Project CPSC 6030 – Data Visualization of Netflix OTT Platform

Group - 14:

Avinash Komatineni (akomati@clemson.edu) - C13513954

Harika Banda (hbanda@clemson.edu) - C11992641

Pandu Ranga Avinash Srikhakollu (psrikha@clemson.edu) - C37573694

Project Links

GitHub Page link: https://psrikha.github.io/

GitHub Repository link: https://github.com/psrikha/psrikha.github.io

Video Recording link:
https://clemson.zoom.us/rec/share/iligY9RRnEHuAHG9ujcp7BiGGsuYF4lC6_p8Wah8XI9YSaz80ovaMHw5pQW-LYJJ.XLlGM10XaQC-kiI7?startTime=1671160812000

Overview

In today's world, OTT media service has become one of the major sources of streaming in aspect of movies or TV shows. The impact created by these OTT platforms on people around the world is increasing exponentially. Also, revenue generated by these platforms compensate more than 50% of income earned by the film or series. Hence analyzing and visualizing the data of these OTT platforms based on certain factors can help people understand what to watch and also help the content creators to choose between releasing dates.

So, the dataset we have considered is related to Movies and TV shows streaming on OTT platforms: Netflix. This is existing data collected from WWW. Data Analytics on data based on OTT platforms plays a significant role in marketing of the TV Shows and movies.

Dataset

As a part of Dataset selection, we have taken Dataset which contains 11 attributes and over 3000+ items. The Dataset we have chosen clearly helped in our design process without much preprocessing.

Attributes list for our Dataset:

1. *Type*
2. Title
3. director
4. cast
5. *country*
6. *date_added*
7. *rating*
8. *released year*
9. *duration*
10. *listed_in*
11. *description*

Items list for our Dataset:

8808 unique show ids

Preprocessing of the Data:

The dataset we have taken needed some modifications to our problem statements so we have done basic Data processing and Data cleaning. As a part of Data cleaning, we have removed null values of the attributes. Though we have removed all of the null values it didn't affected the quality of our dataset and we are able to extract all data we want for our visualizations.

After visualization, we soon realized that attribute 'rating' in our dataset is not a kind of rating based on the performance of the show but the rating related to what age group are preferred to watch the show so we have modified the attribute as we desired by cloning with similar dataset.

Design Process

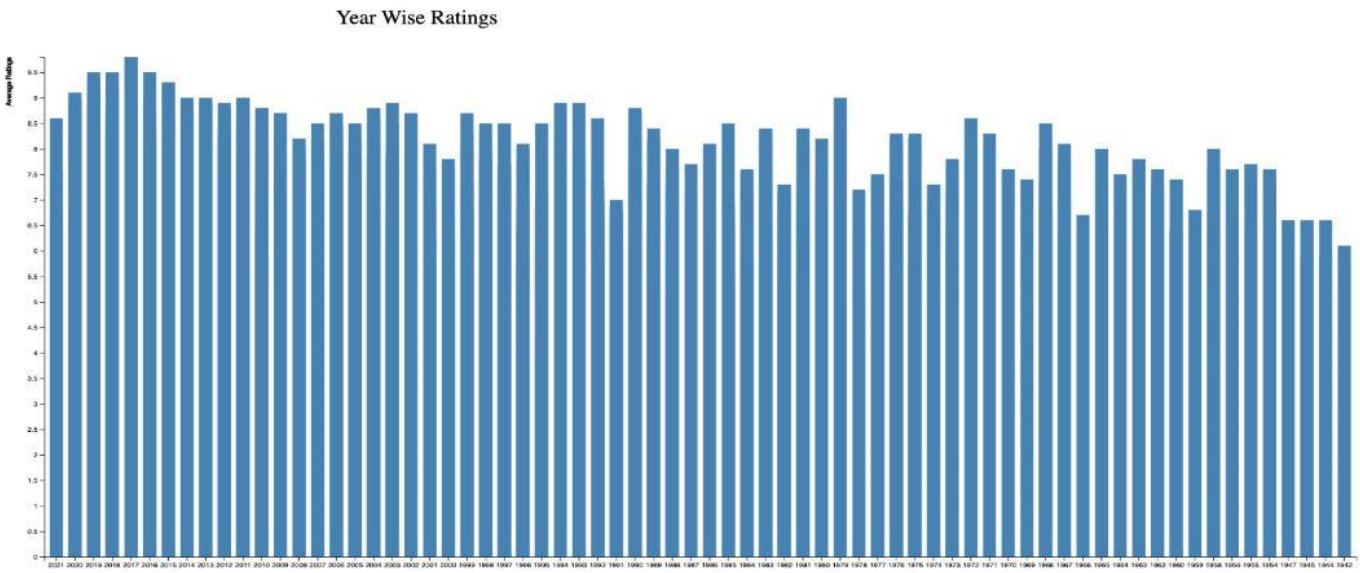
As a part of design process, we have drafted four visualizations. In each of the visualizations, we have used two or more than two attributes from our dataset and presented the design for all the items in the datasets.

