INF 552 Report

Data Visualization for Netflix TV Shows and Movies

Chenwei WAN[1]
Wen YANG[2]
Berenice JAULMES[3]
Yang LIU[4]
[1]chenwei.wan@telecom-paris.fr [2]wen.yang@telecom-paris.fr
[3]berenice.jaulmes@telecom-paris.fr [4]yang.liu@telecom-paris.fr

IP Paris
France
20/12/2023

1 Introduction

Digital streaming platforms such as Netflix captivate audiences from worldwide with a diverse array of content. As we navigate through the vast expanse of Netflix's content library, we noticed that there are data are stored as tabular data, covering a wild range of information about TV shows and movies. However, the tabular data is only a raw representation of data, without any global or statistical description of the data, and it is hard to interpret the connection between different data attributes or different entries of the dataset, this requires to involve data visualization method to provide potential for description of the data attributes, such as the distribution of genres and release years, as well as global viewing of data statistics information associated to their geological location.

For our project, we chose to visualize data about the topic that is familiar to most people nowadays: online streaming. Netflix is the most popular movie streaming platform in the world, with more than 240 million subscribers in over 190 countries. It hosts existing movies and creates its own. Netflix dataset [1] is available on Kaggle. The advantage of this dataset is its collection of variety of data on the movies and TV shows available on Netflix.

The objective of this project is to analyze this dataset and to provide different types of visualisations, including maps, scatter plots, histograms, graphs, etc. We choose to use d3 Javascrip library for the visualizations for its flexibility and power. It is suitable for any kind of data visualization, from simple charts and graphs to complex maps and networks.

1.1 The Dataset

This dataset has 5489 rows for unique movie title IDs and 15 variables containing their information. It includes information about the genre, score on IMDB, the release year and production countries, etc. It also over 77000 credits of actors and directors on Netflix titles with 5 variables containing their information, including the actor or director's name. This dataset was acquired in July 2022 containing data available in the United States. The raw data are provided in csv files.

1.2 Motivation and Major Challenges

Motivation. Our motivation for undertaking this project comes from the drawbacks of tabular data that the raw data remains a narrative way to present numbers or categories, rather than using raw data, we propose to integrate visualization method to provide a vivid description of the data attributes, such as their statistical distribution, the correlation between data, and to project the data statistics on the corresponding geological location. The design of our visualizations is realized using d3 JavaScript library.

We believe that visualizing the statistical information of these attributes will not only provide a comprehensive overview of Netflix's content, but will also offer insights into the evolution of the entertainment industry over time. By employing data visualization methods with d3 library, we aim to transform these distributions into interactive, intuitive representations that allow our readers to navigate the vast expanse of these attributes effortlessly.

In addition, we are motivated to plot a global map of streaming content production country by associating data statistics with geographical locations. Understanding how content produced across different regions is also a key aspect of our project. Through geological visualization, we want to show different ways of viewing aggregated numerical data. This is aimed at providing an understanding of how Netflix content are produced throughout the world.

Besides, our motivation extends beyond the confines of data analysis, it is also rooted in the belief that visual exploration of the data can give insights to the readers. In this project, we also offer an interactive platform where users can personalize their exploration of Netflix data attributes.

Furthermore, we are motivated by unrevealing the hidden dependencies and correlations within the dataset, we recognized the dynamic relationships between the data entries in the Netflix dataset through our work on visualization. Our project brings these discovered connections to diverse forms of data visualization. Connected graphs and hierarchical graphs illustrates the relationships between actors and directors, and the connection between different genres. Scatter plots capture the correlations between scores Heatmap provides a vivid representation of the correlation among different attributes.

[1] https://www.kaggle.com/datasets/victorsoeiro/netflix-tv-shows-and-movies

Challenge. The first challenge is aggregating the diverse array of information in each row of the CSV file. With each entry representing a unique movie or TV show, we want to aggregate data by different attributes such as release year, genre, country and then load the data to the d3 visualization element.

The second challenge is projecting aggregated data into a visual representation on a world map. While the raw data resides in a structured CSV file, the geographical mapping for an intuitive and meaningful mapping system adds complexity of our project, we chose TopoJSON library for the data projection on a world map.

To enhance the user's experience and provide a better understanding of the data, our third challenge is the incorporation of animation into visualizations. Implementing animations will not only make the visualizations good-looking, but also enable users to gain insights at different scales and resolutions.

The final challenge is to add interaction. We want to provide personalized exploration of Netflix data. This involves the implementation of interactive methods that allow users to adjust parameters, choose specific attributes. The challenge here is to find a balance between flexibility and accuracy of the data representation, ensuring that the interactive features enhance user experience without loosing accuracy of the data statistics.

1.3 Outline of Visualizations

In a short summary, the following visualizations were created in our project, and the detailed explanation are given in Section 2 of this report.

