VRIJE
UNIVERSITEIT
BRUSSEL

INFORMATION VISUALIZATION

PROJECT REPORT : Netflix

Brenda Ordonez Lujan | Student ID : 0571129

Deepika Chandrababu | Student ID : 0572230

Madina Myrzaliyeva | Student ID : 0572777

Erdogan Abaci | Student ID : 0584957

29 May 2022

Sciences and Bio-Engineering Sciences

Contents

1 Introduction

One of the ways to deal with big data is using "Information visualization" from which the user can see and find meaning insights from large data. The aim of this project is to analyse and visualize a large dataset following the background information and regulations learnt in the lectures of this course. We chose Netflix as the OTT platform is trending in recent days. Based on the popularity of a movie/series, the content can be changed in order to increase the box office, writers can come up with relevant ideas with respect to top hits or simply to see the top directors and movies of a particular year. There are millions of movies released in Netflix over years, and to interpret all of them is tedious. The data we used was collected from three different sources in kaggle and combined them to make it more reliable. Netflix data has information about movies and series being released in different genre or countries, how profitable they are, ratings of each, popular actors and directors.

Using these information, We wanted to create an interactive dashboard for producers who is looking for a crew to hire. Suppose if the producer has his favourite genre and he wants to hire a writer to build up the story and a director to create the movie, he/she can look up the dashboard and choose the options they want over the past 10 years. For a particular genre, the user can see the best year of that genre, top movies and the top directors and writers according to the ratings from the public. The user can also see the difference between movie/series affecting the boxoffice or other relevant factors.

2 Framework

Tools and Libraries used:

- Node.js: Environment used to run the project.
- Pandas: Library used for the preprocessing.
- Recharts v2.1.9: Framework chosen for the development of the dashboard front-end.
- D3.js: Library used to read CSV files.
- Template Creative Tim: Template used for the dashboard display

3 Dataset

This section discusses the metadata used for the visualisations

3.1 Dataset 1: Analysis of the quality of movies/series on Netflix

The first dataset was taken from Kaggle [1].

Meta Information: This dataset mainly discusses the ratings using imdb scores and box office based on the information from youtube and other API's but not from Netflix. This is one of the reasons why we added an original filter in our dashboard. This dataset contains 15071 rows and 29 columns describing Title, Genre, Tags, Languages, Series or Movie, Hidden Gem Score, Country Availability, Runtime, Director, Writer, Actors, View Rating, IMDb Score, Rotten Tomatoes Score, Metacritic Score, Awards Received,Awards Nominated For, Boxoffice, Release Date, Netflix Release Date, Production House, Netflix Link, IMDb Link, Summary, IMDb Votes, Image, Poster, TMDb Trailer and Trailer Site

3.2 Dataset 2: Netflix Original Movies

The second dataset was tacken from Kaggle [2].

Meta Information: This data discusses original movies from netflix. Original content stands for the movies that were produced by Netflix productions. This dataset has 523 rows and 48 columns describing the movies with respect to Title, Directed by, Produced by, Screenplay by,Based on, Starring, Music by, Cinematography, Edited by, Production,companies, Distributed by, Release date,Running time, Country,Language, Budget, Box office,Running Time, Budget, Release date, imdb,metascore, rotten tomatoes, Written by, Production company, Story by, Narrated by, French, Spanish, Italian, Portuguese,Animation by, Hangul, Revised Romanization, Japanese,Literally, Turkish, Indonesian, German, Norwegian, Polish,Music, Lyrics, Book, Basis, Productions.

3.3 Dataset 3: Netflix Original Series

The third dataset was taken from Kaggle [3].

Meta Information: This data is comparitively small with 358 rows and 14 columns. The columns in this dataset represent original series from netflix with the following information; Title, Genre, GenreLabels, Premiere, Seasons, SeasonsParsed,EpisodesParsed, Length, MinLength,

MaxLength, Status,Active, Table, Language. We collected this dataset and combined with the orginal movies dataset to compare the movies and series contributions in netflix through the past 10 years.

