Project Description

My data set contains information about all the TV shows and movies available on amazon prime. The information was collected from JustWatch in March 2023, containing data available in United States.

SOURCE :
https://www.kaggle.com/datasets/dgoenrique/amazon-prime-movies-and-tv-shows?select=titles.csv

Data Dictionary

- **id**: The title ID on JustWatch.
- **title**: The name of the title.
- **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.
- **genre:** A list of genres.
- **country:** A list of countries that - produced the title.
- **seasons:** Number of seasons if it's a SHOW.

Data Dictionary(Continuation)

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

DATA PREPARATION

1.Changed type from “ABC123” to valid data type for for id, name, character and role in credits table

2.Changed type from “ABC123” to valid data types in titles table, removed IMDB related columns, production_countries column(country) and genre because they had transitive dependency and multivalued column respectively.

3.Created a new table under the name “Genre” and used replace value function to separate the multivalued genre column in parallel with the split by delimiter function. And finally unpivoted other columns based on ID.

4.Created a separate table for IMDB as it had dependencies on IMDB_id.

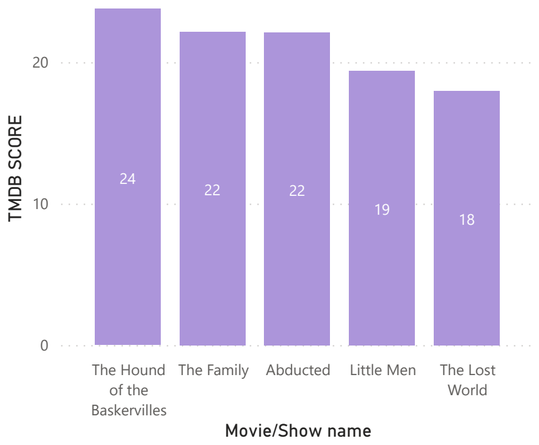
5. Created a separate table for country, used split by delimiter and unpivoted other columns based on id.

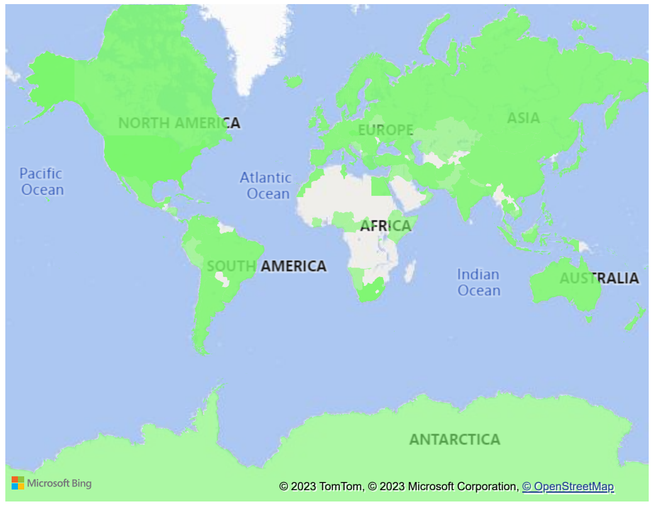
6. For age_certification and seasons as they had higher load for null values, created separate tables for each, thereafter filtering out the null values and linked them in data model view.

7.Renamed to appropriate column names which would be easily understood by users.

Question 1

Provide me a list of top 5 titles by TMDB score, based on the country produced.
Note: Display even the titles with tied scores


TMDB SCORE
20
10
0
24
22
22
19
18
The Hound of the Baskervilles
The Family
Abducted
Little Men
The Lost World
Movie/Show name


NORTH AMERICA
EUROPE
ASIA
Pacific Ocean
Atlantic Ocean
AFRICA
SOUTH AMERICA
Indian Ocean
AUSTRALIA
ANTARCTICA
Microsoft Bing
© 2023 TomTom, © 2023 Microsoft Corporation, © OpenStreetMap

The overall best rated movie/show is "The Hound of the Baskervilles".