1. **World map of the number of movies:** Visualization of the number of movies produced per country on the world map. (Berenice JAULMES)

2. **Collaborative relationships visualization:** Visualization of collaborative relationships between actors and directors by the nodes and links in a graph. (Chenwei WAN)

3. **Most popular actors:** A simple but interesting metric to look at is the most popular actors on the platform. (Berenice JAULMES)

4. **Review score distribution visualization:** A ridgeline plot visualizing review score distributions, with an interactive feature allowing users to filter genres based on specific criteria. (Chenwei WAN)

5. **Correlation between review score and popularity:** A scatterplot analyzing the correlation between popularity and review scores, incorporated a comparison mode for users to select and compare different media types or years side-by-side. (Chenwei WAN)

6. **Interactive histogram for numerical features:** Histograms for presenting the statistics of numerical attributes, incorporated with a drop-down list and a slider bar for users to personalize the parameter of this histogram. (Wen YANG)

7. **Bar charts for categorical features:** Bar charts for presenting the statistics of categorical attributes, showing the top 30 countries with the most number of productions and genres distributions. (Wen YANG)

8. **Treemap for genres counts:** A visualization for displaying count of each genre in a treemap, the area of each node represent the number of production in this genre. (Wen YANG)

9. **Area charts for genres throughout time:** Using a stacked area chart, showing the percentage of production for each genre throughout time. (Wen YANG)

10. **Chord diagram for genre overlaps:** Providing information on co-appearance of different genres, and presenting the connections in a chord diagram. (Wen YANG)

11. **Heatmap for Feature Correlations:** Visualizing the correlation between data attributes using a heatmap. (Wen YANG)

2 Visualizations

2.1 World map of the number of movies

For the visualization of the number of movies per country, we used an interactive world map. When hovering over a country, the user can see its name and the total number of Netflix movies produced there. We chose to use a blue and red color scheme to make the countries that produce few movies stand out from the countries for which there is no data. This was more difficult to achieve with a one-color scheme. A special library was used for this map, namely TopoJSON.

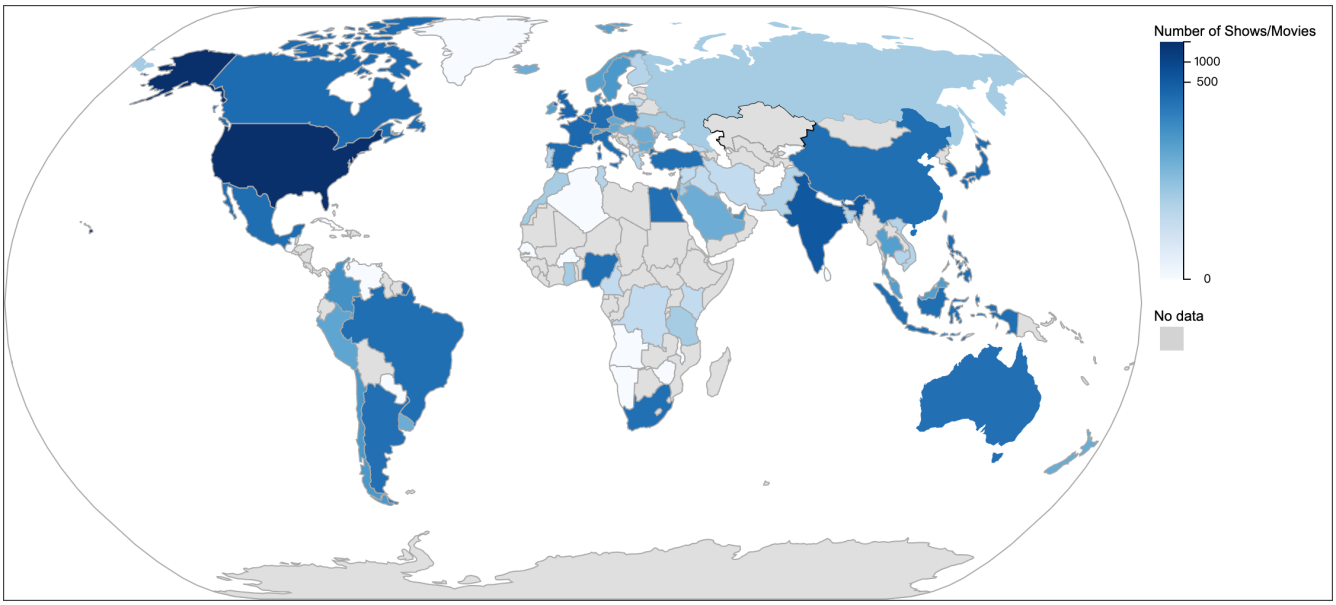


Figure 1: Number of Netflix movies per country

2.2 Collaborative Relationships Visualization

As to the modelling of collaborative relationships between actors and directors, we first introduce our data-mining strategy. We define 'popularity' for actors and directors based on their participation in a minimum of four movies or shows. We then establish a network to map collaboration relationships, exclusively considering 'strong collaborations' characterized by at least four joint projects.

In our network visualization, directors are represented by orange nodes, while actors are depicted with blue nodes. The strength of each collaborative link is quantified by the number of joint projects and visually represented by the stroke of the edge.Given the presence of multiple disjoint subgraphs within the network, we employ a disjoint force-directed graph approach, which enables us to maintain the visibility of separate subgraphs within a single viewport, facilitating comprehensive analysis. Further enhancing the utility of this visualization, we have integrated an interactive feature: when a user hovers over any node or link in the graph, a tooltip window is triggered. This window provides detailed information, such as the name of the actor or director represented by a node, or the complete list of collaborative projects associated with a link between two nodes. This interactive component of the visualization gives users a deeper understanding of the collaborative dynamics within the network.

2.3 Most Popular Actors

A simple but interesting metric to look at is the most popular actors on the platform. Because the United States is the country where most Netflix productions originate, it could be reasonable to assume that the most popular actors would be American. As can be seen in the following visualization, they are all Indian.

