IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

EVALUATION OF DIFFERENT OTT PLATFORMS WITH DATA ANALYTICS TECHNIQUES FOR RECOMMENDING PERSONALIZED CONTENT TO THE USERS.

Dr. Vishwanath Karad MIT World Peace University

Heet Shah, Vaishnavi Jadhav, Tanmay Gharte, Soham Wattamwar, Varsha Naik

Abstract: In this new era of online platforms, there have been many movie content platforms people use to make their busy life relaxed by watching movies online according to their convenience. But because there exists large datasets of movies it becomes very difficult for the user to search and select a movie. In this paper, we discuss movie recommender systems and its techniques and try to analyze the in-depth problem of why some movies get more fame and what all the factors responsible for a movie or a web-series to gainer recognition by performing Exploratory data analysis and visualizing the pre-processed using Tableau tool which is widely used in this domain. Visualizations depict significant relations and the sentiment of the users behind every movie as well as the parameters being considered and evaluated will eventually be a prominent key for any individual who is involved in this field. Factors mainly the geographic location, genre of the film, recognition of actor and director, the past history of the films directed by that director and many more are to be paid much attention while working in this and which we have very well demonstrated through our research.

INTRODUCTION

In the modern world there is a need for a recommendation system as users have been exposed to a large number of multimedia contents with the rise of OTT platforms such as Netflix, Disney Plus and many more. Many recommendation systems are available along with a large number of algorithms that are used for data forecasting for specific results that the user needs. Collaborative approach is divided into 2 methods i.e Model-based and Memory-based. Neural network generation happens in model based methods which understands the facts and shows the required results/recommendation. The well-known recommendation strategy of commercial OTT services is to use user’s rating data, which is then fed for collaborative filtering.

Movie recommendation systems are very popular. They will gain more popularity in the coming years as the filmmaking industry is getting vast and there will be numerous options for the user to choose from the huge collection1][2]. So in this vast collection, users only wish to watch a certain language, theme, genres or some different constraints. In this situation the recommendation system works. It uses the concept of machine learning and works on it by learning the datasets. After that the recommendations can be based on search history, language, same genre, etc. This technique is used to meet the customer needs and getting benefitted by delivering the contents according to the user.

We have studied and analyzed 3 major OTT platforms i.e. Amazon Prime, Netflix and Disney Plus. We studied IMDB movie dataset along with the datasets of each platform to analyze the distribution of various constraints like movie genre, age limits, etc to get the tags representing each platform.

To investigate different OTT platform datasets, analyze the utilization of platforms to provide users with accurate movie recommendations. To personalize a platform for the user the system will search content that the user is interested in and will accordingly show the results. Now the simplest way possible is by recommending the top trending content to the user but to make it more personalized for the user, dedicated recommendation systems are needed.

Literature Survey

Recommender systems is a convenient technology that can solve the problem of vast amounts of information provided to users by providing personalized, specialized building recommendations and services. Collaborative filtering process is widely used which recommends items by identifying other users with similar interests. We have evaluated data mining techniques for this system. We have studied various research papers and the overall flow of research is analyzed. We have used different types of datasets collected from kaggle. In the movies dataset we have downloaded credits.csv and movies_metadata.csv files separately as they were large in size. In the IMDB dataset we have all hollywood movies till 2018 along with all OTT platform movies.

Hyeyoung Ko and team [3] makes the use of Hybrid Recommendation model to solve the sparsity that means a set of numbers is considered to be sparse when a high percentage of the values are assigned to the constant default value. Their algorithm was used to satisfy the absence of the rating data by integrating the information. For this algorithm Content-Based Filtering and Collaborative Filtering models were used.

Abbasi-Moud et al. [4] came up with the tourism recommendation system in which the data from user’s reviews will be considered as data on tourist destinations. The system specifies the circumstances of the user's data review on tourist destinations. The system comes with the data that includes time, location, and weather data. The tourist destination system uses Text Mining, performing text analysis and clustering to review the data and to find out user preferences.

Sneh Srivastava and team [2] says that the Multi criteria recommendation system integrates fondness information on the top of multiple criteria. This algorithm describes the complete fondness of the user for the unit, the system attempts to predict a rating for the trackless units by imposing fondness information operating multiple criteria that may cause a change of this overall preference value. Authors came up with another commendable approach as Location recommendation Mobile system which uses internet-accessing mobile devices to provide custom or factor recommendations. The system make use of GPS data of user’s device for routing taxi drivers to take fare, that take in the location, current time and working status for the passengers.

RESEARCH METHODOLOGY

I. Dataset Description -

We have considered our few dataset from Kaggle and few of them we have scrapped from the Internet. In our project there are 7 datasets which are listed below -

Amazon - In our Amazon dataset we have considered 8 columns to provide our user all shows n related to that information like it consists Name of show, Year of release, Number of seasons, Language, Genre, IMDB rating, Age of viewers and the average vote. This platform has a subscription for streaming. User is able to select a show according to his choice and he will be able to see all the seasons and episodes for that show which will be extracted from our dataset.

Data - In this dataset we have considered 4 columns which consist of director_name, actor_name, genres and the movie_title. Basically this dataset consists of cast and crew names of all the shows.

credits - In this database we have considered 3 columns which consist of cast, crew, ID.

final_data - In this database we have considered 4 columns which consist of director_name, actor_name, genres, movie_title and comb(combination). In this dataset all the data is combined after preprocessing and data extract.

movie_metadata - In this dataset we have considered all information about the show like total actors, reviews, language, country, budget, year, IMDB ratings and aspect_ratio etc.

movies_metadata - In this dataset we have considered all information about show like gender, age(for adult), production countries, revenue, popularity, production countries, status, and average vote for that show

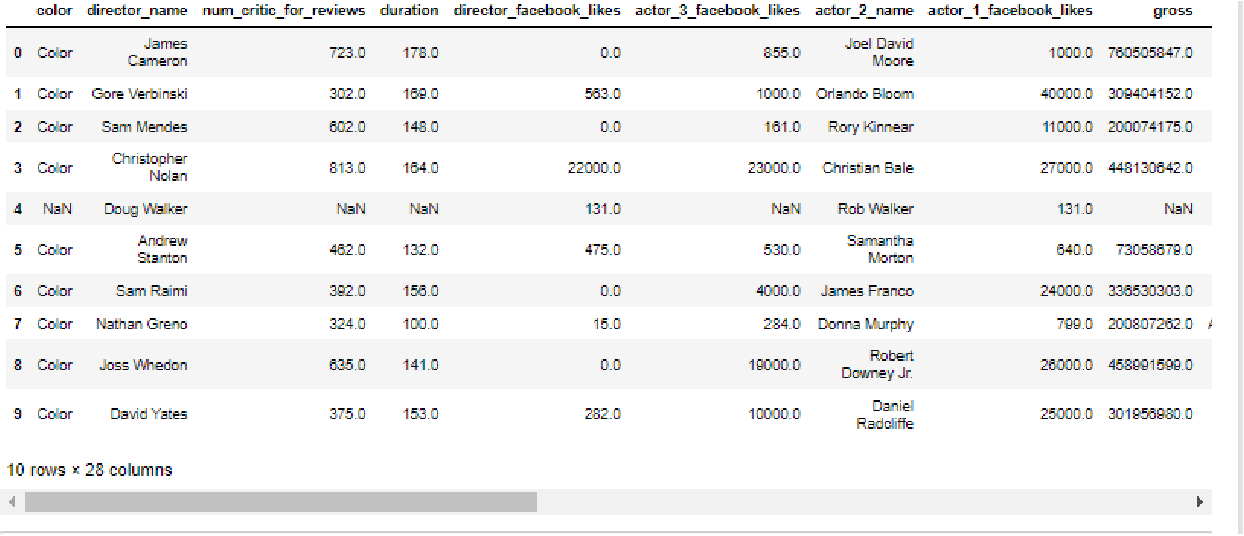
Netflix - In this dataset we have considered 12 columns which are related to the show like show_id, type, duration, listed_in(Reality shows), and description about the show.

