International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

Predictive Modelling on IMDB's Movie Data

Dhruv Joshi
Compute Science and Engineering
SRM Institute of Science and Technology
Chennai, India

Pranjal Sharma
Compute Science and Engineering
SRM Institute of Science and Technology
Chennai, India

Dr. Jothikumar C
Assistant Professor, Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

1. Abstract:- **Movies, also known as films, are a type of videography or cinematography which uses moving pictures and sound to portrait a story about something or tell/teach people something. People in every part of the world watch movies as a type of entertainment, a way to enjoy their day and have fun with their loved ones. Movies can be of different genres, for example, some people like funny movies, so they can watch hilarious movies, whereas some people like suspense movies, so they watch thriller movies and so on. Movies are made to be shown or projected on big screens at movie theatres. Movies are kept in theatres for about two weeks or a month, after that time period, movies are removed from theatres and marketed on several media platforms.**

2. INTRODUCTION

Internet has become universal now days, especially in these recent times of covid-19 pandemic. People are staying at home due to the pandemic and making use of the internet more frequently than before. Nowadays, electronic devices are available at affordable prices. People are using multiple online platforms for their entertainment purpose, such as IMDB, Instagram, youtube and so on. Due to this, there's a heavy demand or popularity of online database which can provide with a database of movies, directors and actors. IMDb is a database, which is widely used all over the world and is the largest database which shows the data in relation to video games, movies, directors, actors, casts and TV programs. People can access to the database and take a look at the information about movies, TV programs and video games. People can also give their personal votes to their favorite movie or actor. People who registered themselves in IMDb have all the access to rate the TV program or the movie according to their likes or dislikes. They are given the option to rate them from a scale of 1 to 10. After all the votes given by the people all over the world, IMDb calculates and tallies all the votes and rates the movie according to that.
In this project we are pulling the data from the IMDB's database and predicting the best movie, director and actor by analyzing the votes given by the people and Facebook likes all over the world.

3. RELATED WORK

First, in these recent times of Covid-19 pandemic we all need a source of entertainment unit or a platform to let go of our boredom at home. IMDB is one of such database that can provide us with an enormous information of every movie or TV show.

Data Analytics is a process that is providing software tools to analyze, process and extract data from a very large data set with which normal working tools cannot deal easily. This has been made possible because of easy access to the large and packed data which can be accessed in a very secure manner.
In these recent times of Covid-19 pandemic, people are staying at home and looking for an entertainment unit to pass their time and have fun without getting bored at home. Looking into this matter, IMDB has gained lot of popularity and has been successful for providing the reviews to the users.
Our project takes the raw and tarnished data, cleans it up and makes sure to provide us with a clean and structured data for easy analysis. The project then examines the data set of IMDB and predicts the box office success of the movie. Our project also does the analysis on IMDB to predict the best director, actor, and movie.

To work on IMDb, we first pulled the raw data from the large and enormous dataset of IMDB, the data which we extracted was raw and unclean, so the next step was to clean the data and make it structured to be used easily for analysis.
After cleaning and making the data unstructured, we arranged the data in the desired and suitable format for the analysis to be performed easily.

After the data has been cleaned and making the data unstructured into the desired format, we used algorithms to perform linear regression analysis and correlation analysis on the files generated.

4. PROPOSED SYSTEM

4.1 Data Fetching, Cleaning the data and converting it into structured data

IMDb list files are converted to .list files, the format of these files are not preferred for data analysis purposes. Hence its is easy to put these files into Relational Databases. We used MySQL Community Database to drop those individual files into the Databases as almost all the data analytics tools have MySQL plugins available. Because of this we can use Relational Databases

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

as a source for the analytics to be performed on the files. While importing the data into Database of MySQL, there's a relation developed between the entities.
After importing the data into the database of MySQL, we cleaned the data of all the unwanted data. The data imported was raw and unstructured, so we cleaned the data and converted it into structured data to be analysed easily. After converting the data into the desired format of analytics, we performed mathematical expressions on the Data Analytics.

