National College of Ireland

Project Submission Sheet – 2020/2021

Student Name: Aanchal Singh, Deepak Kumar Swain, Sai Prasanna Gontyala and Sweta Kumari ………

Student ID: 19221771, 19216769, 19233388 and 19240848 ………

Programme: MSc. Data Analytics …………………………………………………………………… **Year:** 2021 ………………………

Module: Data Analytics and Programming ………

Lecturer: Anu Sahni ………

Submission Due Date: 08-01-2021 ………

Project Title: Data Analysis and Visualization of different trends in movies, TV shows ………

Word Count: ………

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: *Aanchal Singh, Deepak Kumar Swain, Sai Prasanna Gontyala and Sweta Kumari* ………

Date: 08-01-2021 ………

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Data Analysis and Visualization of different trends in Movies and TV shows

Deepak Kumar Swain
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x19216769@student.ncirl.ie

Sweta Kumari
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x19240848@student.ncirl.ie

Sai Prasanna Gontyala
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x19233388@student.ncirl.ie

Aanchal Singh
MSc Data Analytics
National College of Ireland
Dublin, Ireland
x19221771@student.ncirl.ie

***Abstract*—The aim of this research is to analyze the different trends and distribution of popular movies and TV shows with the reviews. Considering the different movies and tv shows as per the popularity, sales. Ratings, genres and various different factors we have created different visualizations for a better understanding to have the trends. Along with sentiment analysis of different positive and negative reviews to compare the different movies and entertainment. We also carry sentiment analysis for popular movies and tv shows. A key note that can be observed is that there is no correlation between the reviews and popularity of the movies. The datasets are related to movie and show ratings and is also web scrapped from website. Once the data is collected data cleansing is done in MongoDB and PostgreSQL is used to structure the data which can later be visualized.**

Keywords—Analytics, visualization, entertainment, trends, movies, shows

I. INTRODUCTION

The most common modes of entertainment these days are different movies, shows which are easily available on different OTT platforms. To take off mind from mundane lives and during the pandemic times people are more used to watching movies and shows available on different platforms. We being students are also avid movie lovers and it always interests us to know how these movies and shows are doing. Do dive deeper we have taken datasets which are related to different movies so that we can check which movies are doing well, which directors or producers have done a better job and how the sales are doing for all the movies.

The above questions intrigued us in trying to find the different trends and see what are the factors which are affecting the people's interests. We gathered data for various popular movies and shows through the years and have tried to visualize different aspects and trends found in the data. Another way of finding what people think and what aspects they have for a movie are through different reviews and ratings given by the people to them. To understand all these perspectives, we have taken in consideration four datasets and came up with the notion of trying to find answers to the research questions: "What are the different movies and shows which have generated the highest revenue, along with, who are the famous actors and directors which are giving hits? Which are the trending genres, movies and the feedback for the different shows and the does age define the people watching the shows?"

To answer the research question, we must have a thorough understanding of the different movies and shows which are considered to be popular among people. The factors which affect these can be the actors, directors, genres, production house and profits. The factor of reviews and ratings has an impact on deciding the popularity of shows and movies.

An extensive study of the different data sets was carried on data similar to above description. It helped in getting an insight on what has already been done and what we can address in case similar findings were made and to keep a note of these from the research material. The below sections describe the complete methodology which is used for conducting the research and all the different insights which are gained while looking for answers for the research question. Let us now go ahead and see how we have found the data, extracted the information related to it and visualized this data.

II. RELATED WORK

The budget of the movies is of the order of hundreds of millions of dollars, making the box office success absolutely essential for the survival of the industry [1]. Advertising campaigns contribute heavily to the total budget of the movies. Sometimes the investment results in heavy losses to the producers [1]. Web portals like IMDB and TMDB become one of the important factors in getting profit for the movie producers [2]. Although most traditional business intelligence tools are clearly aimed at market analysts or a few decision - makers, data visualization is seen as a way to make business analytics available to a wider audience [3]. From Ms. Sushmita Roy's analysis [3] a brief understanding of the importance of data visualization is gained and implemented using python in the current analysis.

