor2 and actor3 are also somehow correlated to the total. So, it has been modified into two variables like actor_1_facebook_likes and other_actors_facebook_likes.

There are high correlations among num_voted_users, num_user_for_reviews and num_critic_for_reviews. The num_voted_users have been retained and the ratio of num_user_for_reviews and num_critic_for_reviews are taken. On this account, the above variables are dropped and found that none of the variables are much correlated to each other and can observe that all the values are below 0.7.

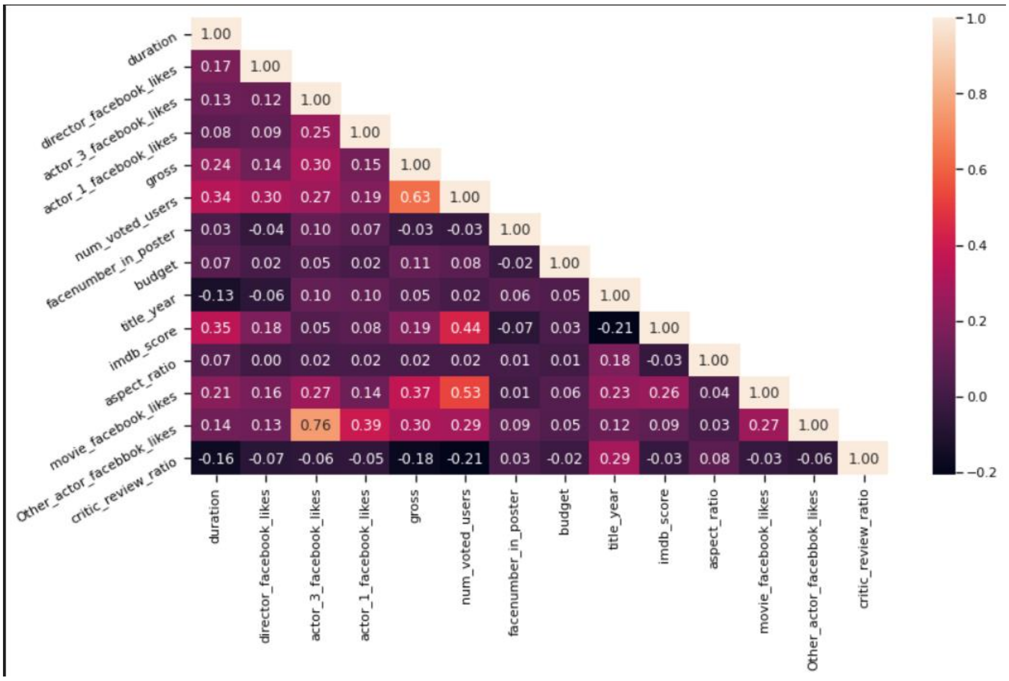
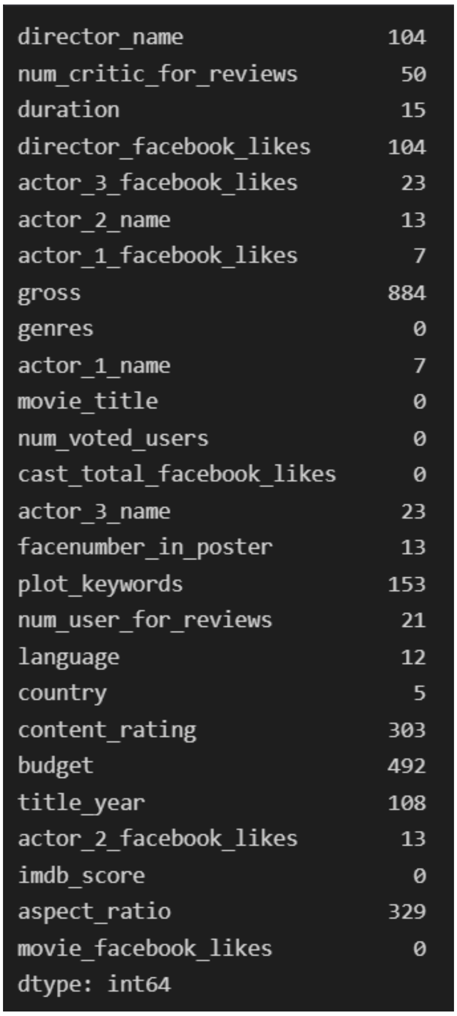


Fig 1: Correlation Matrix



```
director_name                104
num_critic_for_reviews        50
duration                      15
director_facebook_likes      104
actor_3_facebook_likes        23
actor_2_name                  13
actor_1_facebook_likes         7
gross                        884
genres                         0
actor_1_name                   7
movie_title                    0
num_voted_users                0
cast_total_facebook_likes      0
actor_3_name                  23
facenumber_in_poster          13
plot_keywords                153
num_user_for_reviews          21
language                      12
country                        5
content_rating               303
budget                       492
title_year                   108
actor_2_facebook_likes        13
imdb_score                     0
aspect_ratio                 329
movie_facebook_likes           0
dtype: int64
```

Fig.2: Number of Missing values

5. Technologies Used

5.1 Python

Python is used for pre-processing the collected data before importing it into the Logstash. Python has been used and through which data has been preprocessed by removing the missing values in the dataset and further plotted a correlation matrix for the movie dataset.

