DATA101 Final Project Report
Group 5

Luo, Monaliza
Ong, Andrei Lawrence
Penaflor, Neil Andrei
Sta. Cruz, Jose Miguel

---

*Problem Definition and Dataset*

**Target audience**

The present project aims to provide effective and informative data visualizations to three primary groups of people—movie analysts, film enthusiasts, and industry professionals. A common characteristic among these three demographics would be their interest in cinema or film, with the level of immersion being what sets them apart from each other.

We aim to produce a project beneficial to movie analysts by allowing them to identify trends in movie production—populer genres, themes, and audience preferences over time. This may help them in whatever purpose they might be analyzing movies for, may it be market research, comparative analyses, cultural and historical analyses, etc.

The Letterboxd community, which is the source of information and data found in the dataset, is primarily composed of film enthusiasts—our second primary target audience. We want film enthusiasts to use this present project to discover new films, compare their tastes with trends and overall audience preferences, etc. Furthermore, they may also use this project to dive deeper into their already-favorite genres, themes, or even production companies.

Lastly, we also aim to produce a project that would be of use for industry professionals. Similar to the case with movie analysts, we want professionals to be able to identify key trends and market gaps that could be seen by visualizing aspects of the dataset. Additionally, they would ideally be able to use the project as reference for how other movie producers/companies succeeded or failed—possibly seeing strategies based on trends identified.

**Visualization Problem**

People love movie nights, but sometimes a question like, "Have you had moments when you can’t decide which movies or what kind of movies to watch?” pops up and bothers some

people. These are moments when indecisiveness kicks in for movie watchers, which causes struggle and sometimes frustration to them. However, data visualization on movie trends can help movie analysts, film enthusiasts, and industry professionals altogether for film reviews, analysis, or decisions for simple movie nights with friends, families, loved ones, or just alone. Focusing on movie trends is not much of a big issue but an opportunity to visualize the ongoing trends in the film industry based on specific parameters. With numerous movies being released every year, movie watchers would have a difficult time finding the right movie for their interests. It would be a hassle for analysts to monitor observable trends especially with numerous sites giving different information regarding certain movies. Data visualization on movie trends aims to summarize the information or mainly the important and necessary data into a simple graph so that it may help movie watchers in their indecisiveness on which movie to pick.

Having plenty of the most recent movies than old movies not later than 1874 causes the uneven number of movies representing each year, which makes it difficult to produce a visualization on trends or changes over time unless a limitation on the time period being looked at is in place. On the bright side, it is possible to have an interactive map showing which country released the most high-rated movies or interactive bar graphs that indicate the rating from the chosen parameters such as but not limited to genre, themes, and studio, which can be adjusted. Based on the ratings of movies in the dataset, movie analysts may be able to predict which theme will be the next trend for the succeeding years. Film enthusiasts can possibly determine which genre of movies has produced good or bad films. Industry professionals may identify the next genre for their new films based on which genre gets the least-rated movies to utilize the expectations of the public and start a new trend in that specific genre.

**Dataset Description**

The Letterbox(Movie Dataset) by Simon Garanin will be taken from Kaggle. The dataset contains 10 different csv files for different features. Each file is unified by the ID column which shows the ID representing each movie from the site. From the movies.csv file which is the main file for reference, there are around 941,597 values. The data was taken from data retrieval features of Letterbox.com.

Out of the ten present datafiles, the following are selected for this project:

i. movies.csv will contain all basic information about each movie. This file will be the basis of identifying the movies corresponding to each ID for the next csv files. As mentioned above, there are 941,597 values. For this dataset, each row pertains to each movie present in the site. As shown in Figure 1, the columns that would be relevant for our analysis would be “id” for the ID reference for other files, “name” for the movie title, “date” for the year of release, and “rating” for the average ratings of each movie from the site.

