Visualization of Netflix movie dataset

Nataliia Holubtsova†
Faculty of Math
University of Waterloo
Waterloo, Ontario, Canada
nholubts@uwaterloo.ca

ABSTRACT

This article presents a report of final project of course CS 631 – Data-Intensive Distributed Computing. As the topic for the project, it was chosen the visualization of dataset, which describes the information about tv shows and movies, which were uploaded to the Netflix streaming platform in the period from 2014 to 2021. The paper describes 1) relevance of chosen topic and appropriacy of chosen dataset, 2) the methodology used in order to visualize data and get new insights from the given dataset, 3) evaluation of the results which were obtained during the work on the project.

1 Relevance

In the growing world of data, it is crucial to be able to get valuable information from the coming data and to be able to apply it in a beneficial way. Especially, it is important if it concerns insights, which can improve user experience in using various services. Netflix is one of the largest subscription-based streaming services, which is presented in 190 countries and at the first quarter of 2022 it had 221.64 million subscribers [1]. The platform has more than 9000 movies and TV series in various genres, such as comedy, documentary, horror, etc. Moreover, Netflix is known for producing its own shows, which are independent from the influence of large companies and often are created in countries which are considered to be lagging behind in film production. TV series Squid Game is a relevant example of such a production. It was made in South Korea and it became the most-watched show on the platform, and at a production cost of 21 million dollars, brought in about one billion in profits for the company [2]. However, it would not have been as successful as it was, if it not resonated with the public demand for the genre. Thus, the success of a television product depends on the public demand for it. The purpose of this project is to analyze data about the films and programs that are available on the platform to determine which shows can attract more viewers, thereby increasing the profits of the platform.

2 Methodology

2.1 Overview of used tools

This project was made with the usage programming language Python and Spark, which were installed locally in virtual environment in Anaconda. In order to perform aggregation function more quickly instead of resilient distributed datasets, which were mainly used during the course assignments it was decided to use Spark dataframes, where data is organized into named columns.

For creating visualizations such libraries were used:

- plotly – a graphing library, which creates interactive plots and diagrams;
- pandas – to create datasets, which plotly can use as the information for graphs;
- wordcloud – to visualize text data in order to find some ideas, which could be useful in terms of finding some insights.

Overall, these tools are sufficient to analyze and visualize the selected dataset.

2.2 Overview of dataset

Netflix movie dataset, which contains 8806 records, was chosen for the purposes of the project. Let us take a closer look to the origin and properties of the dataset. The data set was taken from the website Kaggle, which is known for having a massive number of data sets for studying data analysis and machine learning [3]. The raw data was web scraped from the Netflix platform through Selenium tool. It contains unlabeled text data of Netflix movies and TV shows. Data was saved as csv file, which contains 13 columns, namely show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, and description, as presented in Figure 1.

```
root
 |-- show_id: string (nullable = true)
 |-- type: string (nullable = true)
 |-- title: string (nullable = true)
 |-- director: string (nullable = true)
 |-- cast: string (nullable = true)
 |-- country: string (nullable = true)
 |-- date_added: string (nullable = true)
 |-- release_year: string (nullable = true)
 |-- rating: string (nullable = true)
 |-- duration: string (nullable = true)
 |-- listed_in: string (nullable = true)
 |-- description: string (nullable = true)
```

Figure 1: Schema of Netflix movie dataset

In order to understand purpose of columns better, it is necessary to describe each of them. Column "show_id" represents unique code of show; column "type" represents kind of show, it contains two values "movie" and "tv shows", which vary from record to record; "title" contains information about title of the show; column "director" contains names of the directors; "country" is a country of show origin; "date_added" is date of uploading show to the platform; "release_year" is the year when show was released; "rating" contains data about motion picture content rating which classifies shows based on their acceptance by the different age groups; "duration" represents duration of product – for movies it shows minutes, while for tv shows it is measured in seasons; "listed_in" is a string which represents a list of genres, which a product has; finally, "description" is the description of the show.