If selected for particular regions such as "CA" or "IN", there would be more than 5 titles appearing, as those appeared would stand shoulder-to-shoulder with other movies/shows, as they have the same score.

One can choose to watch the best titles based on a region to know what the audience likes in a specific country.

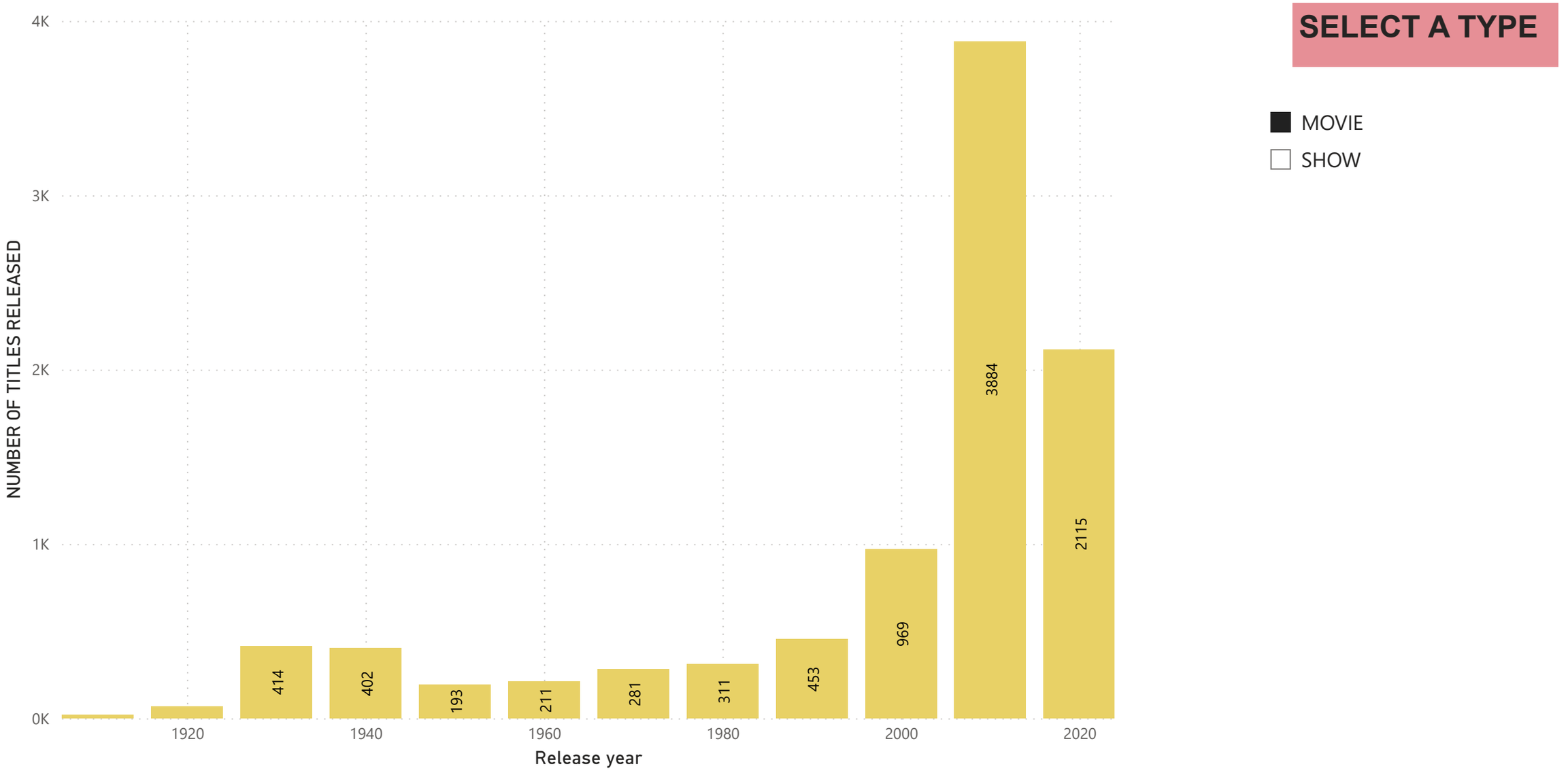
And, as for the business side, Amazon has started producing its own in-house projects so they could concentrate on similar titles to gain traction and increase their subscription base.