4.2 Data Analysis

After converting the data into proper format to perform data analysis, we performed mathematical expressions on the Data Analytics. We performed multiple linear regression and correlation techniques. Following is the mathematical expression for the same –

$$Gross\ Revenue_i = \beta_1\ Num\ Ref_i + \beta_2\ Num\ Follow_i + \beta_3\ Num\ Remake_i + \alpha_t + \gamma_j$$

In this regression equation - The dependent variable is movie gross revenue and independent variables are number of follows, remakes of movie and references.

$$Num\ Movie_{kt} = \beta_1\ Nomination_{k,t-1} + \beta_2\ Win_{k,t-1} + \beta_3\ X_{kt} + \alpha_t + \gamma_j$$

In this regression equation - Dependent variable here is number of movies directed and the independent variables here are number of movies that were awarded Oscars and number of movies that got nominated.

$$Total\ Grosss_{kt} = \beta_1\ Nomination_{k,t-1} + \beta_2\ Win_{k,t-1} + \beta_3\ X_{kt} + \alpha_t + \gamma_j$$

In this regression equation – The dependent variable is the total gross earned by the movie and the independent variables here are number of movies that were awarded Oscars and number of movies that got nominated.

This project proposes a system that uses Data Analysis with Python to predict the following
- Most Popular Director
- Most popular actor
- Best movie of the year
- Most financially successful Movie

5. SYSTEM DESCRIPTION

5.1 Python Libraries used
- **Numpy**: NumPy (NUM-pee)) is a library for the Python programming language, added to help out with enormous, multi-dimensional arrays and matrices. It also helps us with the large collection of mathematical functions that are used to work on these arrays.
- **Pandas**: Pandas is a flexible, fast, powerful and very easy to use data structures manipulation tool. It is used in open source data analysis.
- **MatPlot Lib**: matplot lib is a plotting library for the Python programming language. It is a comprehensive library for creating animated visualizations in Python and its numerical mathematics extension NumPy.
- **Seaborn**: Seaborn is a Library of Python which is based on MatPlot lib. It helps in making or drawing smooth and unique statistical graphics.

5.2 Hardware and Software requirements
Processor clock Rate - 1.5 GHz or Above Hard Disk Space -100 Mb free , RAM - 4 Gb or Above , Jupyter notebook Installed, Python installed and Inbuilt Integrated graphics support

.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

6. LITERATURE OF ONLINE PLATFORM

IMDb.com

Fig 1: logo of IMDb (taken from google images.com)

The Indian Movie Database (IMDb) was introduced to the world in the year 1990. It is the enormous online dataset or the information related to movies, TV serials,cast and the directors of the movie, reviews, box office collection, summary and the plot. As of December 2020, 83 million people have registered in IMDb. The Indian Movie Database has 10.4 million personalities in its database and approx. 7.3 million titles.