II. Data preprocessing and exploratory analysis:

Data preprocessing is the technique where the data is dropped before it is used to enhance the performance. We have preprocessed different datasets of movies which we have scrapped from amazon, movie metadata etc. Data cleaning is the technique that is used for deleting or altering inaccurate data for review. In data analysis this data is not necessary as it can slow the process or show incorrect results. Data cleaning is also used to maximize the detailing of the data collection without deleting anything.

Exploratory data analysis evaluates the datasets to summarize its characteristics. It involves the use of statistical graphics along with different data visualization techniques. In this project we have used various datasets like movie_metadata.csv, amazon.csv, Netflix.csv which we have downloaded from Kaggle.

Images of Pre-Processing and cleaning the data



	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes	gross
0	Color	James Cameron	723.0	178.0	0.0	855.0	Joel David Moore	1000.0	760505847.0
1	Color	Gore Verbinski	302.0	169.0	563.0	1000.0	Orlando Bloom	40000.0	309404152.0
2	Color	Sam Mendes	602.0	148.0	0.0	161.0	Rory Kinnear	11000.0	200074175.0
3	Color	Christopher Nolan	813.0	164.0	22000.0	23000.0	Christian Bale	27000.0	448130642.0
4	NaN	Doug Walker	NaN	NaN	131.0	NaN	Rob Walker	131.0	NaN
5	Color	Andrew Stanton	462.0	132.0	475.0	530.0	Samantha Morton	640.0	73058679.0
6	Color	Sam Raimi	392.0	156.0	0.0	4000.0	James Franco	24000.0	336530303.0
7	Color	Nathan Greno	324.0	100.0	15.0	284.0	Donna Murphy	799.0	200807262.0
8	Color	Joss Whedon	635.0	141.0	0.0	19000.0	Robert Downey Jr.	26000.0	458991599.0
9	Color	David Yates	375.0	153.0	282.0	10000.0	Daniel Radcliffe	25000.0	301956980.0

10 rows × 28 columns

Fig.1 Overview of the dataset

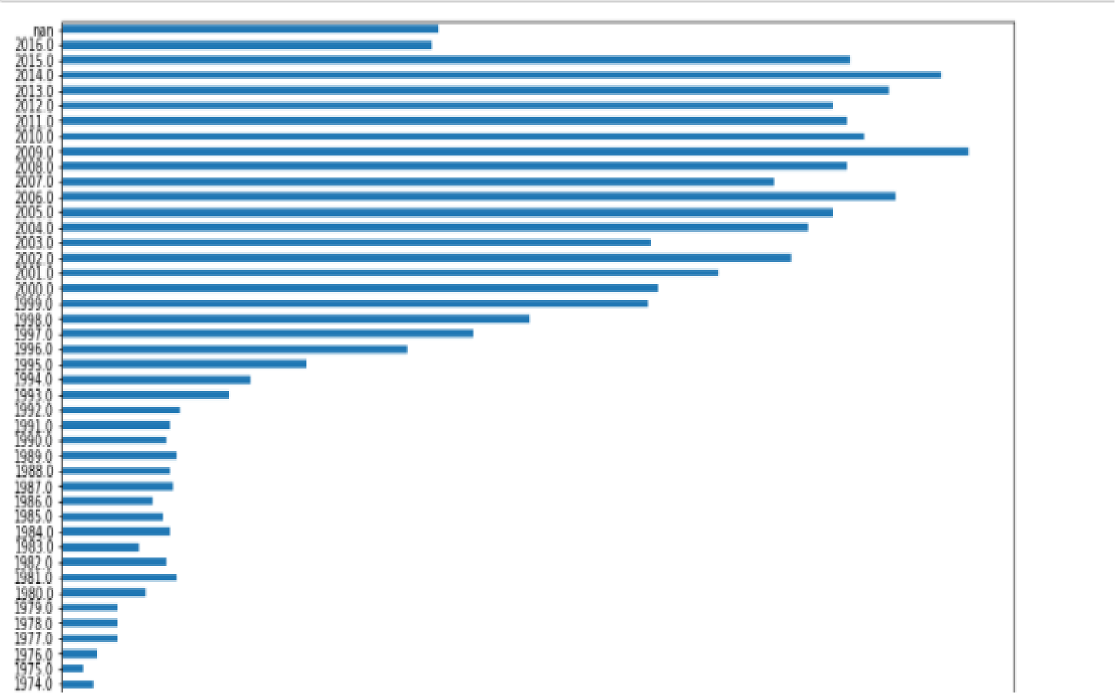
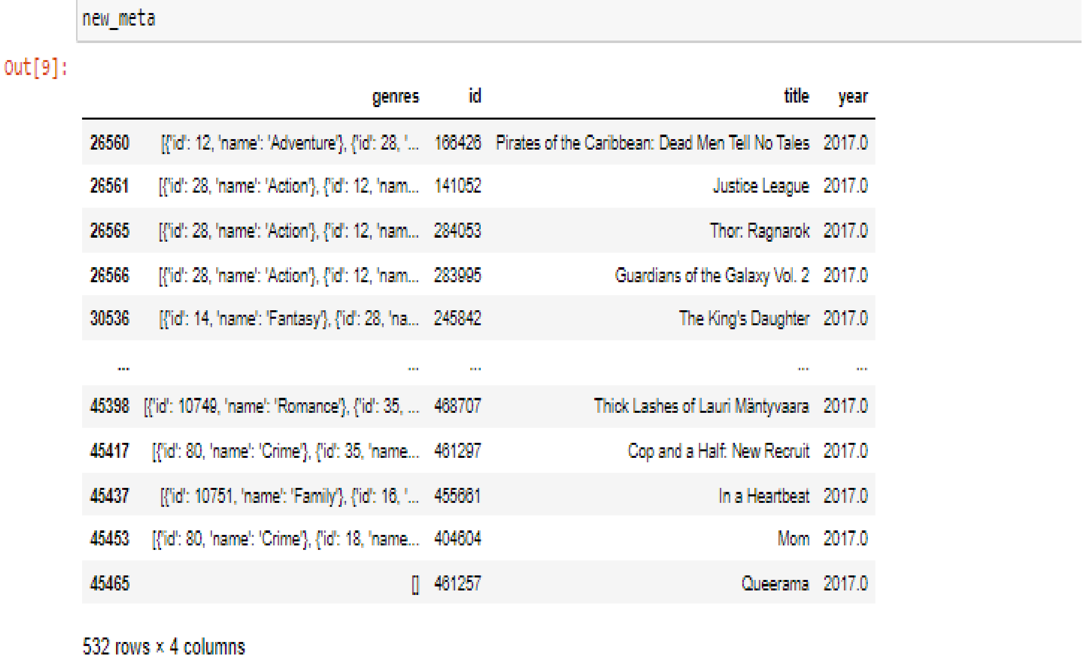


Fig.2 Graph plot for movies for 2016

We have scrapped data for the year 2017 movies**.** We have extracted it from Wikipedia using pandas. After scrapping we got the raw data in text format and then converted it to csv format, we have also used TMDB website to extract the data and eventually merged and pre-processed this raw data.

