**IMDB MOVIES DATA ANALYSIS – 1986 TO 2016**

TANVI GAWADE

CALIFORNIA STATE UNIVERSITY LOS ANGELES

SUBMITTED TO: DR. SHILPA BALAN

**TABLE OF CONTENTS**

**TOPIC pAGE NO**

**Dataset URL** ------------------------------------------------------------------------------------------- **3**

**Web URL ----------------------------------------------------------------------------------------------- 3**

**Objective------------------------------------------------------------------------------------------------- 3**

**Dataset Description** **----------------------------------------------------------------------------------- 4**

**Data Cleaning ------------------------------------------------------------------------------------------ 5**

**Data Visualization------------------------------------------------------------------------------------- 9**

**Dashboard --------------------------------------------------------------------------------------------- 14**

**Story Telling--------------------------------------------------------------------------------------------16**

**References---------------------------------------------------------------------------------------------- 20**

**DATASET URL:**

[**https://www.kaggle.com/**](https://www.kaggle.com/)

[**https://www.kaggle.com/danielgrijalvas/movies**](https://www.kaggle.com/danielgrijalvas/movies)

**WEBLINK:**

[**https://www.boxofficemojo.com/movies/alphabetical.htm?letter=L&page=3&p=.htm**](https://www.boxofficemojo.com/movies/alphabetical.htm?letter=L&page=3&p=.htm)

[**https://www.boxofficemojo.com/alltime/world/**](https://www.boxofficemojo.com/alltime/world/)

[**https://www.imdb.com/chart/top**](https://www.imdb.com/chart/top)

**OBJECTIVE OF THE STUDY:**

Movies, also called as films, are a kind of visual communication which utilizes moving pictures and sound to tell stories or educate (assist individuals with learning). Individuals in all aspects of the world watch movies as a kind of diversion, an approach to have a ton of fun. Most films are made with the goal that they can be appeared on huge screens at movie theaters and at home. After films are appeared on motion picture screens for a time of weeks or months, they might be showcased through a few other media. They are appeared on pay TV or digital TV and sold or leased on DVD circles or videocassette tapes, with the goal that individuals can watch the films at home.

The making and showing of motion pictures became a source of profit almost as soon as the film industry emerged. In the United States, much of the film industry is centered around Hollywood, California. Other regional centers exist in many parts of the world, such as Mumbai-centered Bollywood, the Indian film industry's Hindi cinema which produces the largest number of films in the world. Though the expense involved in making films has led cinema production to concentrate under the auspices of movie studios, recent advances in affordable film making equipment have allowed independent film productions to flourish. Profit is a key force in the industry, due to the costly and risky nature of filmmaking; many films have large cost overruns, an example being Kevin Costner's Water world. Yet many filmmakers strive to create works of lasting social significance. Revenue in the industry is sometimes volatile due to the reliance on blockbuster films released in movie theaters. The rise of alternative home entertainment has raised questions about the future of the cinema industry, and Hollywood employment has become less reliable, particularly for medium and low-budget films. In the national and global entertainment media production center of ‘Hollywood’, the fiscal crisis punctuated a long-term trend towards declining employment and production capacity. In general, these works can be divided into two categories: academic criticism by film scholars and journalistic film criticism that appears regularly in newspapers and other media. Film critics working for newspapers, magazines, and broadcast media mainly review new releases. Normally they only see any given film once and have only a day or two to formulate their opinions. Despite this, critics have an important impact on the audience response and attendance at films, especially those of certain genres. Mass marketed action, horror, and comedy films tend not to be greatly affected by a critic's overall judgment of a film. According to me the plot summary and description of a film and the assessment of the director's and screenwriters' work that makes up most of film reviews can still have an important impact on whether people decide to see a film. By doing analysis I would be able to figure out how the users tastes changes over time with respect to genres, what is the growth rate of the movie industry per decade, which are the top companies as well as movies that have earned maximum profits in the movie industry. Probably my findings would be helpful to build future recommender systems for certain metrics that based on the IMDb ratings and user votes.

**MOVIES DATASET DESCRIPTION:**

The Movies Dataset has over 3 decades of movies data extracted from the Internet Movies Database (IMDb) for the years 1986 to 2016. There are about 6820 movies in the dataset (220 movies per year, 1986-2016). Each movie has the following dataset attributes:

Budget - The budget of a movie. Some movies don't have this, so it appears as 0

Company - The production company

Country - Country of origin

Director - The Director of the Movie

Name - Name of the movie.