5.2. Logstash

It collects logs and events data and it is a server-side data processing pipeline that ingests data from multiple sources. Using Logstash, the data has been processed to the Elastic search. The data has been indexed in the index pattern as moviedata*. With this index, the data can be accessed further for creating the dashboard.

5.3. Elastic Search

It is a search and analytics engine that stores and indexes transformed data. The fetched data from the Logstash is collected as logs and is fed into the Elastic Clusters.

5.4. Visualizations

It is done in Kibana to visualize the data with charts and graphs in Elasticsearch and for creating dashboards. Kibana creates interactive dashboards and using which we could connect two different dashboards and analyze the results through the visualization.

6. Dashboard Features

Central Dashboard helps users to explore and shortlist movies using user-friendly interactive visualizations such as filtering for country of movie, duration of the movie, IMDB rating, genre, year of release etc. We can quickly see the shortlisted movies meeting the user preferences at the bottom of the dashboard. For every dimension and measure except Movie URL we are going to create an individual global filter and show them as a quick filter with appropriate style.

- Top Director based on the highest average IMDB ratings
- The genre which has been hyped in different countries.
- The best genre in each country
- The Actor who is being hyped in each country so that their movies will be viewed
- The best genre, director and actor in respective countries
- Histogram of Movie Released
- Top movies based on its Profit
- Commercial Success vs Critical Acclaim
- Relation between number of facebook likes and imdb_score

Figure 4 represents the Highest number of movies that have been released with respective years, from which it is found that the Year 2014 has the highest number of movies.

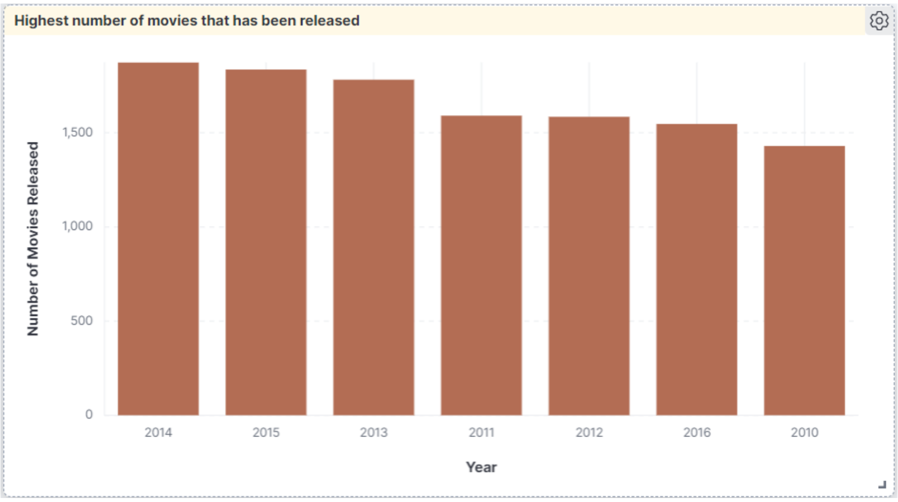


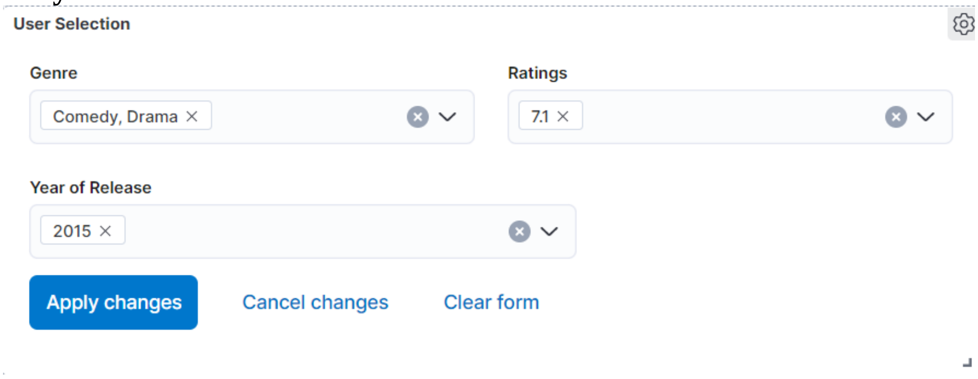
Fig 4. Highest number of movies released in respective years

7. Results and Discussions

The Dashboard created consists of 2 different dashboards, where 1 dashboard is for the users to view the ratings of the movies, the best genre and the year of release through which they could choose their favourite movie. And another Dashboard is for the people from the Film Industry, they could view the budget and the revenue generated for the movies with highest ratings and from which they could opt in making those kinds of movies in the future.