To the best of our ability, we tried to make all the visualizations interactive with each other. We went back and forth during design phase and made lot of changes in the process which we have discussed further in detail.

First Draft Phase

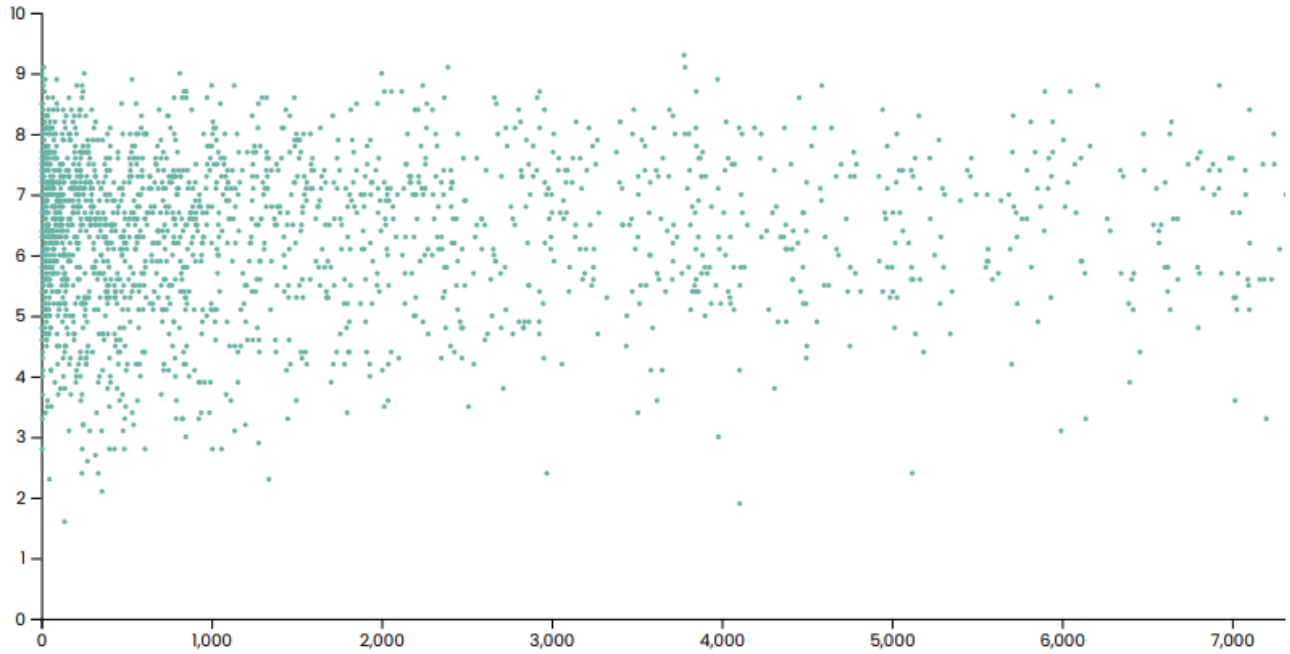
In Draft 1 phase, we have designed four visualizations. Though we were able to successfully get four visualizations, we have changed the visualizations in Final Draft Phase as visualizations are not interactive and not conveying much valuable information to end user. Personally, we felt with little more effort, we can make more appealing visualizations.

Visualization 1:



Visualization 1 shows the Average ratings of the total shows in a year. We have taken attribute 'Average rating' and 'released year' from our dataset. We have dropped this design as we realized that it is not giving any useful message to the end user.

Visualization 2:



Visualization 2 gives information about number of votes registered for a rating to a specific movie. In draft 2 phase, we have improvised it by adding interactions such that when we hover the circle it gives information about the movie, average rating and number of votes.