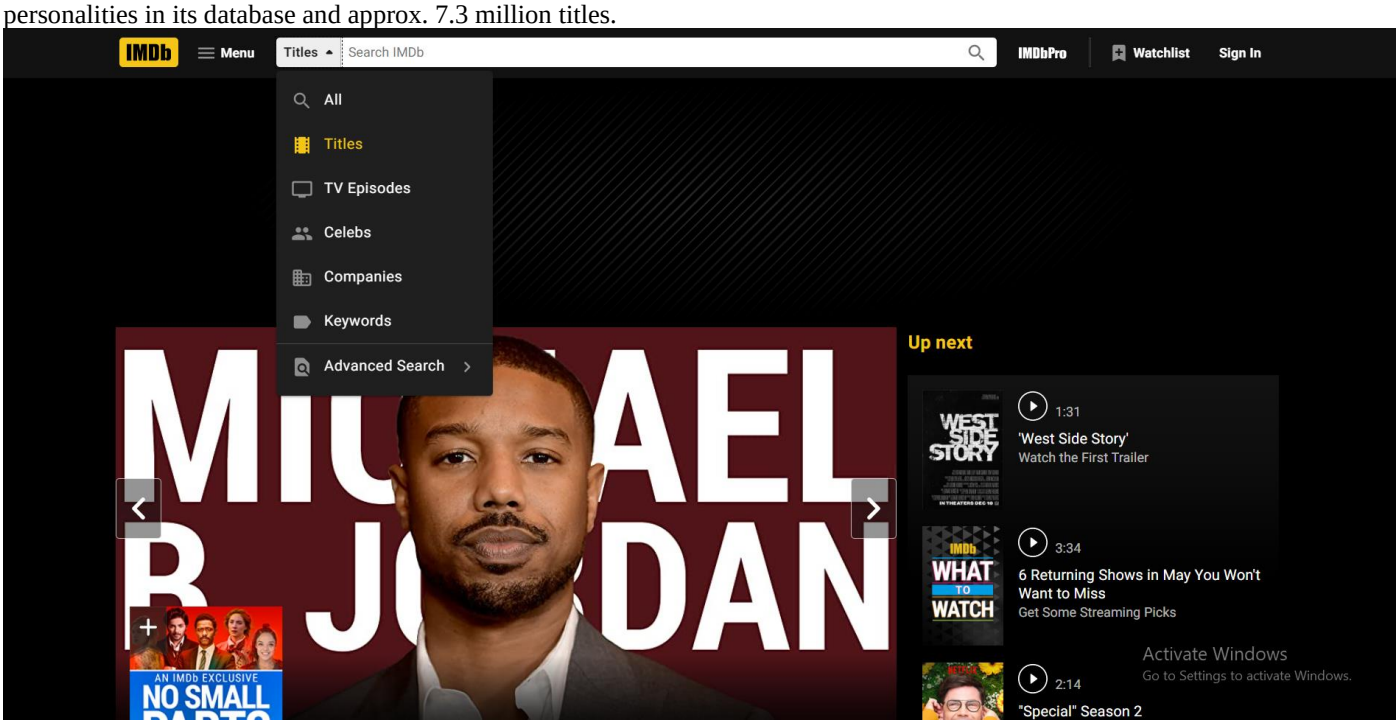

Fig 2: A screenshot of the online website (IMDb.com).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

