Assignment 3

Topic: Netflix Tv Shows and Movies Dataset
This dataset comprises of programs and motion pictures accessible on Netflix starting at 2019. The dataset is gathered from Flexible which is an outsider Netflix web crawler.

Purpose:
In this isolated period, I understood part of individuals are spending in viewing the Netflix web series and movies. So, it intrigued me to do information investigation on this dataset. The objective of the data is to break down the intrigue bits of knowledge of TV Shows and Movies as of late. The dataset comprises of 12 columns and 6172 rows.

My variables are

1. **Show_id** which consist unique values
2. **Type** which says whether the content is tv show or movie
3. **Title** which consist name of the shows and movie.
4. **Director**
5. **Cast** includes actor and actress names.
6. **Country**
7. **Date_added** explains when the content was added on Netflix.
8. **Release_Year** includes the data from 1925 to 2020
9. **Duration** explains movie duration in mins and TV shows in seasons format.
10. **Listed_in** talks about the genre of the shows and movie
11. **Rating** explains about TV Ratings like, TV-MA, TV-14
12. **Description** is summary of the shows.

These are my few questions I have analysed for which I have done **EDAs**

1. What are the most searched keywords?
2. What is the most content watch?
3. Which year has the highest release
4. Which country has the highest number content
5. What is the duration of most of the movies?
6. Quadrant analysis on rating, country and genre.
7. Which is the highes rating watched movie and TV Show
8. Top 20 directors with most country with respect to country.
9. Which TVShow has the highest duration with respect to Country?

Aksha Iyer

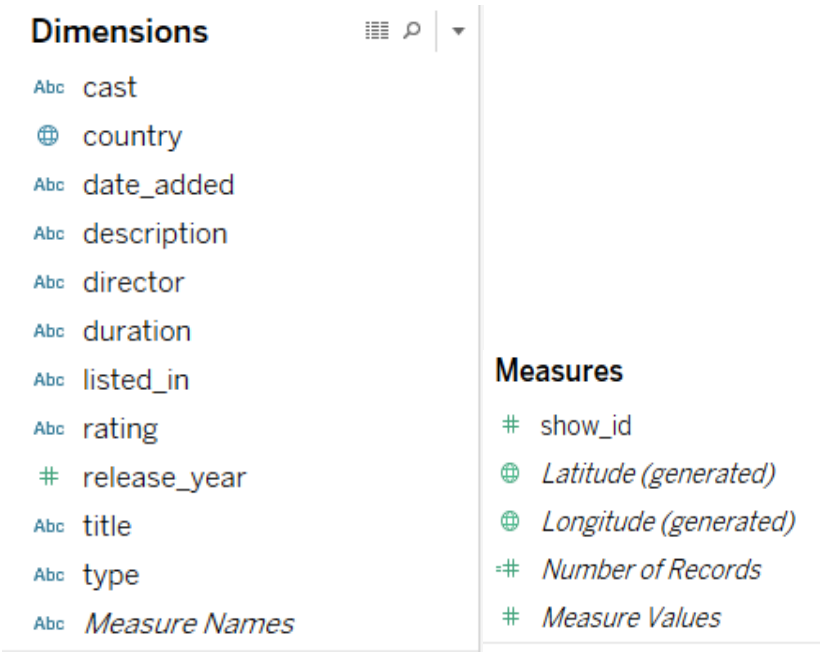
Pre-processing and Overview of the Dataset.