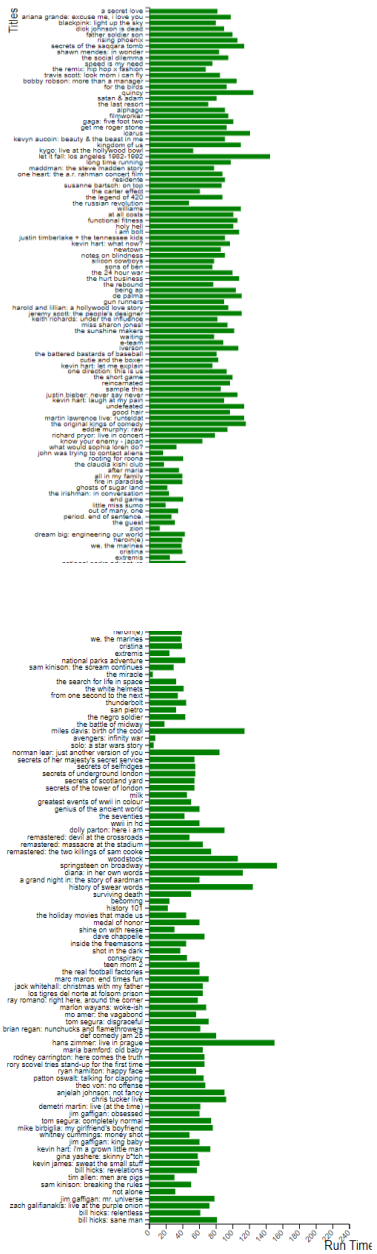
Visualization 3:

Visualization 3 gives information about the run time of the shows in the dataset. We have placed filter on genre attribute so that end user can filter shows based on genre.

It took good amount of time for us to implement filter function in our source code. Below is the code snippet related to filter function in our code.

To make the visualization more appealing, we have added scroll bar to our design.

Running Times of Shows in Documentry Genre



Activate Windows
Go to Settings to activate Windows.

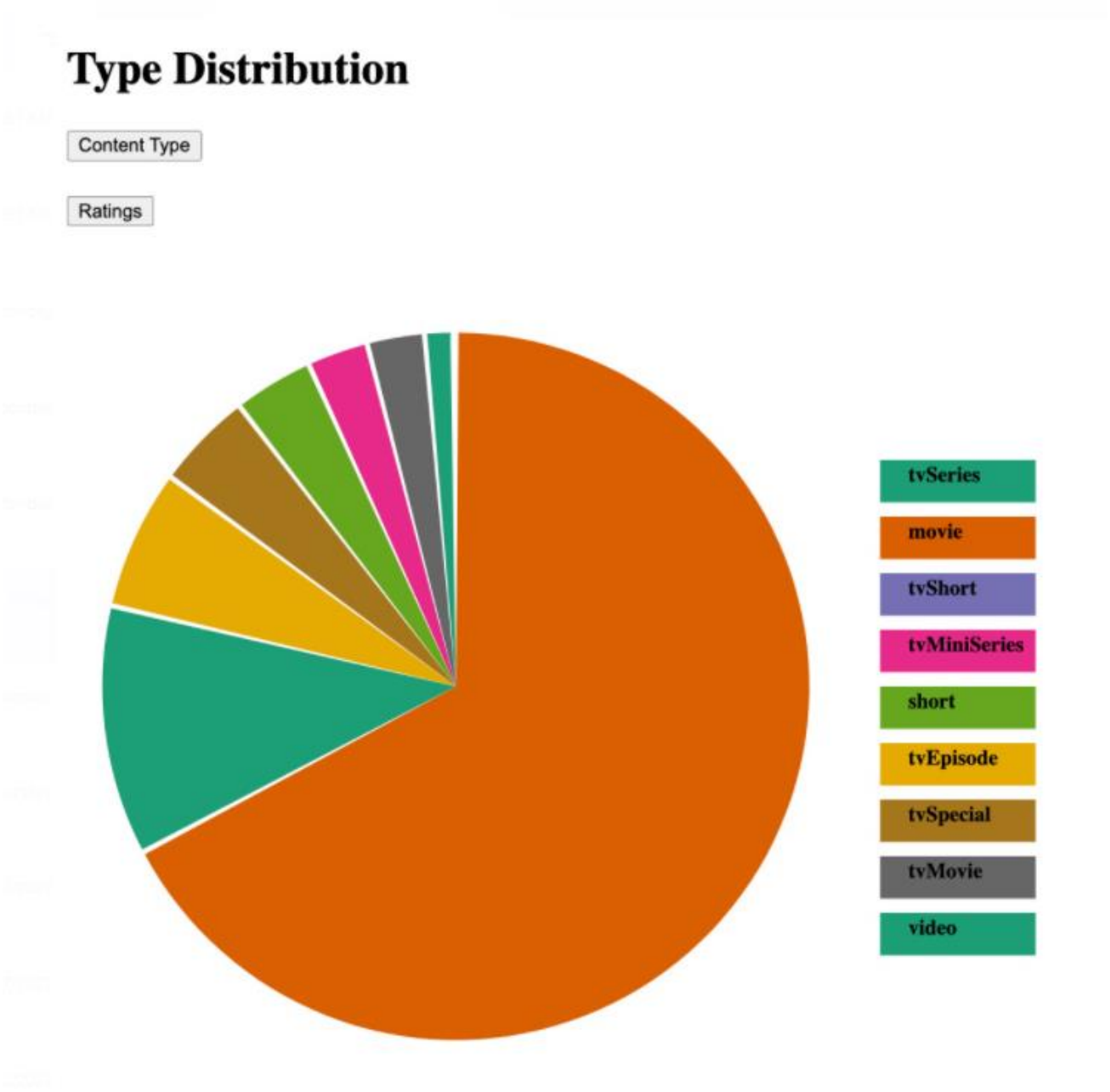
```
var genres="Documentry";

// Parse the Data
d3.csv("clean_df311.csv").then(function(data){
//d3.csv("clean_df311.csv", function(data) {
var filterdata= data.filter(function(d,i){

  if (d["genres"] == genres)
            {
                return d;
            }

});
```