Rating - Rating of the movie (R, PG, etc.)

Released Date - Release date (YYYY-MM-DD)

Runtime - Duration of the movie.

Score - IMDb user rating.

Star - Main actor/actress

Votes - Number of user votes

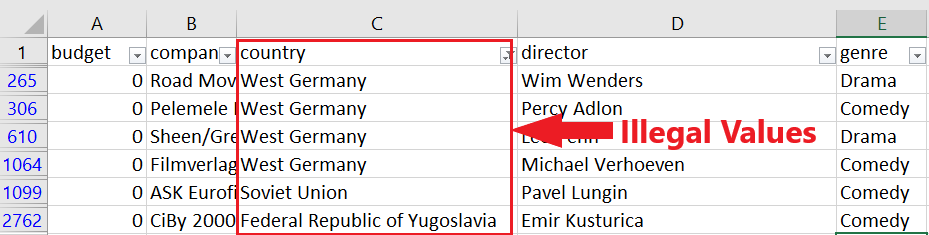
Writer - Writer of the Movie

Year - Year of release

1. **DATA CLEANING**
2. Illegal Values

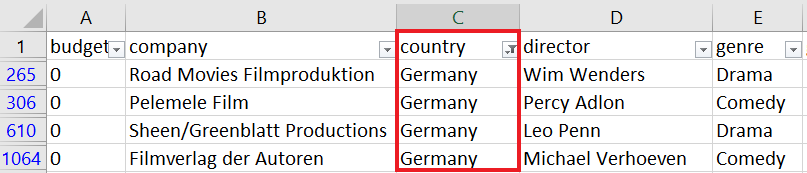
Country/Region Column consists of country values which have been officially dissolved:

* West Germany
* Soviet Union
* Federal Republic of Yugoslavia



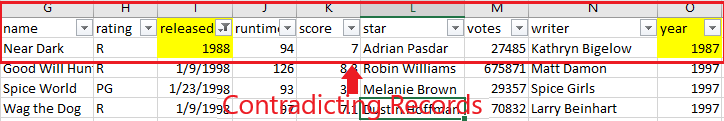
Steps taken for cleaning

Country Field – West Germany has now been replaced with Germany Country Field – Rows consisting Soviet Union and Republic of Yugoslavia has now been removed.



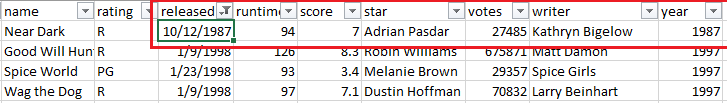
1. Contradicting Records

Column ‘Released date’ and column ‘Year’ have different values for the same entity i.e. Movie Name



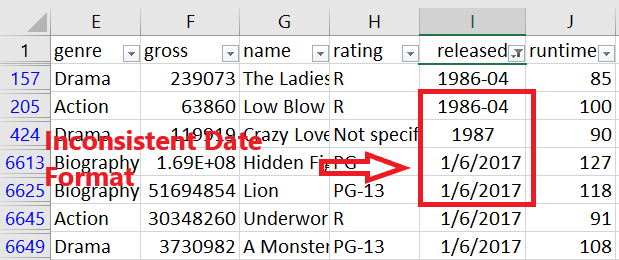
Steps taken for cleaning

Values in the column Released Date corrected by referring to IMDB Metadata



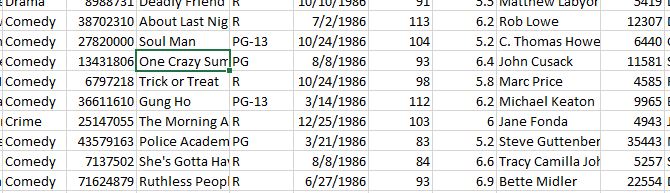
1. Inconsistent Date Format

Column ‘Release Date’ consist of inconsistent date formats



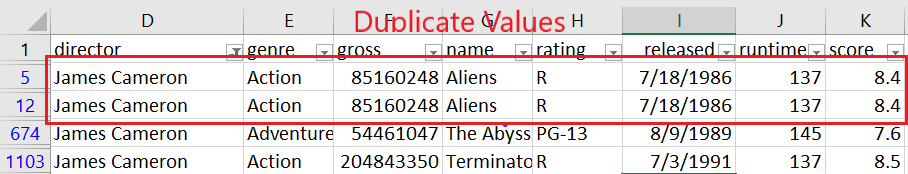
Steps taken for cleaning

Date Formats have been changed and made consistent throughout the datasheet.



