



The Movies Dataset Project

Reham Alnuqayr



To download the data set:

<https://www.kaggle.com/rounakbanik/the-movies-dataset>



The Problem Statement:

Why production companies revenue has not been greater than 245 Billion in the last 10 years?

Executive summary

Elements that impact on the revenue:

- Viewers Popularity
- Release time.
- Budget particularly marketing and advertising.

Based on my analysis, My recommendations are:

- Showing the movies on the main page and publish trailer on social media.
- focus on popularity genres.
- Publish the movies in June.
- Raise and prepare the budget which will improve the revenue

Hypothesis

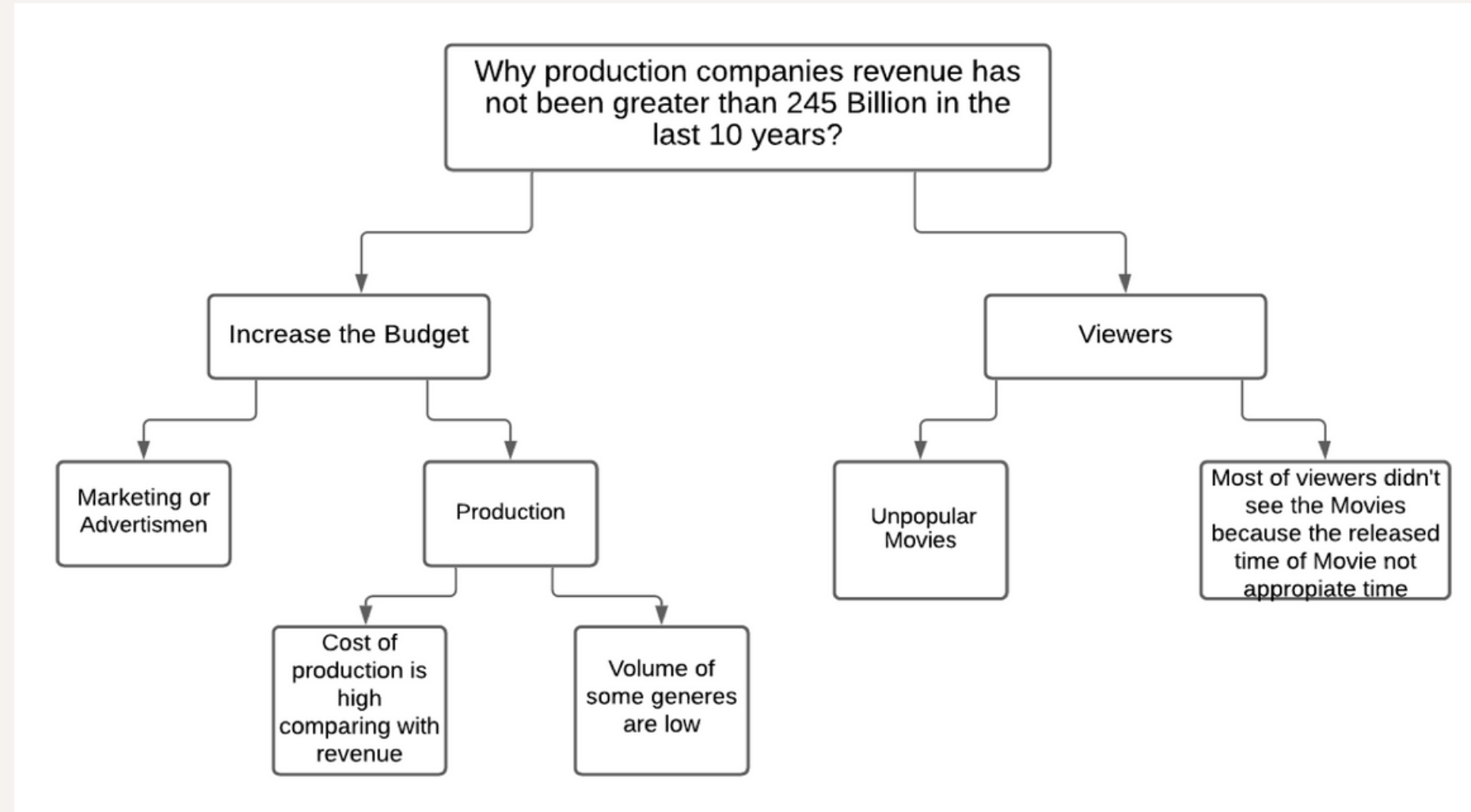
As such, the following hypotheses have been formulated for the given problem statement;