Visualization 4:



As a part of visualization 4, we have designed pie chart based on attributes: Content Type and Title. Design has been improved by switching to Average Ratings attribute from Content type attribute if required by introducing buttons in the source code.

We have taken considerable amount of time to group based on Content type attribute and get the count of Titles attribute. Below is the related code snippet we have coded in our implementation.

```
d3.csv("clean_df3.csv", function(error, data) {
            if (error) {
                throw error;
            }
            const arr = []
            for(var i=0; i<data.length;i++){
                arr.push(data[i].type);
            }
            var data1={};
            for (var j=0;j<arr.length;j++){
                var num = arr[j];
                data1[num]=data1[num] ? data1[num]+1 :1;
            }
            console.log(data1);

            const arr1 = []
            for (var k=0; k<data.length;k++){
                arr1.push(data[k].rating);
            }
            var data2={}
          for (var m=0;m<arr1.length;m++){
                var n=arr1[m];
                data2[n]=data2[n] ? data2[n]+1 :1;
          }
          console.log(data2);
```

Final Draft Phase

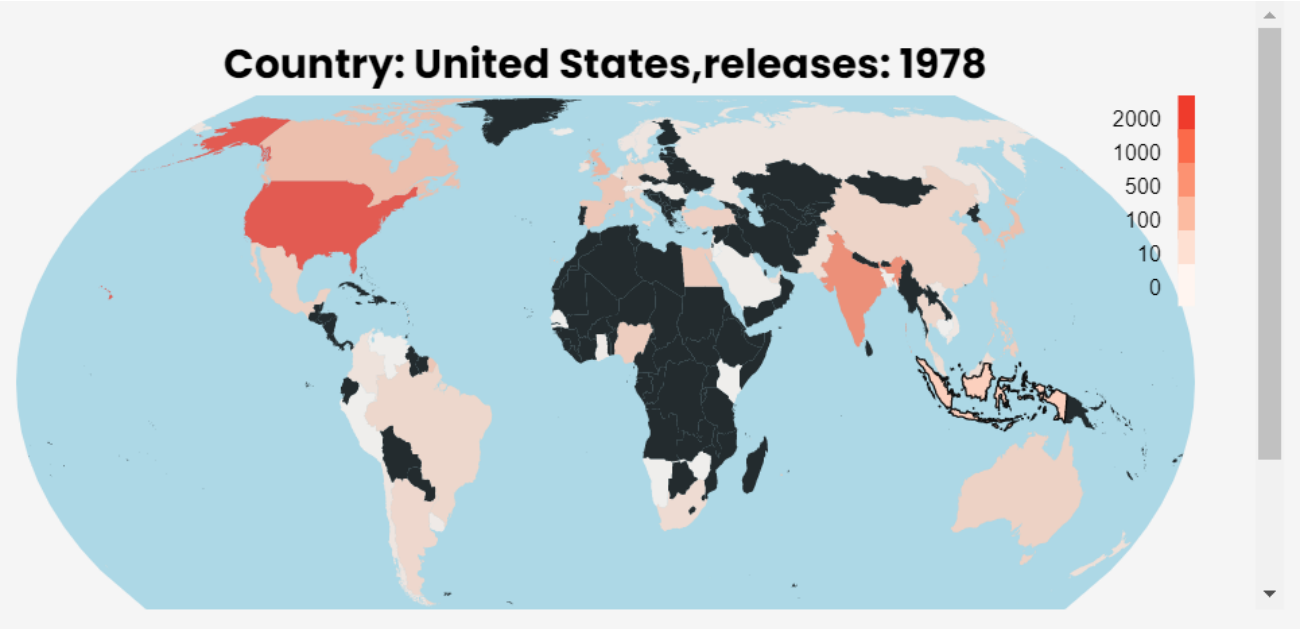
During Project Prototype when we presented our first draft, we got the feedback as follows from the professor which we have taken into consideration and implemented in our final design.