2.3 Cleaning dataset

Next step in exploring dataset is to find out how much records in different columns have null values. For this purpose, pyspark SQL query was written. For instance, for column "type" it works in such a way: with the statement COUNT(*) we count all the records, even if they have null values, with the statement COUNT(type) we count only values distinct from null. As a result, by subtracting one value from another we obtain number of rows with null values. The result of such manipulations is shown in figure 2.

type	title	director	cast	country	date_added
0	0	2634	825	830	11

release_year	rating	duration	listed_in	description
0	4	3	0	0

Figure 2: Number of rows with null values

As it could be seen from the figure 2, the column, which has the biggest number of null values is "director" with total value – 2634, then "cast" and "country" have approximately 830 values, which is also quite a lot. Other columns do not a lot of null values.

There are several techniques to overcome the problem of null values, namely deleting rows with empty values, replacing such values with default values or if it is numerical column – replace null values with average one. All of the columns with massive number of null values have string type, therefore take average value was not an option. At first it was decided to remove all the rows with null, but in case of doing this the representativeness of dataset would be impacted since the number of TV shows which have no director is 2446 and number of movies is 188. Thus, it was decided to replace nulls with default value – "not defined". The result of such replacing is shown in the figure 3.

show_id	type	country	director
s2	TV Show	South Africa	not defined
s4	TV Show	not defined	not defined
s5	TV Show	India	not defined
s11	TV Show	not defined	not defined
s15	TV Show	not defined	not defined
s16	TV Show	United States	not defined
s18	TV Show	Mexico	not defined
s20	TV Show	not defined	not defined
s22	TV Show	Turkey	not defined

Figure 3: Result of replacing nulls with default value

Next step in cleaning data is to deal with genres, which are situated in the column "listed_in". This action is needed for several reasons: 1) there are multiple genres, which are problematic in use since they are presented as one string; 2) another problem is that TV shows and movies are named differently, although they come from the same genre. For example, "Comedies" and "TV comedies" are the same but named differently.

To overcome these issues, it was decided to extract first genre as the main one and to map genres, so that they could show more information about the product.

For these purposes it was decided to use such function as explode which is used to explode or map columns to rows. Then, there was another problem – some strings had spaces, so the function "trim" was used. After that the map dictionary was created, it is shown in the figure 4.

```
genres_dict ={'Documentaries':'Documentaries',
     'British TV Shows': 'International',
     'International TV Shows':'International',
     'Crime TV Shows':'Crime',
    'Docuseries':'Documentaries',
     'TV Dramas':'Dramas',
     'Children & Family Movies':'Children & Family Movies',
     'Dramas':'Dramas',
     'Comedies':'Comedies',
     'TV Comedies':'Comedies',
     'Thrillers':'Thrillers',
     'TV Thrillers':'Thrillers',
```

Figure 4: Example of dictionary for mapping

Finally, we received a dataset with clean genres, yet there was another problem – because of explode() applying a lot of duplicated records with distinct genres were created, so it was decided to remove duplicated rows by applying the function dropDuplicates() to the column “show_id”. The resulting dataset is shown in the figure 5.

show_id	type	title	genre
s1	Movie	Dick Johnson Is Dead	Documentaries
s10	Movie	The Starling	Comedies
s100	TV Show	On the Verge	Comedies
s1000	Movie	Stowaway	Dramas
s1001	Movie	Wild Dog	Action & Adventure
s1002	Movie	Oloibiri	Dramas
s1003	Movie	Tell Me When	Comedies
s1004	TV Show	Zero	International

Figure 5: Cleaned dataset

To summarize, we received a dataset, which is clean and can be used to compute analytics and visualization.

2.4 Visualizing information

For understanding what product is better to make, let us start from the simple visualization of how many movies and tv shows were uploaded to the platform in 2021. For this purpose, the resulting dataset was queried in the way, which is shown on the figure 6.

```
import plotly.express as px

pandasDF = spark.sql("SELECT type, COUNT(show_id) num_shows FROM data WHERE"
                     +" date_added LIKE \"%2021%\"  GROUP BY type").toPandas()
fig = px.bar(pandasDF, x='type', y='num_shows')
fig.show()
```

Figure 6: Query to extract information about the type of shows uploaded in 2021

As it could be seen from the figure 6, we counted number of shows with different types and filtered it by year of uploading. We need to know the latest trends, therefor 2021 year was chosen for the query. Then we turned pyspark dataframe to pandas, it was made due to the fact plotly does not work with pyspark. After that, we created a bar plot to show the difference evidently. X axes shows type of TV ptoducts, while Y displays number of shows, which uploaded this year. The resulting graph is presented in the figure 7.