The research deals with analysis of audiences of 4 SVOD Platform: Netflix, Blockbuster, Hulu and HBO, especially the complexity of GRP Metrices of different and video on demand challenges. By end of first decade of 21st Century how online television became more popular over traditional television, this has brought the digital revolution in entertainment with numerous amounts of content. As demand increases with content, broadcasting channelizing medium, Devices choices has become concerns for distributors of SVOD (Subscription to video on Demand) by Netflix or HBO. Netflix is globally famous and has spread.[4]. The movies have an impact on all people like actors, directors and production houses. This will also end up on deciding upon the budget of the movie and how people will watch it and what provider they will continue. The success of movies depends on factors like actor, actresses, crew, cast, production company, release time, story etc [5]. The different approach to find any relation between budget of a movie and the other financial aspects didn't go away.[6]

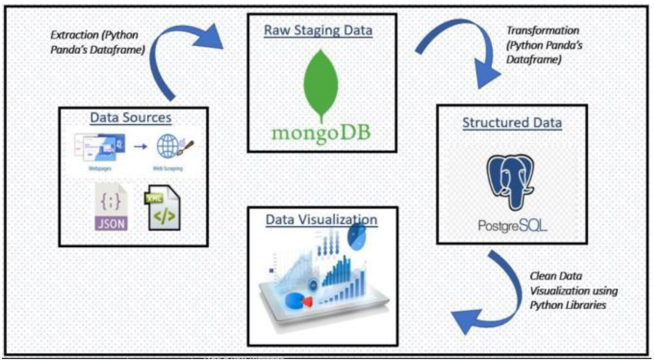
Research by Nemeth, B. et al. states that the region in which these tv shows and movies are released also helps for increasing the loyalty for service provider [7]. The success of these aspects also depends on the reviews given by the audience. The reviews in directly also impact the gross income. We can analyse the reviews and ratings and reach a conclusion on which movies are rated on what basis of genre and are having a good box office. The research checks ways to classify a movie based on different reviews. Each review focuses on some movie and provides a summary of the movies.[8]Genre specific reviews demand special techniques while analysing where we can look for unique meaning based on the context i.e. genre in which they are used.[8]

III. METHODOLOGY

In the project, a demonstration of database and analytics processes is clearly given. To start with the process, four semi-structured datasets were collected based on movies, TV shows. The datasets were as follows.

1. IMDb Dataset in JSON format.
2. TMDB dataset in JSON format.
3. Streaming platform dataset in XML format.
4. Movies and shows data scrapped from IMDb website using python's beautifulsoup library.

To extract the dataset, we have used python as it is one of the most user-friendly programming languages with advanced packages to handle different complex data sources in an efficient manner.


Extraction (Python Panda's Dataframe)
Raw Staging Data
mongoDB
Transformation (Python Panda's Dataframe)
Data Sources
Webpages
Web Scraping
JSON
Structured Data
PostgreSQL
Data Visualization
Clean Data Visualization using Python Libraries

The data were extracted using necessary python libraries and stored in respective lists. Later, these lists were stored in pandas data frames. For each dataset, we had separate data frames.

To store the extracted data, we have used a NOSQL database called 'MongoDB'. Unlike relational databases, Mongodb stores data as document. It stores data in binary JSON format. The reason to choose Mongodb as our staging database is doe to it flexibility in storing data, powerful querying features. Python has an advanced library called 'pymongo' which makes it easier to build a connection between python and Mongodb. With the help of pymongo, necessary connections were made and after that four collections were created. In these collections, the extracted data frames were stored.

Data cleaning process was done in pandas data frame. As we extracted the data from different semi-structured data sources, the quality of data was not good. Hence, following data cleaning activities were performed.

- Dropped unnecessary columns.
- Removed special characters like $, %, dot, characters in numerical columns, +, spaces from necessary columns.
- Handled missing data in all the datasets.
- Data type changes for necessary columns. For example, columns of object type containing numerical variables were converted to integer or float type.
- Renamed few columns to get better understanding

To store the cleaned data, we have used a relational database called 'PostgreSQL'. It is an open-source database which has several advanced features like building schemas, indices. Like other relational databases, it follows the basic principle of 'Atomicity', 'Consistency', 'Isolation' and Durability. With the help of python library 'psycopg2', necessary connections were built between python and PostgreSQL. For each dataset, a table was created in PostgreSQL and respective data were stored in those tables.

Later, the cleaned data were extracted to pandas data frames to perform visualization to get insightful information. Let us see how we have followed the above specified methodology in the below section.