1. Visualizations should be more interactive
2. More flexible should be given to end user while he tries to get the data from the visualization
3. The Prototype would be more compelling if one visualization can interact with another visualization
4. Visualization 4 should be changed as pie chart would be visually good only if number of values are three to five but not more than that as it is hard to convey information about the data which has low number of values

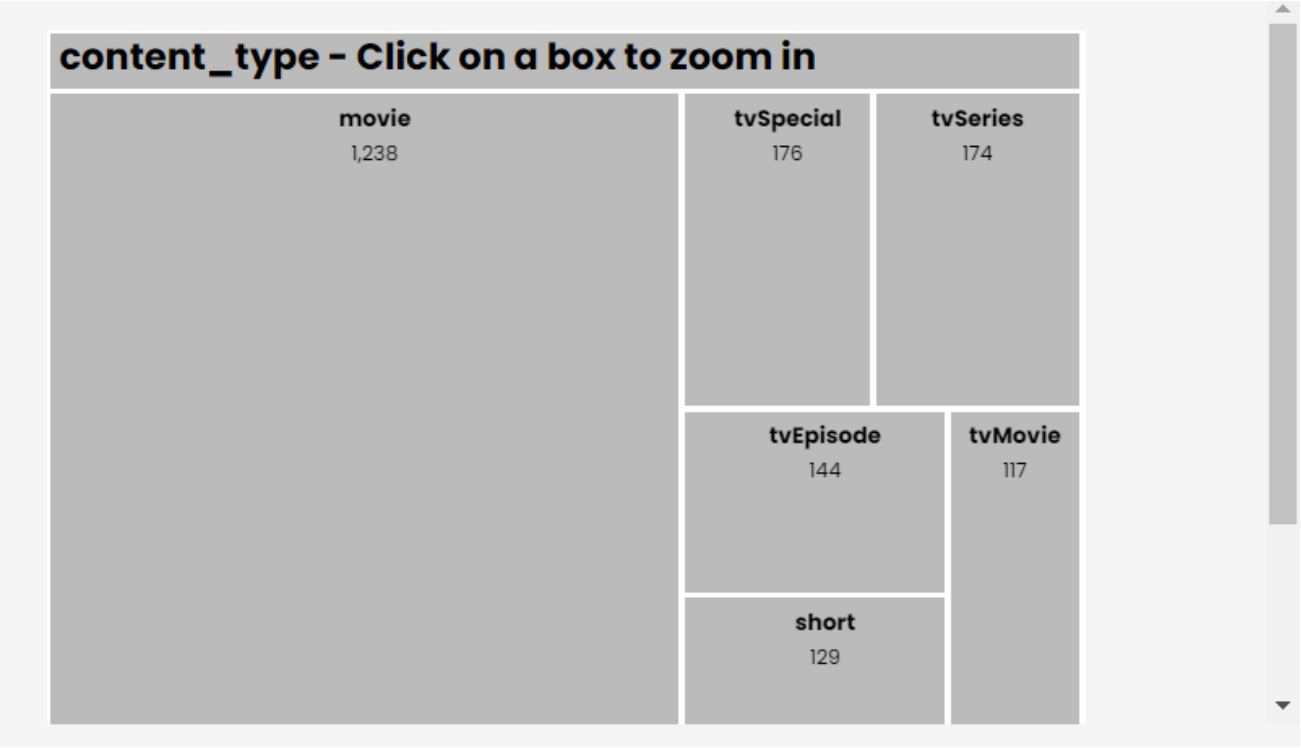
We have designed our final visualizations by taking the feedback and made changes to our visualization such that:

1. All four visualizations can interact with each other such a way that one click in one visualization changes the dynamics of other three visualizations
2. Gives valuable and clear data to the end user
3. Used maximum of attributes in our dataset and used functions like count(), filter(), group(), click() and interactions functions to make near perfect visualizations