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
81,145,628	Movie	Norm of the Nort...	Richard Finn, Tim...	Alan Marriott, A...	United States, In...	September 9, 2019	2019	TV-PG	90 min	Children & Famil...	Before pl
80,117,401	Movie	Jandino: Whatev...	*null*	Jandino Asporaat	United Kingdom	September 9, 2016	2016	TV-MA	94 min	Stand-Up Comedy	Jandino A
70,234,439	TV Show	Transformers Pri...	*null*	Peter Cullen, Su...	United States	September 8, 2018	2013	TV-Y7-FV	1 Season	Kids' TV	With the l
80,058,654	TV Show	Transformers: Ro...	*null*	Will Friedle, Darr...	United States	September 8, 2018	2016	TV-Y7	1 Season	Kids' TV	When a p
80,125,979	Movie	#realityhigh	Fernando Lebrija	Nesta Cooper, Ka...	United States	September 8, 2017	2017	TV-14	99 min	Comedies	When ner
80,163,890	TV Show	Apaches	*null*	Alberto Ammann...	Spain	September 8, 2017	2016	TV-MA	1 Season	Crime TV Shows, ...	A young j
70,304,989	Movie	Automata	Gabe Ibáñez	Antonio Bandera...	Bulgaria, United ...	September 8, 2017	2014	R	110 min	International Mo...	In a dysto
80,164,077	Movie	Fabrizio Copano: ...	Rodrigo Toro, Fra...	Fabrizio Copano	Chile	September 8, 2017	2017	TV-MA	60 min	Stand-Up Comedy	Fabrizio C
80,117,902	TV Show	Fire Chasers	*null*	*null*	United States	September 8, 2017	2017	TV-MA	1 Season	Docuseries, Scie...	As Califor

Data Cleaning:

Using the **Data Interpreter** option, I cleaned the data. It can assist you with cleaning your information if your source is not in an ideal table arrangement. This component distinguishes arranging, for example, void sections, blank area, titles, stacked headers, notes, and so on, and with a tick of a catch, consequently reshapes the information into a 'spotless' table with segments and columns, prepared for analysis. To remove the Null values while plotting the EDAs I have excluded them.

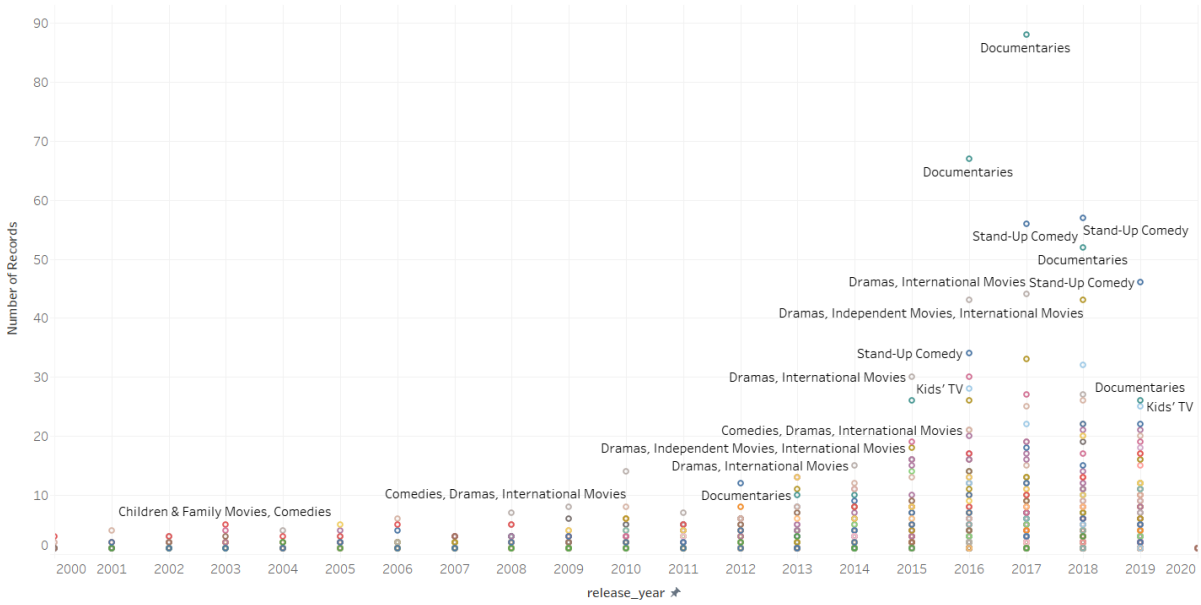
Data Types of my default variables are as given below.