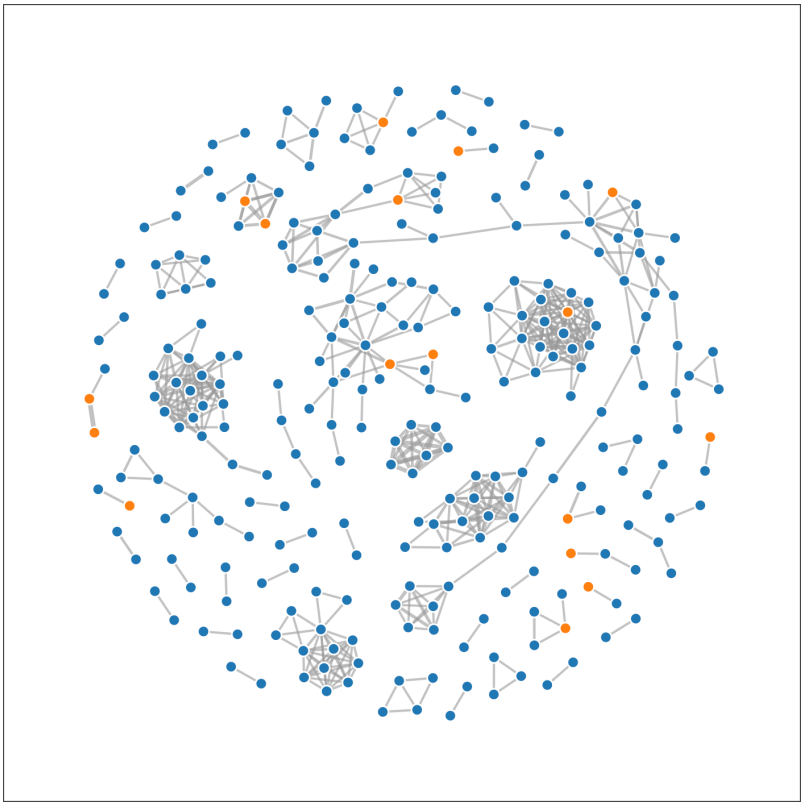


Figure 2: Collaboration

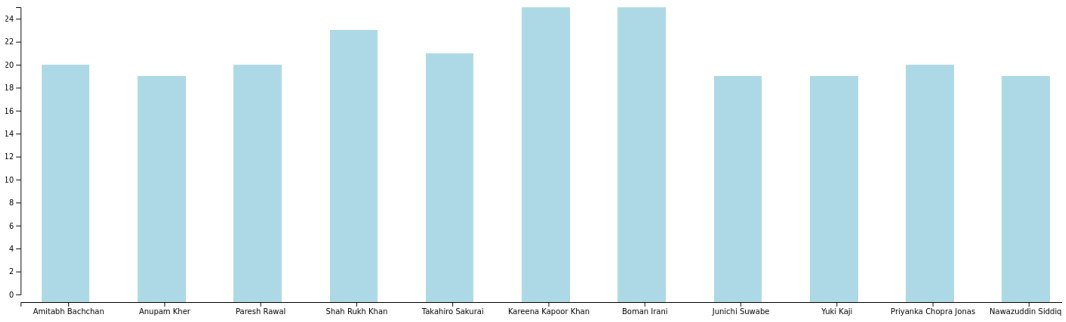


Figure 3: Most popular actors
The y-axis is the number of movies per actor.

2.4 Review Score Distribution Visualization

In this section, we employ a ridgeline plot to elucidate the distribution of review scores across various genres. This visualization is enhanced by a diverging color scheme, which is aligned with the mean review scores to distinctly highlight genres with comparatively higher or lower ratings. These ratings are aggregated from two prominent film review platforms: IMDb and TMDb.

Furthermore, the plot offers two distinct sorting options for users: an alphabetical arrangement of genres and an ordering based on their mean scores. Enhanced interactivity is a pivotal feature of this visualization. Specifically, when users hover over any section of the density plot, the corresponding area is accentuated through a change in opacity, an emboldening of the ridgeline, and a display of the genre's mean score. This interactive element is designed to provide an informative user experience, facilitating a deeper understanding of the data.