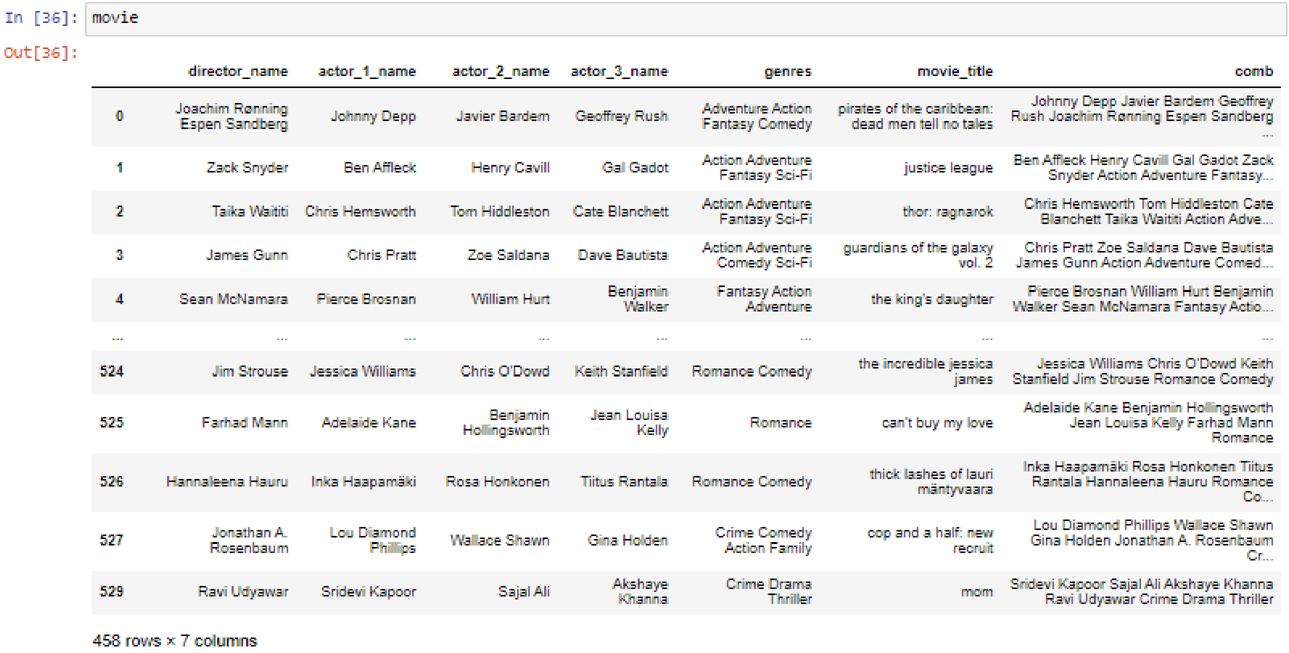


```
new_meta
Out[9]:
```

	genres	id	title	year
26560	[{'id': 12, 'name': 'Adventure'}, {'id': 28, '...	166426	Pirates of the Caribbean: Dead Men Tell No Tales	2017.0
26561	[{'id': 28, 'name': 'Action'}, {'id': 12, 'nam...	141052	Justice League	2017.0
26565	[{'id': 28, 'name': 'Action'}, {'id': 12, 'nam...	284053	Thor: Ragnarok	2017.0
26566	[{'id': 28, 'name': 'Action'}, {'id': 12, 'nam...	283995	Guardians of the Galaxy Vol. 2	2017.0
30536	[{'id': 14, 'name': 'Fantasy'}, {'id': 28, 'na...	245842	The King's Daughter	2017.0
...	...	...	...	...
45398	[{'id': 10749, 'name': 'Romance'}, {'id': 35, ...	468707	Thick Lashes of Lauri Mäntyvaara	2017.0
45417	[{'id': 80, 'name': 'Crime'}, {'id': 35, 'name...	461297	Cop and a Half: New Recruit	2017.0
45437	[{'id': 10751, 'name': 'Family'}, {'id': 16, '...	455661	In a Heartbeat	2017.0
45453	[{'id': 80, 'name': 'Crime'}, {'id': 18, 'name...	404604	Mom	2017.0
45465	[]	461257	Queerama	2017.0

532 rows × 4 columns

Fig.3 Extracting information for 2017 movies



```
In [36]: movie
Out[36]:
```

	director_name	actor_1_name	actor_2_name	actor_3_name	genres	movie_title	comb
0	Joachim Rønning Espen Sandberg	Johnny Depp	Javier Bardem	Geoffrey Rush	Adventure Action Fantasy Comedy	pirates of the caribbean: dead men tell no tales	Johnny Depp Javier Bardem Geoffrey Rush Joachim Rønning Espen Sandberg ...
1	Zack Snyder	Ben Affleck	Henry Cavill	Gal Gadot	Action Adventure Fantasy Sci-Fi	justice league	Ben Affleck Henry Cavill Gal Gadot Zack Snyder Action Adventure Fantasy...
2	Taika Waititi	Chris Hemsworth	Tom Hiddleston	Cate Blanchett	Action Adventure Fantasy Sci-Fi	thor: ragnarok	Chris Hemsworth Tom Hiddleston Cate Blanchett Taika Waititi Action Adve...
3	James Gunn	Chris Pratt	Zoe Saldana	Dave Bautista	Action Adventure Comedy Sci-Fi	guardians of the galaxy vol. 2	Chris Pratt Zoe Saldana Dave Bautista James Gunn Action Adventure Comed...
4	Sean McNamara	Pierce Brosnan	William Hurt	Benjamin Walker	Fantasy Action Adventure	the king's daughter	Pierce Brosnan William Hurt Benjamin Walker Sean McNamara Fantasy Actio...
...	...	...	...	...	...	...	...
524	Jim Strouse	Jessica Williams	Chris O'Dowd	Keith Stanfield	Romance Comedy	the incredible jessica james	Jessica Williams Chris O'Dowd Keith Stanfield Jim Strouse Romance Comedy
525	Farhad Mann	Adelaide Kane	Benjamin Hollingsworth	Jean Louisa Kelly	Romance	can't buy my love	Adelaide Kane Benjamin Hollingsworth Jean Louisa Kelly Farhad Mann Romance
526	Hannaleena Hauru	Inka Haapamäki	Rosa Honkonen	Tiitus Rantala	Romance Comedy	thick lashes of lauri mäntyvaara	Inka Haapamäki Rosa Honkonen Tiitus Rantala Hannaleena Hauru Romance Co...
527	Jonathan A. Rosenbaum	Lou Diamond Phillips	Wallace Shawn	Gina Holden	Crime Comedy Action Family	cop and a half: new recruit	Lou Diamond Phillips Wallace Shawn Gina Holden Jonathan A. Rosenbaum Cr...
529	Ravi Udyawar	Sridevi Kapoor	Sajal Ali	Akshaye Khanna	Crime Drama Thriller	mom	Sridevi Kapoor Sajal Ali Akshaye Khanna Ravi Udyawar Crime Drama Thriller