7. DATA FLOW DIAGRAM

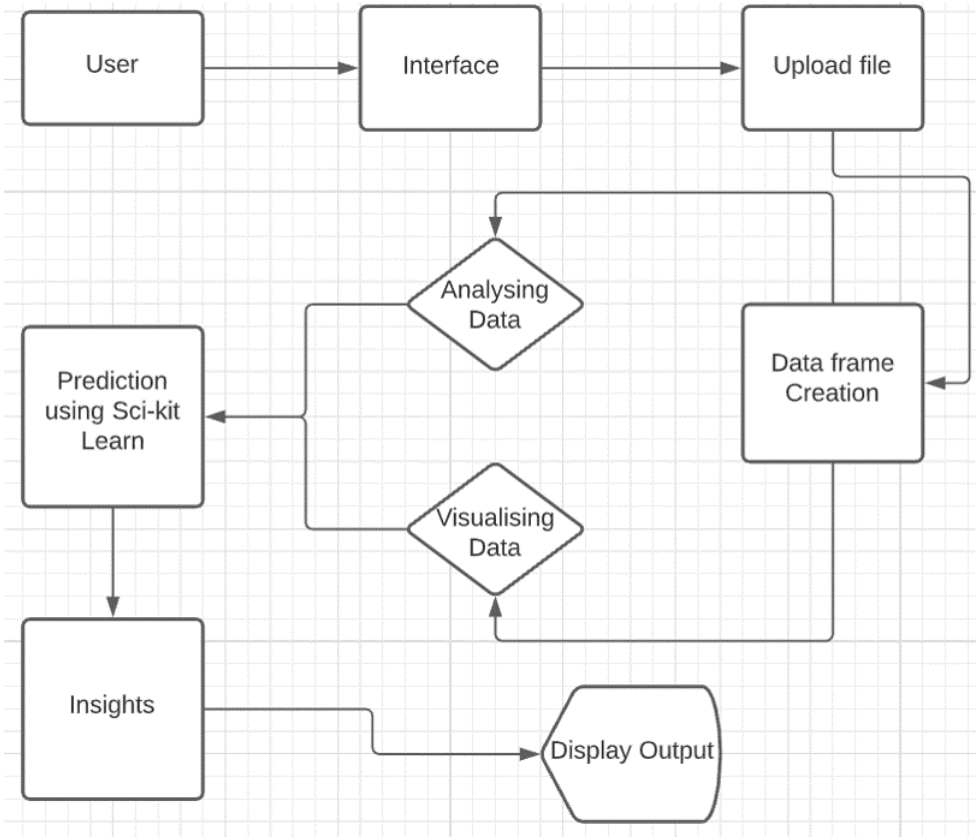

Fig 3: Data Flow Diagram

8. RESULTS –

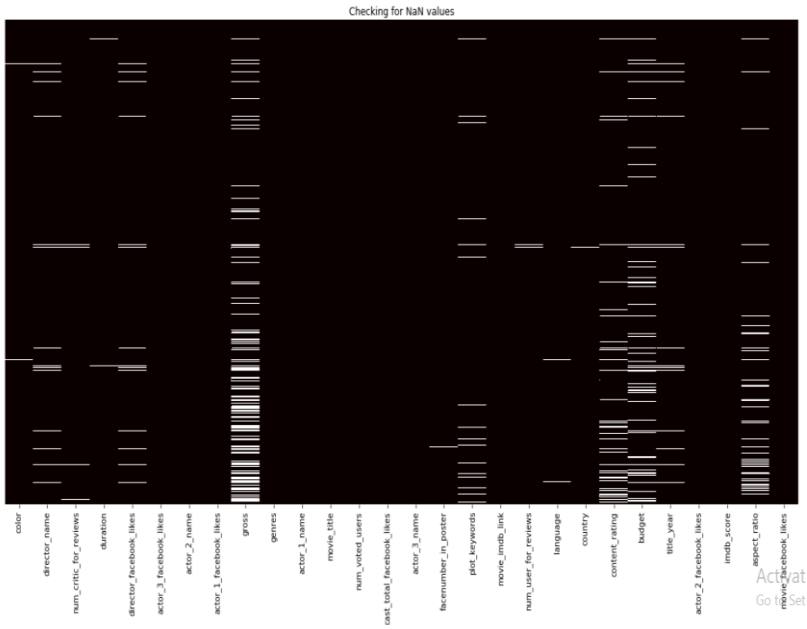

Fig 4: Checking for NaN values

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

In the above figure, there is raw and unstructured data. This data was straight away fetched from the IMDB dataset. In the next figure, you will see the data has been cleaned.

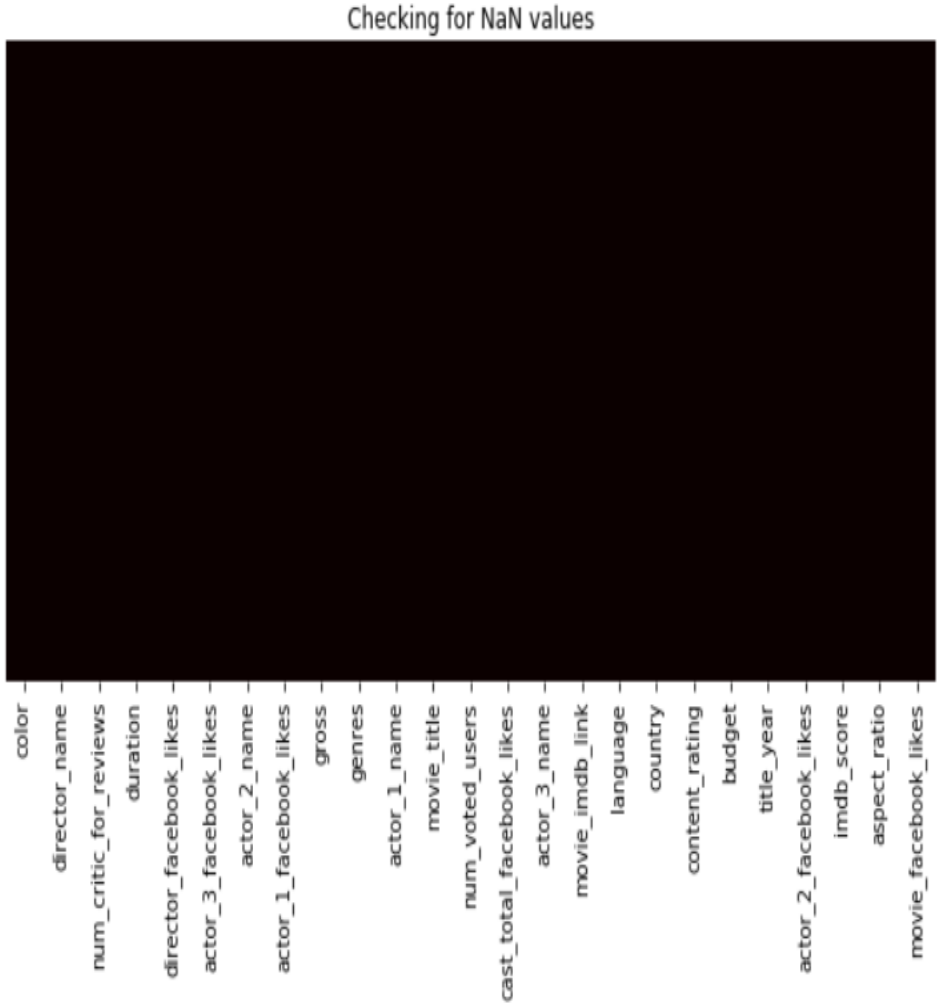

Fig 5: Now here, you can see the data has been cleaned with all the unwanted lines and converted into structured data.