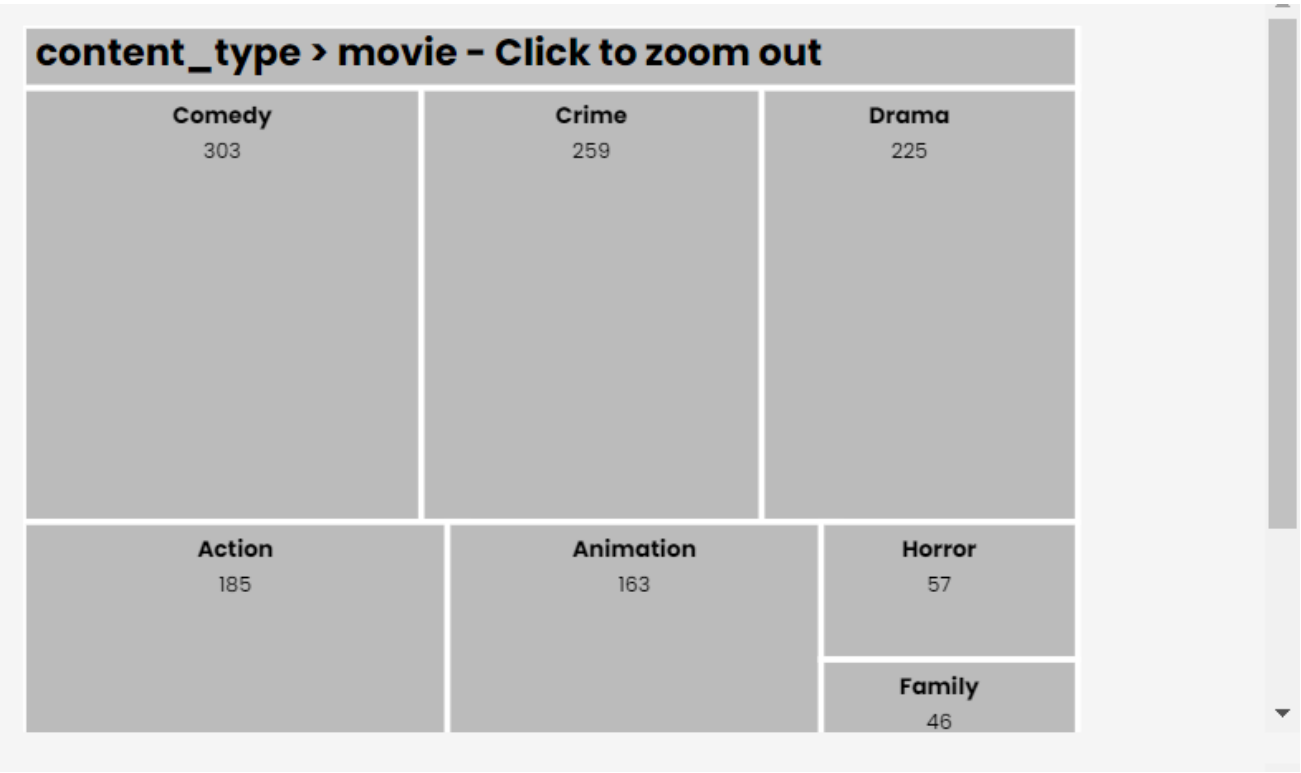
Visualization 1 is map which represents number of shows released in a particular country. We have used attributes Country and Title. When we click on a specific country, it gives the details about number of releases in that country.



On the other side, **Visualization 2** is a heat map which displays the count of type attributes like movie, tvSeries, tvSpecial and short etc., and when clicked on desired type it transforms and displays the count of genre attribute. So using visualization 2 end user get details like number of type of shows present in a country and upon one click, gives number of shows in each genre.