4 Design Decisions

Because the dataset presented in the previous section allows for a wide range of analyses, we started by defining our target user and task to narrow our approach. While deciding on the target user, we considered the main responsible for the creation of movies and/or series on the Netflix platform. As a result, our target user is: “A producer of Netflix original content”. And our target task is as follows: “Provide information about Netflix content ratings by genre and category to future projects”. Also, we added Netflix logo on the left corner to give a quick overview of what the dashboard is about and not for a "fancy" look.

As our target task indicates, we want to provide information to the producer so that he can make decisions about the film’s direction. To do so, the producer should have already considered what sort of film genre he wants to make and if he wants to focus on movies or series. Furthermore, we also want to show the influence of this genre on box office contribution in case there’s a possibility of release to the cinemas. The subtasks below provide more detail regarding the type of information to be provided. The subtasks are as follows:

- Know the rating trend of a specific genre.
- Know if Netflix is releasing films to the cinema.
- know the box office contribution of each film.
- Find the most popular directors and/or writers based on ratings.

5 Data Preprocessing

Two types of preprocessing were performed for the dashboard. The first one required combining the two datasets and the second one required rearranging the data for each visualization.

5.1 Combining the dataset

The main reason why it is required to combine the two datasets with the main dataset is to distinguish which movies and series were produced by Netflix. For that reason, a field called

"original" is added to the main dataset, where the value "1" means that the content is produced by Netflix. The only possibility we have for this combination is through the "title" field. To reduce possible typing differences as it is a text field, all titles were converted to uppercase and spaces were removed. See the complementary material for more details of the code.

5.2 Rearranging the data for each visualization

The preprocessing is not performed in the same framework; instead, all processed data is exported to a csv file that can then be imported into the framework. It is also necessary to mention that all the operations were performed in the dataset resulting from combining the dataset and the following is an explanation of the considerations taken into account for each visualization. See the complementary material for more details of the code.

5.2.1 Graph 1: Netflix content rating trend by year

The number of films and the average rating of each genre were calculated and sorted by genre, year, original, and category. In the process, the only inconvenience was the classification by genre. Because the genre type of each content in the dataset could be composed of many values separated by commas. Therefore, the "genre" column is decomposed by all its possible values and duplicates are removed to know the individual genres. Only contents from the year 2010 onwards are considered because it is the year where Netflix started producing original content.

5.2.2 Graph 2: Box office contribution

All titles, ratings, and box office profits were listed and grouped by genre, year, original, and category. Similarly, only content from 2010 onwards was filtered as previously described, with the exception that all Nan values in the "rating" and "box office" variables were converted to 0. Despite the fact that Nan is an unknown value, it was considered that such input would not add value to such fields for this graph.

5.2.3 Graph 3 and 4: Popular directors and writers of the year based on the rating

From the combined dataset, category(Movies/series), Original content from netflix and genre were grouped together with respect to the average ratings the directors and writers received for each title based on the imdb score. The ratings were ordered from maximum to minimum and the year was filtered to be above 2010. Also, there were a maximum of four writers for some movie

or series where its possible that each writer would have wrote the story for an episode/season in case of series. Because of the space limitations, we extracted the first writer's name and showed one writer's last name on the y axis and chose to show the contribution of all the writers on the hover. The same applies for the directors.

6 Dashboard

The dashboard shows the Netflix platform's rating for each genre over the course of a year. The subtasks mentioned in section 3 define the information it offers. Figure 1 represent the dashboard mentioned, which is composed of initial filters and four graphs in total, the first of which is the main graph and the other three are the result of the interaction with it. The dashboard is also available on the internet at the following address: http://infoviz.teotse.com/.

Figure 1: Dashboard

The preprocessing of the datasets for the dashboard will be explained first, followed by the setup instructions, and lastly the details and purpose of each graphs.

6.1 Dashboard Filters

- **Category:** You can choose whether to see Netflix movies or series.
- **Genre:** You have the option of selecting the genre of the content you want to see.

- **Original:** You can choose whether to show all Netflix content or simply original content.
- **Year:** You have the option of viewing a specific year. The previous filters will influence this field.