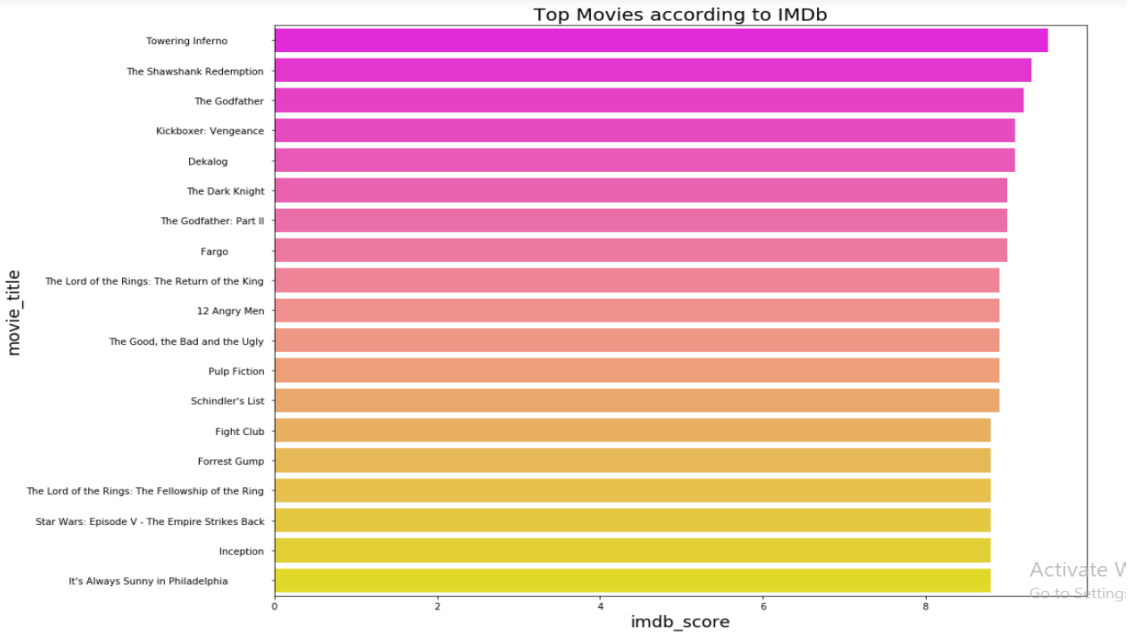

Fig 6. Top Movies according to IMDb

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

In this figure, we have printed out a graph which shows the Top movies according to IMDb.
Top movie according to IMDB is – Towering Inferno with an IMDB score of more than 8.5

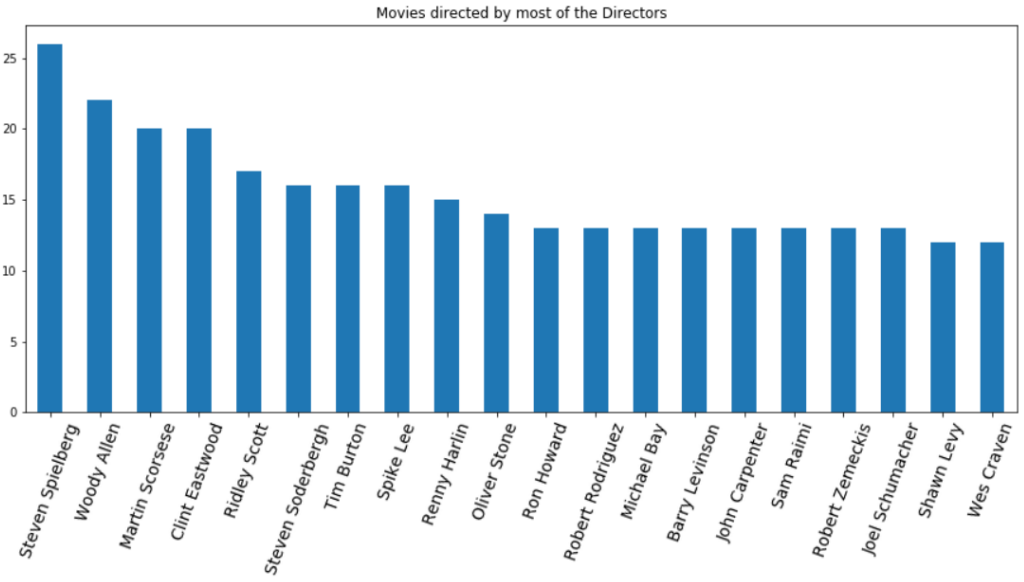

Fig 7: Movies Directed by most of the Directors

The figure above shows a histogram/graph of movies directed by most of the directors.
The most popular director here is “Steven Spielberg” with more than 25 directions.