Dimensions **includes Cast, description, director, listed_in, rating, title and type** are clearly in **string data type** as they contain alphabets and here **date_added** is also **string datatype** because it contains mixed value of numeric and alphabet, it can be changed into date datatype if needed. While **release_year** is in **numeric data type** if needed it can change it do date datatype as per EDA required.

While my aggregated field are **show_id, Number of records** which are in **numeric datatype**. While **latitude and longitude** are **geographic datatype**.

Outliers



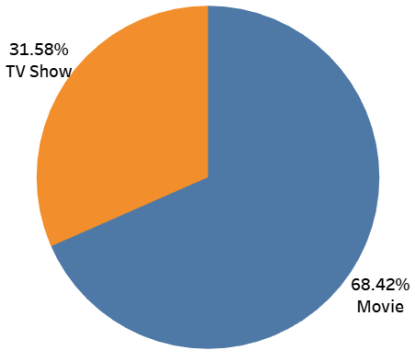
An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. We can clearly see that in all three years that is from 2017 to 2019, Documentaries falls on outliers. Which means most of the people did not prefer watching those kinds of genre. While 60% preferred watching Stand-up-Comedy, International Movies and Dramas.

Exploratory Data Analysis

1. **What are the most searched keywords?**

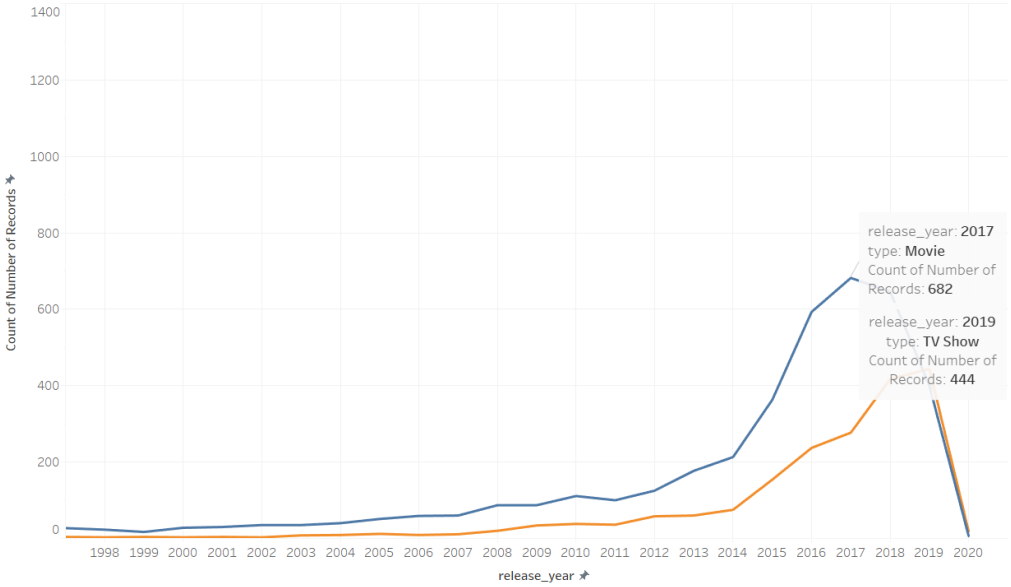
I utilized the Title column from the dataset and pivoted it to the Actual Text using pivot function which gives me unmistakable words. The Actual Text segment is in string datatype. I made a parameter name Top which will review the main 30 watchwords looked by User. I used Actual Text section in Filter, Colour and Label and Count of number of records in Size. We can clearly see that Love, Christmas, Man, Black, American, Girls are the most used search word.