| | id | name | date | tagline | description | minute | rating |
|---|---|---|---|---|---|---|---|
| 0 | 1000001 | Barbie | 2023.0 | She's everything. He's just Ken. | Barbie and Ken are having the time of their li... | 114.0 | 3.86 |
| 1 | 1000002 | Parasite | 2019.0 | Act like you own the place. | All unemployed, Ki-taek's family takes peculia... | 133.0 | 4.56 |
| 2 | 1000003 | Everything Everywhere All at Once | 2022.0 | The universe is so much bigger than you realize. | An aging Chinese immigrant is swept up in an i... | 140.0 | 4.30 |
| 3 | 1000004 | Fight Club | 1999.0 | Mischief. Mayhem. Soap. | A ticking-time-bomb insomniac and a slippery s... | 139.0 | 4.27 |
| 4 | 1000005 | La La Land | 2016.0 | Here's to the fools who dream. | Mia, an aspiring actress, serves lattes to mov... | 129.0 | 4.09 |

Figure 1. Overview of movies.csv

ii. countries.csv will contain information regarding the countries the movies where produced. Some movies involve more than one country for production. This leads to some rows of the file having duplicate IDs to accommodate all countries involved. For this dataset however, there are only 693, 476 observations considering some movies didn't report their country of production in the website.

| | id | country |
|---|---|---|
| 0 | 1000001 | UK |
| 1 | 1000001 | USA |
| 2 | 1000002 | South Korea |
| 3 | 1000003 | USA |
| 4 | 1000004 | Germany |

Figure 2. Overview of countries.csv

iii. genres.csv will describe each movie with different genre tags; movies can have multiple genres assigned. This dataset has a total of 1,046,849 observations where duplicate IDs were given to each genre pertaining to each movie.

| | id | genre |
|---|---|---|
| 0 | 1000001 | Comedy |
| 1 | 1000001 | Adventure |
| 2 | 1000002 | Comedy |
| 3 | 1000002 | Thriller |
| 4 | 1000002 | Drama |

Figure 3. Overview of genres.csv

iv. releases.csv pertains to the release type of each movie in different countries. Since some movies have different kinds of releases in different countries, there were a total of 1,332,782 observations.

| | id | country | date | type | rating |
|---|---|---|---|---|---|
| 0 | 1000001 | Andorra | 2023-07-21 | Theatrical | NaN |
| 1 | 1000001 | Argentina | 2023-07-20 | Theatrical | ATP |
| 2 | 1000001 | Australia | 2023-07-19 | Theatrical | PG |
| 3 | 1000001 | Australia | 2023-10-01 | Digital | PG |
| 4 | 1000001 | Austria | 2023-07-20 | Theatrical | NaN |

Figure 4. Overview of releases.csv

v. themes.csv describes all the themes related to each movie. Just like genres, there are multiple themes that can be associated with a movie. This dataset also has fewer observations involved with only 126, 641.

| | id | theme |
|---|---|---|
| 0 | 1000001 | Humanity and the world around us |
| 1 | 1000001 | Crude humor and satire |
| 2 | 1000001 | Moving relationship stories |
| 3 | 1000001 | Emotional and captivating fantasy storytelling |
| 4 | 1000001 | Surreal and thought-provoking visions of life ... |

Figure 5. Overview of themes.csv

**Connections**

This dataset is extremely valuable in film, particularly for film enthusiasts, critics, and industry professionals interested in analyzing movie trends. It can be utilized by a film enthusiast to identify the highest-rated movies, explore popular genres, or track the evolution of film preferences over time. Streaming services and production studios can utilize the insights to identify audience preferences, optimize recommendations, and make data-driven decisions on which types of films to produce or promote. Researchers and analysts can also examine movie production's geographical distribution to understand global cinema trends better. Whether for personal enjoyment, business strategies, or academic research, this dataset offers valuable insights that will improve decision-making in the industry.

*Data Merging and Data Cleaning*

From the ‘movies.csv’ we will be extracting only the ‘id’ and ‘ratings’ column. This will then be renamed to a different file ‘Ratings.csv’

The data merging process aimed to combine `countries.csv` and ‘Ratings.csv’ datasets with the datasets of each category. In order to systematically combine datasets, the `.merge()` function in Pandas was implemented. This function utilized a left join (`how='left'`) to retain all movie entries from the primary dataset and incorporate available information from supplementary datasets. This results in the creation of three different merged dataframes for each category. Three separated dataframes were necessary to avoid the duplication of variables from the columns ‘genres’, ‘themes, and ’types’. Merging all the dataframes at once to one merged dataset would lead to inaccurate counts caused by the duplication.