H1: The highest generated revenues of between \$700M and \$2.1B can highlight the top 50 movies preferable for streaming.

H2: The higher popularity of between 24 and 185 points can highlight the top 50 movies preferable for streaming.

H3: The higher vote count between 5000 and 15000 can highlight the top 50 movies to stream

Issue Tree



Overview of Analysis

The problem statement was to determine:

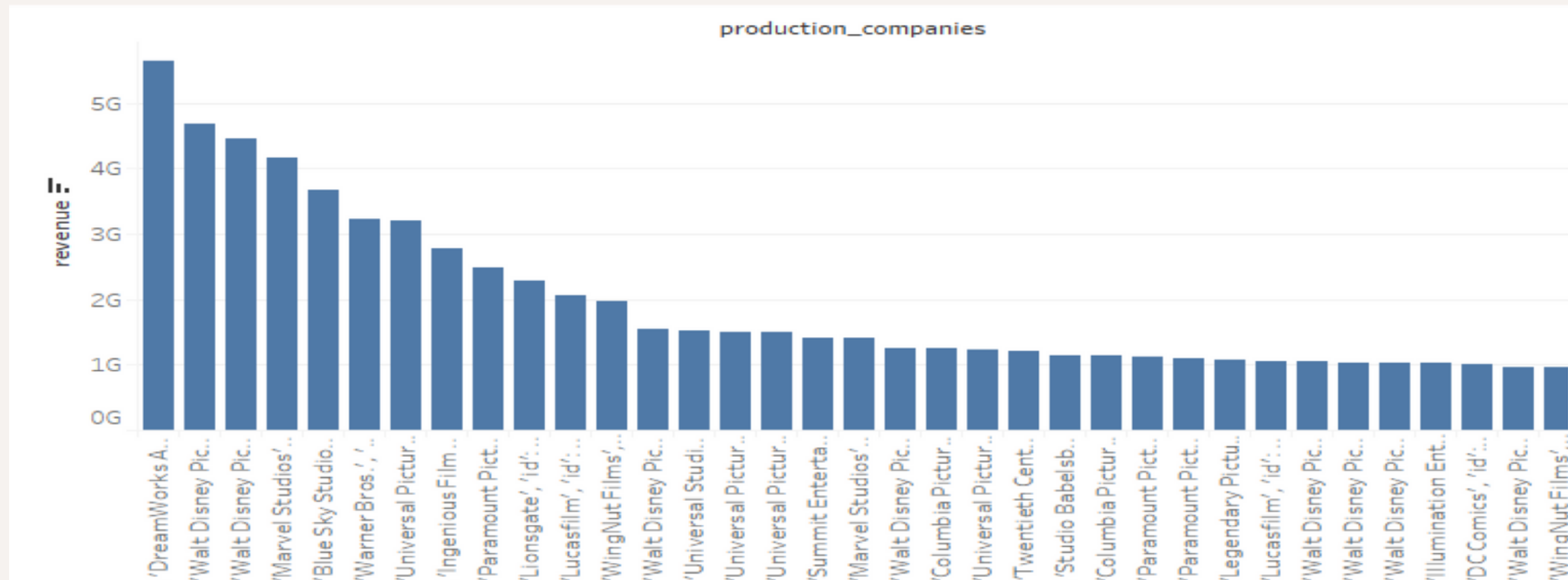
Why production companies revenue has not been greater than 245 Billion in the last 10 years?