2. What is the most content watched?



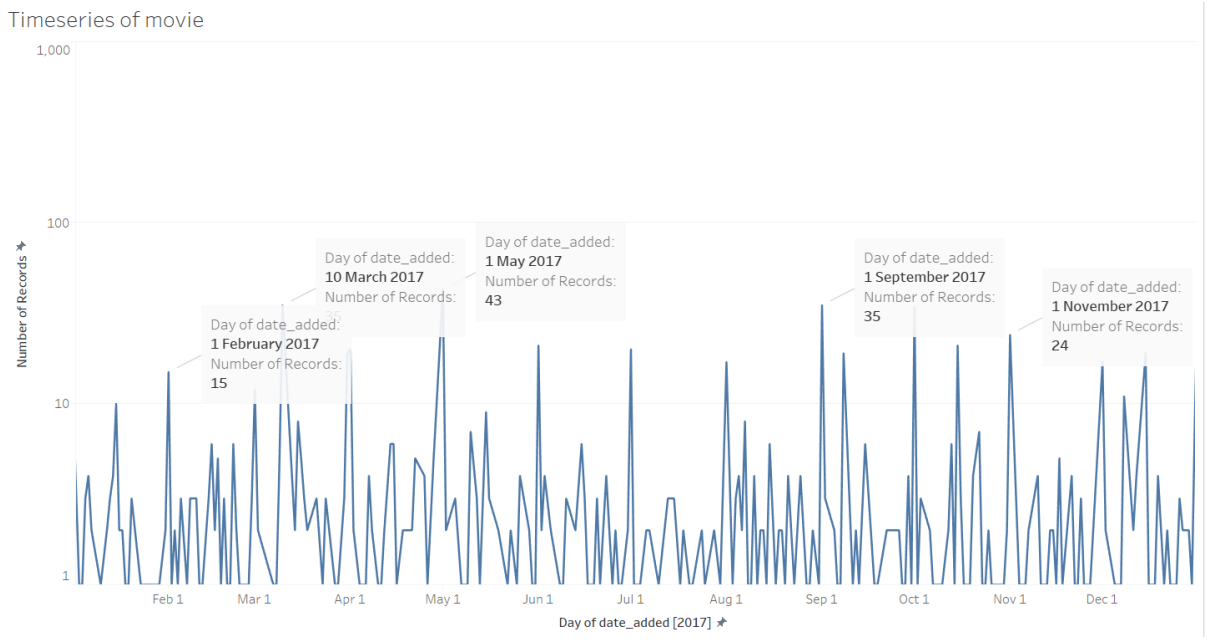
I used pie chart to visualize as it gives clear understanding of percentage category. And clearly, we can see 68.42% of users watch movie while 31.58% users watch TV show. I implemented this using by Type and Number of Records columns. I used the Type column in colour and label and measured the sum of the Number of the Records and used it in Size and Angle.