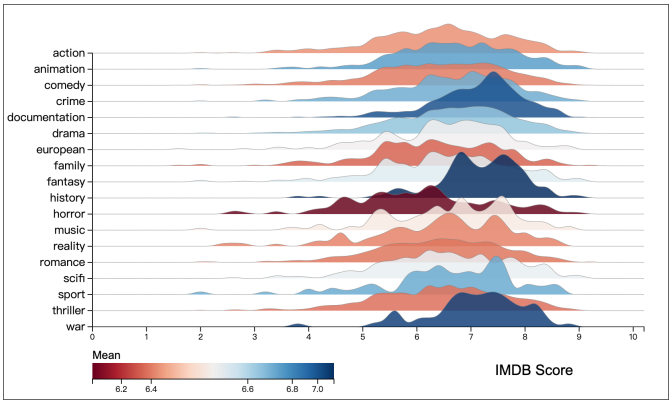


Figure 4: Ridgeline plot on IMDB score distribution, ordered by mean score

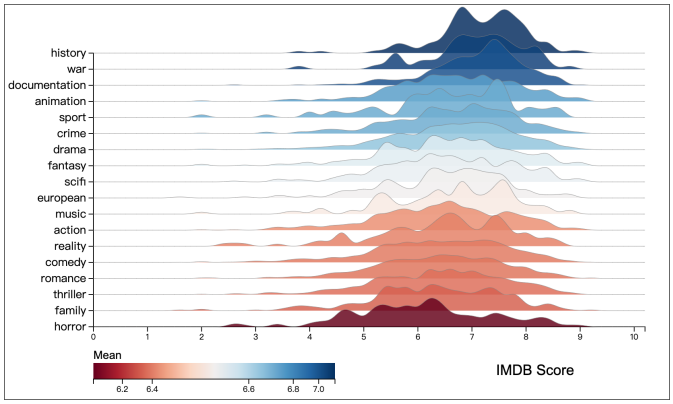


Figure 5: Ridgeline plot on IMDB score distribution, ordered by alphabet

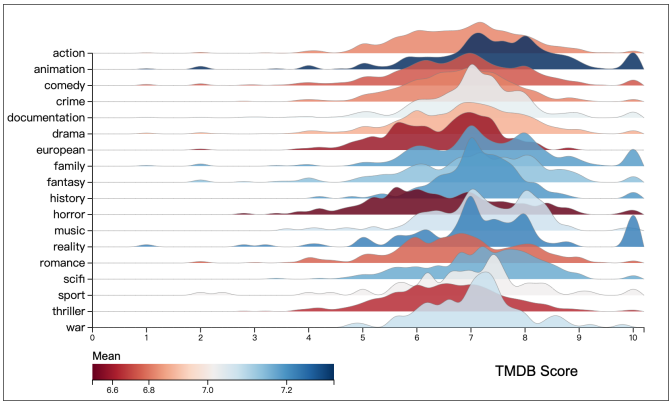


Figure 6: Ridgeline plot on TMDB score distribution, ordered by mean score

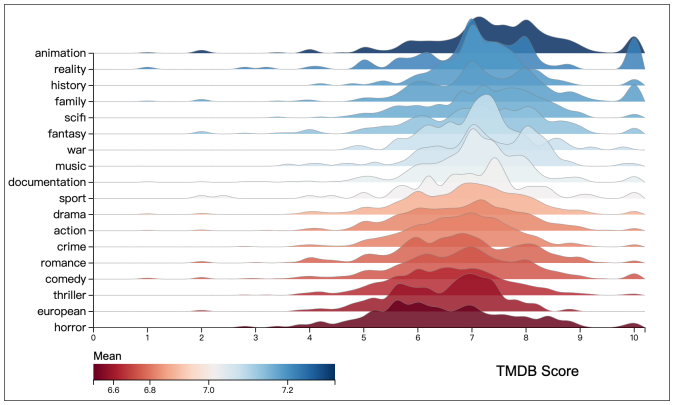


Figure 7: Ridgeline plot on TMDB score distribution, ordered by alphabet

2.5 Advanced Features in Correlation Analysis: Correlation between Review Score and Popularity

In this section, we present a scatterplot to analyze the correlation between popularity and review scores of media content. Considering the significant variation in the distribution of popularity data, the scatterplot is equipped with a scale that users can toggle between linear and logarithmic, enhancing the visual interpretability of the data. The scatterplot is enriched with a linear color scheme, which corresponds to the release year of the media content. Distinct shapes are utilized to differentiate between movies and shows.

To further aid in the comparative analysis of movies and shows, a histogram is displayed beneath the scatterplot. This histogram shows the count of different Netflix media types. Interactivity is a key feature here: when users select either of the bars in the histogram, the scatterplot dynamically updates to display only the data points corresponding to the selected media type. Moreover, when users hover over a data point in the scatterplot, a tooltip window appears, providing the name of the media. This design not only enhances user engagement but also offers insightful and tailored views of the data.

2.6 Interactive Histogram for numerical features

Netflix dataset contains both numerical attributes and categorical attributes, in order to study the showing general distributional features for those attributes, we proposed to create histograms for numerical attributes. Histograms reflects the count of occurrences in the dataset for each logical range or bins.

To deal with numerical attributes, the visualization for histogram is created with d3.histogram() for setting the parameters and with d3 “rect” element for building the chart. The height of rectangles represents the count

