**Analysis on Movies**

**California State University Los Angeles**

**By**

**Divya Pakhale**

**Table of Contents**

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

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

**Data Description -------------------------------------------------------------------------------------- 3**

**Physical properties of the data** **--------------------------------------------------------------------- 4**

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

**Relationship between the tables--------------------------------------------------------------------- 9**

**Measure / Formula ------------------------------------------------------------------------------------ 10**

**Visuals---------------------------------------------------------------------------------------------------- 11**

**Dashboard---------------------------------------------------------------------------------------------- 20**

**Storytelling--------------------------------------------------------------------------------------------- 20**

**Purpose Map------------------------------------------------------------------------------------------- 23**

**References---------------------------------------------------------------------------------------------- 25**

**DATASET:**

<https://www.kaggle.com/danielgrijalvas/movies>

**WEBLINK:**

<https://www.boxofficemojo.com/movies/alphabetical.htm?letter=L&page=3&p=.htm>

<https://www.boxofficemojo.com/alltime/world/>

<https://www.imdb.com/chart/top>

**DATA DESCRIPTION:**

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. For a few people, fun movies can mean films that make them laugh, while for other people, it can mean movies that make them cry, or feel afraid. 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. You can likewise download or stream motion pictures. More seasoned motion pictures are appeared on TV broadcasting stations. Movies can be [fictional](https://simple.wikipedia.org/wiki/Fiction), or true, or a mix of the two. Most movies lose money but some make profits in the hundreds of millions, be they dollars, euro or pounds.  The industry has always been dominated by a few major movie studios like MGM/UA, Warner Bros, Columbia, Lucasfilm, Paramount or Disney. By doing analysis on movies, I was able to figure out which company made the highest budget, which movie had highest number of votes in that particular year.

The movie dataset is extracted from Kaggle. The dataset size is about 953.22 KB. It has 6820 rows and 15 columns.

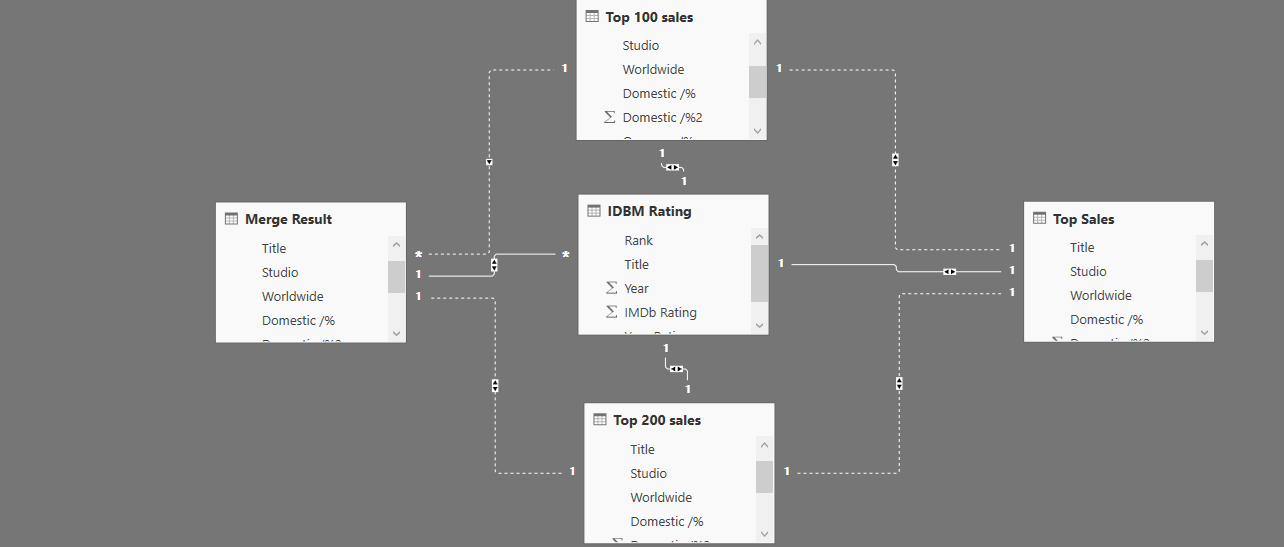
**PHYSICAL PROPERTIES OF THE DATA:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Field Name | Type | Sample Value | Range of values | Attributes/Format | Comments |
| Budget | Quantitative | $8000000 | $0 - $300000000 | Whole Numbers | Some values missing |
| Company | Textual | Studio Canal | Different types of Studios | Max 50 characters | Some values  missing |
| Country | Textual | USA | Countries all over the world | Max 30 characters | Some countries are with the same name. |
| Director | Textual | François Ozon | 50 unique directors | Max 30 characters | Most of the movies have same directors |
| Genre | Categorical  (Nominal) | Comedy | Types of Genre | Max 15 characters | Different types of Genre |
| Gross | Quantitative | $41715860 | $0-$50000000 | Whole Numbers | Some values missing |
| Movie Title | Textual | Aliens | 3000 unique movies | Max 111 characters | Some movies have the same name |
| Rating | Textual | PG-13 | 13 Unique ratings | Max 10 characters | Ratings based on parental guidance |
| Released Date | Textual | July 18 2005 | 22/8/1986 to 28/7/2017 | mm-dd-yyyy | Some dates are missing |
| Runtime | Quantitative (Ratio) | 92 | 50-366 | Whole Numbers | records the duration in minutes of titles in the database. |
| Score | Quantitative (Ratio) | 6.3 | 0-10 | Whole Numbers | Based on how many like the movie |
| Star | Categorical  (Nominal) | Wil Wheaton | 60 Unique Actors | Max 21 characters | Could argue for textual type but the role here is categorical |
| Votes | Quantitative | 23717 | 0 - 2000000 | Whole Numbers | Based on how many like the movie |
| Writer | Categorical  (Nominal) | Nia Vardalos | 60 Unique Writers | Max 30 characters | Could argue for textual type but the role here is categorical |