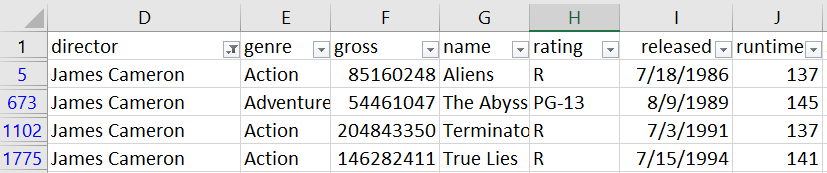
1. Duplicate Records

Dataset consists of few duplicate values for example: Movie details for the Movie Alien occurs more than one in the dataset



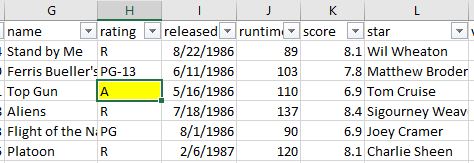
Steps taken for cleaning

Rows with duplicate values are discarded after cleaning



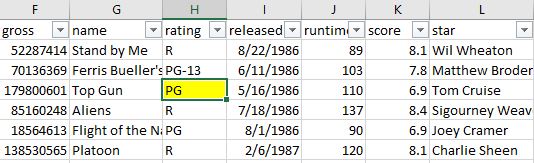
1. Referential Integrity Violation

Column ‘Rating’ consists of an undefined rating value ‘A’. Movie Rating are of the specific type: G, NC17, PG, PG-13, R.

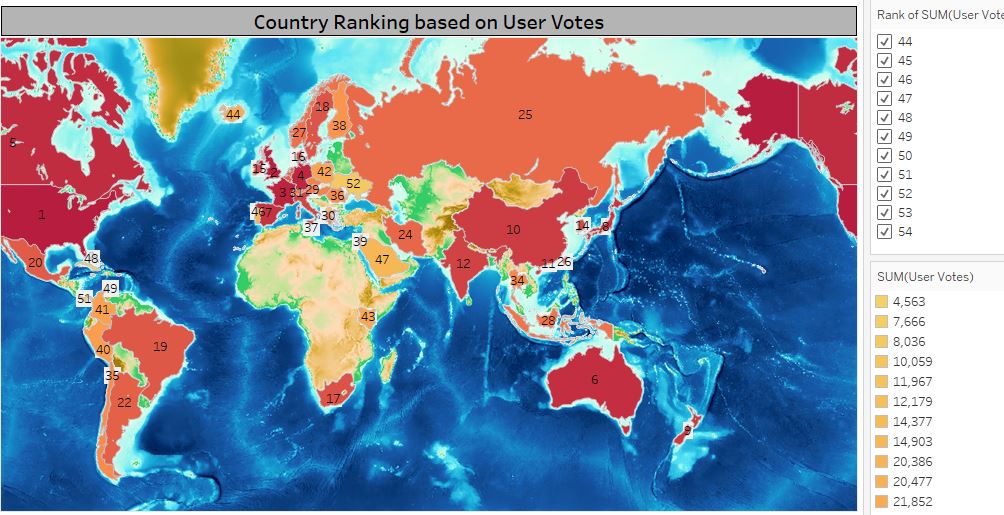


Steps taken for cleaning

With reference to the IMDB Repository dataset, the correct rating for the movies were identified and replaced in the datasheet.



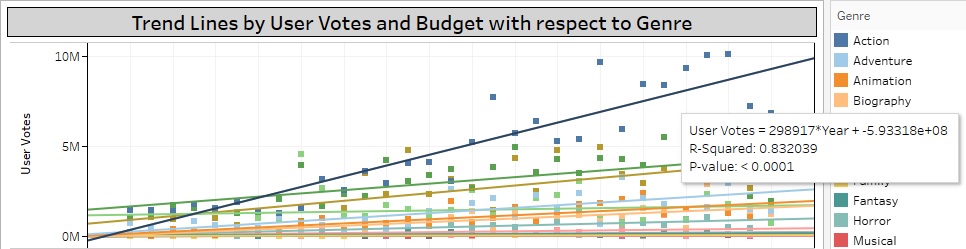
1. **DATA VISUALIZATION**
2. **Country Rankings based on Number of User Votes**