After the merging, some values in the columns of ‘countries’, ‘genres’, and ‘themes’ were needed to be merged or binned either for similarity and visualization purposes. The .replace() function was used to efficiently rename the values of the selected columns.

Each of the merged dataframes will first be grouped based on counts using .groupby(). The goal here is to count the values of genres, releases types, and themes based on ‘country’. Another .groupby() function will be used to find the average ratings by still basing on the same columns used in count.

*Exploratory Data Analysis*

In order to acquire preliminary insights into the dataset, exploratory data analysis (EDA) was implemented to recognize trends, distributions, and patterns among critical attribute categories. The cleansed and merged dataset was subjected to various statistical and visual techniques to guarantee its suitability for further analysis. In order to investigate the distribution of numerical features, summary statistics were computed using.describe() in Matplotlib and Pandas. Furthermore, the entirety of the data was verified by rechecking absent values using.isnull().sum(). Histograms, bar charts, and value counts were implemented to investigate critical trends and emphasize patterns in movie ratings, release years, genres, and country distribution.

Seaborn's histplot() was employed to generate a histogram of movie ratings for numerical attributes, which offered a glimpse into the frequency distribution of ratings. A bar chart of movie release years disclosed the tendencies in film production over time. Similarly, categorical attributes were examined, including identifying the top 20 most prevalent movie

genres and the top 20 countries with the most excellent density of movies utilized, plotted as bar charts and value_counts(). The dataset was well-prepared for in-depth analysis as a result of the intuitive comprehension that these visualizations facilitated. Early identification of critical trends was achieved through EDA, which facilitated the subsequent decision-making process in the analytical workflow and verified the dataset's integrity.