**DATA CLEANING:**

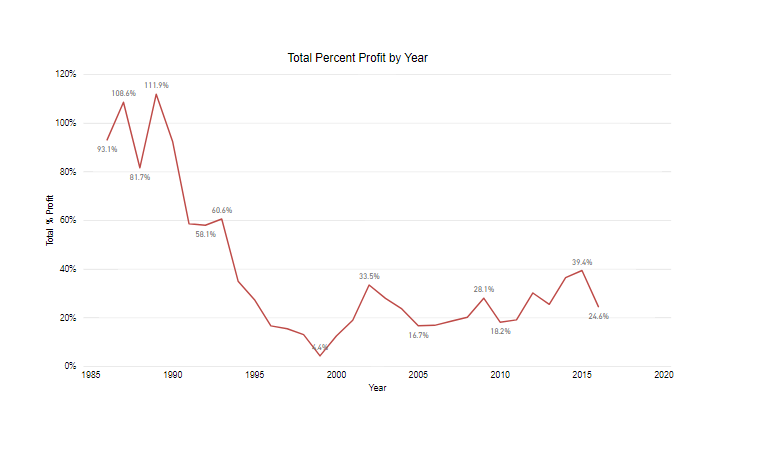
|  |  |
| --- | --- |
| **Problem** | **Un-cleaned and Cleaned Data** |
| Missing values | There were “n/a” values for Score and Gross Column. They were replaced with the help of replace function |
| Split title and year column | Rank and Title columns are separated with the help of split column by delimiter |
| Leading and trailing | This column has some extra spaces at the beginning and end of the values, so these spaces are removed with the help of Transpose using trim function |
| Erroneous values | Year column has extra closing brackets which was removed with the help of replace function |
| Inconsistencies | Column names were repeated and hence it was removed by “remove top row function” |
| Date format | Date format was text and it was changed to whole numbers. Year column had special character “^” hence it was removed with the help of transpose and replace function |
| Unnecessary columns | Unnecessary columns were removed. |

**RELATIONSHIP BETWEEN TABLES:**



Relationships between the tables are necessary in order to accurately calculate results and display the correct information in the reports. Power BI Desktop makes creating those relationships easy. I have combined 3 web datasets and 1 csv dataset for my visualization. Rank is the primary key for the IDBM rating dataset set and Studio is the foreign key.