A. *Scrapping The Data From Webpage*

The first dataset is scrapped from IMDB which is an online database of information related to movies, TV shows and various entertainment contents. With the help of python library 'BeautifulSoup', the data were scraped from the official website of IMDb. To scrape the complete dataset, 237 URLs were used. Three different lists were used to store the URLs for Movies, TV shows and Anime. Initially, the data were stored in three different data frames (Movies, TV shows and Animated series) and later these data frames were merged into one data frame before dumping into Mongodb. The extracted dataset contains the details of top-rated movies, TV series and Animated series based on user ratings.

B. *IMDB Data from JSON*

The dataset that we have chosen here contains 10000 rows and 24 columns. The dataset contains information about different movies and shows along with the date and year when it was published. It has information related to different actors, directors, genres, the budget and gross incomes for US and world. It also has reviews and ratings for all the different shows. We are using all these factors to analyse and create visualizations.

C. TMDB Data from JSON

The TMDB dataset contains data of 10,866 movies and 21 fields. This mainly has the information of budget and revenue against each movie and the respective production house and genre. This fields helps in analyzing and visualizing cases like which movies made highest revenue, which production house makes the highest money and which genres demand the highest budget. Such part of research questions is addressed analyzing this dataset.

D. OTT shows and movies from XML file

The OTT dataset contains data of 16739 observation and 16 fields. This mainly has the information of 4 most popular OTT platform worldwide, movies, Ratings, Genres wise details, Runtime, Language, Country etc. Using this dataset, I have found which Platform channel is most used, which Age category has subscribed most, which Year more users have gone high and also compared ratings from IMDb and Rotten Tomatoes. I have found Prime Video users are high in numbers, second comes Netflix and Adult users are more. USA has highest number of users and 2000 – 2020 has highest number of movies made

IV. RESULT

A. Scrapped dataset from IMDb

The cleaned data were extracted from PostgreSQL using pandas data frame for visualization.

- The following bar plot gives an overview of number of shows with respect to certificate types. These certificate values vary with countries. Since, the project work is done in Ireland, following types of values were extracted from the website.

  G - Suitable for children of school going age

  PG - Parental guidance suggested for younger children

  12A - Children under 12 may attend only if accompanied by a parent or guardian (2005-)

  12 - Same as 12A for videos: suitable for persons 12 and older

  12PG - Same as 12A (-2004)

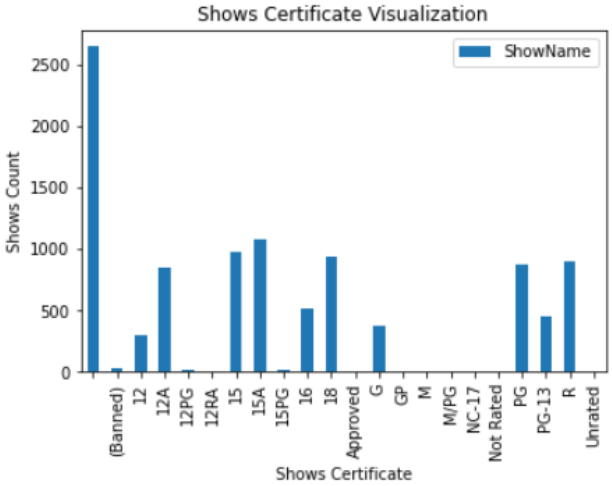
  15A - Children under 15 may attend only if accompanied by a parent or guardian (2005-)

  15 - Same as 15A for videos: suitable for persons 15 and older

  15PG - Same as 15A (-2004)

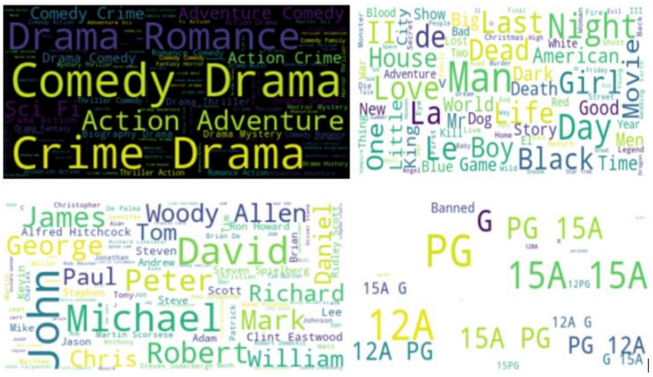
  16 - Persons under 16 not admitted to cinemas (2005-)