3. Which year has the highest release?

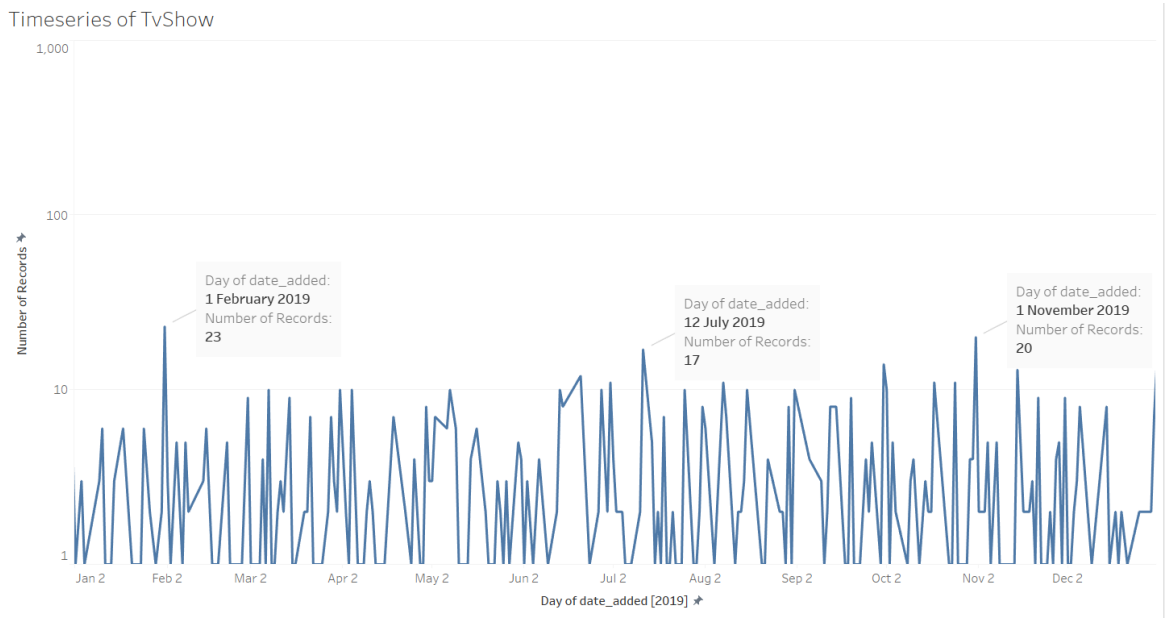


We can see that highest Movie content was released on 2017 and TvShow content on 2019. While 2020 has the lowest content due to Covid-19 majority of releases has been in hold. Reason of releasing highest movies on 2017 could be the newly emerged Amazon Prime which gave strong competition to Netflix. Also, the information shows Netflix buying and investing more in motion picture licenses for the clients. I did this analysis by using type in colour where Blue line is for Movies and Yellow line for TV Show. Comparison is done between release year and Number of count where I kept Unit as Year for release_year column.

Timeseries analysis for Movies on 2017



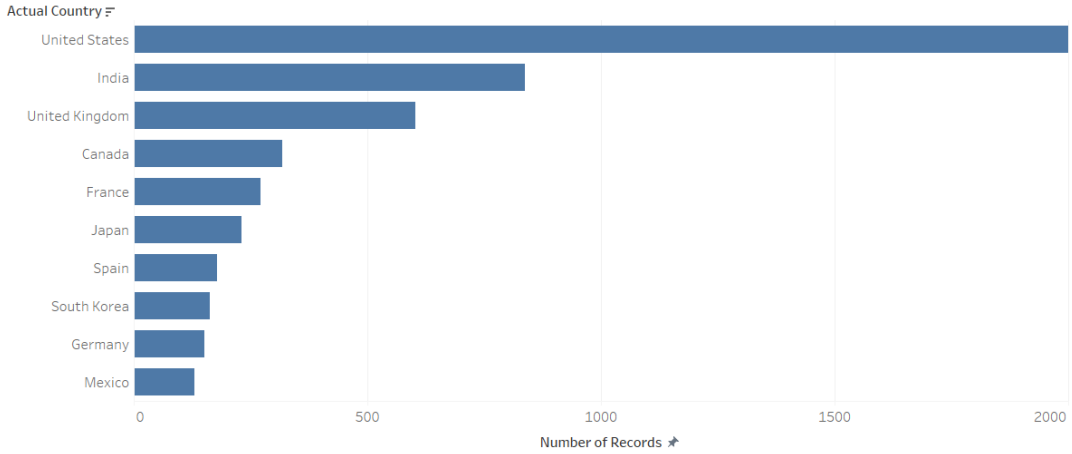
Timeseries Analysis for TV Shows in 2019.



We can see that February month has slightly high contain as its Valentines month. Similarly, majority of the content is added in May month as its summer break. While in November we can see high records reason could be thanksgiving special.