My analysis focused into three workstreams:

- 1- If the budget affects revenue.
- 2- If popularity genres affect on revenue.
- 3- If release time affects revenue.

Analysis 1

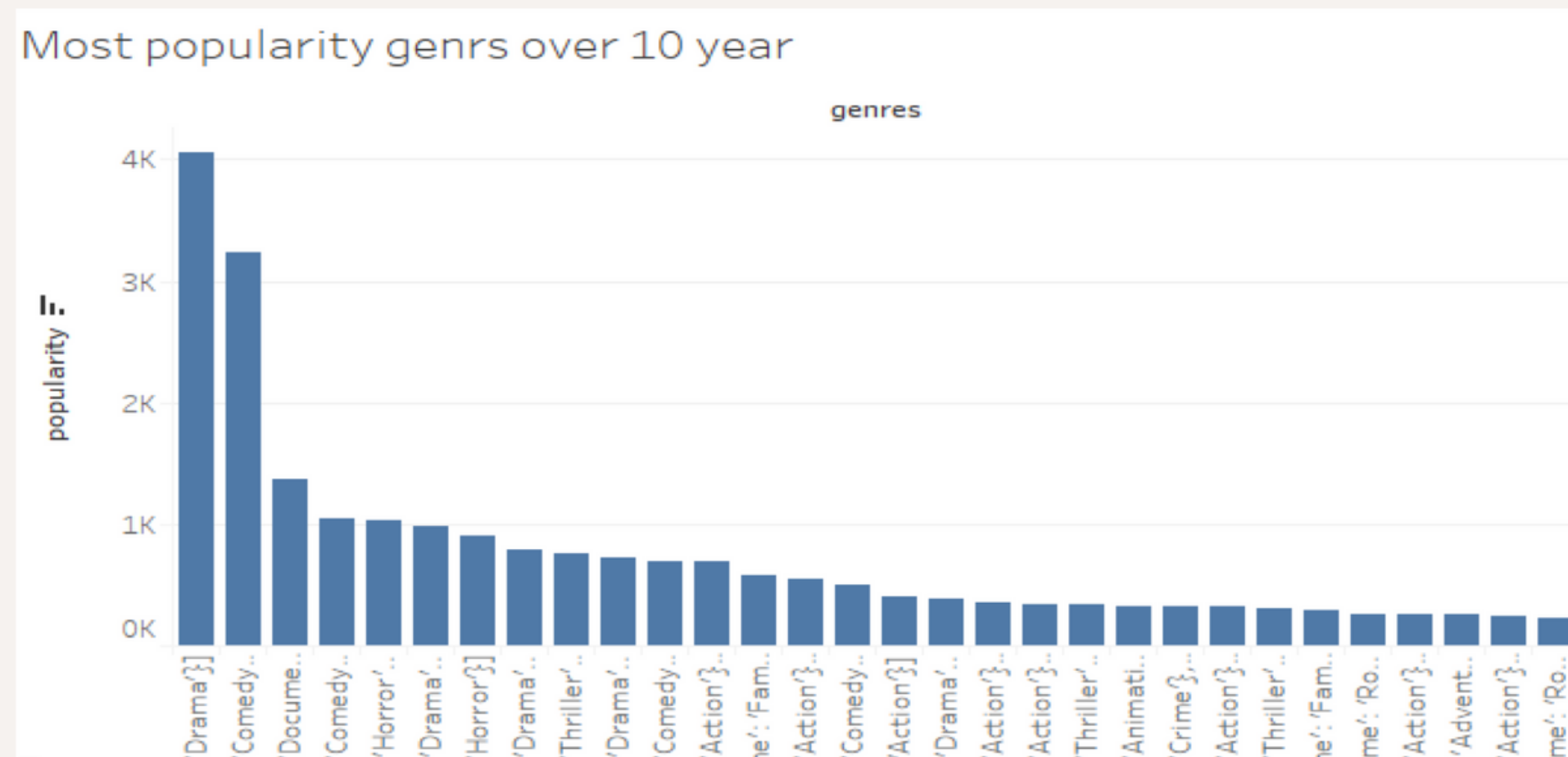
Revenue Of Production company Last 10 year



