**Data Visualization on Movies Datasets**

ISM 6419.001S22.23777

**Vivek Gollapudi**

U68784650

**Introduction:**

Movies are something that we all love, and they have become an unavoidable part of life for many people more often than not. Personally, for me movies are the go-to thing when I find some time or want myself to be stress free. They are a very strong medium of entertainment and the reach that movies bring to any content is just unimaginable. It even crosses borders like language, region, religion, and creed. We are attached to this movie culture since ages. May be the way of entertaining the audience but it all boils down to story telling at the end of the day. Be it the drama, plays or movies, all of them do the same thing in different ways through different medium. Considering movies from Revenue perspective, they are a vey huge prospect and the largest grossing movie Avatar has a lifetime gross of $2,847,379,794. This is nearly 0.1% of hugely populated country like India. With this stated stat, we can imagine how big a market the movies are.

There are several databases where we can find the information, or the data related to the movies. Internet Movie Database which also popularly known as IMDb is the most prominent one among them. Apart from movies or films, IMDb contains data related to television series, home videos, video games and online streaming content. It also contains tons of data related to the cast and crew for all the above-mentioned categories. The very famous IMDb rating is nothing, but the rating given by the registered users on a scale of 1 to 10. This rating is the basis for many people in deciding to watch a movie on online streaming platforms or in other words many people check the IMDb rating for a movie before they decide on whether to watch it or not on the streaming platforms.

Another very famous database for movies is Rotten Tomatoes, an American review aggregation website for movies and Television related content. One interesting fact about Rotten Tomatoes is that it was started in 1998 by 3 undergraduate students at the University of California, Berkeley. This database has a very famous score called Tomatometer score which represents the percentage of professional critic reviews that are given positive for a given film or television show. People generally consider a movie, or a TV show to be good if it has a Tomatometer rating of above 75%.

**Research Questions:**

The research questions that are discussed in this report are below:

1. Average IMDb rating and Audience rating on Rotten Tomatoes for movies based on the country in which they are produced.
2. Target audience for movies available on different online streaming platforms namely Disney, Hulu, Netflix, and Amazon’s Prime Video.
3. Several comparisons like Average IMDb rating, Average Tomatometer rating, Average Audience Rating and Average Runtime for movies based on different languages namely English, French, Hindi, Japanese and Russian.

**The ambitiousness of the Project:**

The visualizations in this report are mainly focused on bringing out the hidden details of movies with respect to different factors such as Language, Country, and Streaming Platform.

One major ambitiousness of this project would be the target audience movies that are being streamed on different Online streaming platforms. We can dig more into this if we have more access to the data. We have enough data to pull insights, but more analysis can be done, and much more can be learnt by scraping more and more data about the movies being streamed on Online Streaming Platforms. It even helps the business model to be much precise and clear.

**Methodology**

There are different datasets used for this project. They are a combination of data that is pulled from several official website as well as the web scraped data.

1. *MoviesWithIMDbRating.csv*

This table/Dataset contains columns namely Movie Title, IMDb rating for the movie, Runtime, Primary Language, Year in which the movies has been released, Genres that are associated to the movie, Directors’ details, Country in which the movies has been produced majorly, Available Languages details, and the age group that the movie is made for.

Data Source: <https://datasets.imdbws.com>

1. *AvailabilityOnStreamingPlatforms.csv*

This table/Dataset contains columns like Title of the movies along with One hot encoded value for availability of the movie on various streaming platforms namely Disney+, Hulu, Netflix, and Amazon’s Prime Video.

Data Source: (IMDb Scraped Dataset) <https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset>

<https://www.primevideo.com/>

<https://www.netflix.com/>

<https://www.hulu.com>

<https://www.disneyplus.com>

1. *Rotten\_tomatoes\_critic\_reviews.csv*

This table/Dataset contains columns like Critic Name, Publisher Name, whether he is a top critic or not, Review Date, Review Content, Review score and Review Type along with the Movie Title.

Data Source: <https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset?select=rotten_tomatoes_critic_reviews.csv>

1. Rotten\_tomatoes\_movies.csv

This table/Dataset contains columns like Movie Title, Movie Info, Genre, Director, Content Rating, Actors, Audience Rating, Tomatometer rating, Audience Count and many more details about the movie. This is a huge dataset about the movie details.

Data Source: <https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset?select=rotten_tomatoes_movies.csv>

**Analysis**

Chart, line chart

Description automatically generated

The above visualization shows the movies count available on different online streaming platforms. I have filtered the data to only have visualization between 1970 and 2020 because prior to 1970 there was not much difference between them. One more point to note is that this shows the movies that are available on a particular platform and their release years and not that these platforms existed in 1970s. I had to combine the columns from two different tables (1&2) for this visualization.

The insight from this visualization is that Amazon’s Prime Video has movies that release in 2012-2013 in most number and the number certainly went drastically in 2019 and 2020 with the outset of the Covid-19 Pandemic. The movie industry was something that went through a lot of struggle during the unprecedented Pandemic time. Disney+ has the least number of movies among the selected ones and Prime Video being the highest holder.

**Note:** I have also added an Animation that shows the number of movies on Streaming platforms against the year in which they have released. This animation is also between the timeframe 1970-2020.

Chart, bar chart

Description automatically generated

This explains the Average IMDb rating for movies based on the Primary Language they are made in. There is not much insight observed from this visualization except that the values range from 5 to 8 odd change when the ratings are sorted based on the Language.

Chart, bar chart, waterfall chart

Description automatically generated

This visualization is focused on visualizing the English Language movies count that are available across different Online Streaming Platforms. There is not much insight apart from the part that Amazon’s Prime video has the highest number of English Movies followed by Netflix and Hulu, with Disney+ being the least contributor. This is also a combination of columns from two different datasets 1 and 2.