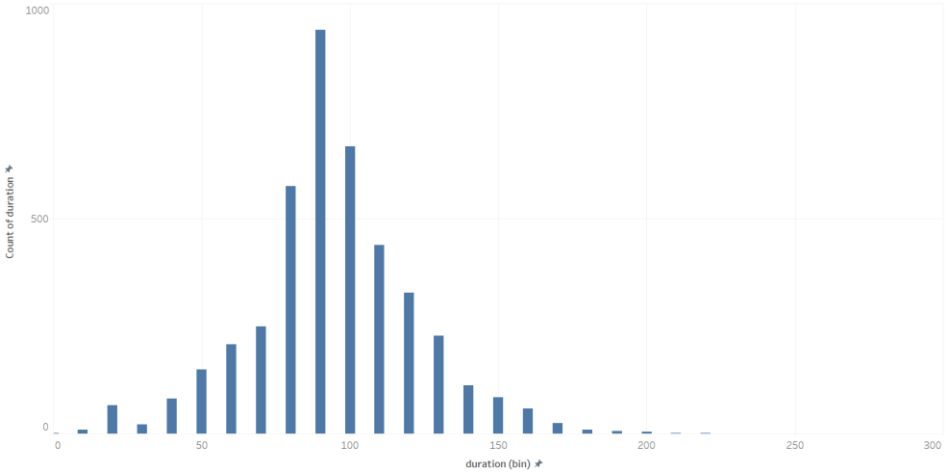
Here the date_added dimension was in string data type which was changed it to Date datatype in order to perform timeseries analysis. I implemented this analysis by filtering it with type and date_added columns. I changed the date_added column to measure and compared with Sum of the Number of Records. Units of date_added column used was Day.

4. **Which country has the highest number content ?**



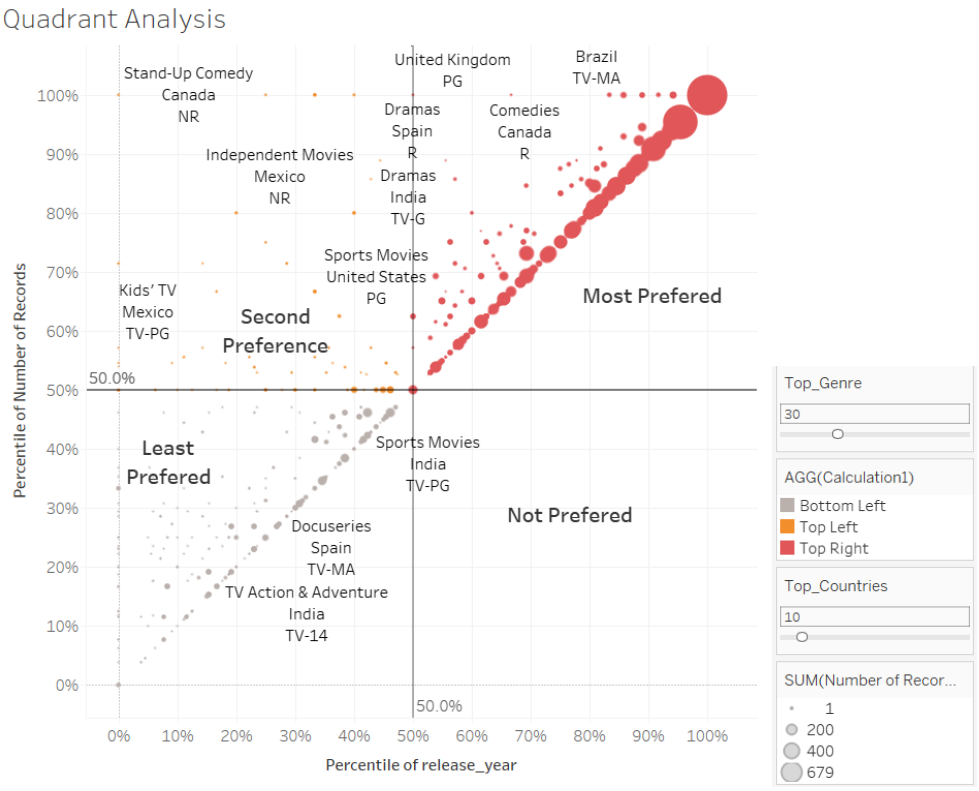
The majority of the content is discharged solely for USA (considerably more than the entirety of top 10 nations) This may be on the grounds that Netflix has been very popular in the USA from a Long time while in different nations like India its Popularity began to Increase from 2018. I implemented this by pivoting the country column and giving it as new column as Actual_Countries where visualised with Number of records by creating Top_Countries parameter.