Figure 2: Netflix Dashboard filters

Figure 2 shows the “category”, “genre”, and “original” filters. The "original" filter's purpose is to give the user more directional alternatives as well as information about the genre's overall contribution or simply the Netflix content. Figure 3 shows the “year” filter, which adds interaction to the graph. When the “year” filter is selected, the following graphs will update. This update was done in orange, which is the same color as the rating because all of the graphs are based on it, and the blue color indicates the initial values, which by default belong to the year 2013. If there are no values to examine for that year, then It will appear a text with the phrase “No data”.

7 Visualizations

Before looking at the dashboard, a producer should have thought about what kind of movie and/or series he wants to make, as indicated in section 3. This is important in order to gather particular information regarding his upcoming project. Therefore, the function of each filter will be discussed first, followed by the graphs in detail.

7.1 Netflix content rating trend by year

Figure 3 shows the first and most important graph, which is used to determine the rating trend of a given genre in relation to the number of movies/series produced. This comparison is important since it allows the user to evaluate which year was the best for that genre. It is preferable to know the directors and writers who took part in the most optimal year based on their own judgment.

Graph characteristics:

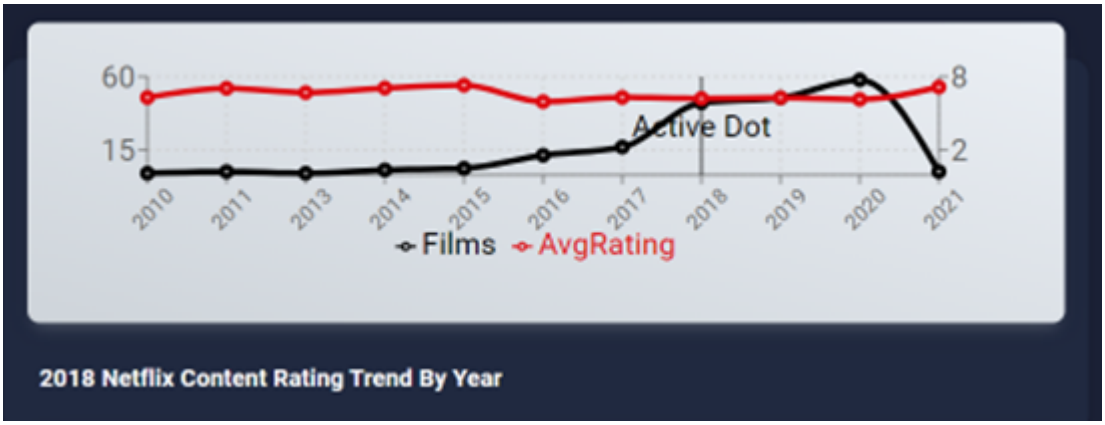

Figure 3: Netflix content rating trend by year of drama movies

- Selection of the graph: Biaxial Line Chart. It allows to combine and compare the two values close to one another.
- Color choices: The red color was used for the rating, while the black color was used for the number of films. To distinguish the values, these colors were preferred.
- Interaction: When the user hovers the mouse over a year, a tooltip showing the average rating and number of films values appears. The graph allows the user to select a year for further interaction; once the year is selected, a message "Active Dot" appears to inform the user that the year was chosen; if no information is available, the graphs are updated with the phrase "No data" to inform the user that an interaction occurred but no values were found.

7.2 Box office contribution

Figure 4 shows the graph of box office contribution by genre, in this case 2018 drama films. The importance of the graph is to know if Netflix is releasing films to the cinema in that year and how well the film is going; according to the data, Netflix has been producing films for the cinema since 2010 so today it becomes a decision whether to release it only on the platform or to invest in theaters first. With this visualization, the user can see if a certain genre has a good box office in the cinemas, either Netflix's own content or in general, which can help the user's decision.

Graph characteristics:

- Selection of the graph: Scatter chart. It lets each film to be displayed as a dot, allowing the size to support hundreds of films, and it was chosen to see the film's profit.