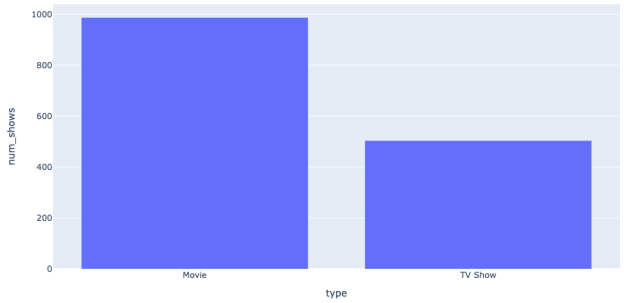

Figure 7: Movies and TV shows difference

As we can see from the plot, number of movies is almost two time bigger than TV shows. Therefore, it could be concluded that films are made more often than TV series.

Now let us take a look what countries were the main gamers in the field during 2021. This information is beneficial to know since producers are interested in creating contracts with those states who are able to make films and tv series in a stable, productive rhythm. To create such a query firstly we explode countries and clean them from spaces, then we filter by year 2021,then we count shows which came from these countries. Netflix is a worldwide company which distributes films from all over the world, yet there is a large number of countries which produce few films or shows per year, therefore it was decided to group countries with number of films lower than 30 per year in the category “others” by using construction when()…otherwise(). The resulting pie chart is shown in the figure 8.

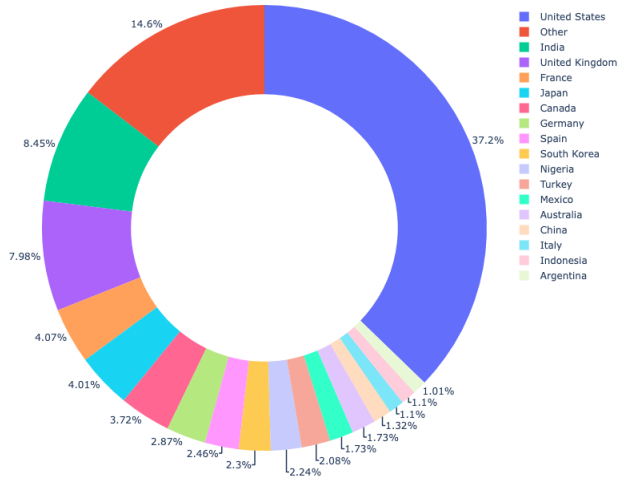

Figure 8: Most productive countries in the film industry

As it could be seen from the figure 8, most productive country is the USA. The U.S. has always been considered a leader in film production, but the films produced in this country illuminate one point of view. To make the content more inclusive, it might be worth looking at countries like India or Japan, as they also produce movies and TV shows steadily and have a large audience.

Now let us look at the relation between product rating and its type. This information is useful for two reasons: firstly, we will understand which rating was the most popular in 2021 and secondly, we will see what type of product suites best for this rating. For this aim a bar chart was created and it is presented in the figure 9.

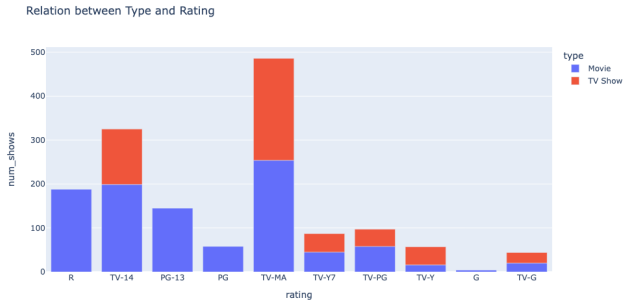

Figure 9: Relation between type and rating

As it could be seen from the figure 9, the most popular rating is TV-MA, which is assigned to films for people over the age of 17; next most popular rating is TV-14, which is assigned to shows for people over 14 years old. In both these categories TV series lead. Interestingly, that R-rating has only movies.

It could be concluded, that if a producer wants to make a movie or TV show for Netflix, it is better to make it with rating TV-MA or TV-14.

Another important fact to consider is what genre is most popular on the platform. This information is needed to decide what film or series could be commercially popular on the platform. Again, we will use bar chart to show what genre is most common on the platform. The result is presented in figure 10.