The Figure 5 represents the Slicer generated for the Genre Dashboard from which the users can select their desired Genre, Language, the ratings, the year of release of the movies so that they could view the movie they wish to watch.

7.1 Country-wise Dashboard



User Selection

Genre: Comedy, Drama

Ratings: 7.1

Year of Release: 2015

Fig 5. Slicer for the Genre Dashboard

Figure 6 represents the Genre all over the world. From the above figure, it is found that the Genre of Drama has the highest percentage and it is mostly in the USA.

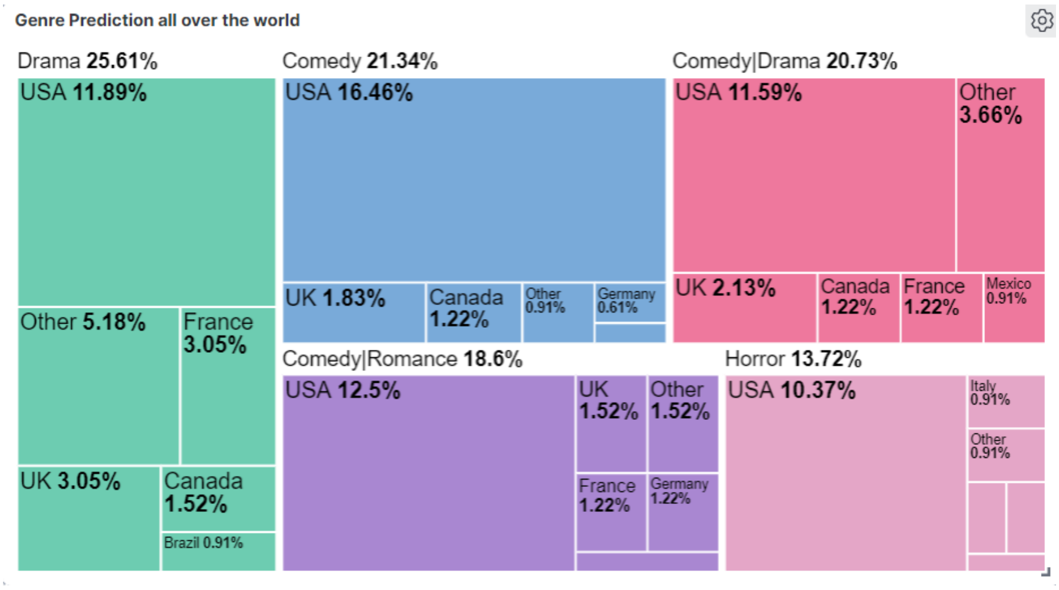


Fig 6. Genre all over the world

Figure 7 depicts the languages of the movies that have higher ratings, from which the users can get to know the ratings of the movies based on the language and opt for the movie in their preferred language as well as the higher rated language.