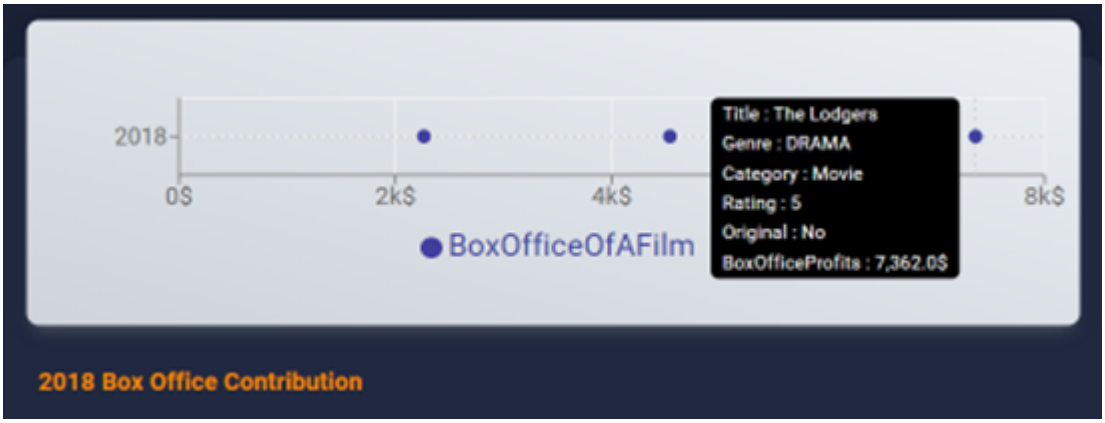

Figure 4: Box office contribution of 2018 drama movies

- Color choices: The blue color was used to display the values and the orange color was used for the title as a result of the interaction with the principal graph.
- Interaction: When the user hovers the mouse over a dot, a tooltip showing the film's detail appears such as the title, genre, category, original, rating and box office.

7.3 Popular Directors of the year based on the rating

This section is important because it affects users ratings if the direction is not good. The most popular directors of the given genre and year are shown in figure 5, in this case for 2018 drama films. The graph's significance is that it provides potential candidates for the film or series direction; the directors were picked because they are a significant factor in the film or series success; the filter "Original Netflix content" allows for more director options, in case the producer wish to continue creating links with directors who have already been involved on the Netflix platform.

Graph characteristics:

- Selection of the graph: Horizontal bar chart. It is chosen because we only need to compare one value, the rating, and give priority to the director's name being legible.
- Color choices: The blue color was used to display the values and the orange color was used for the title as a result of the interaction with the principal graph.
- Interaction: When the user hovers the mouse over a bar, a tooltip showing the detail appears such as the complete name of the director, his average rating, category, genre,

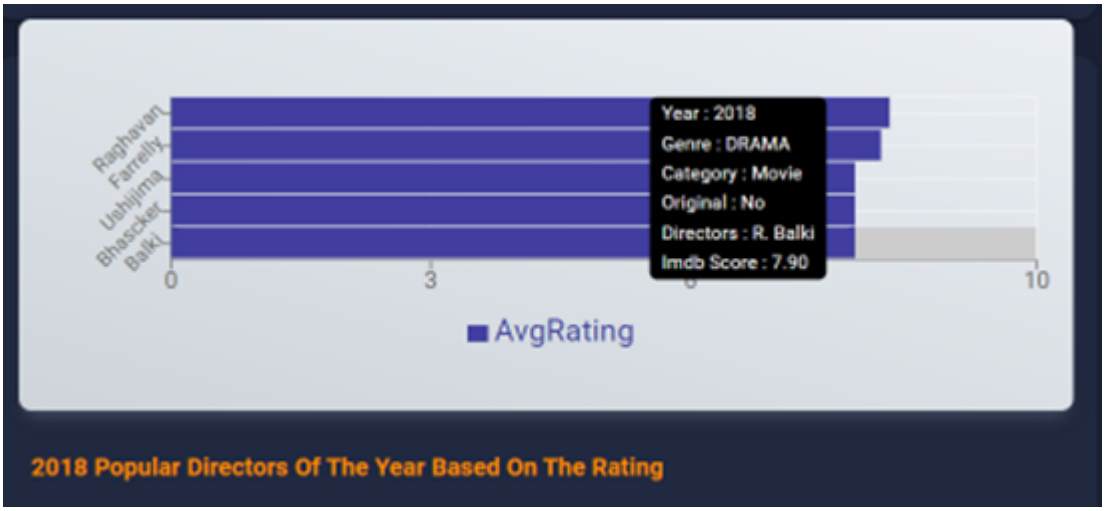

Figure 5: Popular Directors of 2018 based on the rating of drama movies

original and year.