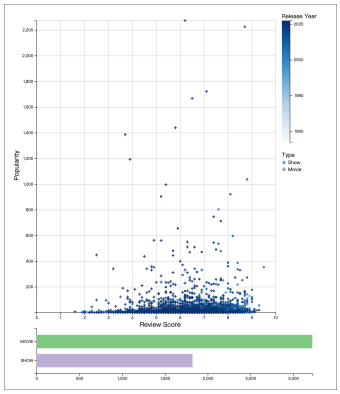


Figure 8: scatterplot, linear scale

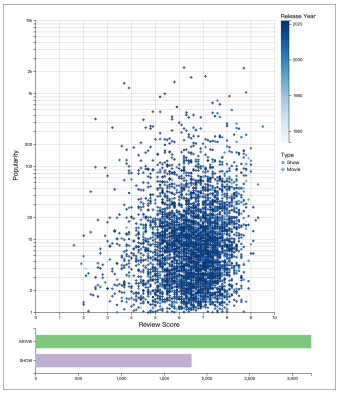


Figure 9: scatterplot, log scale

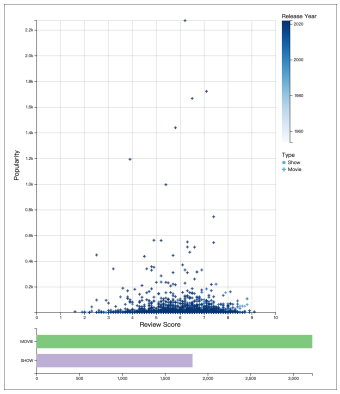


Figure 10: scatterplot of movie type, linear scale

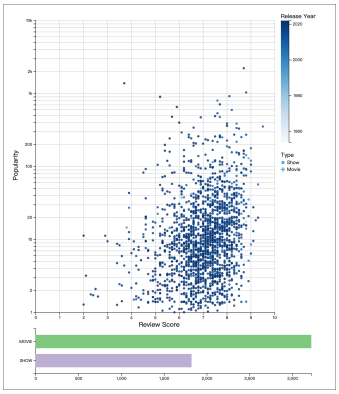


Figure 11: scatterplot of show type, log scale

of occurrence and the width represents the value of the attribute. The video on Netflix could be TV shows or movies and the different natures of shows and movies, such as different runtimes, our histogram is designed in an interactive way by varying the colors of the rectangles representing each type, when chosen one type of the video, the corresponding histograms' color can be highlighted. Also, by stacking the bars of the histogram, this visualization could provide a global description for the numerical attributes in this dataset. In addition, we integrated a slider and a drop-down menu list for users to choose the attribute and choose the desired number of bins in the histograms.

Fig. 12 illustrate the histogram for runtime for different types of content.Netflix offer two kinds of streaming content, namely movies and TV shows. Distribution of the content on this platform is skewed towards the movies, which occupy 64% of all the contents, two times more than the number of TV shows. This may due to its users prefer to spend limited amount of time rather than bench watching episodes of TV shows. This feature might be important while analyzing other features.

In general, movies have longer runtime than TV episodes. From the histogram, it is noticed that most movies are longer than 80 minutes while most TV shows are around 30 minutes or around 50 minutes, this is due to the fact that TV shows are often within 30 minute and hour show is within one hour, in Netflix dataset, they are both categorized as TV shows.

Fig. 13 present the histogram of season, imdb score and release year. In Netflix database, only TV shows have seasons, most of them are less than 3 seasons. Release of contents was happening from way back in 1945, and the number of productions increased through time. Since 2019, entertainment industry significantly produced more than past decades. For imdb score, most of the productions get 5 to 7 points, indicating that the content in Netflix is generally appreciated by the public.

2.7 Bar Charts for Categorical Features

For the categorical attributes, bar charts are plotted to show the count of occurrences in the Netflix dataset for each category. To deal with categorical attributes, the visualization is realized by creating a bar chart using d3

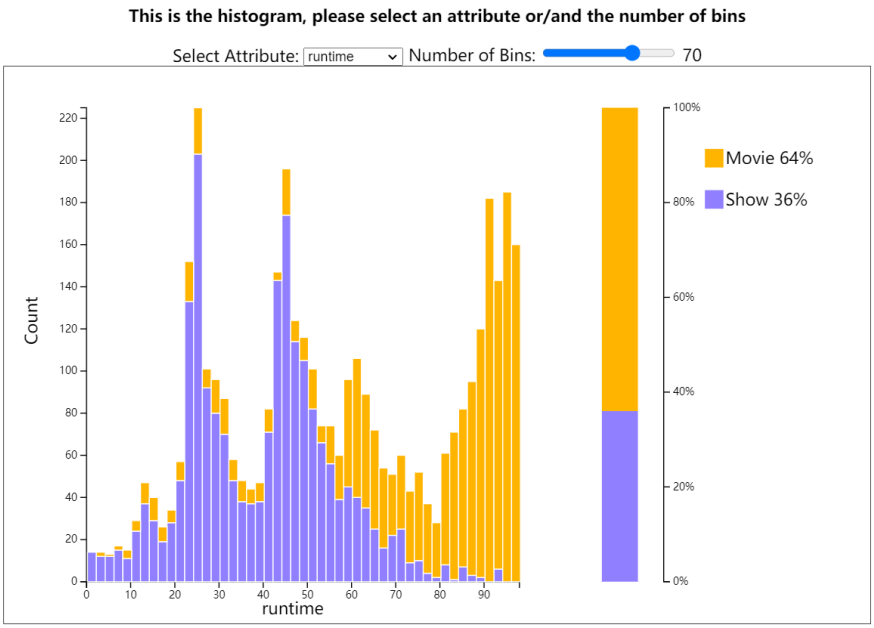


Figure 12: Histogram of runtime for different types of production

"rect" element, with the height representing the count of occurrences of different categories. The position of each bar is ordered by their count in descending order. This sorted bar chart allow us to know which category has the most count, and provide a comparison between different categories.

We calculated the number of productions in each country and show the top 30 countries in terms of their number of productions (Fig. 14a). The United States produced most, more than 3 times than the second one (India), and the United Kingdom (Great Britain) ranked number 3. Since Netflix is an American company, it is expected that it covers more productions from US and English-speaking countries.

Fig. 14b presents the number of productions in each genre. The most popular genres of productions are comedy and drama, both of them have more than 2000 production in our Netflix dataset. The genres with the least number of productions are war, sports and western, it can be concluded that Netflix is more focused on comedy or drama rather than western, war and sport type of content. This could be the result of the influence from costumers' interest and from the producers' working field.

2.8 Treemap for Genres Counts

Treemap is a visualization method for displaying hierarchical data. It uses nested rectangles to represent the branches of a tree diagram. Each rectangle has an area proportional to the amount of data it represents. In our case, we have limited hierarchical structure in our dataset, as a result, each genre is represented as one leaf node, and they all have the same level of hierarchy, all of them are connected to the root node in this graph (shown in Fig. 15).

The rectangles are filled with different colors to differentiate each genre. With a text element on each rectangle showing the genre and its count, treemap allows to easily identify the most and the least popular genres. Some genre with a very small count is represented by a very small rectangle, and the text will exceed the border of the rectangle, therefore, a clip function is used to constrain the location of the text element. This type of could be even more useful when illustrating hierarchical data.