**MEASURE / FORMUAL:**

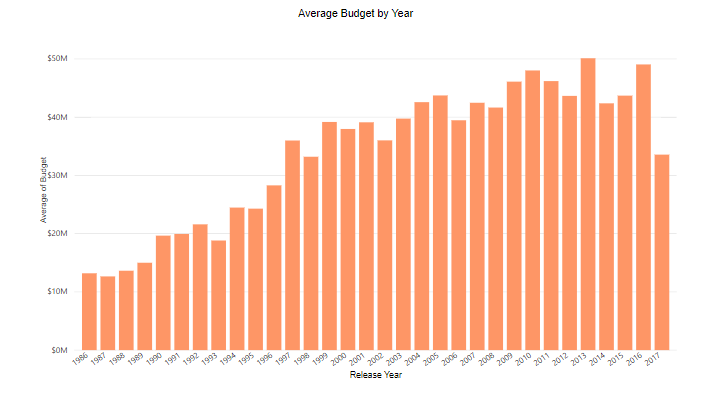


Formula used: Total % Profit = DIVIDE ([Total Profit], [Total Budget],0)

For this visualization, I have used measure formula to calculate the total percentage by year. A line graph is commonly used to display change over time as a series of data points connected by straight line segments on two axes. The line graph therefore helps to determine the relationship between two sets of values, with one data set always being dependent on the other set. The above line chart depicts about the total percent profit by year. As per the chart, we can make out that total percent profit was highest in the year 1989 and lowest in 1999 which was about 4.4%. From this, we can conclude that profit rate was reduced.

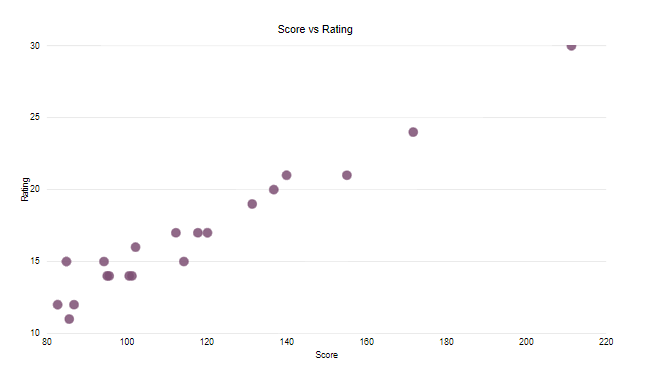
**VISUALS:**

1. AVERAGE BUDGET BY YEAR



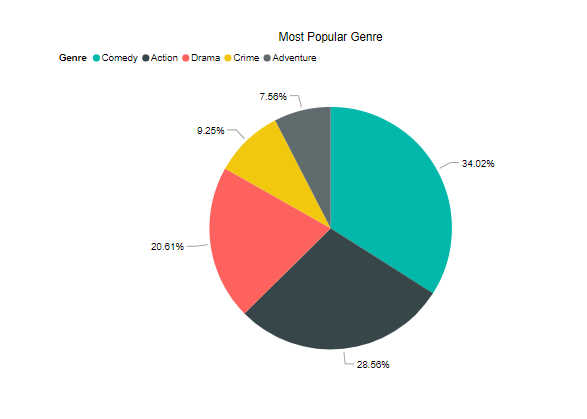
The line and stacked column chart describe about how the budget of the movie changes every year. From this visualization, we can see that average budget in the year 2013 was around $50M and in 2016 it was close to $45.5M. Average budget started increasing from the year 2000 onwards. This could be because of the fact that people might have more free time to watch movies.