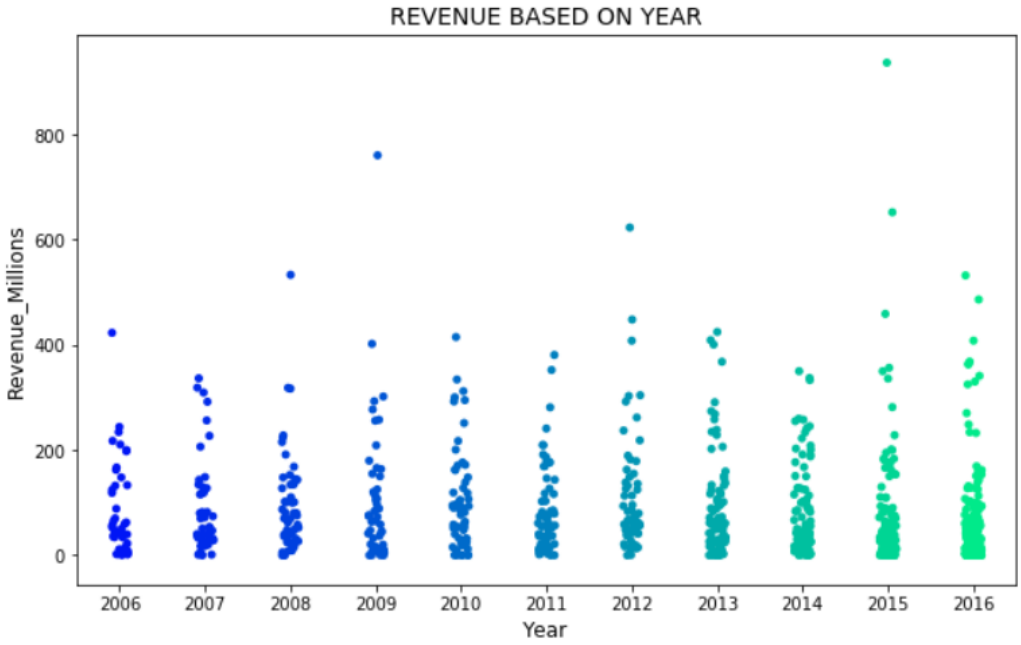

Fig 8: Revenue Based on year

The above graph shows the Revenue produced by the movies per year.
We have made the graph for the movies from the year 2006 to the year 2016.
We can clearly see that in the year 2015, movies made the most revenue of more than 800 million.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