The results on counts for each genre in this treemap is the same as is detailed explained in bar charts for genre counts.

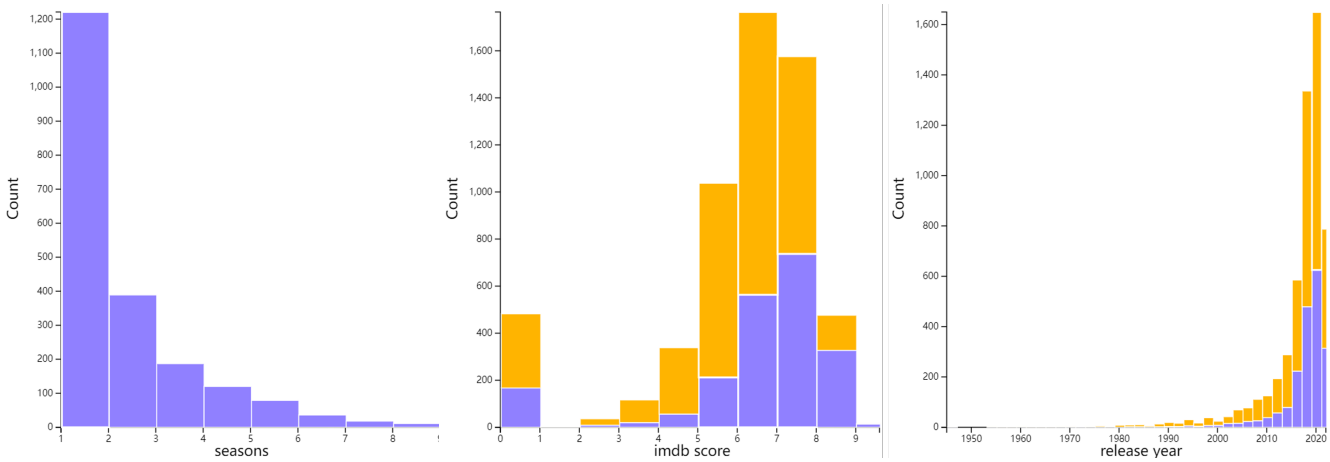


Figure 13: Histogram of season, imdb score and release year

2.9 Area Charts for Genres throughout Time

Stacked area chart displays the evolution of the value of several groups on the same graphic. The values of each group are displayed on top of each other, which allows checking on the same figure the evolution of the importance of each group.

In this figure, the percentage of production of each genre throughout time is analyzed. The fully-stacked height of the topmost line corresponds to the total when summing across all genres, in our case is 100%. Comparing the heights of each segment of the curve provides a general idea of how each genre compares to the other in their contributions to the total. In practice, this chart loses information about the trend of the absolute counts but helps to bring out the comparison of relative contributions between genres. The data are prepossessed before plotting the chart. Firstly, an aggregating by genre is performed, then we compute the sum of each time period and obtain the percentage of each genre during this period. Secondly, area charts are generated with d3.stack() and d3.area() function to draw the area and stack areas together. At last, each genre is assigned with a different color by a customized color scale to differentiate them between each other.

The horizontal-axis shows time period. Time range is from 1953 to 2021, since there isn't any data for 1949 to 1953 this period is ignored when we process the data. Some genres are not very visible because they only take up very small proportion, as a result, we decided to highlight the color and corresponding area when mouse pointer moves on this area.

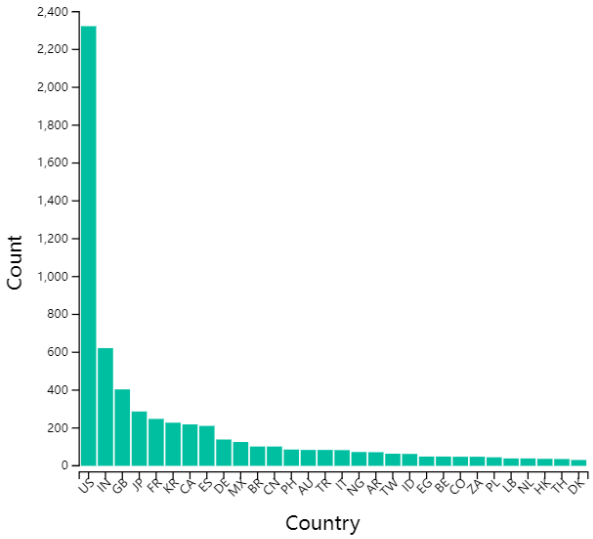
Results are given in Fig 16. Some genres have been produced throughout time, such as crime or drama; some genres started to be produced later, such as animation or science fictions, they begin to make this type of productions after the application of computers in film industry or TV series studios. Some genres only appears during some specific time period, for example, most western genres are added to Netflix database around 1960s due to their representation of cultural sentiments.