5. **What is the duration of most of the movies?**



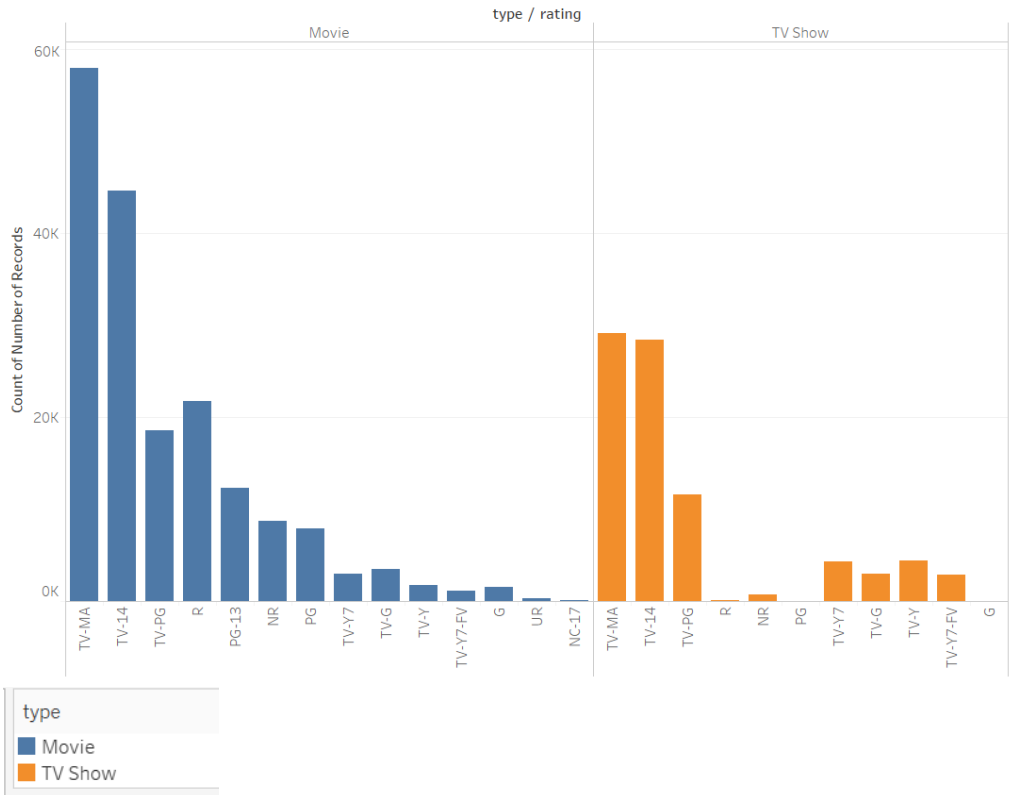
The Normal distribution, otherwise called the Gaussian dissemination, is a likelihood appropriation that is symmetric about the mean, demonstrating that information close to the mean are more continuous in the event than information a long way from the mean. We can see that majority of the movie duration is 90 mins in Netflix. Also it's clear normal distribution as it forms the proper bell curve.

6. **Quadrant analysis on country, rating, and genre.**



Quadrant charts are bubble charts with a background that is divided into four equal sections. Quadrant charts are useful for plotting data that contains three measures using an X-axis, a Y-axis, and a bubble size that represents the value of the third measure. From above start we can see most prefered. So from above we can see that most prefered genre of the users TV-14 are International Movies, Dramas, Action Adventures, Comedies and Sports. While their least preference are Thrillers, Horro Movies, Music and Musicals. For TV-MA most prefered genre are Dramas, International Movies, Horror Movies and Comedies while least prefered is Romantic Movies, Thriller and SciFi.