7.4 Popular Writers of the year based on the rating

Figure 6 shows the most popular writers of the given genre and year in this case for 2018 drama films. The only difference between this graph and the one mentioned previously is that in addition to directors, we also want to provide candidates for writing the film or series script.

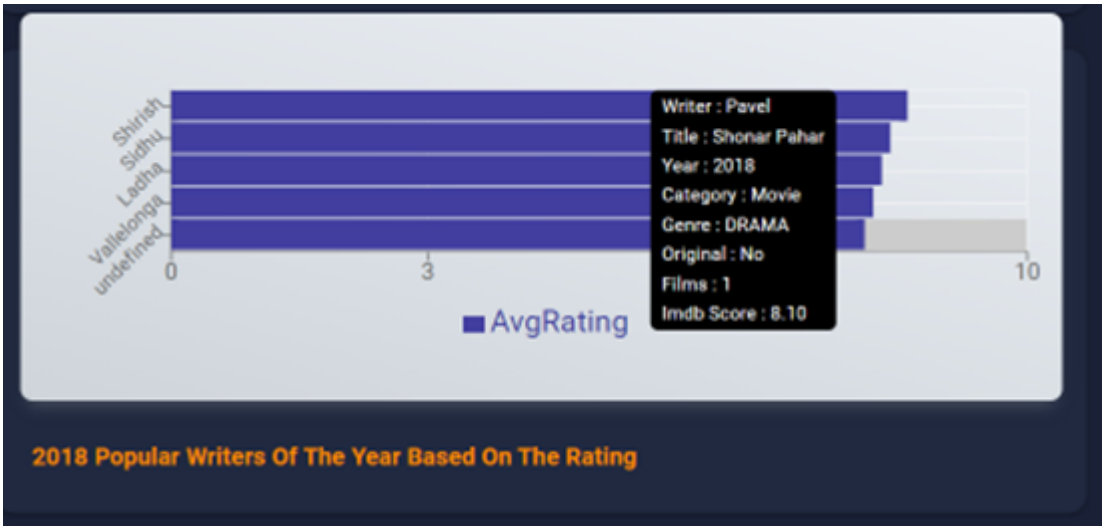

Figure 6: Popular Writers of 2018 based on the rating of drama movies

Graph characteristics:

- Selection of the graph: Horizontal bar chart. It is chosen because we only need to compare one value, the rating, and give priority to the writer's name being legible.
- Color choices: The blue color was used to display the values and the orange color was used for the title as a result of the interaction with the principal graph.
- Interaction: When the user hovers the mouse over a bar, a tooltip showing the detail appears such as the name of the writer, his average rating, category, genre, original and year.

8 Setup Instructions

To run the project, go to the folder where the project is located and follow these commands in order:

- npm install
- npm install react-scripts
- npm run start

9 Evaluation

In order to validate our project top-down design or problem driven approach was followed starting from the domain situation. In the first place the data-set to be used was chosen, based on which target user and the possible tasks were identified. All the visual encoding and interaction were justified. After all the features of the dashboard were implemented, the system's time complexity was measured. As our chosen target user is Netflix producer, no field study could be done. Nevertheless, several people were asked to share their opinion on the dashboard by filling the user experience questionnaire.

9.1 Time complexity:

To measure the time complexity of the dashboard the Google page speed tool was used. According to the tool, the longest time for the dashboard to update is 1.3 seconds. Total blocking time is

0.260 milliseconds and time to interactive is 1.5 seconds. Overall, the dashboard's performance is rated by 85%, which is defined to be above average.