2.10 Chord Diagram for Genre Overlaps

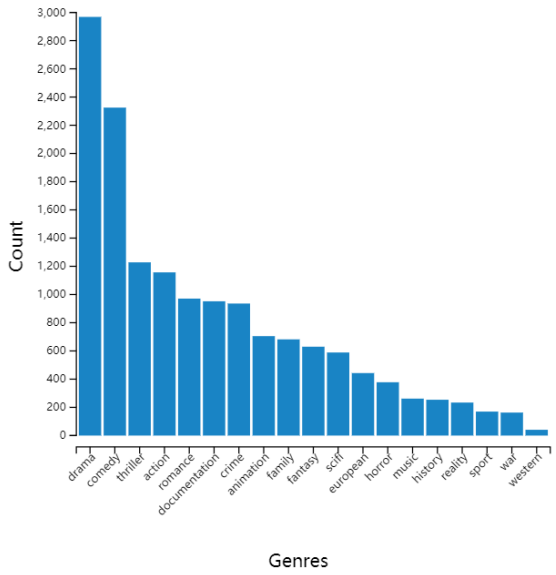
The co-appearance of different genres is also studied. One movie or one TV show may belong to several genres at the same time. For this kind of analysis, we propose to use a chord diagram. A chord diagram represents connections between several nodes. In our case, each genre is represented by a node. And it is shown as a fragment on the outer part of the circular layout. The length of this circular fragment corresponds to the percentage of the count for this genre among the summation of all counts. And each genre has a different color, which makes them distinguished from each other. Then, arcs are drawn between each node. The width of the arc is proportional to the count of co-appearance of the between two genres.

The data is processed by d3.chord() function to set parameters for our chord diagram, the circular fragment and connected arcs between nodes are created with d3.arc() d3.ribbonArrow() function respectively.

From Fig 17, it is noticed that comedy and drama, romance and drama or action and crimes are often appears together, while some genres (for example, horror) are less overlapped with other genre. This is due to the nature



(a) Count of production for top 30 countries



(b) Count of genres

Figure 14: Bar charts for categorical attributes

of content. Take drama as an instance, dramatic movies often provide a heightened emotional experience, and it could be categorized both as drama and romance at a time. Besides, multi-labelled contents are more likely to be recommended to the users, since a genre information is an important feature for recommender systems.

2.11 Heatmap for Feature Correlations

Some attributes are correlated with each other, in order to show this kind of correlation, we use Pearson correlation coefficient to describe the correlation. Pearson correlation coefficient is computed by Eq. 1

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{1}$$

It is a number between –1 and 1 that measures the relationship between two variables. Negative values suggest a negative correlation, while positive values suggest a positive correlation. We proposed to use a heatmap to visualize the correlation coefficients. Heatmaps show relationships between two variables, one plotted on each axis. By observing how cell colors change across each axis, we can observe if there are any patterns in value for one or both variables. In our visualization, we studied 6 different attributes, and the correlation coefficients are represented by squares with different color. The squares are created using d3 "rect" element. The color scale defined as dark blue for -1, gray for 0 and dark red for 1.

The heatmap is presented in Fig. 18. Release year is slightly negatively correlated to both IMBD score and runtimes. It is reasonable that consumers prefer recently released films or shows, and due to the technical constraints, very old films and show has less runtime than recent productions.

3 Conclusion

The aim of our project is to utilize various visualization techniques to present the statistical distribution and the connections among attributes in this dataset.

Through the creation of scatter plots, histograms, bar charts, and area charts, we've not only presented data but also provided users of our webpage with the ability to interact and adjust these visualizations to meet their specific requirements.

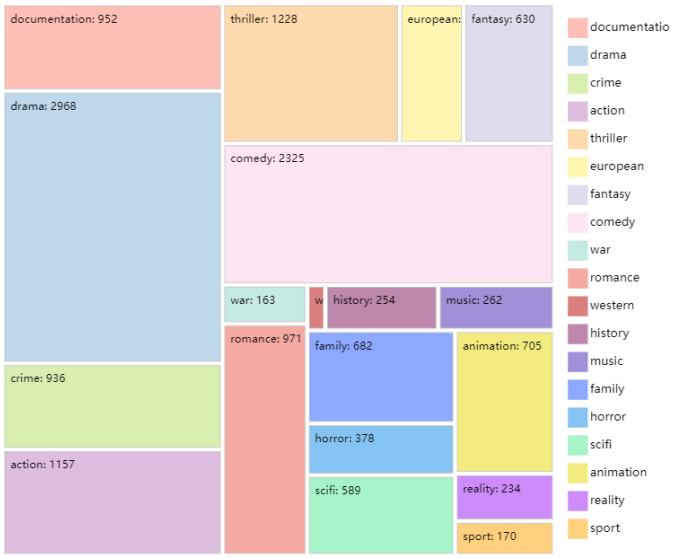


Figure 15: Treemap for Genres

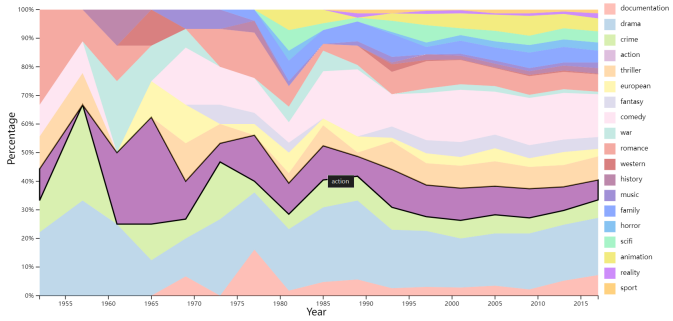


Figure 16: Stacked Area Chart for Genres over Time

In addition, the advanced graphical representations, such as graphs, treemaps, and chord diagrams, has allowed us to present data connections in a visual manner. These visualizations could serve as powerful tools for users to extract meaningful patterns from complex datasets.

Besides, the inclusion of a world map to present data geographically adds an extra layer of richness to our visualizations.