2. SCORE vs RATING FOR TOP 20 DIRECTORS



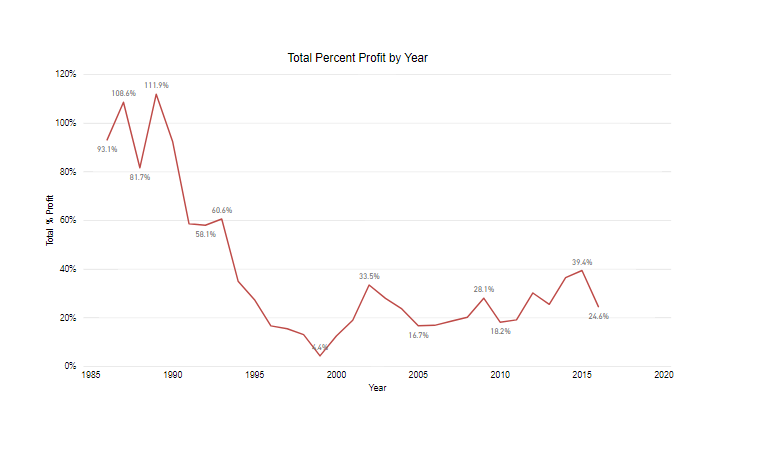
From this scatter plot, we can examine how score and ratings are related. A film score (also sometimes called background score, background music, movie sound track, film music or incidental music) is original music written specifically to accompany a film and ratings are a film's suitability for certain audiences based on its content. So, from this visualization we can make out that for the score of 170 rating is 24. This visualization is based on top 20 directors and so we can make out from this which director had more ratings and score.

3. MOST POPULAR GENRE



The Pie Chart reveals information about preference given to 5 different categories (genres) of movies. As it can be clearly seen that maximum preference has been given to comedy movie genre/ type, which accounts for 34.02%. Followed by quarter percentage of those who like to watch action movies. Further followed by percentage of people who prefer watching drama movies corresponding to 20.61%. On the contrary, least preference was given to crime and adventure category accounting only for 9.25% and 7.56%. Overall, it can be concluded that comedy movies made a lead among all others whereas, smallest slice was shared by that of the crime movies category.

4. TOTAL PERCENT PROFIT



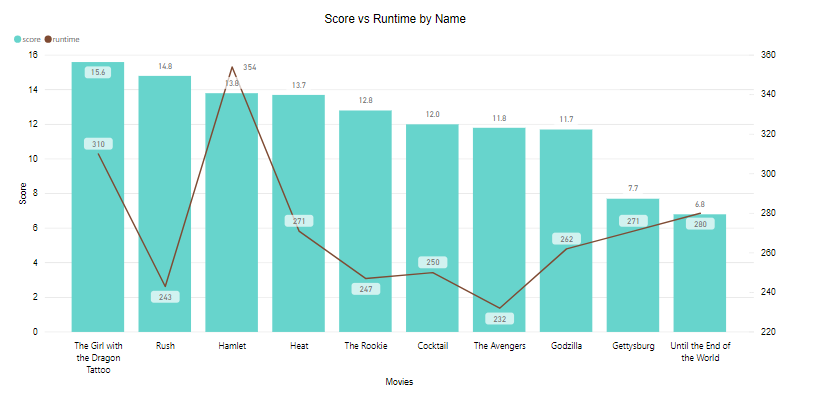
A line graph is commonly used to display change over time as a series of data points connected by straight line segments on two axes. The line graph therefore helps to determine the relationship between two sets of values, with one data set always being dependent on the other set. The above line chart depicts about the total percent profit by year. As per the chart, we can make out that total percent profit was highest in the year 1989 which accounts for about 111.9% and lowest in 1999 which was about 4.4%. From this visualization, we can conclude that profit rate keeps on varying per year depending on the factors such as audience review, budget etc.