458 rows × 7 columns

Fig.4 Concatenating extracted info to new column

IV. RESULTS AND DISCUSSION

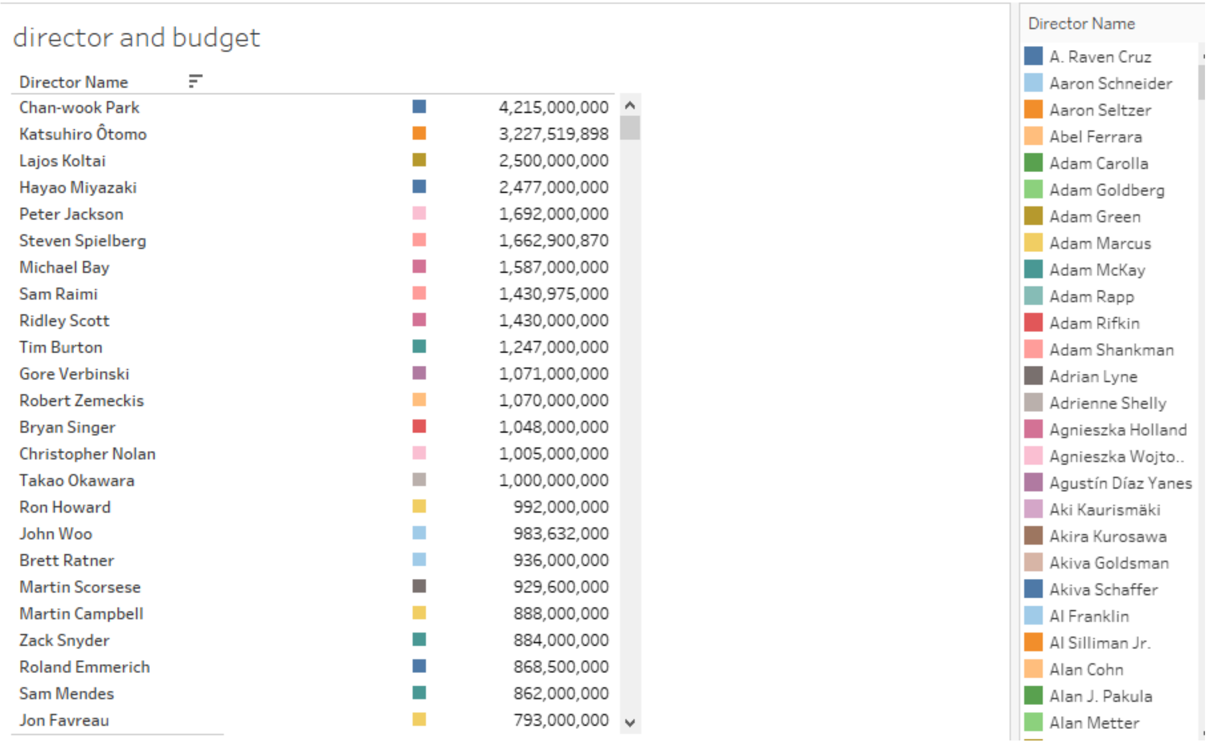
Data Visualization

Tableau is a widely used data visualization tool and is an extremely powerful tool and has an easy UI for interaction with data. Tableau has different key products and the first stage of this process is transaction processing where it ensures data is collected and stored in a database. Second is Data analyzed here data cleansing is done where in- appropriate data is handled. Tableau allows us to connect to any form of data such as excel, web APIs etc.

Tableau can be easily handled with just dragging and dropping dataset columns and selecting suitable charts or graphs required for them. It allows us to build reports and dashboards that can be shared across any organization. Another key feature is insights shared and checks safe and secure ways to share data and give access. Tableau writes optimized sequel queries in order to fetch data from a dataset or database. Tableau desktop and tableau online are used to design insights and tableau server, reader and tableau public are used for publishing reports. It divides the data into two parts that is dimensions where string, dates and geographical values data is considered and second is measure where all the numeric, float or figures data is present.

Few reasons why tableau is used is because of its speed, scalability and powerful integration. Less time is needed in tableau to create graphs compared to other software's. Algorithms are used to format the graph precisely which eventually benefits the heavy lifting in data analysis. It is an all in one tool that is integrated which helps create sophisticated presentations for the data. Controllable dashboards, automatically updates to show the data the user has queried. With its powerful integration it helps/allows it to work with the raw data in multiple formats i.e. text files, excel, tables from pdf which then can be extracted with tableau. It allows connections to the organization database as well.

Results:



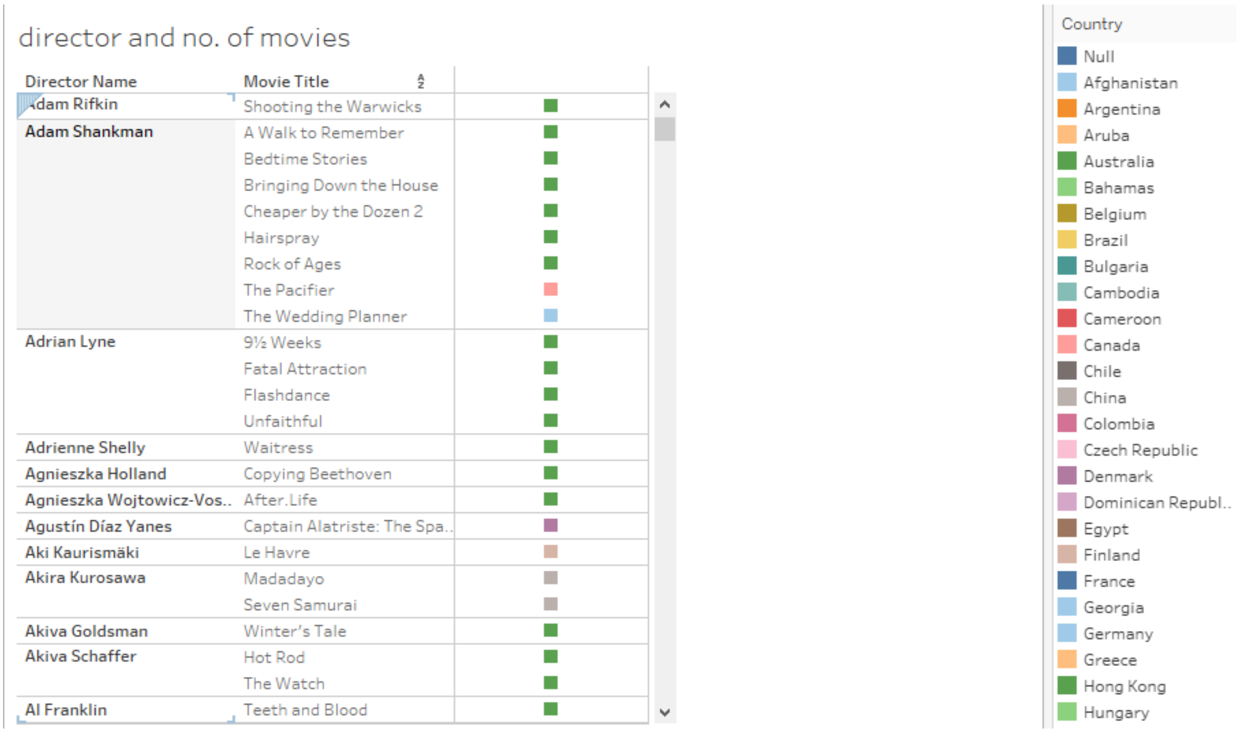
director and budget

Director Name		
Chan-wook Park	🟦	4,215,000,000
Katsuhiro Ôtomo	🟧	3,227,519,898
Lajos Koltai	🟨	2,500,000,000
Hayao Miyazaki	🟦	2,477,000,000
Peter Jackson	🟪	1,692,000,000
Steven Spielberg	🟥	1,662,900,870
Michael Bay	🟥	1,587,000,000
Sam Raimi	🟥	1,430,975,000
Ridley Scott	🟥	1,430,000,000
Tim Burton	🟩	1,247,000,000
Gore Verbinski	🟪	1,071,000,000
Robert Zemeckis	🟧	1,070,000,000
Bryan Singer	🟥	1,048,000,000
Christopher Nolan	🟪	1,005,000,000
Takao Okawara	⬜	1,000,000,000
Ron Howard	🟨	992,000,000
John Woo	🟦	983,632,000
Brett Ratner	🟦	936,000,000
Martin Scorsese	⬛	929,600,000
Martin Campbell	🟨	888,000,000
Zack Snyder	🟩	884,000,000
Roland Emmerich	🟦	868,500,000
Sam Mendes	🟩	862,000,000
Jon Favreau	🟨	793,000,000

	Director Name
🟦	A. Raven Cruz
🟦	Aaron Schneider
🟧	Aaron Seltzer
🟧	Abel Ferrara
🟩	Adam Carolla
🟩	Adam Goldberg
🟨	Adam Green
🟨	Adam Marcus
🟩	Adam McKay
🟩	Adam Rapp
🟥	Adam Rifkin
🟥	Adam Shankman
⬛	Adrian Lyne
⬜	Adrienne Shelly
🟥	Agnieszka Holland
🟪	Agnieszka Wojto..
🟪	Agustín Díaz Yanes
🟪	Aki Kaurismäki
🟫	Akira Kurosawa
🟫	Akiva Goldsman
🟦	Akiva Schaffer
🟦	Al Franklin
🟧	Al Silliman Jr.
🟧	Alan Cohn
🟩	Alan J. Pakula
🟩	Alan Metter

Fig.5 Director and Budget Graph

The director and budget plot allows us to analyze the relation between director and the budget planned for each movie. We have arranged the budget in descending order in the above text bar. We have taken the sum of the budget of each director which is represented by unique colors.



director and no. of movies

Director Name	Movie Title	
Adam Rifkin	Shooting the Warwicks	🟩
Adam Shankman	A Walk to Remember	🟩
	Bedtime Stories	🟩
	Bringing Down the House	🟩
	Cheaper by the Dozen 2	🟩
	Hairspray	🟩
	Rock of Ages	🟩
	The Pacifier	🟥
	The Wedding Planner	🟦
Adrian Lyne	9½ Weeks	🟩
	Fatal Attraction	🟩
	Flashdance	🟩
	Unfaithful	🟩
Adrienne Shelly	Waitress	🟩
Agnieszka Holland	Copying Beethoven	🟩
Agnieszka Wojtowicz-Vos..	After.Life	🟩
Agustín Díaz Yanes	Captain Alatriste: The Spa..	🟪
Aki Kaurismäki	Le Havre	🟫
Akira Kurosawa	Madadayo	⬜
	Seven Samurai	⬜
Akiva Goldsman	Winter's Tale	🟩
Akiva Schaffer	Hot Rod	🟩
	The Watch	🟩
Al Franklin	Teeth and Blood	🟩

	Country
🟦	Null
🟦	Afghanistan
🟧	Argentina
🟧	Aruba
🟩	Australia
🟩	Bahamas
🟨	Belgium
🟨	Brazil
🟩	Bulgaria
🟩	Cambodia
🟥	Cameroon
🟥	Canada
⬛	Chile
⬜	China
🟥	Colombia
🟪	Czech Republic
🟪	Denmark
🟪	Dominican Republ..
🟫	Egypt
🟫	Finland
🟦	France
🟦	Georgia
🟦	Germany
🟧	Greece
🟩	Hong Kong
🟩	Hungary

Fig.6 Director and number of movies graph

This plot helps us to understand which director has directed the kind of movies Director and number of movies along with the country to which the movie belongs and also which movie was directed by respective directors. Each movie title is also represented by square symbols in different colors to better understand the country in which the movie was produced by the respective director.