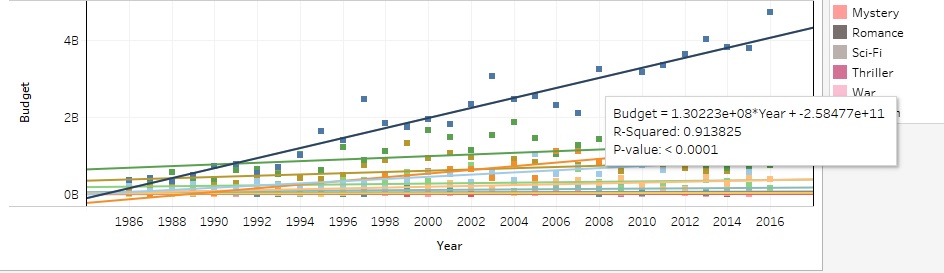


[Application Used: Geographic Maps & Rank]

Visualizations depicts the rankings of the countries based on the number of User Votes received for various types of movies produced in each country as well as the intense concentrations of the number of users on a global scale. This is demonstrated with the help of intensity of color code given to each country as well as shows the rank in numbers annotated for each country. We can see that United States of America holds the first position and United Kingdom the 2nd position. This visualization helps us understand what concentrations of active movie audience are present globally and what could be the factors for the major concentrations of audience from a country, is it the large-scale movie industry from a country, is it the strong social marketing strategy applied by the movie industry to engage the audience. The visualization also gives us a hint in the level of advancement in technologies for a country used by the movie industry.

1. **Change in the number of User Votes with respect to Genre Type and the Budget allocation to each these Genre type over 3 decades 1986 – 2016.**

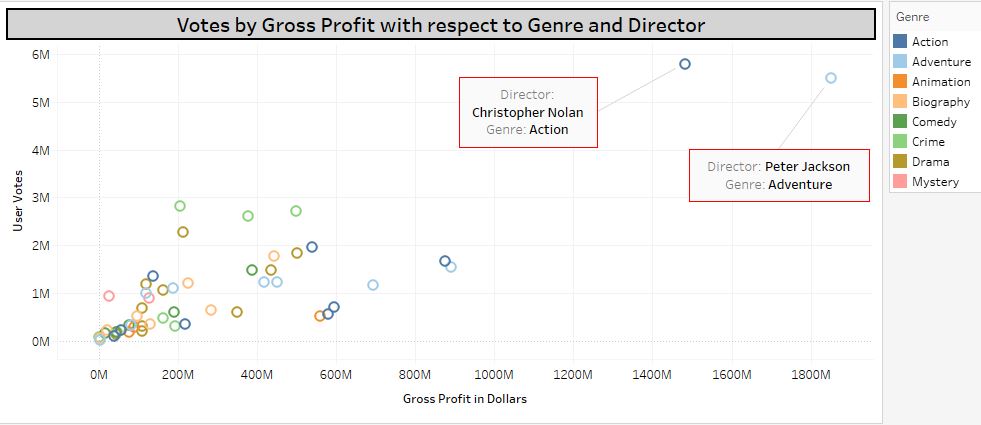




[Application Used: Trend Lines, Marks Card]

Visualization shows the continuation of the trend over 3 decades for variables User Votes and the Budget invested in the production of a movie and helps us anticipate the future market. We observe the correlation between the User Votes over the years with respect to Genre types and can observe the data trends using lines of best fit based on statistical model. For the upper visualization we can evaluate based on the best fit line where R squared value of 0.83 for genre type Action has received highest number of User Votes continuously for 3 decades and one can anticipate that theme-based movies on Action would receive large number of User Votes in the future as well. This would help the movie industry focus more based on the audience likings and can continue making Action based movies. Similar for the lower visualization we observe that the best fit line with R squared value of 0.91 which indicates based on the history of fact that budget invested in making the movies has been continuously increasing again for the Action genre type and hence we can assume that the movie industry is on the positive path as they seem to be considering the User Votes on the genre type Action and investing more on the same theme.