5. TOTAL VOTES BY DIRECTOR NAME



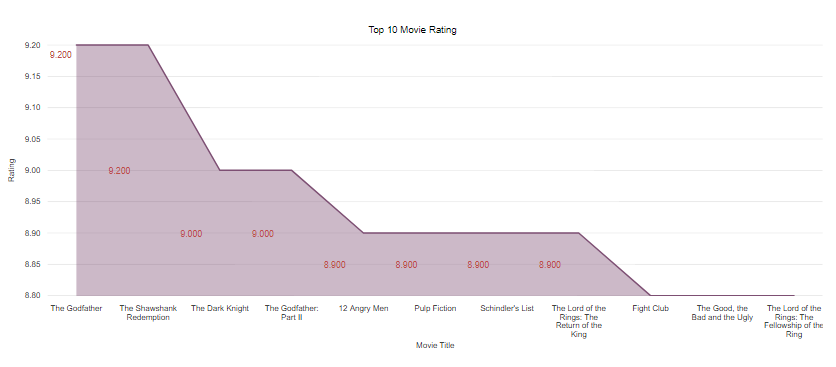
The tree map reveals information about votes for the movie for top 5 directors. As per the visualization, we can make out that maximum votes of 499 was for the movie Straight out of the Brooklyn which was directed by Matty Rich. Next maximum votes of 489 was for the movie Armed Response directed by Fred Olen Ray. Least votes of 183 was for the movie Sky Bandits directed by Zoran Perisic. From all these, we can conclude that director Matty Rich was most famous director as compared to other directors.

6. SCORE vs RUNTIME



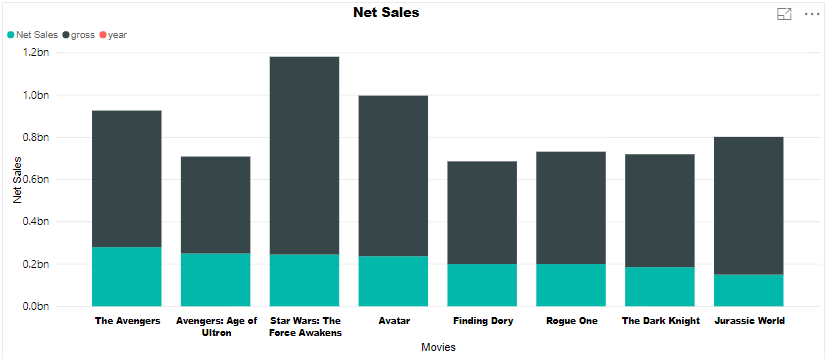
The above visualization is of line and clustered column chart which is describing about the relationship between score and runtime for the top 10 movies. In this visualization, I have merged two graphs line graph and bar graph. The line graph depicts about the runtime for the movie and the bar graph is describing about score. As we can see, the runtime for the movie hamlet is highest among others and score is highest for the girl with the dragon tattoo. The least runtime is for the movie Avengers and least score is for the movie named as Until the end of the world. The runtime is expressed in minutes. Concluding the above visualization, we can make out that the movie with longest runtime has maximum length of duration and the one with highest score has very good background music, film sound track, that is original music written specifically to accompany a film.

7. TOP 10 MOVIE RATINGS



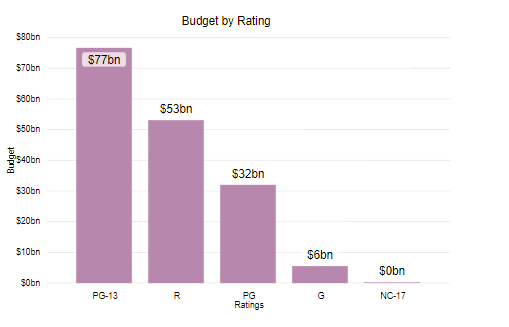
The stacked area chart reveals the information about top 10 IDBM movie ratings. It can be clearly seen that maximum rating of 9.2 was for the movie the godfather. Next preference was given for the dark knight and the godfather part2 which accounts for 9.0. Least rating movies were the fight club, the good and the bad and the ugly. From these ratings viewers can decide which movie are good and which are not.