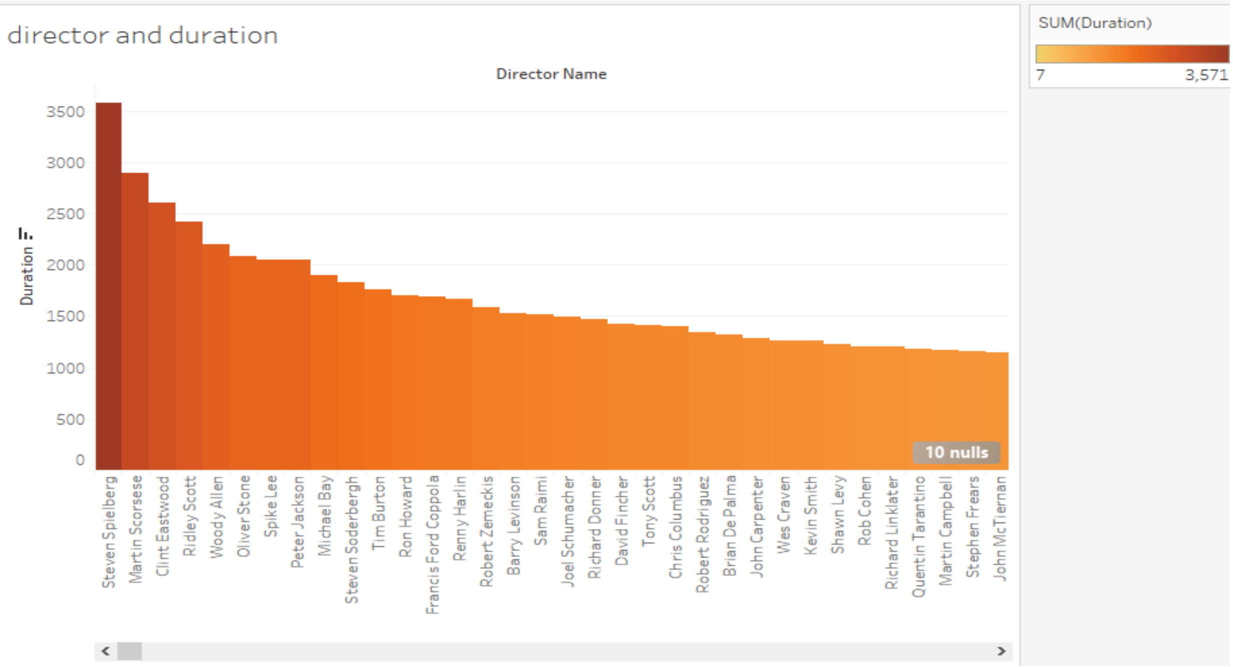


Fig.7 Director and Duration graph

This bar graph gives information of duration of each movie and their producer's relation in the graph using horizontal bars in descending order within a specified range. The duration column is filtered to a maximum range of 3500 which descends gradually.

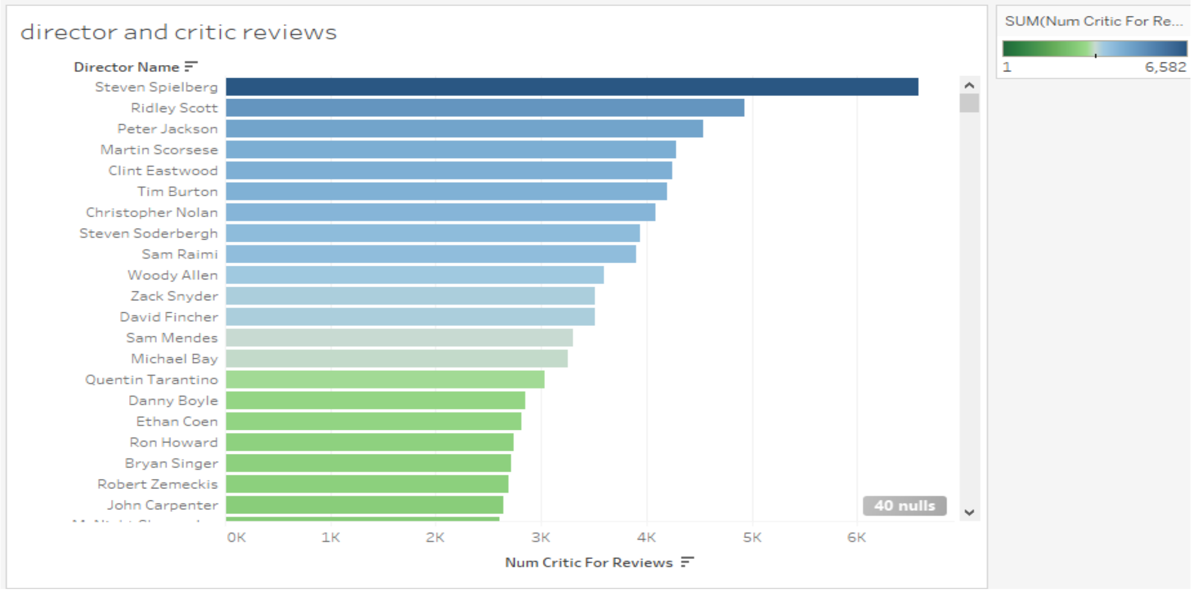


Fig.8 Director and Critic Reviews graph

Based on the user reviews with their respective countries, directors are plotted in this graph, which is an effective way to predict which movie is better. Here we have plotted the graph using horizontal bars to show the critic review along with the director’s name in descending order.

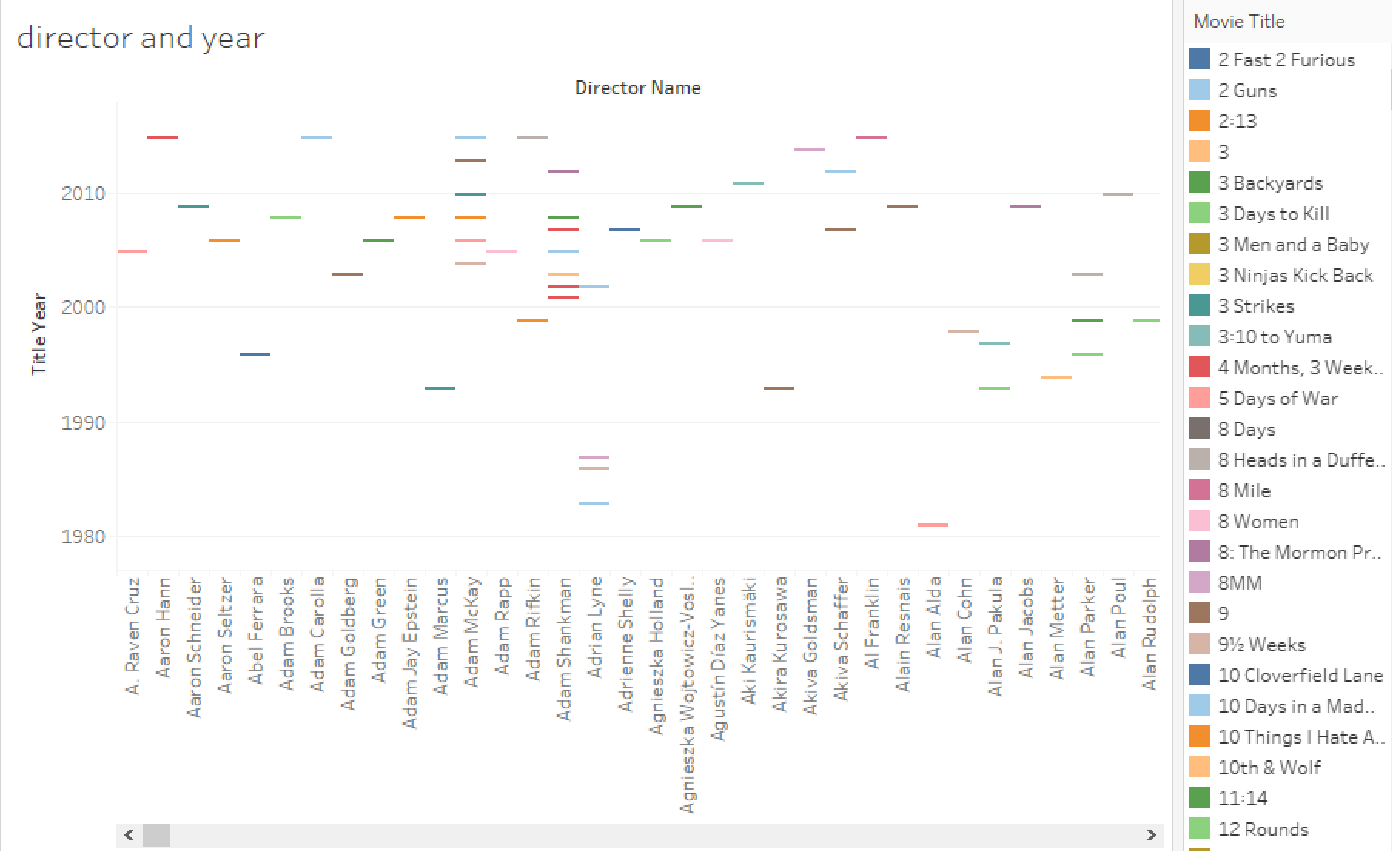


Fig.9 Director and Year graph

The gantt chart gives the information about the year in which the movie was released and the scatter plot is plotted in different colors with movie names.