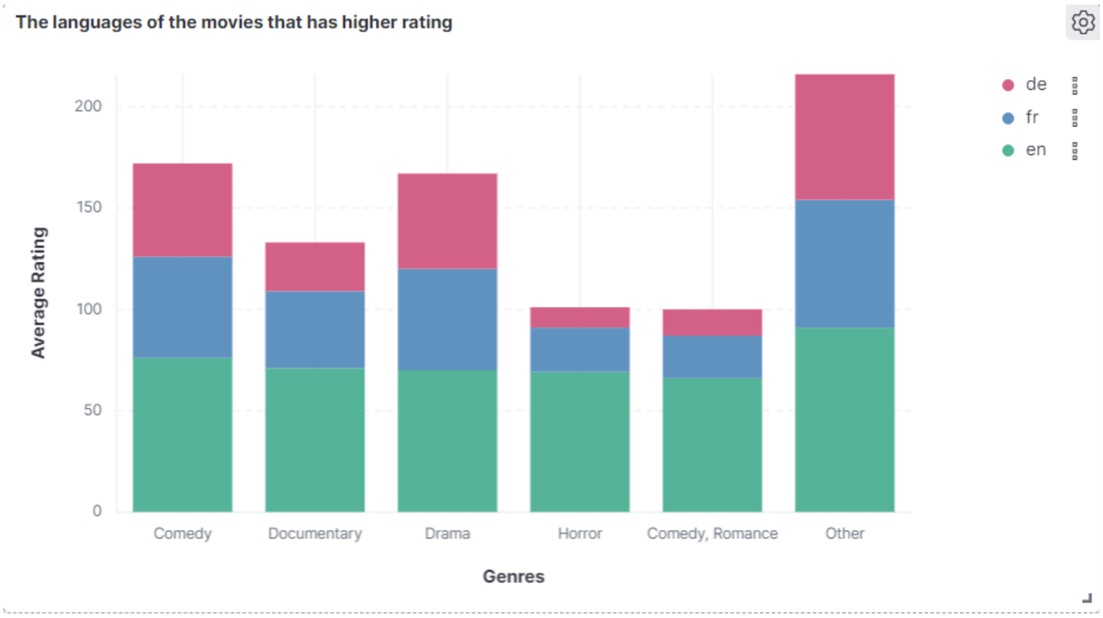


Fig 7. The Languages of the movies that has higher ratings

The Figure 8 represents the highest rated genres based on the IMDB Scores for the respective years. From which the users can check the IMDB scores and the genre and opt to watch a movie of their preference considering the year of release.

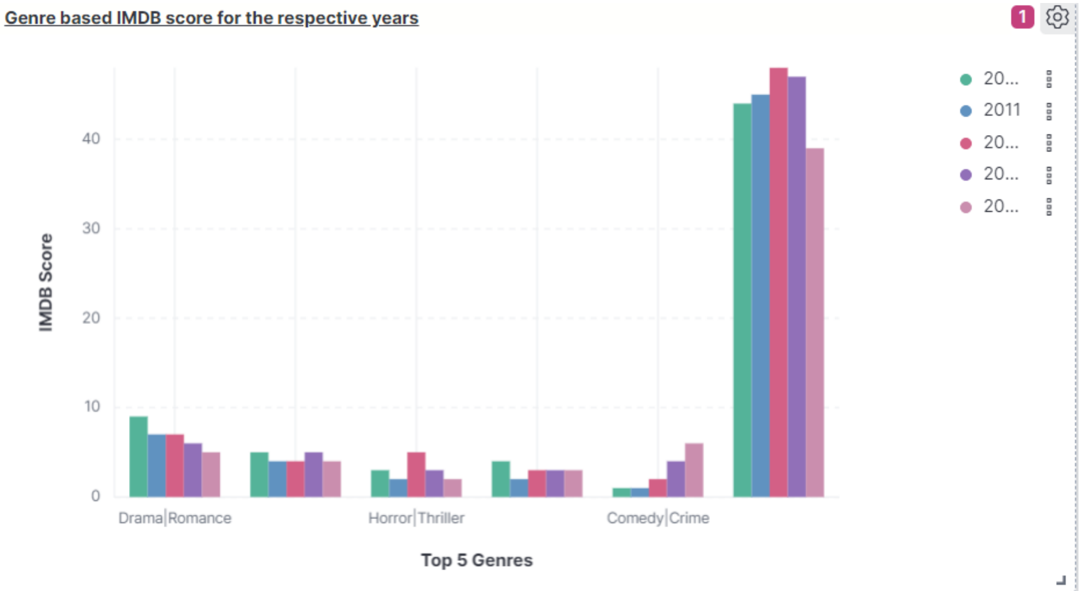


Fig 8. The Genres based on the IMDB Score for the respective years

7.2 Movie Dashboard

The Figure 9 represents the Ratings of the Content based on the year. From which it is found that the quality of the content was higher in the year 2013. This can be visualized by the users so that they choose the movie with good content.

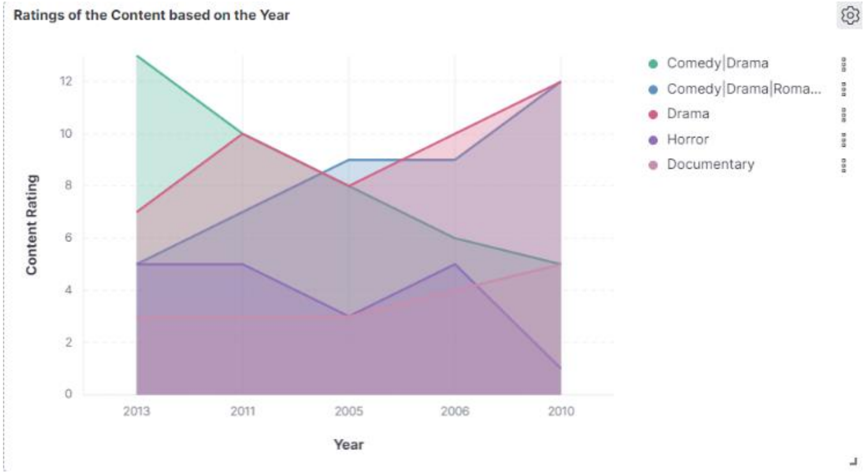


Fig 9. Ratings of the Content based on the year

The Figure 10 represents the Average Ratings in the respective years in different Countries. From which it is being analyzed that the year 2013 has more movies and it is higher in the USA.