8. NET SALES



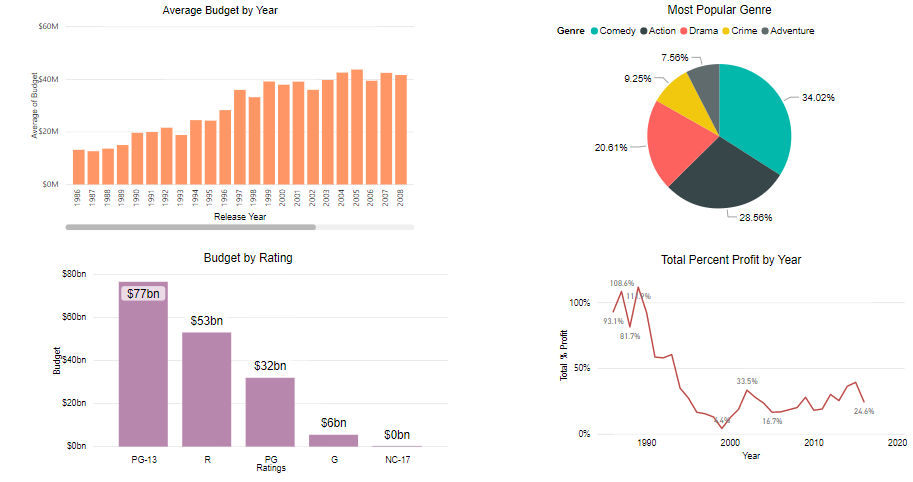
The above visualization is describing about the net sales and gross for top 10 movies. In this visualization, I have used measure formula to calculate the net sales with the help of using budget column. As we can see that gross amount for the movie Star War was the highest among others but the net sales were less. In contrary to this, the net sales were highest for the movie avengers but the gross amount was less. This visualization will help directors as well as movie makers to decide about the profit or loss for the movie.

9. Budget by Rating



Movie ratings are based on parental guidance. The highest budget and ratings were given to PG – 13 section. It means most of the movies were under parental strongly cautioned. PG-13 material may be inappropriate for children under 13. Parents are urged to be cautious. Some material may be inappropriate for pre-teenagers. Next preference was given to R section followed by PG. R rating movies are restricted movies under 17 requires accompanying parent or adult guardian. Contains some adult material. Parents are urged to learn more about the film before taking their young children with them. G rating movies were for General audience all ages admitted. Nothing that would offend parents for viewing by children.

**DASHBOARD:**



**STORYTELLING:**

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. In spite of the fact that, 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 as though they are for the most part similarly favored. The three best films of the year were the satire, Meet the Parents, the dramatization, Ghost, and the thrilling, Final Destination. Here the analysis done in this project gives the insight on four main research questions.

1. Most popular Genre

Genre is simply categorized due to the same or similar conventions being used to portray a particular genre over and over again. However, this hasn’t meant that these conventions have stayed the same since film began. At the start of film production, the first ever genre was “Western”, followed by the “musical” genre. Now today, there are many different genres of movies available to watch. Genre is important for producers because they need to make sure that they engage their target audience. To do this the producers need to know exactly what kind of film they aim to produce in order to use actors, filming techniques, etc. to make sure they attract the audience they want to. For audience genre is important to inform them of what the film will be like. The audience will be attracted to films based on genres and most iconic features from that genre. From the analysis above, overall, it can be concluded that comedy movies made a lead among all others whereas, smallest slice was shared by that of the crime movies category.

2. Average budget by Year

Film budgeting refers to the process by which a line maker, unit creation chief, or generation bookkeeper readies a financial plan for a film creation. This record, which could be more than 134 pages in length, is utilized to anchor financing for and prompt pre-generation and creation of the film. Numerous drafts of the financial backing might be required to whittle down expenses. A financial plan is regularly separated into four segments: over the line (innovative ability), beneath the line (coordinate generation costs), after creation (altering, visual impacts, and so on.), and other (protection, fruition security, and so forth.). Movie budget keeps on varying over the years. From the above visualization, we can see that average budget in the year 2013 was around $50M and in 2016 it was close to $45.5M. Average budget started increasing from the year 2000 onwards. This could be because of the fact that people might have more free time to watch movies.

3. Budget by Rating

Most guardians are worried about the things their kids see and hear in the media. It is an appalling world out there, and keeping in mind that it might be unavoidable that they will in the end find out about those things, most guardians don't need their children to see realistic savagery or develop content or hear hostile dialect at too young age. One of the most seasoned and most natural rating systems is the one made by the MPAA in 1968. The MPAA's ratings arms is known as the Classification and Ratings Administration (CARA). CARA is comprised of a gathering of unknown leaders, with fluctuating family and expert foundations, that observe each motion picture and make an assurance of the ages for which they feel it is generally proper. They at that point allocate a rating going from G to NC-17 to the motion picture to help gathering of people individuals decide if the film's substance is appropriate for them and their families. According to the analysis, the highest budget and ratings were given to PG – 13 section and least was given to NC-17 section.