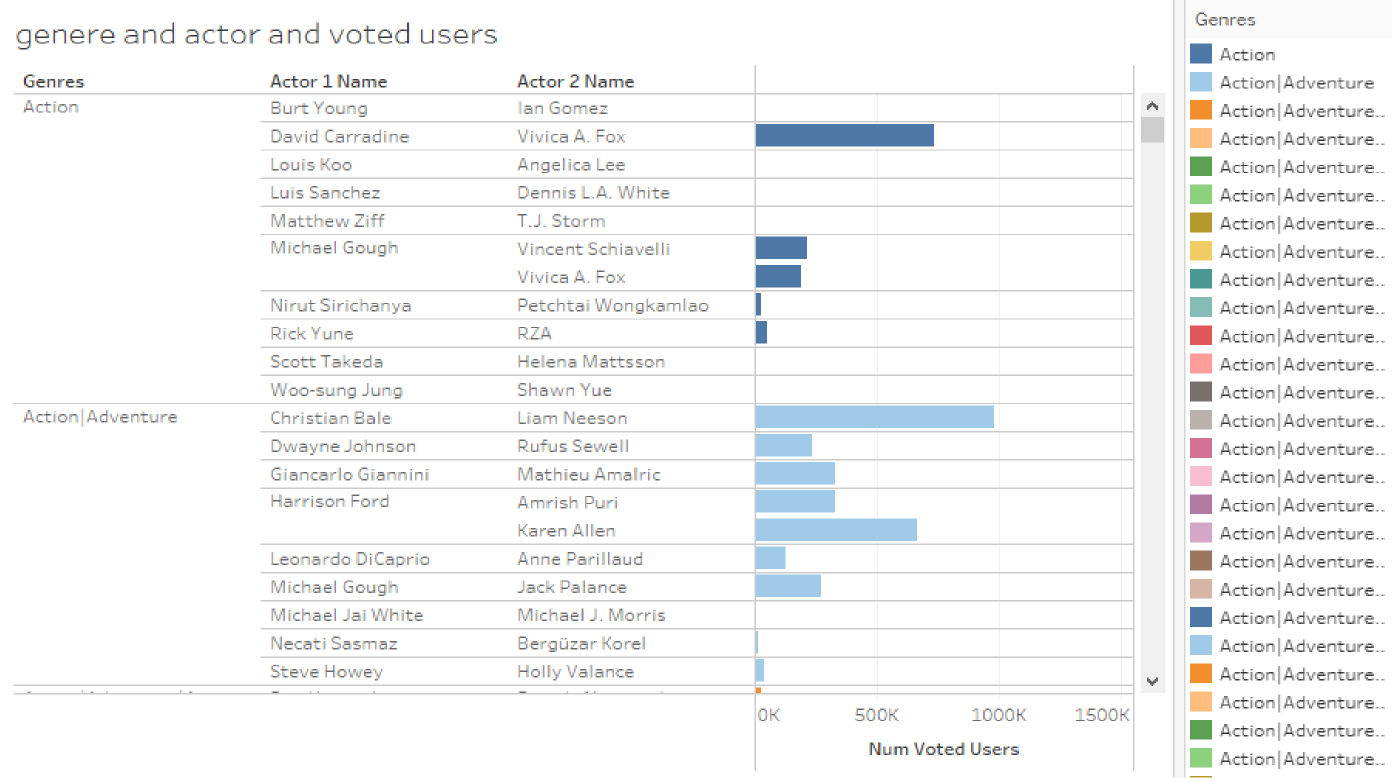


Fig.10 Genre, Actor and Voted users graph

The insight above shows the plot between genres and the cast members of all the movies in the dataset along with the users voted for that particular cast and movie.

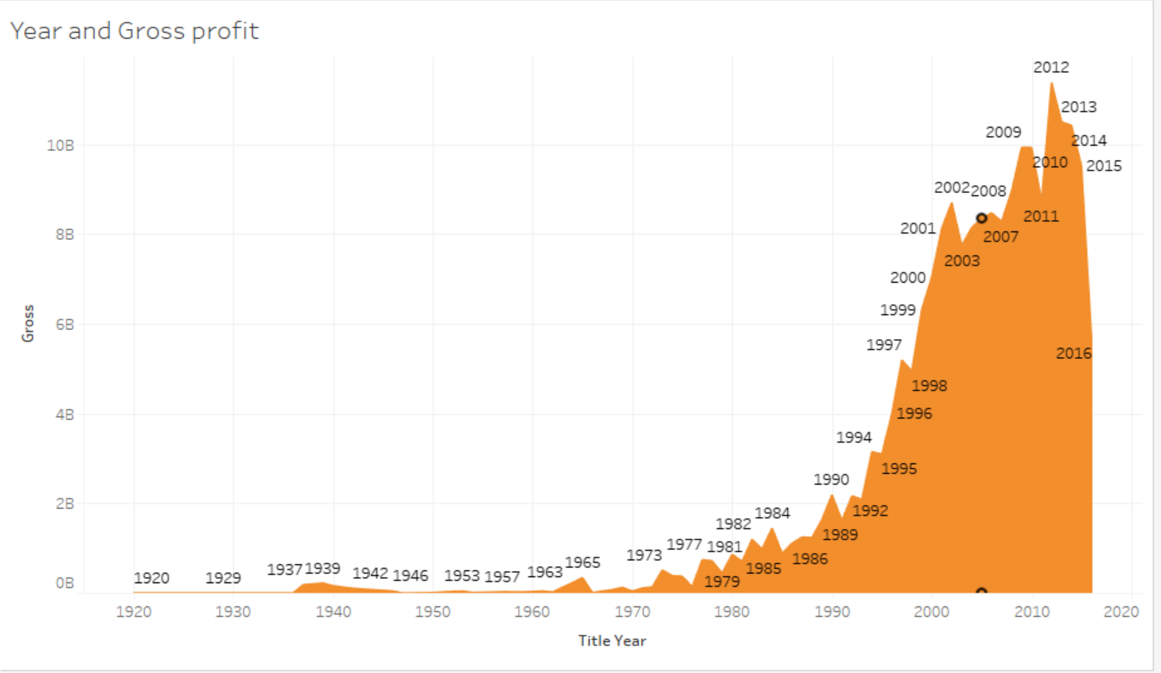


Fig.11 Year and Gross profit graph

The Year and Gross is an analysis of the overall profit earned in the year of release and the max and min value for each label of that year and we can clearly see that there has been a significant rise in the profit in this industry.