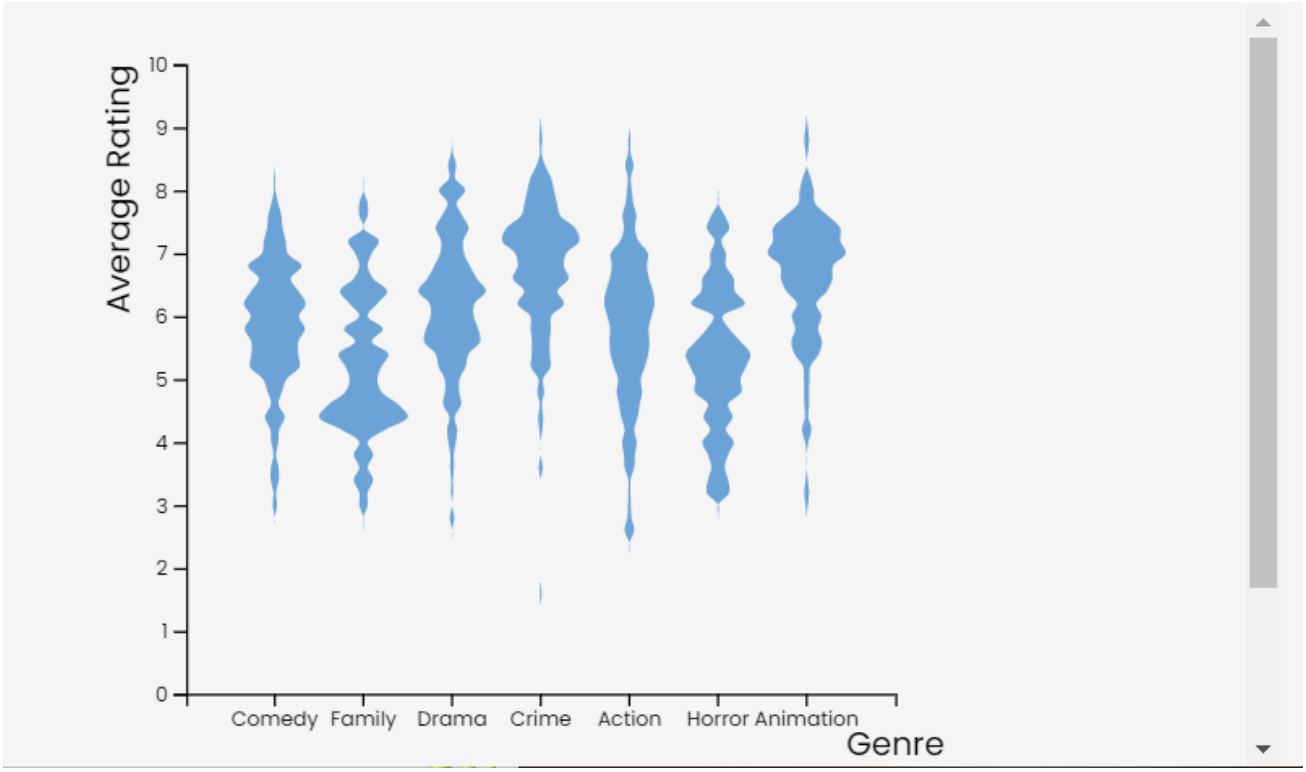


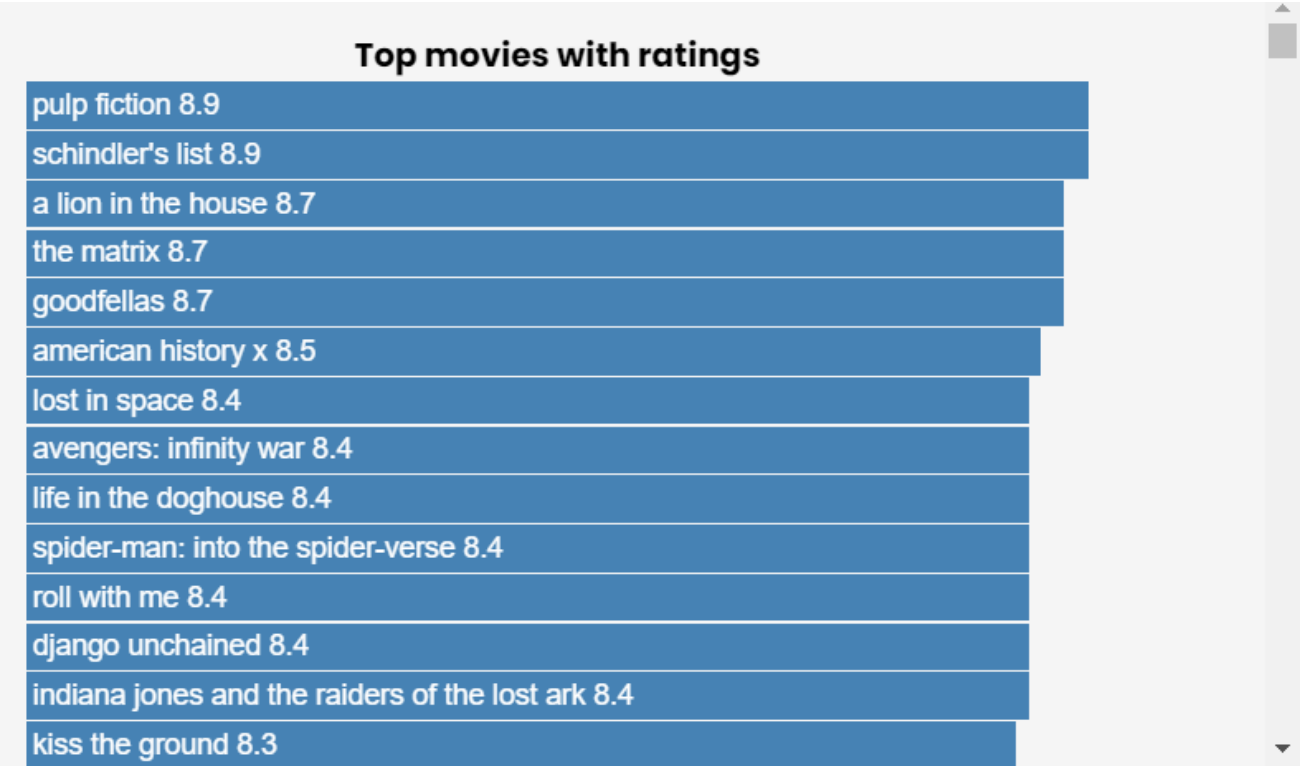
Upon click,



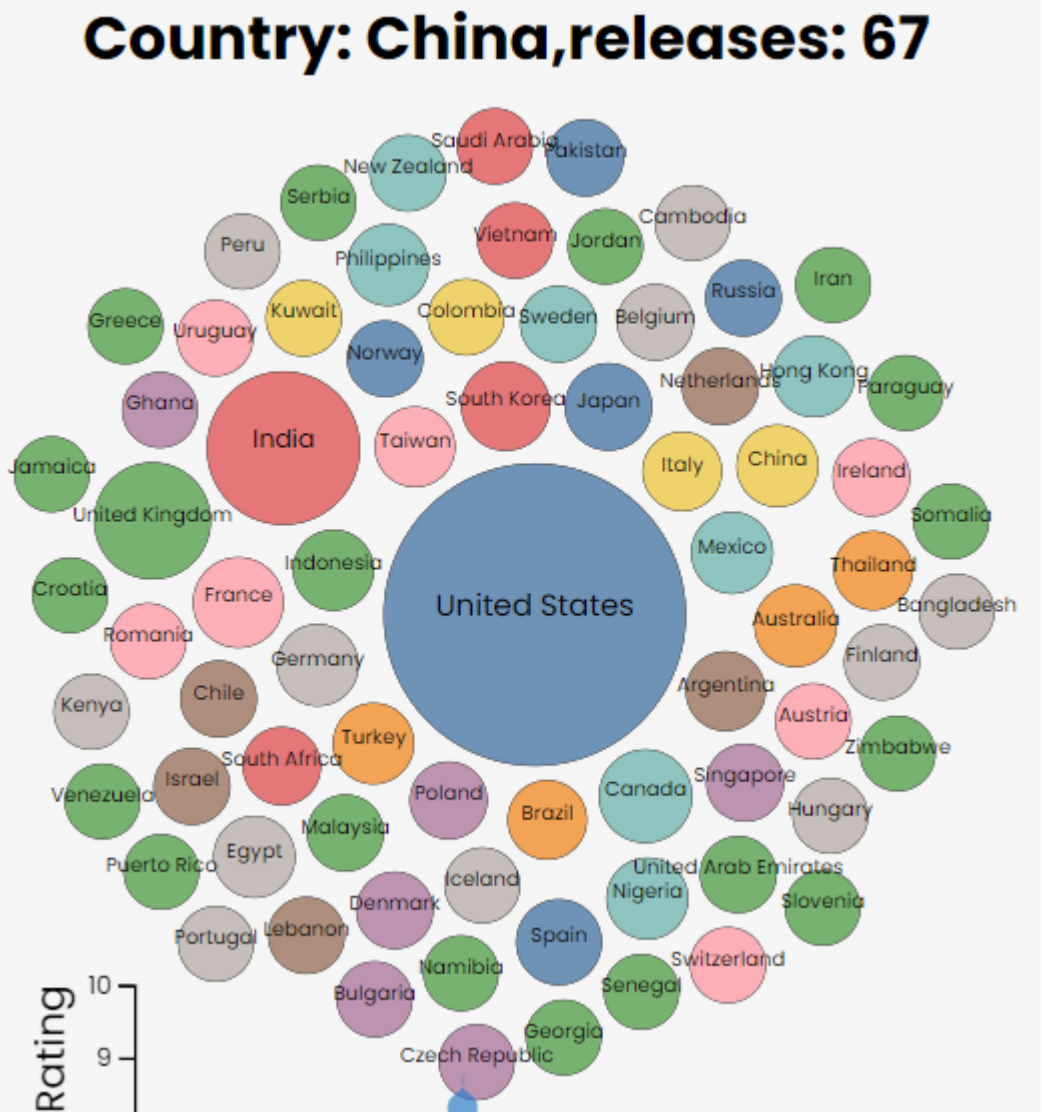
Visualization 3 is a simple violin chart gives details about the average rating on the basis of genre utilizing the attributes average rating and genre.

On selecting the desired country, type and genre from visualization 1 and visualization 2 by the user, Visualization 4 gives the top-rated show names with ratings. **Visualization 4** contains Title and performance-based ratings attributes.





After Final Presentation, we have taken feedback from Professor that if first visualization is changed to something like Bubble chart then it would be perfect from design perspective. So, we have changed the visualization from map to Bubble chart in last hour.



In conclusion, End user without the knowledge about the dataset and design process can easily navigate through the design flow and get the data.

Credits

Thanks to professor Federico luricich for regular feedback😊

https://www.kaggle.com/datasets/shivamb/netflix-shows?resource=download

Contributors: *Subin An, Niharika Pandit and Raenish David*