As the fields are ranked by the top scores, there is ranking data relationship that is being explored, as we are wanting to have the top 5 titles

Question 2

Based on release year(per decade), what can you say about the trend of movie/show released, also add a drill down under the decade to get an in-depth analysis.

Include a filter to to check the trend for movie or show


4K
3K
2K
1K
0K
NUMBER OF TITLES RELEASED
414
402
193
211
281
311
453
969
3884
2115
1920
1940
1960
1980
2000
2020
Release year
SELECT A TYPE
MOVIE
SHOW

The report shows that a time series relationship is being explored, in which the overall trend of movies/shows are increasing on a decade-to-decade basis.

Further, one may be deceived by the number of titles released in the decade on or after 2020, this is due to the incomplete data, as we are currently in 2023, and we cannot compare the current decade with the previous one until 2030.

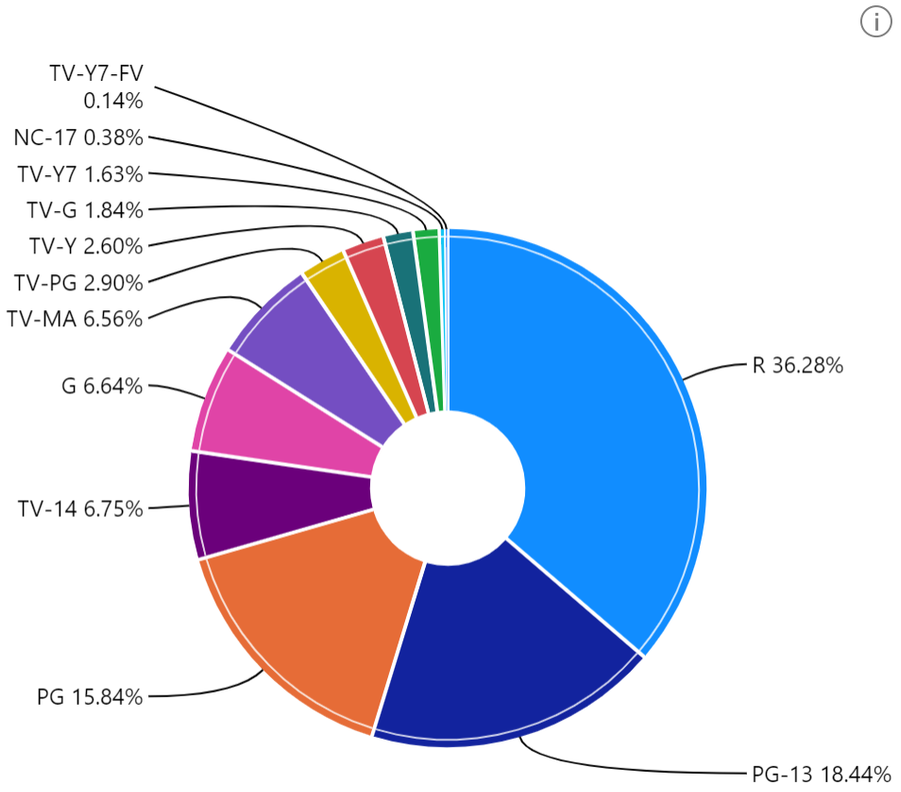
When we drill down, we can see study increase in the number of titles released in almost every decade, except for the 2020, from 2020 to 2021, we can see a rapid increase, this might be due to everyone being at home and demanding more shows to be released, due to pandemic.