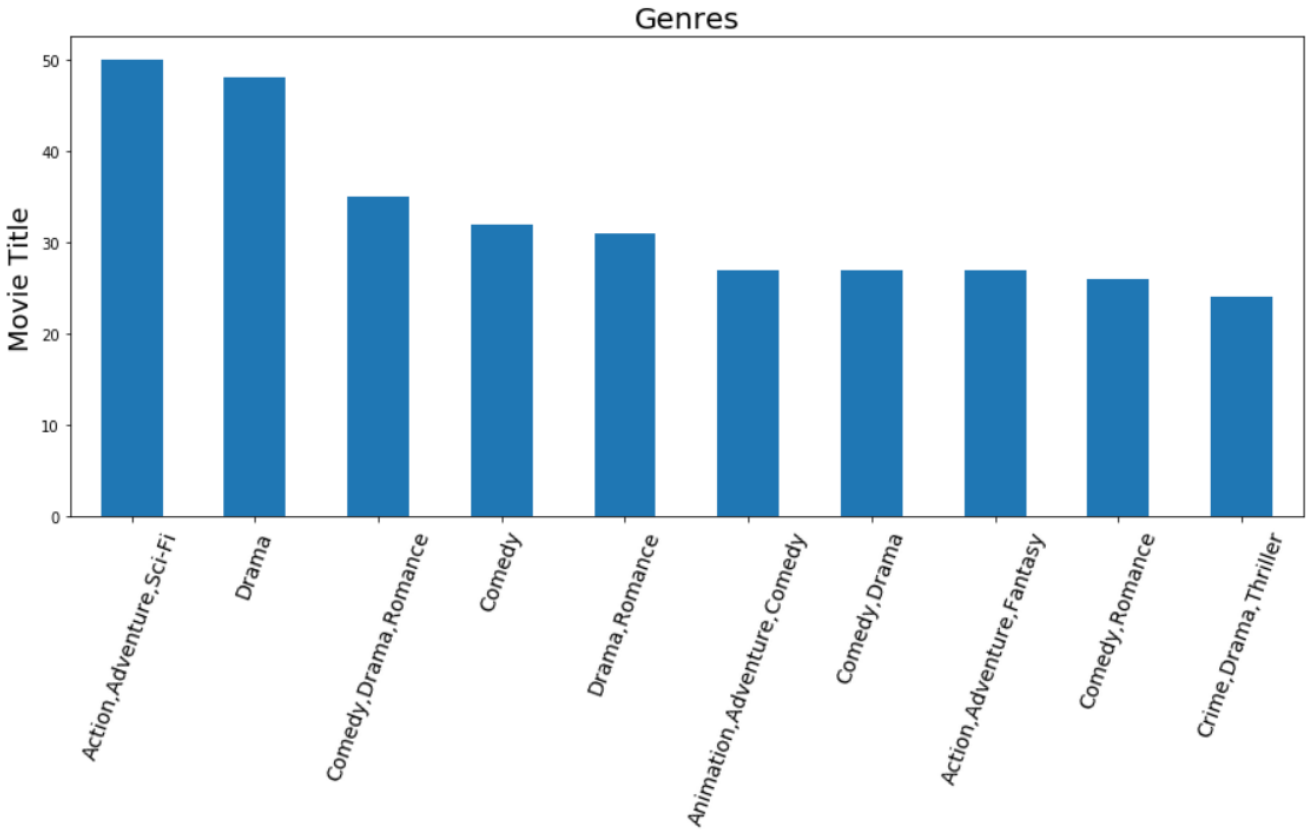

Fig 9: Genres of the movie

Fig 9 shows the different genres of different movies.
Genres are basically a particular category divided into one's choice, for example – Drama, Rom-Com, Crime, etc.
Here, most movies are form the genres, Action, Adventure, Science fiction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