## *Prototyping and Planning*

### Prototype Design

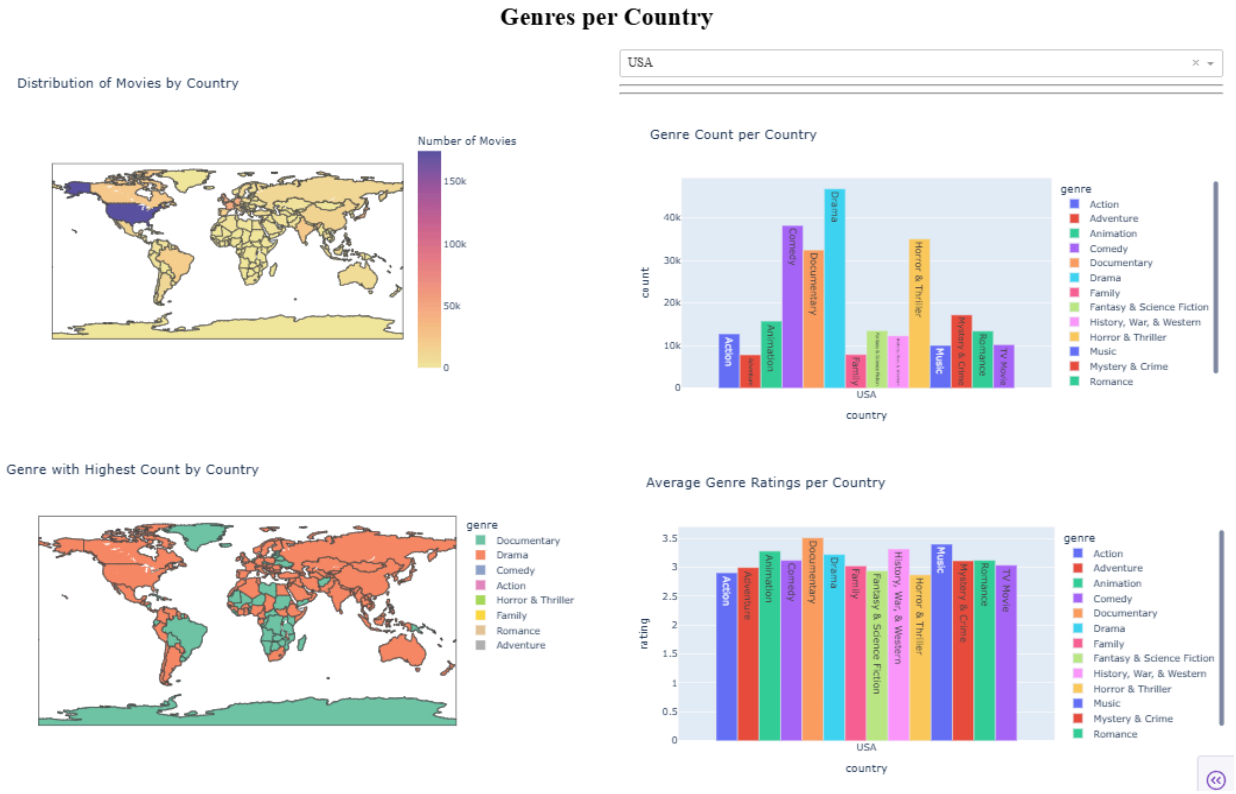


*Figure 1. Prototype of Visualizations.*

The Choropleth Map will be used to see the top movie trends in each country depending on the category they would choose. The categories that can be selected are Movie Distribution, Genres, Themes, and Release Types. Two Bar Charts will be used to completely see the trends and ratings in genres, themes, or release types of the chosen country.

**Visualization Idiom Choices**

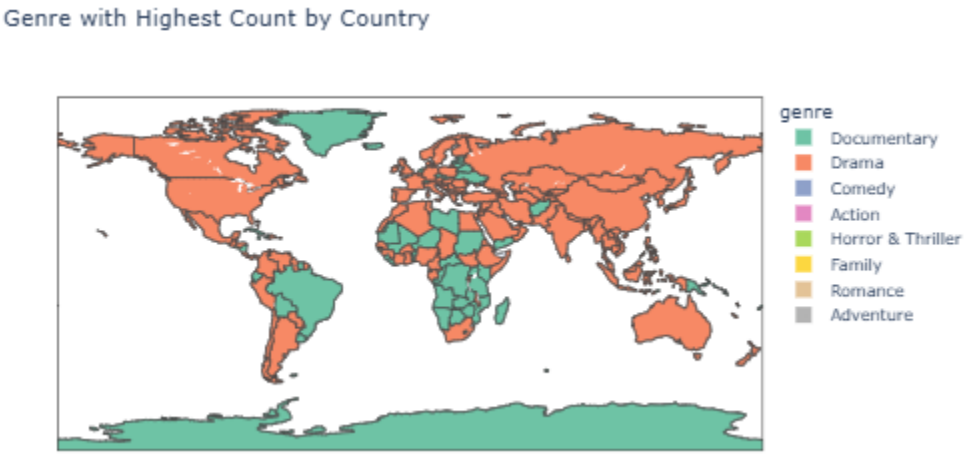


*Figure 2. Sample Choropleth Map of Worldwide Movie Trends. Image retrieved from: http://visualizingrights.org/kit/charts/choropleth-map.html.*

| | |
|---|---|
| **Idiom** | Choropleth Map |
| **Data** | Movie counts for Distribution, Genres, Themes, and Releases |
| **Channels** | Color<br>● For the case of Movie Distribution, Magma Gradient will be used to differentiate movie counts for each country<br>● Color scheme for other categories will use the default assigned colors from python |
| **Task** | Locate Trends |