And from 2021 to 2022, we can see a drop this might be due to funding issues, as the economy started to regain from pandemic but not much money would be poured in filming industry, and also the trend has moved more towards releasing the titles in parts, which require much production and design time.

There is also an another insight that can be gained by the movie or show filter, when clicked on show almost no titles appear in the early 1900's, this is not same for the movies, as the audience were unaware of the concept on the shows.

Question 3

Which top movies/shows would you suggest to watch based on age certification and TMDB score, also sort out on how popular they are among the audience



POPULARITY

0.00 3,187.53

MOVIE/SHOW Name	SCORE
The Hound of the Baskervilles	23.78
The Family	22.16
Abducted	22.10

The relationship that we are exploring is part-to-whole relationship, where the different age-certification groups are being compared.

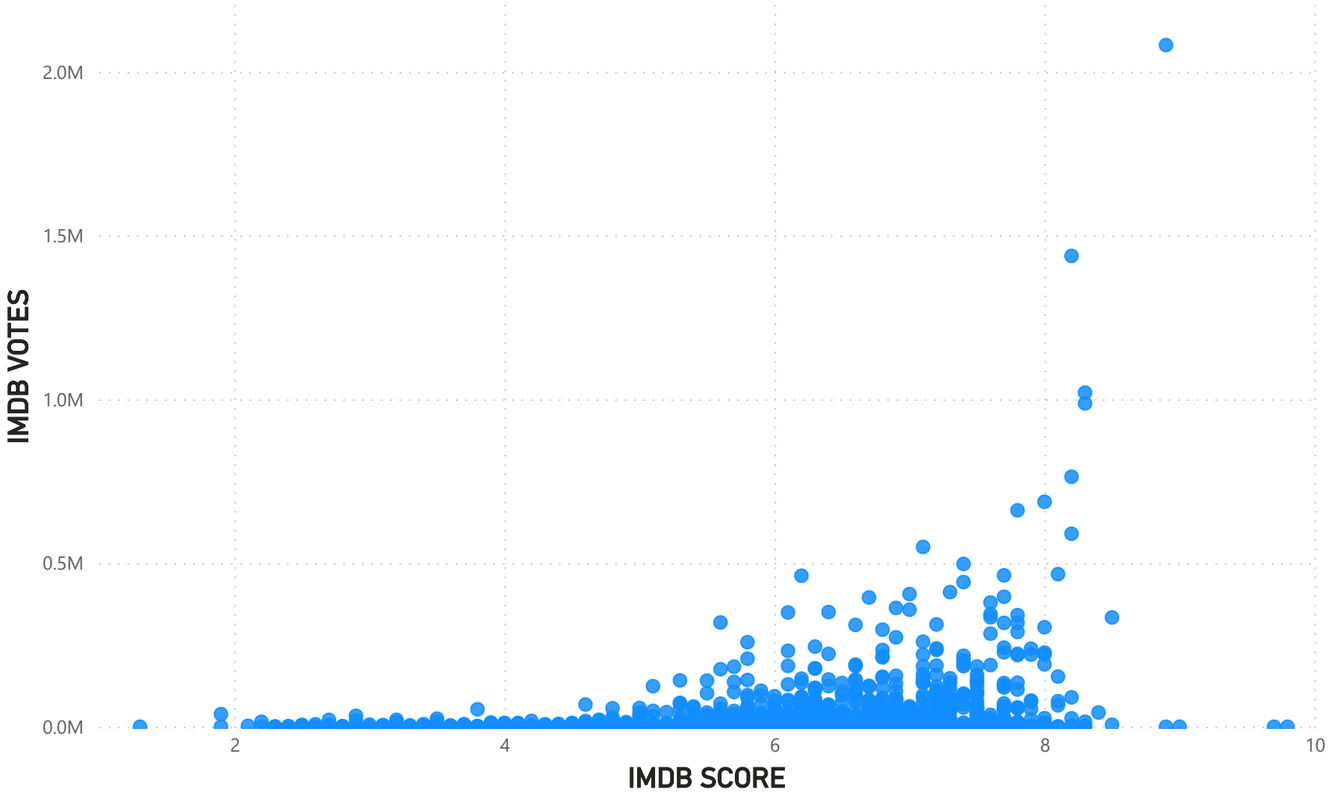
Overall, the R category contributes to the majority chunk of titles produced, and TV-Y7-FV has the least involvement. This means that the audience is accepting more of R- rated movies/shows than the other categories, this also makes sense, as majority of audience would be above 17 years of age to watch R-rated titles.