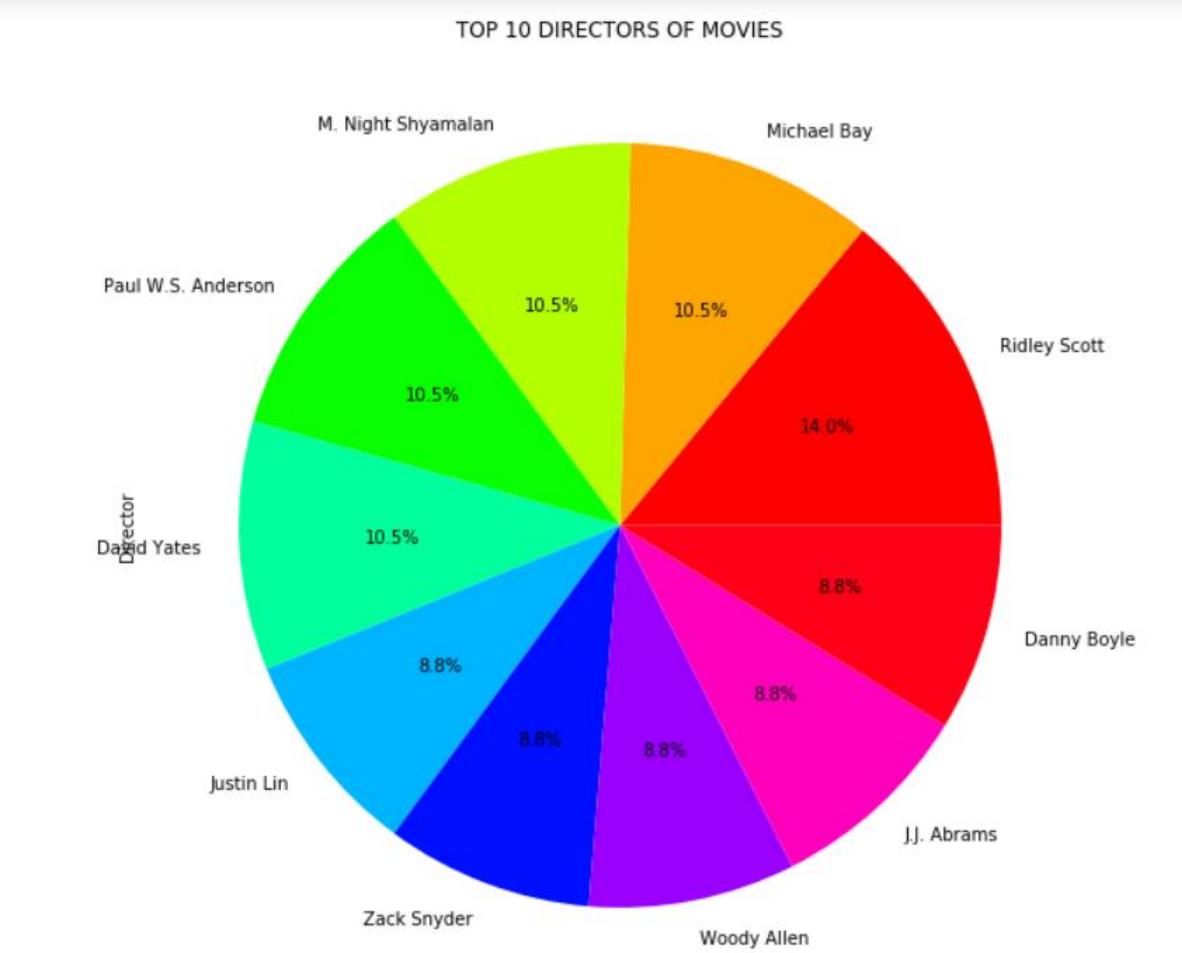

Fig 10: Top 10 Directors of the movies

Ridley Scott is the highest voted Director according to IMDb with 14% votes.

CONCLUSION

Internet have played a very important role in the modern world development. Internet has given people a source of gaining knowledge and entertainment.

Seeing this, we created a project that can give us the clear data of IMDB movie dataset for clear knowledge of who got the maximum votes and which director has topped the charts.

Researchers have always found big data analytics fascinating and have shown special interest to it, because big data analytics always helps in discovering many hidden and anonymous patterns. Big data analytics has always helped in unknown correlations that exists in the large and enormous data. We created a project that can give us the clear data of IMDB movie dataset for clear knowledge of who got the maximum votes and which director has topped the charts.

REFERENCES

[1] Kanika, Almadi, “Quantitative study of the movie industry based on IMDb data”, Published by: Massachusetts Institute of Technology (2017)
[2] Jehoshua Eliashberg, Anita Elberse and Mark A. A. M. Leenders, "The Motion PictureIndustry: Critical Issues in Practice, Current Research, and NewResearch Directions",Published by: INFORMS in Marketing Science, Vol. 25, No. 6, 25th Anniversary Issue (Nov.-Dec., 2006)
[3] Nithin Vr, “Predicting Movie Success Based on IMDB Data”, Published By: National Institute of Technology, Calicut (2014)
[4] Max Wassermana, Xiao Han T. Zengb, and Luis A. Nunes Amaralb," Cross-evaluation ofmetrics to estimate the significance of creative works"
[5] Wenjing Duan, Bin Gu, Andrew B. Whinston, "Do online reviews matter? -An empiricalinvestigation of panel data", Retrieved from journal homepage - www.elsevier.com/ locate/dss
[6] Gerda Gemser,Mark A. A. M. Leenders,Nachoem M. Wijnberg, "Why Some Awards AreMore Effective Signals of Quality Than Others:A Study of Movie Awards"
[7] Shane Greenstein, Feng Zhu, "Do Experts or Collective Intelligence Write with More Bias?Evidence from Encyclopdia Britannica and Wikipedia"
[8] David A. Reinstein and Christopher M. Snyder,"The Influence of Expert Reviews onConsumer Demand for Experience Goods: A Case Studyof Movie Critics", Published in TheJournal of Industrial Economics, Vol. 53, No. 1 (Mar., 2005),
[9] Film industry, Retrieved from https://en.wikipedia.org/wiki/Film-industry
[10] Theatrical Market Statistics 2016, Retrieved from http://www.mpaa.org/wp-content/uploads/2017/03/2016-Theatrical-Market-Statistics-Report-2.pdf
[11] www.IMDb.com.