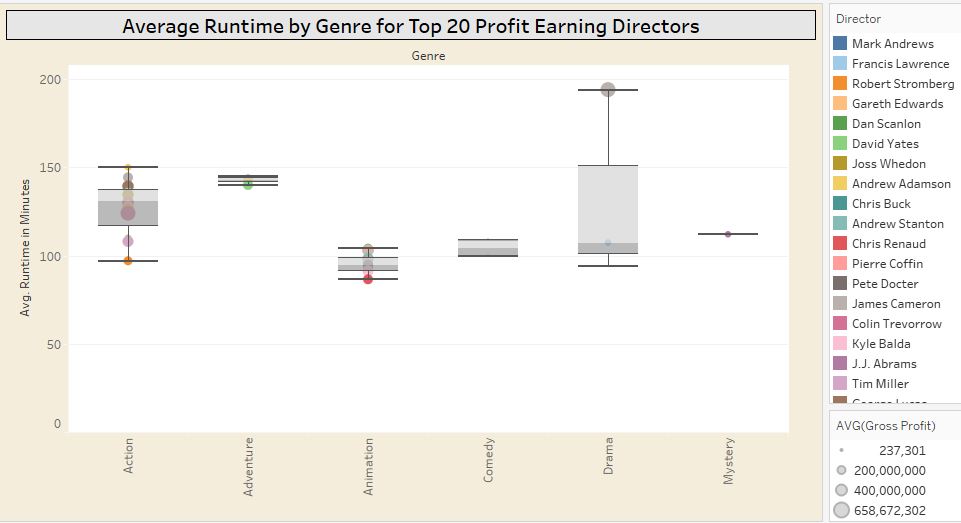
1. **Type of Genre and Director that received the Highest Number of User Votes and High Gross Profit**



[Application Used: Scatter Plot, Filters, Marks Card]

Visualization shows the positive dependency of the 2 variables i.e. User Votes and Gross Profit on each other which depicts a strong correlation between the 2 variables. Increase in the number of User Votes does bring a growth in the Gross Profit earned by the Movies. Dependency is evaluated with respect to the type of Genre for example action, adventure, animation and more and Director of the Movie. This visualization helps us understand that the Genre type Action and Adventure has the highest number of user likings as well as depicts high earned Gross profit for movie directors Peter Jackson and Christopher Nolan. Based on the popularity of genre type and director, audience can choose to watch more movies made on the same genre type and by the same directors. The movie industry can judge based on the above visualization which types of movies and under which director should the movies be produced in the future to earn high profits.

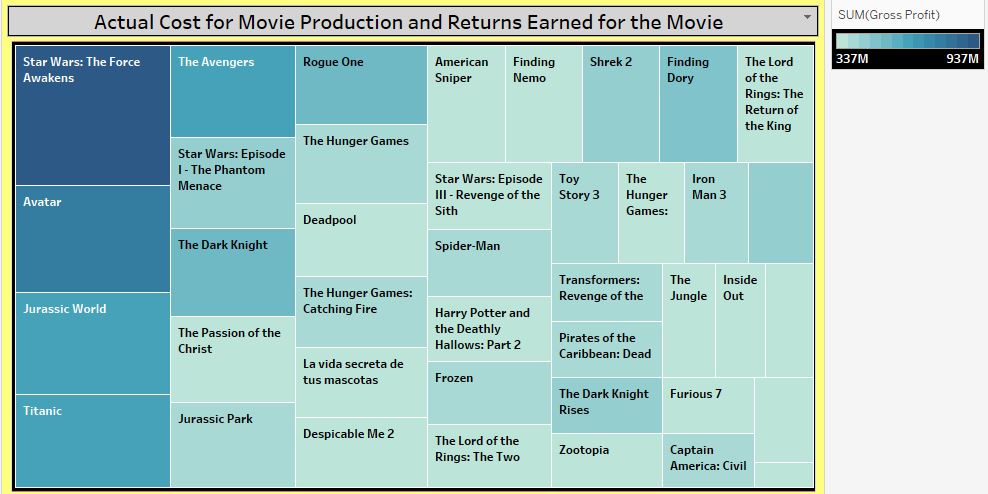
1. **Average Runtime of Movies by Genre for Top 20 Profit Earning Director**



[Application Used: Box and Whisker Plot & Filter]

Visualization illustrates the Average Runtime in minutes for the top 20 directors who have the maximum average profits with respect to the Genre types of the movies they directed. The box plots show the distribution of the average runtimes for the movies produced by top 20 profit earning directors with respect to the Genre type and we observe based on the box plots that the genre type Drama and Action for high gross earning directors have the longer duration of runtime ranging from 100 minutes to 200 minutes. Quartiles 2 and 3 within the box for Genre type Drama have the maximum number of movie runtimes ranging from approximate 100 to 150 minutes with the maximum duration of a movie being of 194 minutes directed by James Cameron who also has maximum earned profits. For the genre type Action maximum number of movies are made within the duration ranging from approximate 117 to 137. This helps us understand duration of runtime for each genre type as well as demonstrated how the successful directors allocate the duration for the theme of their movies. Are the successful directors allocating the right amount of time for their movies which helps keeps the audience engaged?