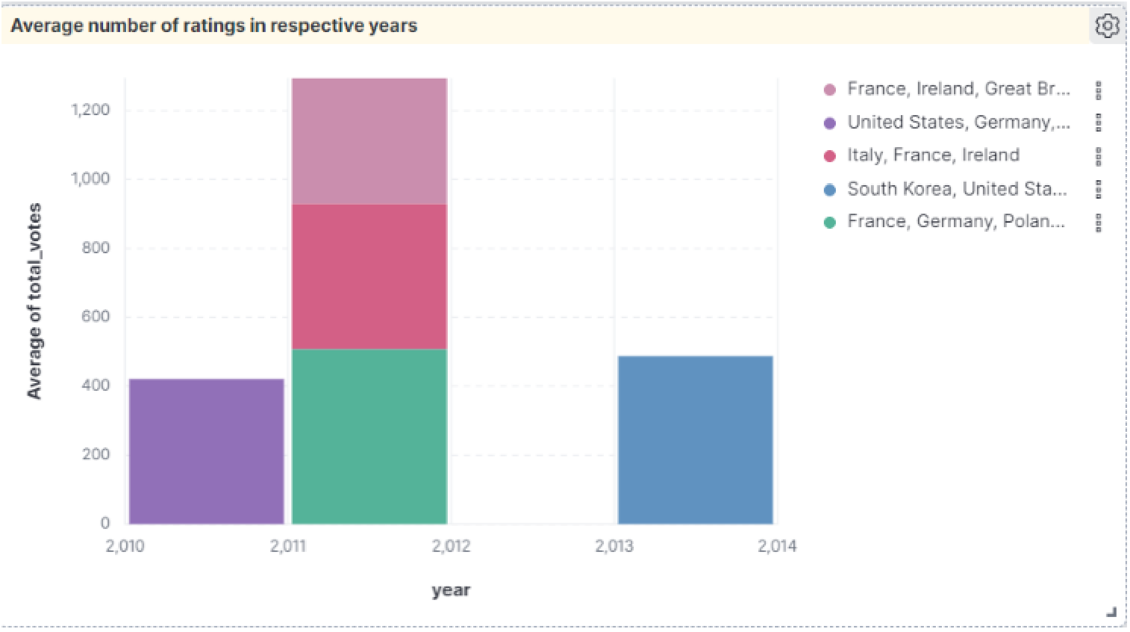


Fig 10. Average ratings in respective Years

Figure 11. represents the Cast Statistics in the respective years. Through which, the directors can analyze the best actor or the hyped actor and can proceed further in making up movies with the particular cast.