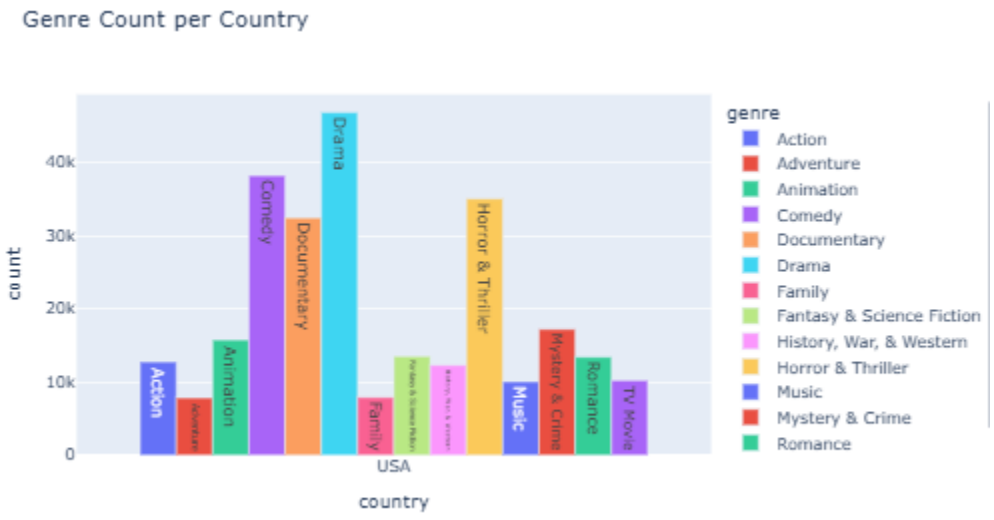


*Figure 3. Sample Barchart for each Country.*

| | |
|---|---|
| **Idiom** | Bar Chart |
| **Data** | Counts and Average Ratings for Movie Genres, Themes, and Releases. |
| **Channels** | Color<br>● Color will be adjusted based on what was assigned in the Choropleth Map |
| **Task** | Identify Distribution (Count) and Compare Trends |

**Interactivity Techniques and Justification**


Figure 5. Visualization of Interactivity

The Bar Graphs will have an interactive dropdown menu containing all countries of the dataset. Users will be able to pick a country to show count distributions and average ratings of each genre, theme, or release type. This is aimed to supplement the highest count choropleth map by giving users more insight to how the other genres, release types, and themes are received in every country.

**Any potential modifications to project specification**

The group is considering potentially having an option to filter out movies from the USA from the visuals. The US holds lots of movies from the dataset and it would help analyze deeper local trends of the other countries.