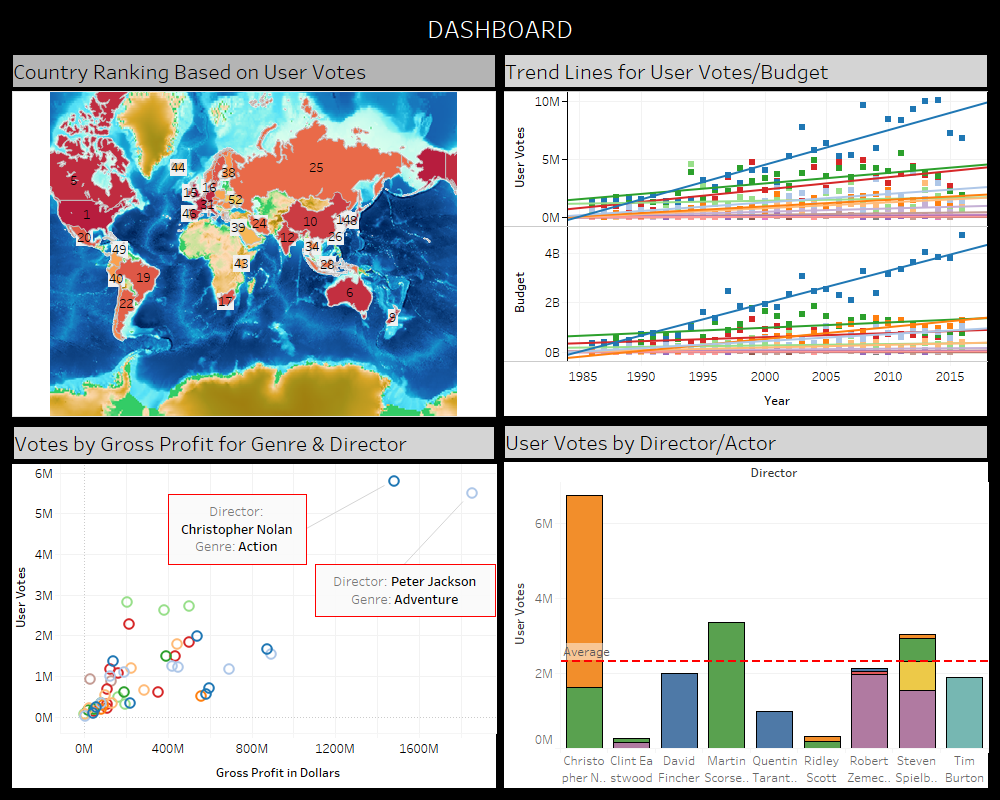
1. **Actual Cost Allocated for the Movie and the returns gained on the Movie**



[Application Used: Tree Map & Calculated Fields & Filters]

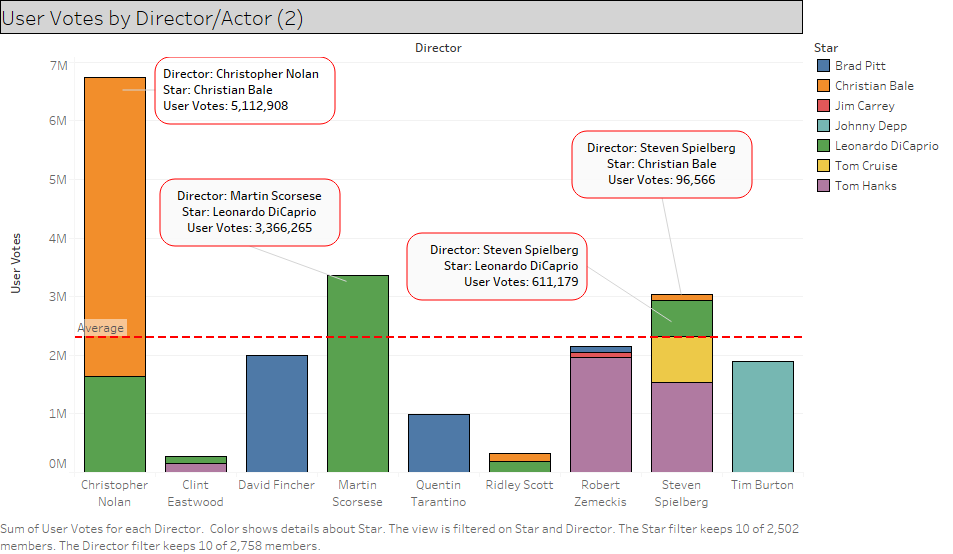
Visualization shows structure of the Tree Map based on size of Calculated Fields derived from evaluating the actual Cost utilized in producing a movie (Budget – Gross Profit). The larger size of the rectangle depicts these calculated field values, larger the size larger are the actual cost allocated in making a movie. Here we observe the actual cost allocated for the movies in the descending order 1. Star Wars 2. Avatar 3. Jurassic World 4. Titanic are the largest in the movie’s dataset. But have movies gained high returns? This can be visualized with the help of the intensity of the color code. Higher the intensity of the color code higher are the profits made by the movie. Hence, we can see that the movies stated above have achieved high returns of profit amongst all in the dataset. This helps us understand the relationship between the high amount allocated in making a movie with the high profits earned for such high cost movies. There might be several factors impacting the high profits i.e. advanced technology utilized in making a film, high talent stars involved which incurs such high cost.