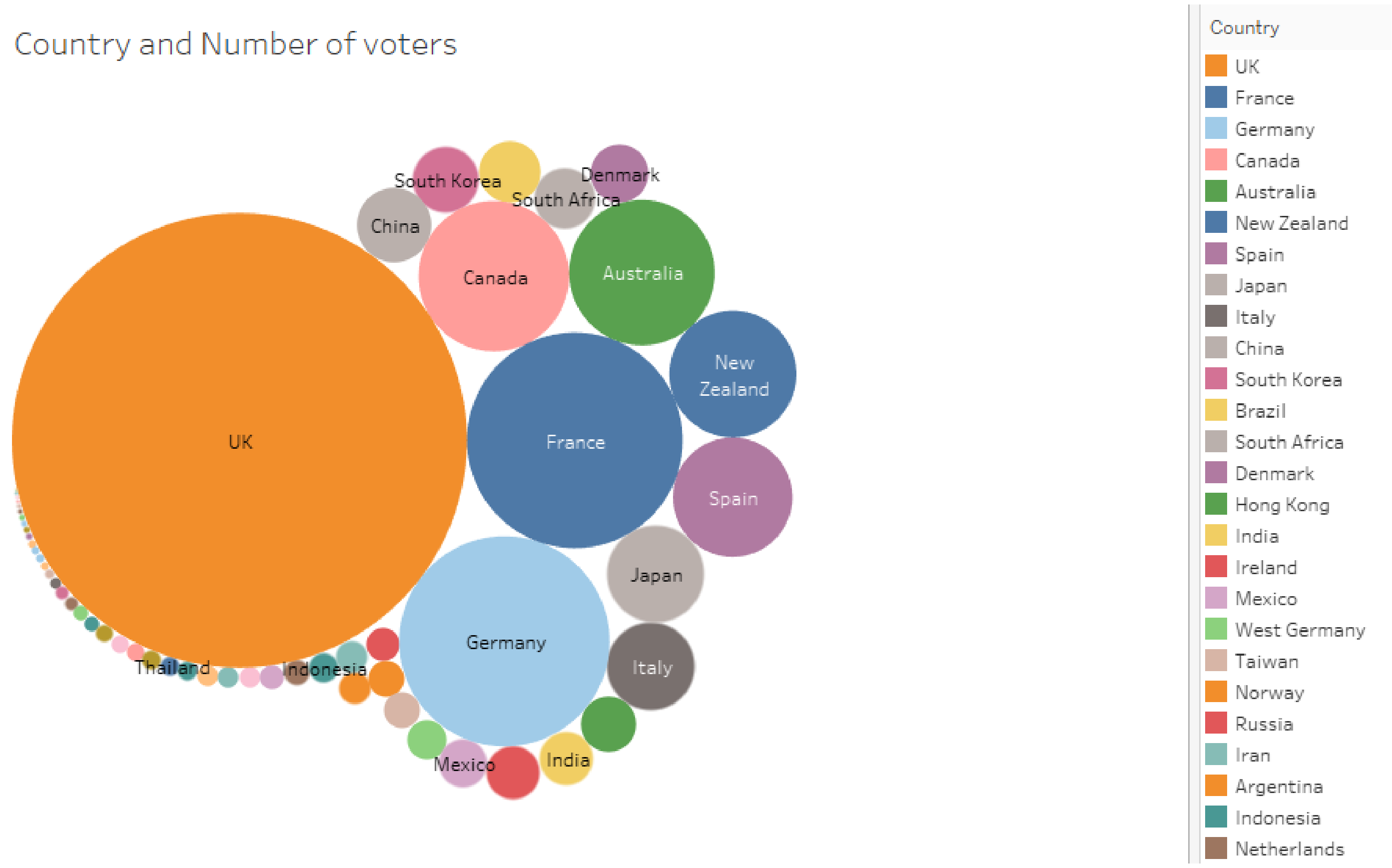


Fig.12 Country and Number of voters graph

The pie chart is the prediction for the number of voters in each country who have voted maximum for each movie and as seen the UK has the most voters.

Conclusion:

We have cleaned the data that we obtained from Kaggle and IMDB. We performed exploratory data analysis on the data like taking valuable insights from the data, removed the missing values and did some preprocessing as well so that we can train our model. So now whenever a user will visit our website and write the movie name in the search bar, the user will see some auto-suggestions related to the movie. As soon as the user presses enter, the page will redirect to the new page where all the information related to the searched movie like cast, genre, ratings, reviews, etc. will be visible to the user.

Future Scope:

Collaborative filtering [7] analyzes the information of the user considering its similar interest and gives probability that the target individual or user will be interested. To give personalized recommendations to the user according to its similar preferences, collaborative filtering uses various algorithms for filtering data from user reviews.

After successfully collecting the required data and then carried out pre-processing techniques on the same which was then followed by data visualization on tableau. Now this data and insights can be further used to build a recommender system.
A recommender system based on Item-based collaborative filtering which will make predictions based on the user ratings vectors and movie id in the dataset. User will be asked to fill a survey form and that data will be passed to our CF-model, along with the movie id there will be user id which will be used by the model to compare with the type of movie current user has requested for and if both have given similar ratings to the movie then model can predict them as similar users.

Once the similarities are calculated for users then we find the top n users and find correlation with the input user. In this way we will carry out the same process for different users. Once calculated for multiple users now we will multiply the correlation and ratings column for that user to get weighted ratings. At the end we sum the weighted rating and correlation and then divide the results to get the top users.

REFERENCES

[1] Keshava, M. & Reddy, P. & Srinivasulu, S. & Naik, B.. (2020). Machine Learning Model for Movie Recommendation System. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS040741.

[2] Kashyap, A. et al. “A Movie Recommender System: MOVREC using Machine Learning Techniques.” (2020).

[3] Ko, Hyeyoung, et al. "A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields." Electronics 11.1 (2022): 141.

[4] Zahra Abbasi-Moud, Hamed Vahdat-Nejad, and Javad Sadri. 2021. Tourism recommendation system based on semantic clustering and sentiment analysis. <i>Expert Syst. Appl.</i> 167, C (Apr 2021). DOI:https://doi.org/10.1016/j.eswa.2020.114324

[5] Lee C, Han D, Han K, Yi M. Improving Graph-Based Movie Recommender System Using Cinematic Experience. Applied Sciences. 2022; 12(3):1493. https://doi.org/10.3390/app12031493

[6] D. Oreščanin, T. Hlupic and I. Sorić, "Predictive models for digital broadcasting recommendation engine," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 1243-1248, doi: 10.23919/MIPRO.2018.8400225.

[7] Breese, John & Heckerman, David & Kadie, Carl. (2013). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. UAI.