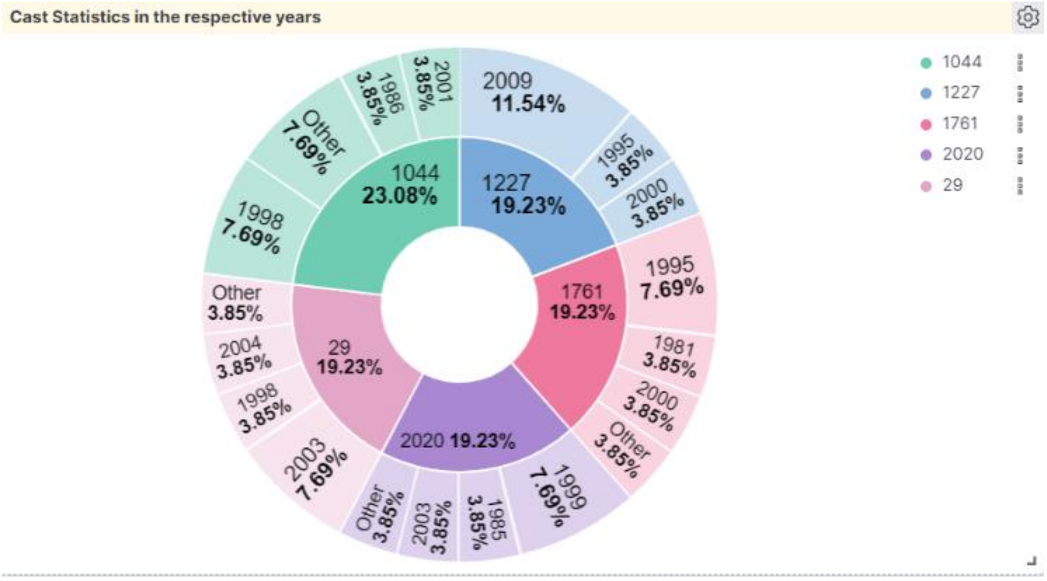
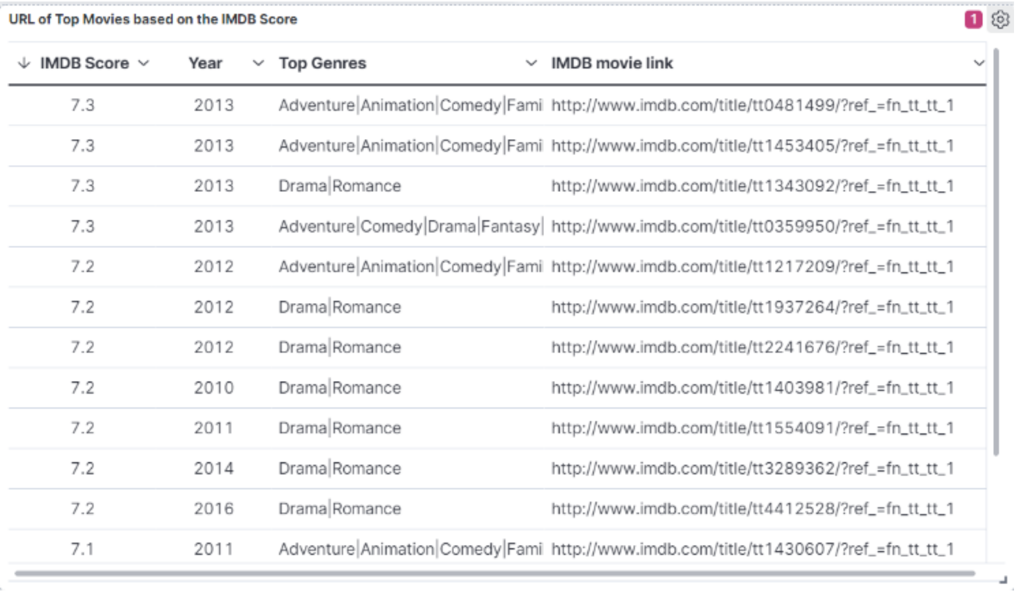


Fig 11. Cast Statistics in the respective years

Figure 12 represents the URL of the top-rated movies, through which the users can check for the highest ratings and select the URL of the Movie.



URL of Top Movies based on the IMDB Score

IMDB Score	Year	Top Genres	IMDB movie link
7.3	2013	Adventure|Animation|Comedy|Fami	http://www.imdb.com/title/tt0481499/?ref_=fn_tt_tt_1
7.3	2013	Adventure|Animation|Comedy|Fami	http://www.imdb.com/title/tt1453405/?ref_=fn_tt_tt_1
7.3	2013	Drama|Romance	http://www.imdb.com/title/tt1343092/?ref_=fn_tt_tt_1
7.3	2013	Adventure|Comedy|Drama|Fantasy|	http://www.imdb.com/title/tt0359950/?ref_=fn_tt_tt_1
7.2	2012	Adventure|Animation|Comedy|Fami	http://www.imdb.com/title/tt1217209/?ref_=fn_tt_tt_1
7.2	2012	Drama|Romance	http://www.imdb.com/title/tt1937264/?ref_=fn_tt_tt_1
7.2	2012	Drama|Romance	http://www.imdb.com/title/tt2241676/?ref_=fn_tt_tt_1
7.2	2010	Drama|Romance	http://www.imdb.com/title/tt1403981/?ref_=fn_tt_tt_1
7.2	2011	Drama|Romance	http://www.imdb.com/title/tt1554091/?ref_=fn_tt_tt_1
7.2	2014	Drama|Romance	http://www.imdb.com/title/tt3289362/?ref_=fn_tt_tt_1
7.2	2016	Drama|Romance	http://www.imdb.com/title/tt4412528/?ref_=fn_tt_tt_1
7.1	2011	Adventure|Animation|Comedy|Fami	http://www.imdb.com/title/tt1430607/?ref_=fn_tt_tt_1

Fig. 12 URL of the top-rated movies

7.3 Revenue Dashboard

Figure 13 represents the analysis of the best director. From which the producers can have an analysis of the director before producing the film and also approach the best director for their production.