1. **DASHBOARD**

****

Dashboard shows the collection of selected views of our data analysis worksheets all in one place and lets us compare and relate each analysis from the view. We have identified appropriate analysis worksheet to build a story and identify what the key incites of our data are, as given below:

* Country Ranking based on User Votes: We start with identifying the value of movie business industry on a global scale by observing the response received from the audience from each of the given country. In this way we could have an overall view of the growth of movie business industry which could help us with further analysis.
* Trend Lines for User Votes and Budget Allocation to different types of Movies: Secondly, we choose the analysis which demonstrated the growth of the Movie industry over 3 decades with respect to the genre and the trends of the budget allocation to the making of movies.
* User Votes by Gross Profit for Genre and Director: Thirdly we observe the likings of the audience based on movies that are made of different genre type. i.e. whether the audience likes action, adventure, comedy, war movies. Based on these likings we can observe the profit made for each of these genres which gives us an idea about the specific likings of the audience.
* User Votes by Director/Actor: Visualization as given below



[Application Used: Stacked Bar Chart, Reference Lines, Filters, Marks Card]

Visualization depicts the audience’s choice for their favorite stars featured by the top 20 gross profit earning directors. With help of the reference line which shows us the average User Votes for all the stars, we can analyze that the actors Leonardo DiCaprio and Christian Bale are highly voted and mostly featured by the successful directors Christopher Nolan, Steven Spielberg, Martin Scorsese who have high gross profit earnings till date. Based on the analyses, we can anticipate that these directors would most likely feature the stars that are mostly liked by the audience as well as the audience would definitely watch more movies featuring these stars and directed by the directors mentioned.

1. **STORY TELLING**



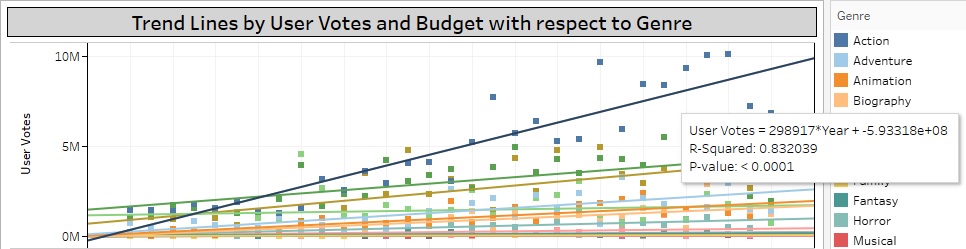
Movies also known as films, are a type of visual communication which uses moving pictures and sound to tell stories or inform. People in every part of the world watch movies as a type of entertainment, a way to have fun. For some people, fun movies can mean movies that make them laugh, while for others it can mean movies that make them cry or feel afraid. A motion picture proposal is imperative in our public activity because of its quality in giving improved diversion. Such a framework can recommend an arrangement of films to clients dependent on their advantage or the popularities of the motion pictures. Although, an arrangement of motion picture suggestion frameworks has been proposed, the vast majority of these either can't prescribe a film to the current clients proficiently or to another client using any and all means. Movies are most loved side interest all through America. There are a wide range of sorts of movies, and individuals lean toward various ones. Regardless of whether it is dramatization, satire, or tension, it appears they are for the most part similarly favored. Here the analysis done in this project gives the insight on four main research questions.