  18 - Persons under 18 not admitted to cinemas

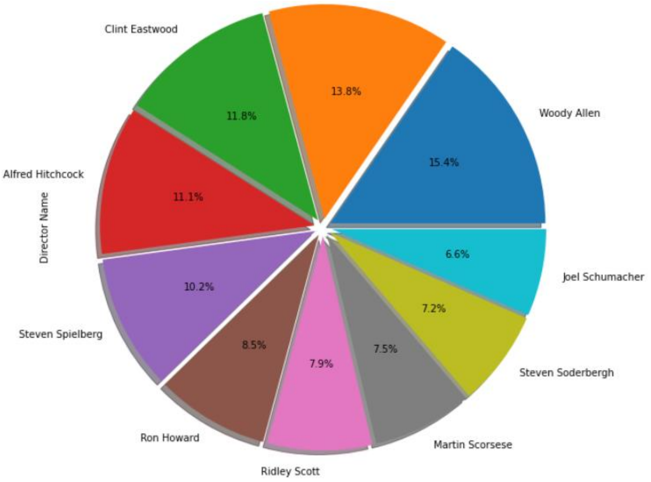


The first bar with no certificate type is indicating the number of shows which do not have any certificate type on IMDb website.

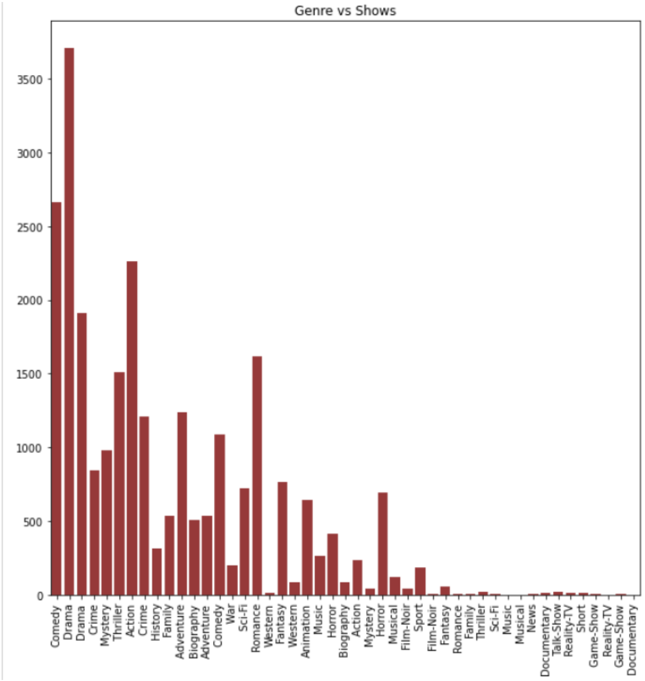
- The following word clouds are the visual representation of text data. The words with bigger size show high importance. Also, it means, these words are used frequently in the dataset.



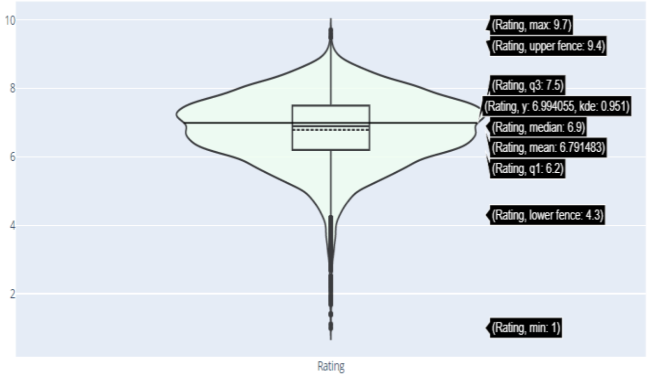
- The following pie chart shows the top 10 directors based on number of shows they made.



- The following bar chart shows the number of shows counts with respect to unique genres. Plotting this plot was a bit tricky as most of the shows were belonging to multiple genres. At first, unique genres were extracted by passing the parameters in a for loop and then the counts of shows were extracted for each unique genre. After this, the following plot was drawn.