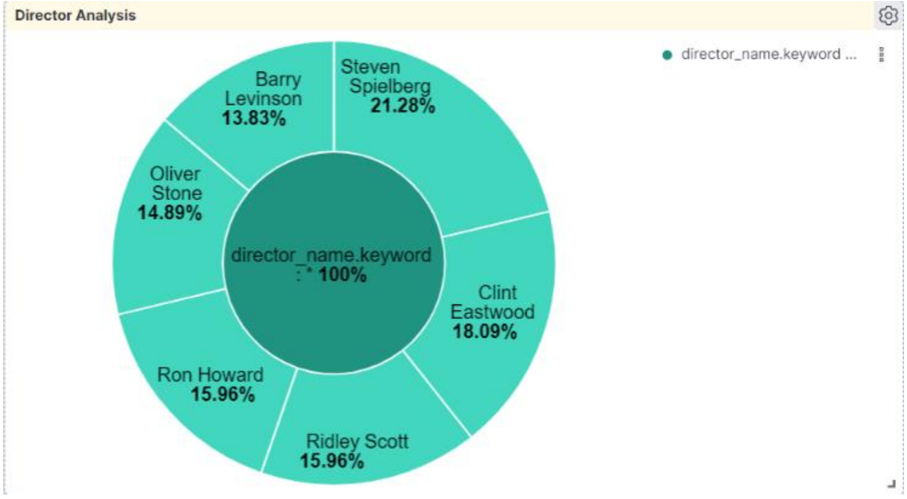


Fig 13. Analysis of the best director

Figure 14 represents the Percentage of Revenue generated by the Movies in the respective years. From which, depending on the revenue generated, the directors, producers etc, can have an analysis on the profit and loss and proceed accordingly.

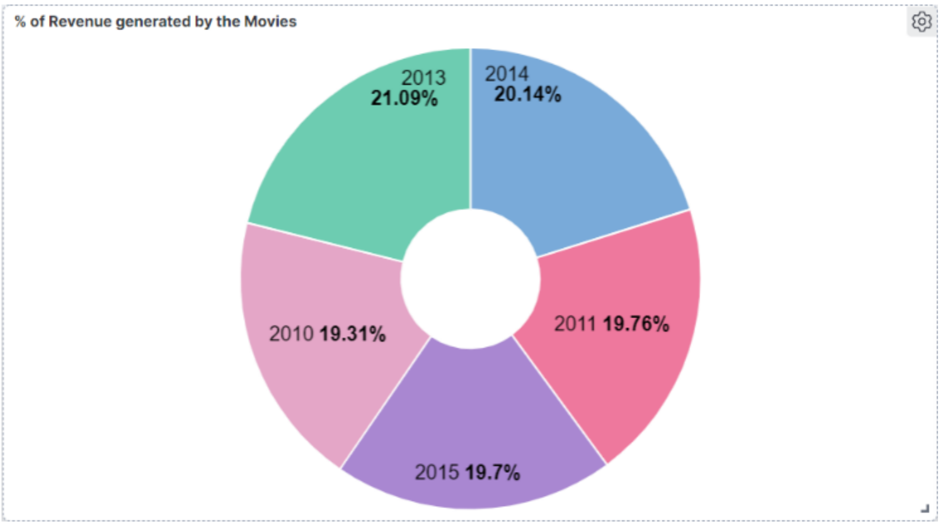


Fig 14. Percentage of Revenue generated by the Movies

Figure 15 represents the percentage of ratings on the highest revenue generated. A comparison of the ratings and Revenue generated could be taken into consideration by the Film Industry.

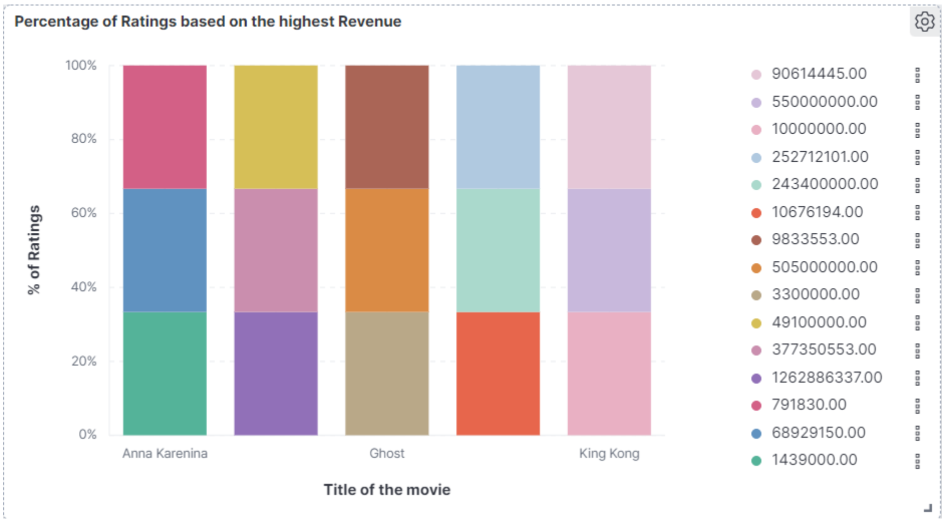


Fig 15. Percentage of Ratings based on the highest Revenue generated

Figure 16 represents the Highest Revenue generated Movies. From which it can be seen the features of the movie because of which the revenue has been generated and can adopt those for their future projects to gain highest revenue.