Chart, line chart

Description automatically generated

This visualization shows all the movies that are available on Rotten Tomatoes based on the year which they are released. This Visualization can be drilled down to see the trend based on the Quarter, month, week, and date of the release as well. The year range of the data is from 1915 to 2020. Once again, the Covid-19 pandemic effect is clearly observed with a dip in the number of movies off late. The peak (or) the maximum number was observed between 2010 and 2013. The curve has been upward since 1920s and the growth was significantly high after 1990s.

Graphical user interface

Description automatically generated

This compares the Tomatometer rating vs the Audience rating on the Rotten Tomatoes website based on the release year of the movie. The year range is from 1915 to 2019.The average Tomatometer rating used to be higher than the Average audience rating for the movies till 1970s and the trend has been reversed since then. Anyhow, since these are just average values, we can’t infer much from these trends.

Chart, bar chart, histogram

Description automatically generated

This visualization is for a particular critic named “Adam Sweeting” who is rated as a Top Critic on Rotten Tomatoes. This compares the ratings of the critic against the audience rating for the movies and we can clearly see that they are aligned with each other. This shows that the critic is in sync with audience mindset and that is exactly what made him a top critic on Rotten Tomatoes.

To explain the visualization a bit more, for all the movies which the critic has given a rating of 1 out of 5, the average audience rating is around 25 out of 100 and for all the movies which the critic has rated 5 out of 5, the average audience rating was around 77-78 which is acceptable. This way we could pull out data for any particular critic and visualize it. There are hundreds of critics on Rotten Tomatoes. This visualization is a combination of columns from datasets 3 and 4.

Chart, pie chart

Description automatically generated

This pie chart depicts the percentage of movies that fall under each category when divided based on the age groups the movie is targeted for. This shows that most of the movies which is around 47.24% of them fall under 18+category. The least number of the movies are under 16+age group category. There are around 11.46% of movies which are made for all age groups and around 19.88% movies for age groups above 7.

The age groups have been color coded so that they can be recognized easily by the consumer when he/she looks at the pie chart.

This chart is based on 7354 movies in total which are collected from the IMDb database.

**Evidence1:**

Map

Description automatically generated

***Justification and Conclusion:***

This visualization helps us to identify the Average IMDb rating of a Country based on the films that were produced in that country. As we can see from the above image which says that the Average IMDb rating ranges from 3.0 to 8.30. As stated earlier in the report, the IMDb rating is on a scale of 10. Few insights from the visualization are below:

Average rating for United States: 5.745

India: 6.310

China: 5.931

Russia: 6.010

Japan: 6.551

France: 6.318

Map

Description automatically generated

***Justification and Conclusion:***

The above visualization depicts the Average Audience Rating of movies based on a country. These ratings are collected from Rotten Tomatoes site, and they are on a scale of 100 unlike IMDb. The Average Audience rating ranges from 30.0 to 87.0. For this visualization, I have combined columns from two different tables/datasets which is Average Audience rating from the “Rotten\_tomatoes\_movies.csv” dataset and the Country column from the “MoviesWithIMDbRating.csv” dataset. I had to filter out many null values and that is expected when we are combining data from two different data sources.

The Average Audience rating for some of the countries are as follows:

United States: 58.60

India: 63.58

China: 54.67

Russia: 72.25

Japan: 71.53

France: 55.44

The above 2 visualizations and their insights answer our first research question about the Average ratings for films made in the Country on two most popular databases or sources for information about movies namely IMDb and Rotten Tomatoes.

**Evidence2:**

Chart

Description automatically generated

***Justification and Conclusion:***

The above visualization is a very important one and this answers our second research question for this project.

For this visualization I had to combine columns from tables/datasets 1 and 2. We have 4 different sections of movies based on the age category they are associated with namely 7+, 13+, 16+, 18+ and all age groups.

The insights are below:

1. Amazon’s Prime Video has the highest number of movies for the age group of 18+ years of age.
2. Prime Video has the highest number of movies in each category/
3. Disney+ has least number of movies in the 18+ category and they have their highest count in all age groups category.
4. Netflix also has their maximum count in the 18+ age group category whereas they have them least in all age group category.
5. Hulu also follows the same trend as Netflix and Prime Video with highest proportion of their movies in the 18+ category but the count is less than both.
6. Hulu has a smaller number of movies that fall under all age groups category.

**Dashboard:Chart, bar chart

Description automatically generated**

***Justification and Conclusion:***

This dashboard answers and provides insights to our research question3 which is comparing several factors like average ratings and runtime based on top5 languages’ movies. This includes 4 different visualizations which combines data from three different datasets/tables.

***Average IMDb Rating:***

This avg IMDb rating is highest Japanese language (6.5335) movies among the selected 5 different languages. On the other hand, English movies has the lowest in the selected ones with average rating of 5.7916. This rating is on a scale out of 10.

***Average Tomatometer Rating:***

This rating is highest for Russian movies, which is 80.1. The least average rating among the selected languages for this rating is for English movies which stands at 60.6. This rating is on a scale out of 100.

***Average Audience Rating:***

The average audience rating is highest for Japanese movies which stands at 70.32 and least being with English movies, which is 58.27. This rating too is on a scale of 100. The average audience rating for Hindi movies stands at 67.15.

***Average Runtime:***

The interesting fact about this visualization is that the average run time for Hindi language movies is almost 1.5 times the average run time for English movies. The average run time for Hindi movies is 136.08 minutes while the same for English movies stands at 89.09 minutes. The Russian movies are a bit longer than English ones and shorter than the Hindi movies. Russian Language movies have an average run time of around 113.86 minutes.