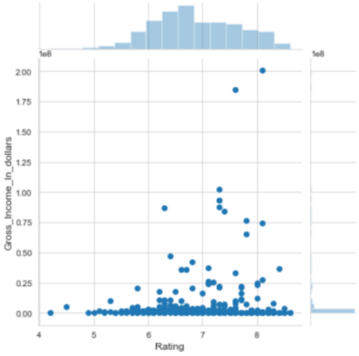


- The following violin plot was drawn with the help of Plotly library. This plot is a mixture of whisker plot and density plot. A seen in the following plot, the density between the ratings 6 an 8 is high. This indicates that large number of movies are rated between 6 and 8.

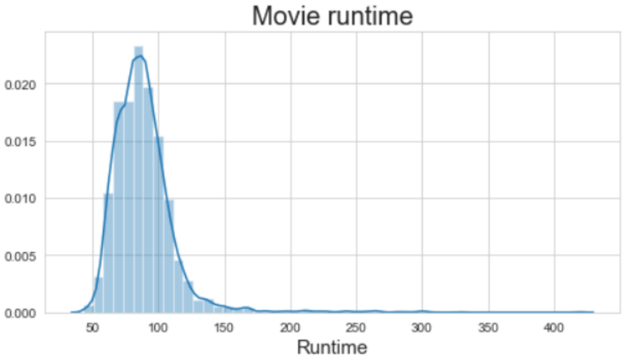


B. *IMDb JSON DATASET*

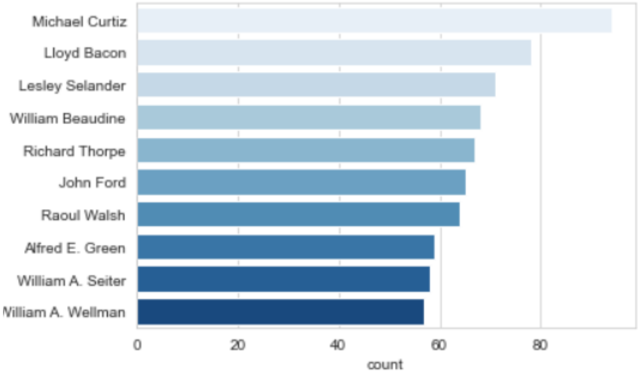
- The below visualization show cases a joint plot where we have taken rating and gross income and creating a merged and scatterplot and bar plot. It shows the rating wise gross income. It covers both these factors together.



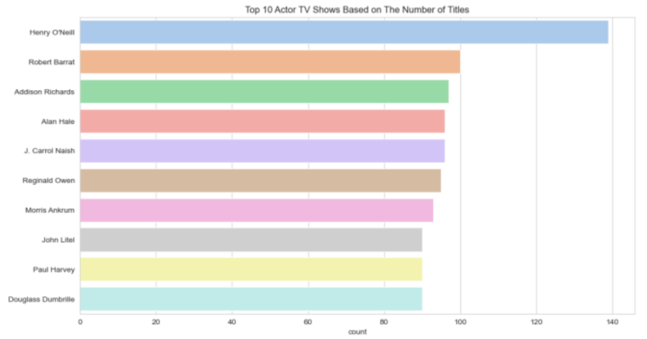
- We have also visualized the data where we check the analysis of the shows that we have and we can observe that most of the shows have their duration between 50 minutes to 150 minutes.

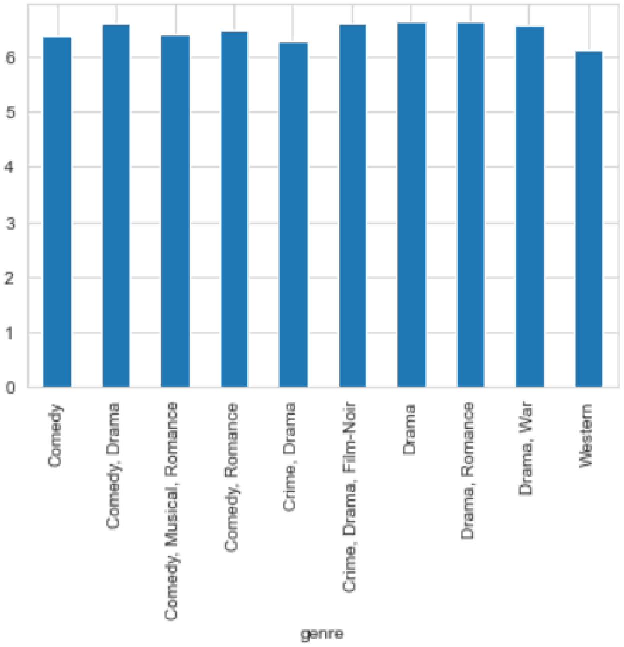


- On the basis of maximum number of movies we have categorized the directors who have directed the maximum number of movies.