Further upon drill-down we can see that most of the titles for each age-certification lie under either 6 or 7 score, meaning average or above average shows and not only that the same titles are the most popular among the audience.

From amazon's business perspective they could focus more on promoting the top age-certified titles which are R or PG-13 on their streaming platform and also could produce more of the similar kind in-house, even if it performs above average, the subscription rates would increase driving the revenue.

Question 4

Based on IMDB score and votes, suggest which is the best watchable movie/show and which is not. Display the title in a card.
Note : Include only the titles with PG-13 or R ratings



MOVIE/SHOW NAME

#cats_the_mewvie

The relationship that we are exploring here is the Correlation, where the title id's are scattered on the basis of scores and votes.

We can see that there is a strong correlation and positive direction for the title "Pulp Fiction", we can experience the same with another title "The Wolf of Wall Street", this obviously makes sense as these are the best titles ever produced in the film industry, as most of the audience are familiar with them and have voted them maybe in excellent category. These would be the top titles one could watch.

If we look at the other end of the chart, the title "The Rest of Us" has a negative strength, which means that this title would be least suggested to watch for the audience.

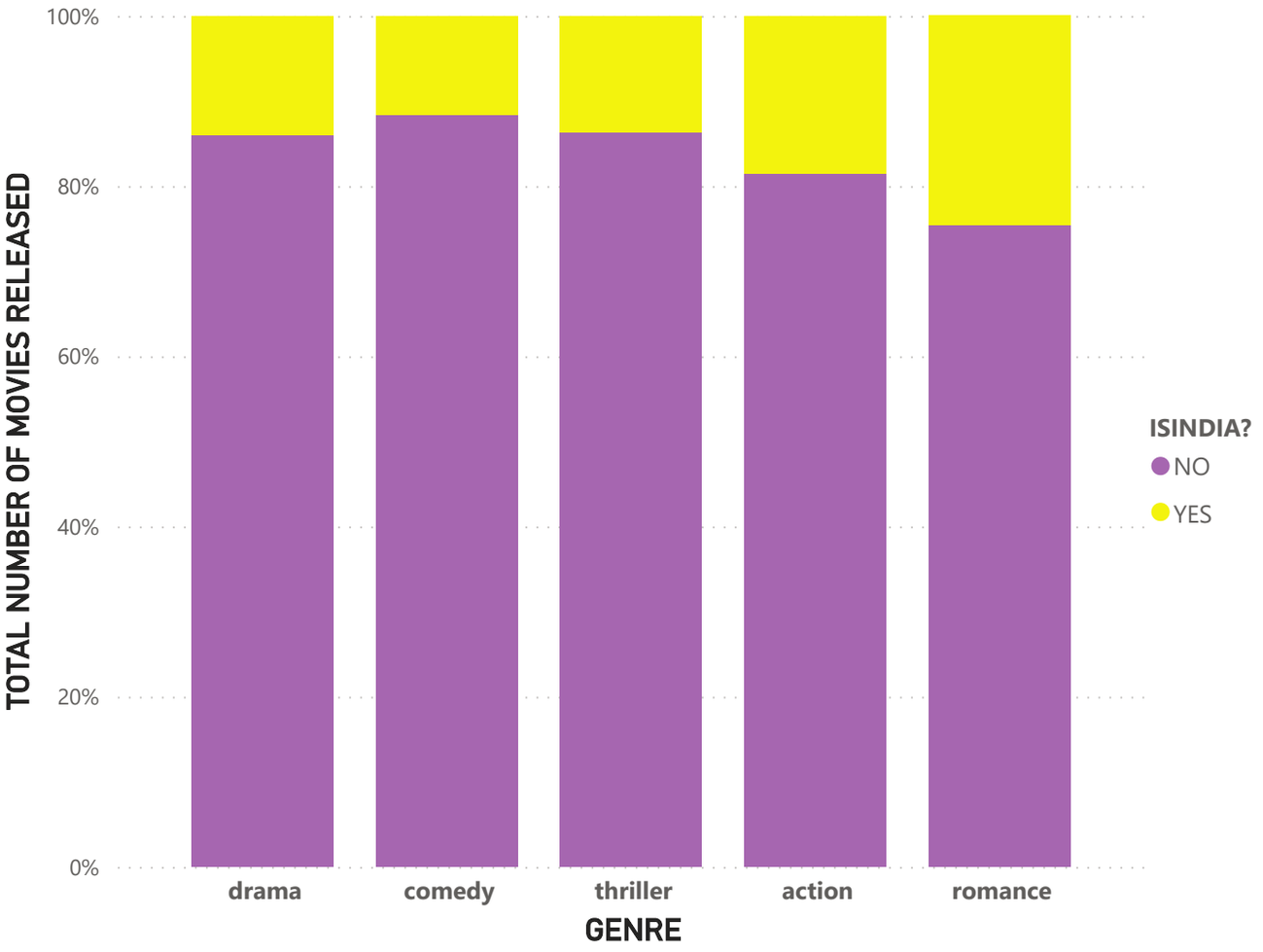
Also note that, a title might be exceedingly well, but would have less number of votes and might not be indicated on the chart as an outlier, the same may be true for negative strength titles as less/more number of people might have voted biased towards a particular topic.