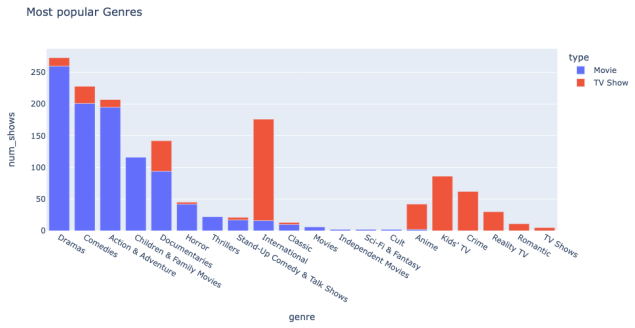

Figure 10: Relation between type and rating

As it could be seen from the figure, dramas, comedies, and action are most common genres on the Netflix, so in order to make a commercially successful show it would be a smart decision to use one of these genres.

We found out most common genres, next step is to find most productive directors in this area during 2021. This data is important because to make a salable product those directors who can present as much ideas as possible needed. Therefore, a query which counts a number of films or shows, which were made by a certain director, among most popular genres, was created. Its visualization presented in figure 11.

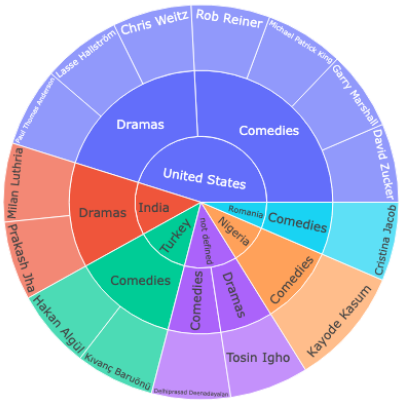

Figure 11: Most productive directors in 2021

As it could be seen from figure 11, most productive drama-makers are Prakash Jha and Chris Weitz, who made two films in 2021. They come from India and United States, which are known for their productiveness in the film industry. Thus, this information could be used, while choosing a candidate for a director role.

Another useful graph which can help producers to learn what topics are most common in the certain genre is a word cloud. Word cloud presents words based on their frequency. The more often a particular word appears in the sentence the bigger size it will have. This information would be useful in brainstorming ideas for new movies or shows. Let us create word clouds from descriptions divided by genres. The result is presented in the figure 12.

Figure 12: Word cloud based on the description divided by genres

As it could be seen, most popular words in drama genre are “love”, “life”, and “young”, while in comedies they are “life”, “new”, “find”. Interestingly, word clouds in these genres are similar. It seems, that it happens since comedies and dramas presents same stories, but from different point of view.

Overall, the dataset was analyzed enough to make next conclusions. To create commercially successful product, it is needed to produce a movie in genre drama or comedy with the rating TV-MA or TV-14. Directors preferably should come from India, Japan or the USA and they should create at least two movies per year.

3 Evaluation of the results

The results visualized a set of raw data in order to identify insights that can help make a commercially successful product. However, there are several drawbacks to this study.

First, the raw data does not include information on user ratings and the number of views of a particular film. This information could improve query results and help select not only the most productive filmmakers, but also those who are the most popular with viewers.

Second, as one of the improvements to the original study, we could suggest the use of a world map to provide a more visual representation of the countries producing films and TV series.

It is worth noting that the use of Spark for the analysis of this dataset is justified, as IMDB and other streaming services have similar data sets. Thus, this code can be scaled for large amounts of data.

As a further improvement, we can propose to add analysis of actors and select the most popular ones, as well as to analyze the average duration of movies and shows in order to select the most optimal values.

REFERENCES

[1] Netflix statistics - truelist 2022. TrueList. (2022, August 5). Retrieved December 14, 2022, from https://truelist.co/blog/netflix-statistics/#:~:text=Netflix%20had%20about%20221.64%20million,compared%20to%20the%20previous%20quarter.

[2] Perez, C. (2021, November 5). How much money did squid game's creator really make from netflix? Looper. Retrieved December 14, 2022, from https://www.looper.com/651415/how-much-money-did-squid-games-creator-really-make-from-netflix/

[3] What is Kaggle, why I participate, what is the impact?: Data Science and Machine Learning. Kaggle. (n.d.). Retrieved December 14, 2022, from https://www.kaggle.com/getting-started/44916