You can see from the chart above the revenue of production companies

Analysis 2

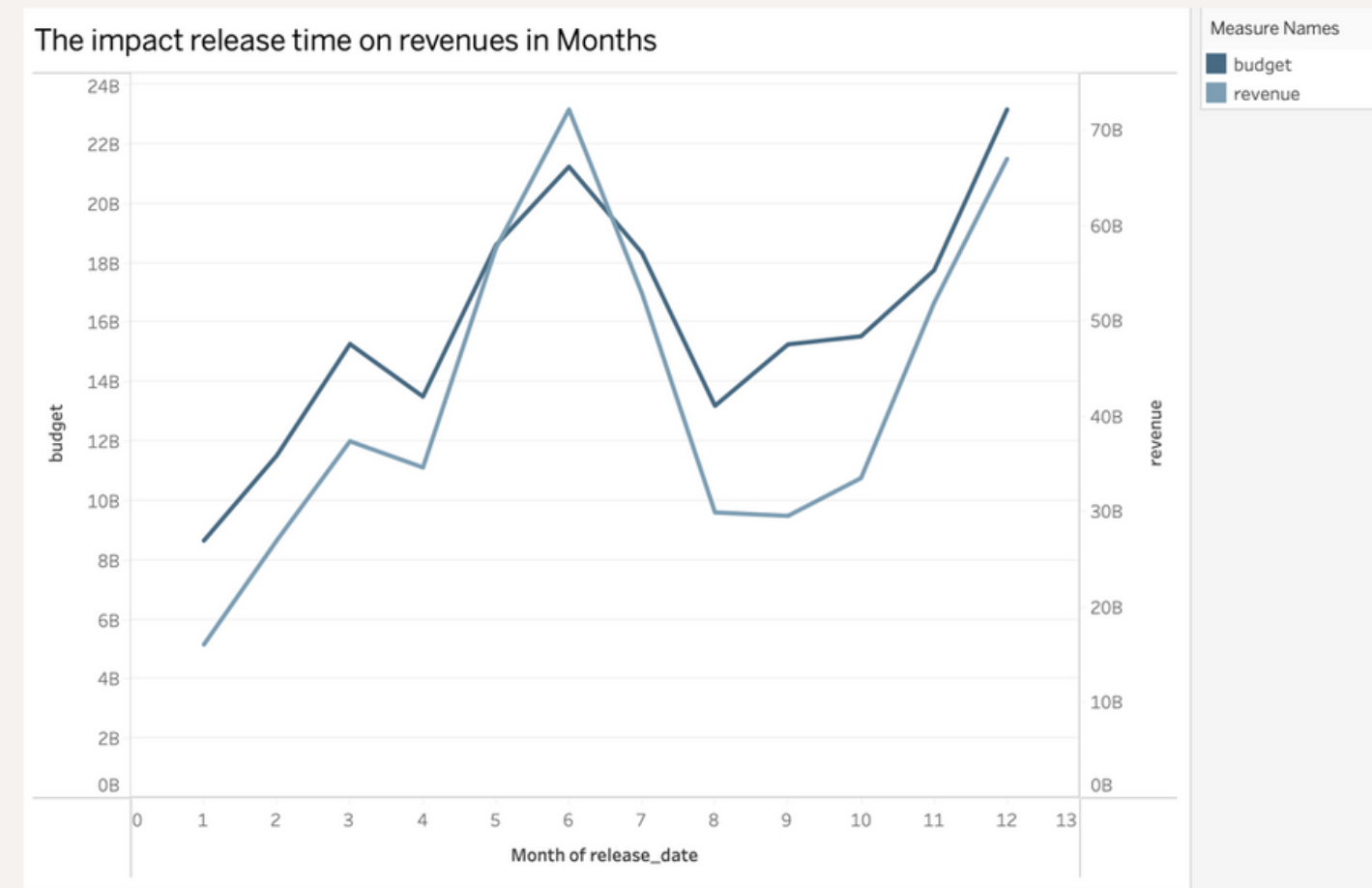
Most popularity genres over 10 year



You can see from the chart above the most popular genres is Drama, so you must choose the appropriate genres in future production to increase the revenue.