From a business perspective, Amazon should possibly try to remove the lower hanging titles as the company(Amazon) would be incurring losses by paying the digital rights stream and there would be no audience to view/stream the title.

Question 5

What percent of movie/shows are from India on the basis of top 5 genres, and list the top 7 titles to watch from each genre of India, include only the latest movies/shows
Note : Latest movies/shows are those which are released on or after 2014

MOVIE/SHOW Name
Jai Bhim
K.G.F: Chapter 1
K.G.F: Chapter 2
Kantara
Pushpa: The Rise - Part 1
Shershaah
Soorarai Pottru



In this report, we are exploring part-to-whole relationship, with the country produced of India vs the rest of world, and also we are doing a ranking comparison based on the genre's produced by India.

Based on the genre's produced, we see that India produces about 25% of the world's romance genre movies, followed by 18% of action movies.

If we see in the table, for the top titles which are ranked and filtered on the basis of above 7 IMDB score and whether or not it being a latest movie/show, and also the IMDB votes, the list makes sense as these were the titles that had registered as blockbusters and have been promoted globally as well.

The analysis can be made that the top titles that one could suggest to watch and be in the know-how of the film industry would be appearing from this list for sure.

From a business perspective, Amazon could launch ad-campaigns and promote these titles in social media, saying that they offer these titles and the audience can take a subscription to enjoy the experience by sitting at the comfort of home. This would drive masses to register for subscription as it is not only dependent on score but also on votes, that means many viewership as well for the company.

In addition to this, they can do the similar promotion for different countries as well, and promote these titles in the home page of the amazon prime video streaming service, so that the audience get hooked to watching, without hesitating or understanding to decide on what to watch on.

For the second question, I had used grouping and binning to create a group based on decade of the release year.

For the third question I had used drill down donut to drill down by the age-certification and view the TMDB scores per certification category.