*Final Product*


Figure 6. App for Genres

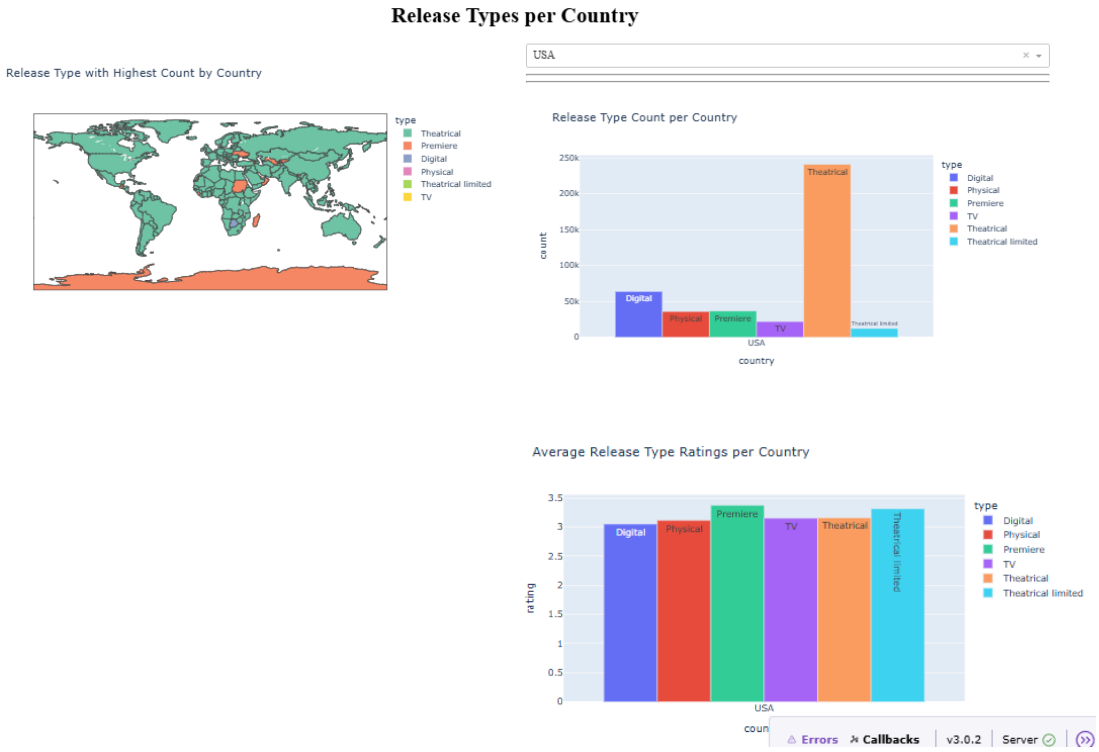


Figure 7. App for Release Types

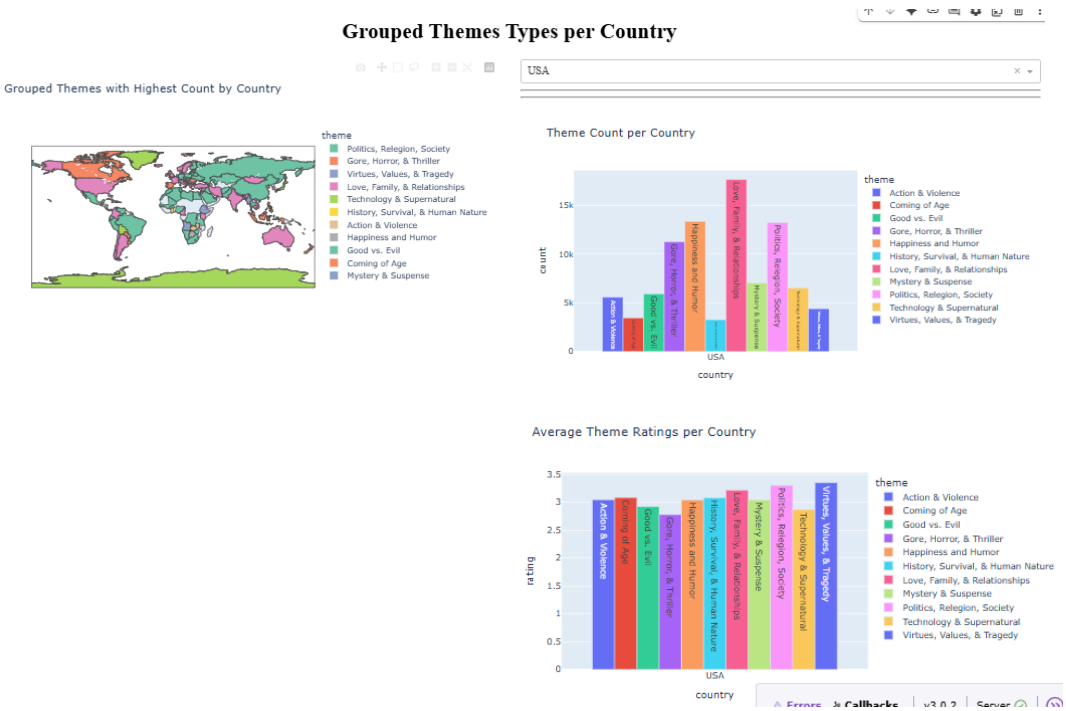


Figure 8. App for Grouped Themes

For the final output, three different apps were created for Genres, Release Types, and Grouped Themes respectively. The interactivity is only present for the bar graphs which allows users to filter what country they want to see for the distribution of movie counts and average ratings. Three different apps were made due to the errors persisting when combining them all into one. There were also some errors with the graphs due to the difference in sizes of the used dataframes for each category.

What was changed from the demo was the color of the basic Distribution of Movie by Country. From using a “Sunset” gradient, we settled with a “reds” gradient to better show the density and intensity of the distributions.