9.2 Questionnaire:

User experience questionnaire (UEQ) was used to gather people's opinion on the dashboard's usability. Questionnaire asks people to rate the product in the range of 1 to 7 for 26 features. Due to the time limit only 18 people could fill the questionnaire. Most of the people who participated are engineering and management students at VUB and their age range is from 20-32. According to the results, the average responses are neutral or positive. However, it was identified by the UEQ tool that three answers were inconsistent and therefore were deleted. Figure 7 shows the analysis of the final results. It could be seen that the dashboard has mostly positive reviews. Although the product is found to be attractive, it lacks a bit of novelty and reliability.

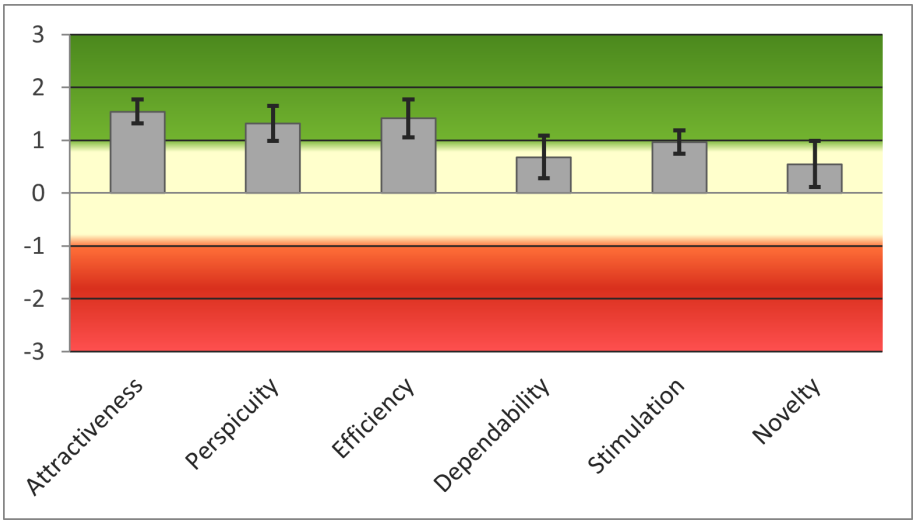

Figure 7: UEQ results

10 Limitations

Although the dataset provided a variety of approaches for performing an analysis, some limitations were discovered with respect to the dataset during the process based on our purpose. The following limitations were discovered:

- The dataset does not have a gender distinction for actors, directors and writers. The

dashboard was therefore limited to not recommending actors and not making a distinction in the case of writers and directors.

- The dataset does not provide information on the country or language of directors and writers.
- The dataset does not include information on whether or not the directors and writers are still active, as well as their country of origin and language. This restricts the dashboard's recommendations in that it can't suggest directors or writers who speak specific languages to avoid problems with the communication among other things.

It's also important to note the limitations of the study; one of the issues is that the evaluation was not conducted with our target user in mind, thus the participants aren't entertainment experts. Some users did not grasp what a box office was, what a good rating was, or what ImdbScore meant, according to the feedback. To avoid misconceptions, the evaluation was carried out while providing them with context and information.

11 Conclusion

We presented an interactive dashboard for Netflix producers in this work. We designed and programmed the dashboard interface with the goal of providing information about the box office contribution of movies, as well as recommendations for the best directors and writers based on the user's choice of genre and year. Although the original goal was to provide information based on a genre, the developed dashboard also provides another point of view, which is a comparison between the genres trend ratings of movies or series. Normally, the user would rely on memory to evaluate this last point, which is not a good method, but it is a positive aspect because it is a secondary scope of the dashboard.

We evaluated the dashboard's attractiveness, perspicuity, efficiency, reliability, stimulation, and innovation using a data analysis questionnaire, as well as testing the application's functionality with Google page speed tool. The dashboard was improved as a result of some user's feedback, although there are still limitations due to the lack of information in the dataset utilized for this project.

References

[1] A. Gupta. Latest netflix data with 26+ joined attributes. [Online]. Available: https://www.kaggle.com/datasets/ashishgup/netflix-rotten-tomatoes-metacritic-imdb

[2] J. Udayakumar. Netflixdataset. [Online]. Available: https://www.kaggle.com/datasets/jayaudaykmar/netflixdataset

[3] S. Bhange. Netflix original movies. [Online]. Available: https://www.kaggle.com/datasets/swapnilbhange/netflix-original-movies