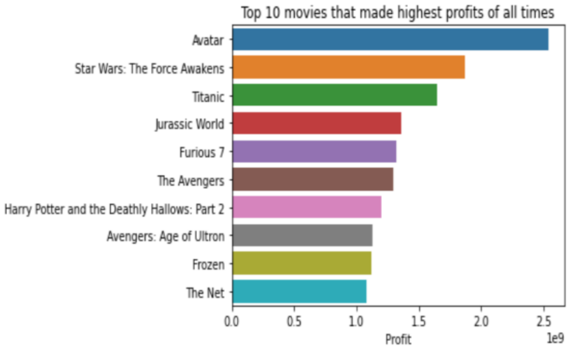
- Another countplot shows the best actors based on the number of shows they have acted in.



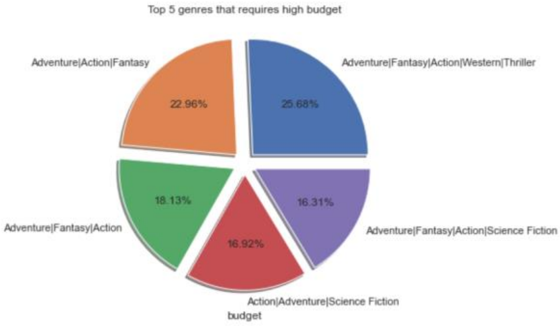


- The above mean plot shows the different genres which are being watched more. Here we have taken mean of the genres and plotted those.

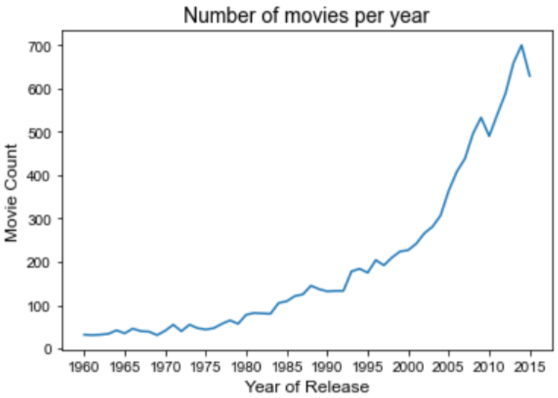
C. *TMDB JSON FILE*



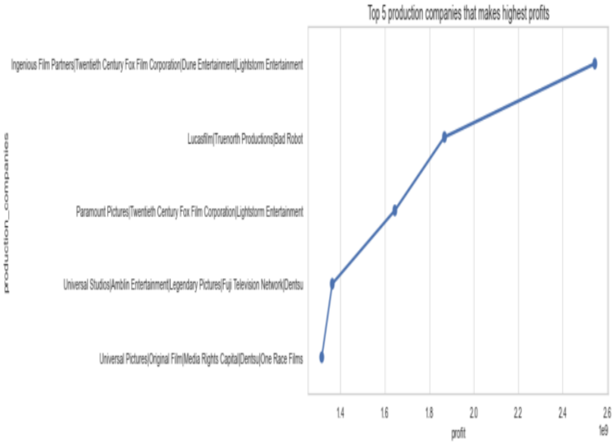
- The above barplot illustrates the top movies that has made the most revenue of all years. Avatar made the highest revenue of all the movies in the dataset.