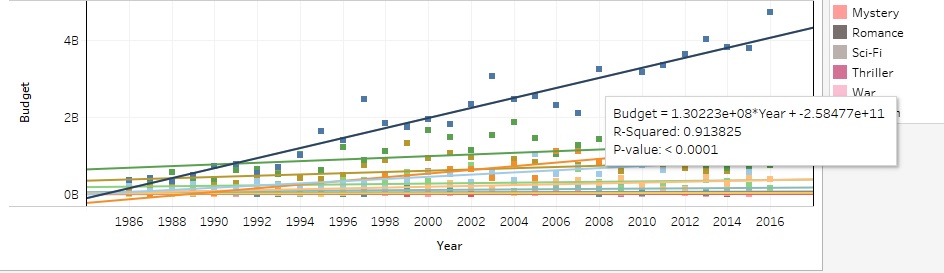
* **Concentration of User Votes by Country**

Report depicts the rank of the country based on the concentration of the User Votes. User Votes are the votes received by an individual movie originating from each of the country. Report shows the top ranks holders based on User Votes i.e. 1. United States 2. United Kingdom 3. France and countries are ranked up to 54. According to analysis we can determine that the top ranked countries have the highest number of movie audience and indicates the level of advancement in technology in movies industry and social networks used for creating movies which not only attracts the audience on a large scale but also given an opportunity to a user to convey his satisfaction level based on votes.

* **Impact of User Votes on the Budget allocated to the Movies based on Genre type over 3 Decades – 1986 – 2016**

Information present in the report depicts the change in the number of User Votes over time based on different Genre types. Trend lines represented depicts the continuous growth of User Votes for the Genre Type Action over 3 decades, this has been evaluated with help of best fit line which confirms positive growth over time. According to the analysis we can affirm that the likings of audience towards genre type Action is strong enough to create a greater number of votes, these evaluations would be helpful for the movie industry to create more such type of movies for each genre type. The report also demonstrates the fact that the budget invested on the Action theme-based movies have been increasing positively with every year. Analysis helps us determine the future investments on budget for the different genre types. The heavy investment on movies for genre type Action would be justified with the fact that the audience are liking Action based movies and would in the future receive an increased growth of Votes.





* **Popular Genre Type and Director along with highly earned profit for Genre type and director**

The information present in the report depicts the way user votes have an impact on the gross profit earned by movies for specific Genre types and known directors. Genre types represent theme on which the movies are based, it can be an action movie, comedy movie, mystery movie and more such types. Report shows that the genre type Adventure and Action are popular amongst the audience. Movies based on Adventure and Action Genre type have received not only high user votes but also earned high profits. The directors that have directed the movies based on these genre type have earned high profits. According to the analysis we can determine the themes that are liked by the audience and based on these likings the directors from the movie industry can choose to direct their movies and hence earn high profits in the movie business industry.

* **Average Runtime of Movies by Genre for Top 20 Profit Earning Director**

Report of the visualization depicts the distribution of runtimes of the movies of different genre along with the details of the successful directors who are involved in the movie making process. Based on the we can judge the movies of the genre type mystery are of an average runtime of 112 and very less successful directors actually make such type of movies.

**REFERENCES:**

1. Cambridge Journal of Regions, Economy and Society, 6(1), 141–157. “Hollywood in decline? US film and television producers beyond the era of fiscal.”

<http://cjres.oxfordjournals.org/content/6/1/141>

1. “Why Genre Is Important for Screenwriters and Filmmakers - 6 Reasons Why It's Important.” Raindance, 5 Apr. 2018, [www.raindance.org/genres-importance-screenwriters-filmmakers/](http://www.raindance.org/genres-importance-screenwriters-filmmakers/).
2. Kenny, Glenn, et al. “Movie Reviews.” *The New York Times*, The New York Times, 8 Nov. 2018, [www.nytimes.com/reviews/movies](http://www.nytimes.com/reviews/movies).
3. <https://www.imdb.com/interfaces/>
4. [https://help.imdb.com/article/contribution/titles/running-times/G8UMSUYEDCN3M2K8?ref\_=helpart\_nav\_26#](https://help.imdb.com/article/contribution/titles/running-times/G8UMSUYEDCN3M2K8?ref_=helpart_nav_26)
5. <https://milnepublishing.geneseo.edu/exploring-movie-construction-and-production/chapter/2-what-is-genre-and-how-is-it-determined/>