7. **Which is the highest rating watched movie and TV Show**

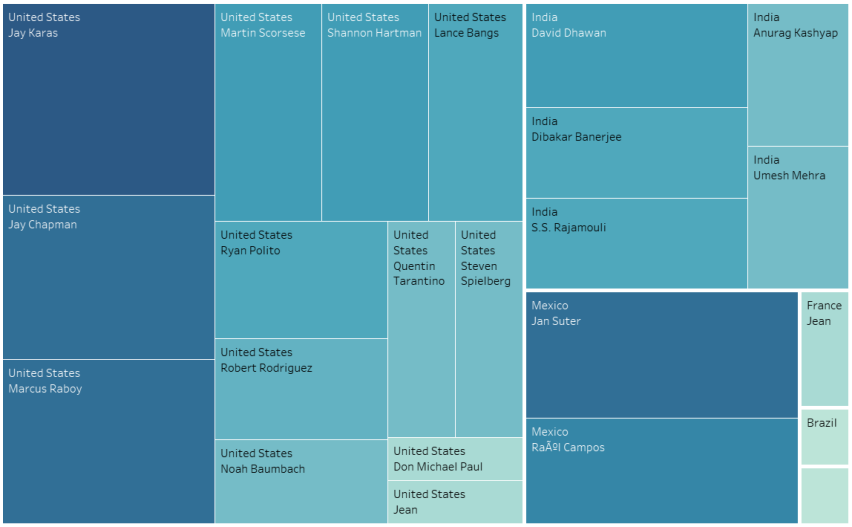


We can clearly see TV-MA and TV-14 are winning in both Movies and Tv Show.

TV-MA: MATURE AUDIENCE ONLY
This program is specifically designed to be viewed by adults and therefore may be unsuitable for children under 17. This program contains one or more of the following: graphic violence (V), explicit sexual activity (S), or crude indecent language (L).

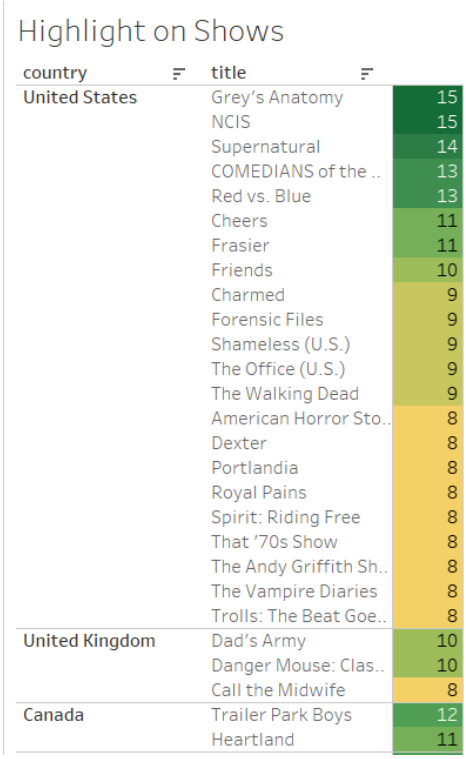
TV-14: PARENTS STRONGLY CAUTIONED
This program contains some material that parents would find unsuitable for children under 14 years of age. Parents are strongly urged to exercise greater care in monitoring this program and are cautioned against letting children under the age of 14 watch unattended. This program contains one or more of the following: intense violence (V), intense sexual situations (S), strong coarse language (L), or intensely suggestive dialogue (D).

8. **Top 20 directors with most country with respect to country.**



We can see Jay Karas from United States and David Dhawan from India has highest amount of content.

9. **Which TVShow has the highest duration with respect to Country?**



Highlight on Shows

country	title	
United States	Grey's Anatomy	15
	NCIS	15
	Supernatural	14
	COMEDIANS of the ..	13
	Red vs. Blue	13
	Cheers	11
	Frasier	11
	Friends	10
	Charmed	9
	Forensic Files	9
	Shameless (U.S.)	9
	The Office (U.S.)	9
	The Walking Dead	9
	American Horror Sto..	8
	Dexter	8
	Portlandia	8
	Royal Pains	8
	Spirit: Riding Free	8
	That '70s Show	8
	The Andy Griffith Sh..	8
	The Vampire Diaries	8
	Trolls: The Beat Goe..	8
United Kingdom	Dad's Army	10
	Danger Mouse: Clas..	10
	Call the Midwife	8
Canada	Trailer Park Boys	12
	Heartland	11

From the above table we can clearly understand that highest duration TvShow is Greys Anatomy as it has 15 number of seasons. I implemented by creating Top_Show parameter which gave me top 20 results.