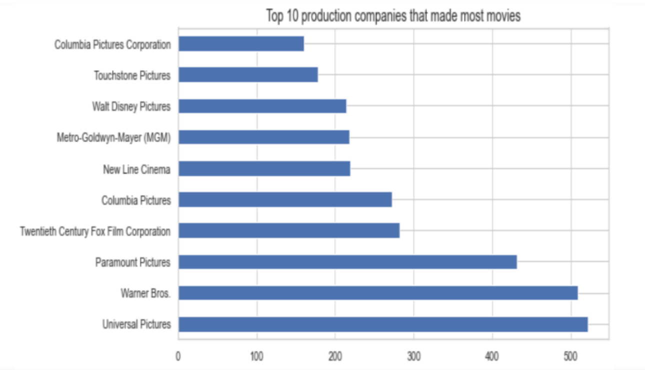
- From the above pie chart it can be deciphered which genre requires the maximum budget to make and release a film. We can observe that the Action, Fantacy, Science Fiction and their combinationsrequires huge budget to make a film.



- The above plot depicts the increasing move count over the years. We can observe that the increase is exponential. However, there is an huge increase in the move count released between 2000 and 2010.



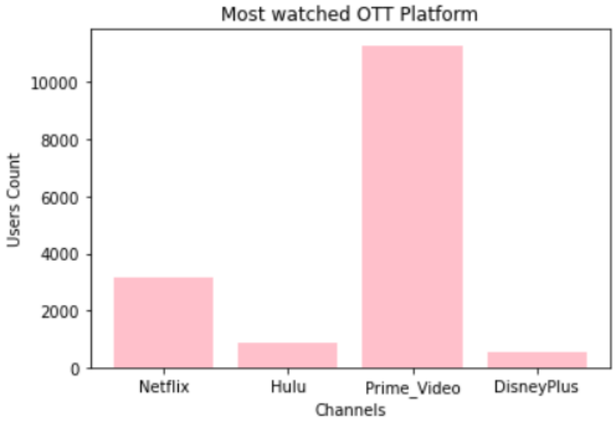
- From the above plot depicts the top five production houses that are in profits based on the TMDB dataset. It can also be deciphered that all the top production houses that made high revenue have worked jointly.



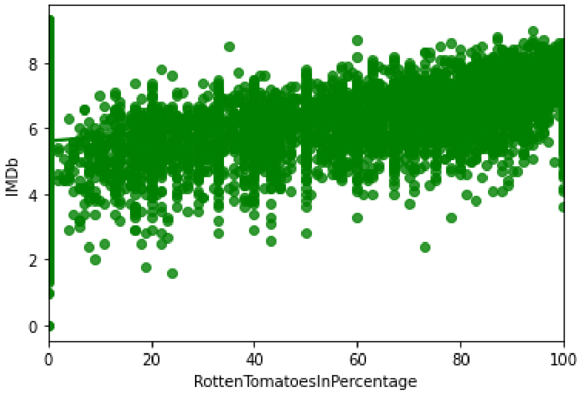
- The above barchart indicates the production companies that made the highest number of movies over the years. It deciphers that major section of movies are made by Universal Pictures is a production house that makes the most movies.

D. STREAMING DATA XML FILE

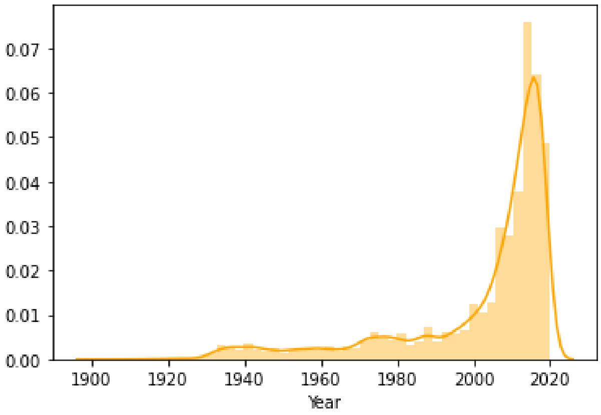
- Considering Four OTT Platform we have figured out Prime Video is the most used platform whereas Netflix comes second, and least subscribed platform is Disney Plus. This helps us to figured out that Prime Video content is more preferrable for all categories of Customer.