Another important highlight of our project is the integration of interactive features into the webpage. Users now have the flexibility to adjust parameters of the charts, providing a personalized experience. This not only enhances user engagement but also ensures that our visualizations are adaptable for a wide range of applications and purposes.

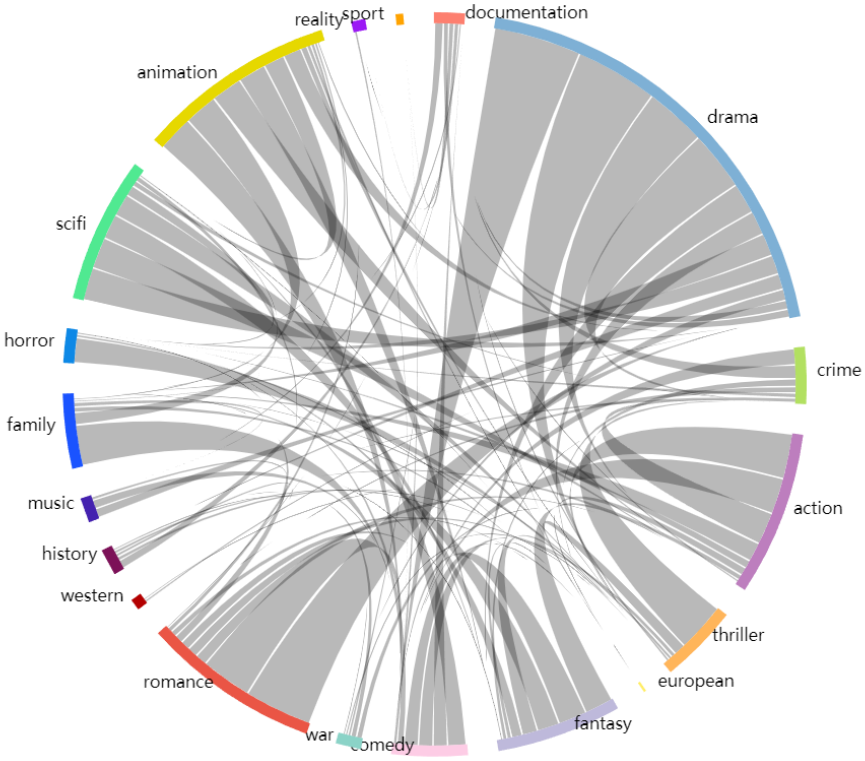


Figure 17: Chord Diagram for Genres

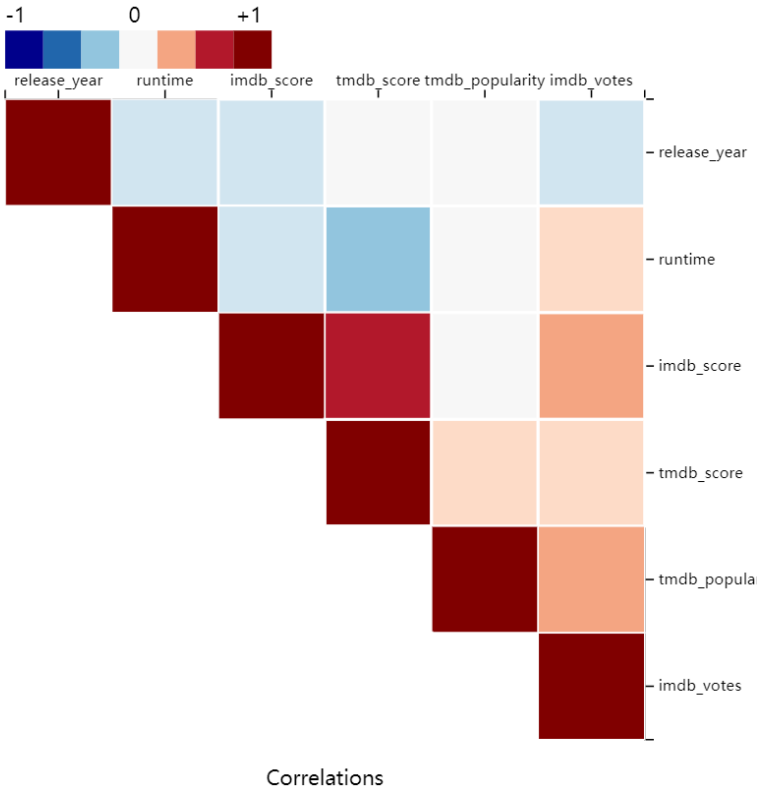


Figure 18: Stacked Area Chart for Genres over Time

A Complete dataset description

"Titles.csv" contains more than 5000 unique titles on Netflix with 15 columns containing their information :

- id: The title ID on JustWatch.
- title: The name of the title.
- show_type: TV show or movie.
- description: A brief description.
- release_year: The release year.
- age_certification: The age certification.
- runtime: The length of the episode (SHOW) or movie.
- genres: A list of genres.
- production_countries: A list of countries that produced the title.
- seasons: Number of seasons if it's a SHOW.
- imdb_id: The title ID on IMDB.
- imdb_score: Score on IMDB.
- imdb_votes: Votes on IMDB.
- tmdb_popularity: Popularity on TMDB.
- tmdb_score: Score on TMDB.

"Credits.csv" contains over 77000 credits of actors and directors on Netflix titles with 5 columns containing their information, including :

- person_ID: The person ID on JustWatch.
- id: The title ID on JustWatch.
- name: The actor or director's name.
- character_name: The character name.
- role: ACTOR or DIRECTOR.