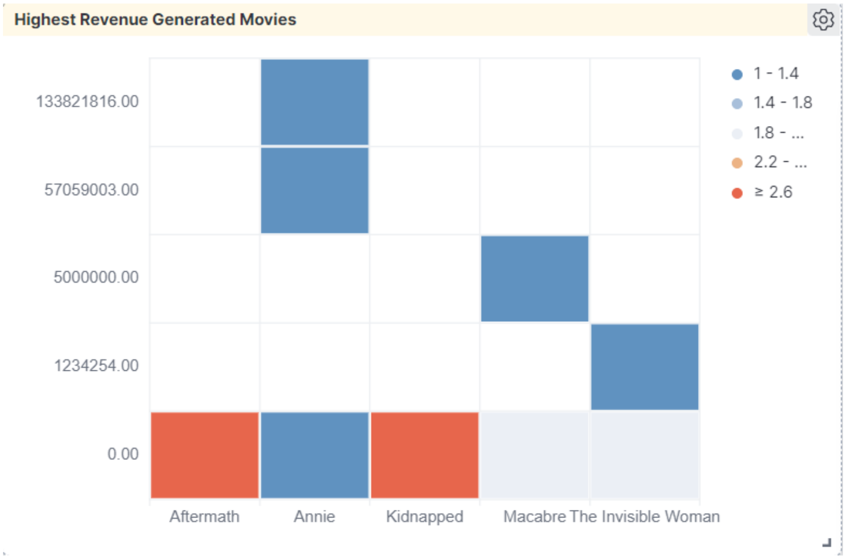


Fig 16. Highest Revenue generated Movies

8. Conclusion

A commercial success of the movie not only entertains the audience, but also enables film industries to gain tremendous profit. A lot of factors such as good directors, experienced actors are considerable factors for creating good movies. However, the famous directors and the actors always bring an expected income in the box-office but cannot guarantee a highly rated imdb score. The data has been analyzed for the top movies around the world from a timeline of 1986–2016, which is three decades. All the essential aspects which make a movie either go big or not. The IMDB rating, the Budget allocated, and finally, the revenue of the movie genre. Everything leads us to the fact that the Drama genre is growing big day by day. The technology will be very soon accepted as a viable performance method by the powers that will be present in the future. Visualizations like these can lead the users to draw some notable insights about the movie and further will help the director, producers etc from the film industry to have a glance on the statistics of the movie and further use it in the future to make movies of that kind.

References

[1] *Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten de Rijke, "Predicting IMDB Movie Ratings Using Social Media" , Advances in Information Retrieval, 503–507,* ***2012****.*

[2] *Bharadwaj Naidu Muthuluru, Anvesh Dudimetla, Varun Patwardhan and Vasanth Reddy, "Exploratory Visual System for IMDB Movies"*

[3] *Abba Suganda Girsang , Arief Handany,, Christopher Edmond, Marcellino, "Business Intelligence for Explore Customer in Internet Movie" , International Journal of Emerging Trends in Engineering Research, Volume 8. No. 5, May* ***2020****, ISSN 2347 - 3983.*

[4] *Nithin VR, Pranav M, Sarath Babu PB, Lijiya A, "Predicting Movie Success Based on IMDB Data", International Journal of Business Intelligent (IJBI), Vol. 3 Issue 2, December* ***2014****, ISSN: 2278-2397, Page(s): 34- 36.*

[5] *Klaus Dodds, "Popular geopolitics and audience dispositions: James Bond and the Internet Movie Database (IMDb)" , Transactions of the Institute of British Geographers, 31(2), ISSN 0020-2754, 116–130,* ***2006****.*

[6] https://www.kaggle.com/code/saurav9786/imdb-score-prediction-for-movies/data

[7] *Academy Award's best actor and actress winners and nominees from 2000-2004 downloaded from http://www.imdb.com/Sections/Awards. Accessed on May* ***2017****.*

[8] *O. Małgorzata, Business Intelligence as a Future, Analysis and Interpretation Of Data In Real Time, International Journal of Advanced Trends in Computer Science and Engineering, pp. 121–126, Jan.* ***2019****.*

[9] *D. W. S. Kusuma, Business Intelligence Infrastructure of Medical Record Data History System to help Doctorin differencing rare and dangerous disease in patients, International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 1, pp. 664–672, Feb.* ***2020****.*

[10] *M. Yulianto, A. S. Girsang, and R. Y. Rumagit, Business intelligence for social media interaction in the travel industry in Indonesia, Journal of Intelligence Studies in Business, vol. 8, no. 2, Sep.* ***2018****.*