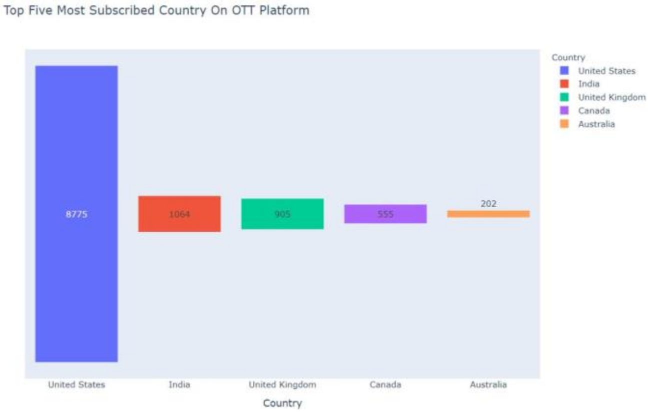
- The feedback for movies is compared in the dataset. We can see IMDb Ratings and Rotten Tomatoes both are almost similar that means top rated movies watched more compare to less rating movies in IMDb.



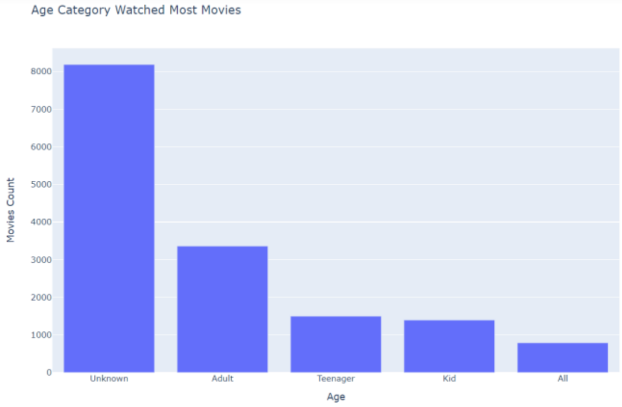
- Frequency of released movies have gone high over last two decade specially observed from 2000 to 2020 as seen on above Figure due to more demand and emergence of more directors. As entertainment is leading specially from last two decades along with digitalization, we can see even spike in more movies being produces over these years.



- Top 4 Countries having more no of users, above we have concluded USA users are high next is India then United Kingdom and Canada. USA brings the most revenue when it comes to use OTT Platform reason could be its high-tech country.



- With the data provided we have tried to find which age category is leading in terms of opting for OTT Platform and here we can see Unknow is high after which adult category is leading then Teenager.



V. CONCLUSON AND FUTURE WORK

By the implementation of data analysis and visualization on the movie related data and OTT platforms insights of the budget that a movie would demand based on the genre and the revenue that the movies of the same category have made are attained. It also provides understanding of the impact of feedback on the success of the movies and the statistical details of the movies and genres. The analysis also emphasis on disclosing the top genres and directors that made the most movies. To summarize, from the results of the data visualization an understanding of the factors and risks for the success of the movie can be attained.

These factors can be further used in creating machine learning and deep learning models to predict the success and revenue of the movies prior its release

VI. REFERENCES

[1] Predicting Movie Success Based on IMDb Data by Nithin VR, Pranav M , Sarath Babu PB , Lijiya As

[2] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Recommender Systems Handbook, Springer Press. 2011.

[3] Data Visualization on Movies Dataset using Tableau by Ms. Sushmita Roy.

[4] CITE :T. Suárez-Cousillas, V. Martínez-Fernández and E. Sánchez-Amboage, "SVOD Platform Audience. The Case of Netflix, Blockbuster, Hulu and HBO," 2019 14th Iberian Conference on Information Systems and Technologies (CISTI), Coimbra, Portugal, 2019, pp. 1-6, doi: 10.23919/CISTI.2019.8760790.

[5] Quader, N. et al.,'A machine learning approach to predict movie box-office success', 20th International Conference of Computer and Information Technology, ICCIT 2017, 2018–January, pp. 1–7.doi:10.1109/ICCITECHN.2017.8281839

[6] Asad, K. I., Ahmed, T. and Rahman, M. S., 'Movie popularity classification based on inherent movie attributes using C4.5, PART and correlation coefficient', 2012 International Conference on Informatics, Electronics Vision (ICIEV), Informatics, Electronics Vision (ICIEV), 2012 International Conference on, pp. 747–752. doi: 10.1109/ICIEV.2012.6317401.

[7] Nemeth, B. et al., 'Visualization of movie features in collaborative filtering', 2013 IEEE 12th International Conference on Intelligent Software Methodologies, Tools and Techniques (SoMeT), 2013, IEEE 12th International Conference on, pp. 229–233. doi: 10.1109/SoMeT.2013.6645674

[8] Genre Specific Aspect Based Sentiment Analysis of Movie Reviews