4. Total Percent Profit by Year

The worldwide film industry demonstrates solid projections for the coming years, as the worldwide film industry income is figure to increment from around 38 billion U.S. dollars in 2016 to almost 50 billion U.S. dollars in 2020. The U.S. is the third biggest film showcase on the planet regarding tickets sold every year, just behind China and India. Just shy of 1.2 billion motion picture tickets were sold in the U.S. in 2016. There are around 5,800 film destinations in the U.S. starting at 2016. As per an ongoing study, 13 percent of Americans head out to the motion pictures about once per month, seven percent go see films in the motion picture theater a few times each month, though 31 percent go not exactly once every year. This is a significant offer considering the 52 percent of American grown-ups who favor watching motion pictures at home.

**PURPOSE MAP**



1. EXPLANATORY

Explanatory as visualizers provides the viewer with a visual portrayal of the subject’s data and will also take some responsibility to bring key insights to the surface, rather than leave the prospect of interpreting the meaning of the information entirely to the viewer. Figure 1 comes under explanatory because it is easily understandable by the user. From the visualization, viewers can make out which movie genre is most popular. Overall, it can be concluded that comedy movies made a lead among all others whereas, smallest slice was shared by that of the crime movies category. The visualizer here is attempting to assist the viewers process of understanding as much as possible, in particular with the interpretation, drawing out the meaning of the data.

2. EXPLORATORY

Exploratory visualizations differ from explanatory visualizations in that they are focused more on helping the viewer or more specifically in this case the user find their own insights. Figure 2 comes under exploratory because the user first needs to understand what the score and runtime mean, what is the relation between the two terms. Basically, the user needs to explore by himself about the related terms and then try to understand what the visualization is. This project is simply a visual window into the analysis of this data that lets users perceive the data values and interact with the different dimensions offered.

3. EXHIBITORY

Exhibitory are neither explanatory nor exploratory. With exhibitory visualizations the viewers have to do the work to interpret meaning, relying on their own capacity to make sense of the display of the data and the context of the subject matter. Figure 3 is exhibitory because the viewer must be able to understand the content of a display as well as the context of the display. Exhibitory rely entirely on and make assumptions about the capacity of and interest among the target audience. From the visualization, firstly user should know on what basis the ratings are categorized and what each type mean. Once the user is able to understand all the concepts about the ratings and how they are classified then it would be very easy for the viewers to interpret the visualization. This helps the user to draw their own conclusion after understanding the visual.

4. READING

At the top of a purpose map the tone of the visualization design choices is geared towards optimizing the ease with which viewers can accurately estimate the magnitude of and the relationship between the values. The reading tone is the best fit approach when the purpose of work requires to facilitate understanding with a high degree of precision and detail. In this scenario there is no need to seduce an audience through aesthetic treatment. Figure 4 comes under reading because the bar chart is easily readable by the audience and also, they are familiar with the terms like gross value and net sales. Hence the visualization is very easy for the viewers to interpret the meaning.

5. FEELING

Feeling tone gives more emphasis on determining the gist of the big, medium and small values and a general sense of the relationships that exist. Sometimes an ‘at-a-glance’, high level view is the most suitable way to portray a subject’s value. Figure 5 comes under feeling tone because it cannot remotely estimate the relative proportions, but can estimate the gist of the scales involved.

**REFERENCES:**

1. *Hg.org*, [www.hg.org/legal-articles/how-do-movies-tv-shows-music-and-video-games-get-rated-for-content-35246](http://www.hg.org/legal-articles/how-do-movies-tv-shows-music-and-video-games-get-rated-for-content-35246).

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. “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/).

4. “Metacritic Reports: Movies.” *Metacritic*, Livemixtapes.com, [www.metacritic.com/feature/movie](http://www.metacritic.com/feature/movie).