Analysis 3

The impact release time on revenues in Month



You can see from the chart above the high revenue is in June so the best time to publish the movies is in June.

Biases and Limitations

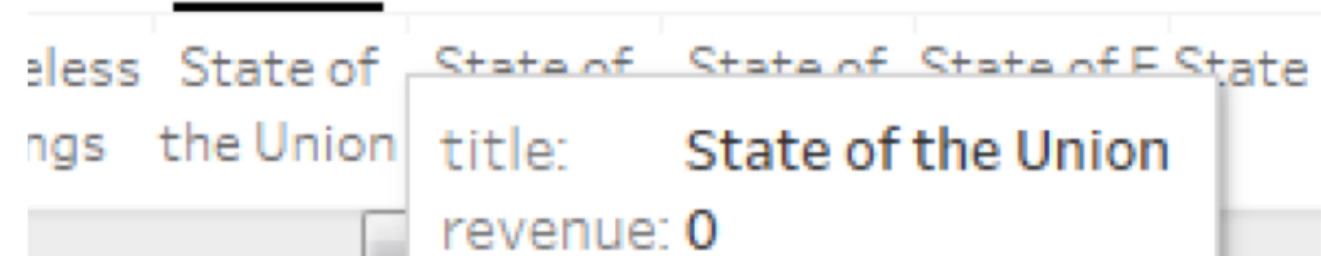
Data Collection:

- The budget is not detailed enough.
- Not knowing how popularity has been calculated.
- The evaluation were only from those who visited the IMDB website.
- Most of data are collected from US so you can't get accurate information.

Biases and Limitations

Data processing:

The distribution of most elements are skewed with outliers because of the missingness values in revenue and budget variables for most of the movies which affect the recommendation. I chose to include it in certain analyzes while removing it from revenue and budget analysis. So that I could compare it with other genres.



A screenshot of a data table with a dropdown menu open. The table has columns for movie titles and revenue. The dropdown menu shows the selected item 'State of the Union' with a revenue of 0.

Movie Title	Revenue
Stateless	
State of the Union	0
State of the Union	
State of the Union	
State of the Union	
State of the Union	

Biases and Limitations

a major limitation was that voting count did not qualify as a credible source for deciding on the exact popularity or success of a movie, as individuals vote based on distinct and variable perceptions.

In terms of bias, there was an involvement of feature biases. As such, biases associated with the reviewer's thoughts, perceptions and objectivity were noteworthy. Besides, it is also important to mention the biases pertaining to IMDB jury that is responsible for movie reviews and rating. Specifically, it is reasonable to argue that the movie reviewers have varying levels of understanding and perception of the movie concept, idea, message it communicates and other attributes. These differences can reflect poorly on the overall movie rating system.

Biases and Limitations

Analysis

The IMDB dataset was retrieved and processed using Tableau, Microsoft Excel and EDA analysis. The results of this analysis highlighted several biases and limitations in the IMDB dataset, which could hinder the identification of top 50 IMDB movies. Based on the analysis, a set of recommendations have been formulated.

Biases and Limitations



Recommendation 1: Include other features / attributes to review movies, such as their genres and storylines. This can help mitigate feature bias.

Recommendation 2: Characterize the voting count to specifically include key benchmarks that can inform further review, scoring and rating of movies.

Recommendation 3: Develop a movie reviewing system or framework that inherently avoids pertaining biases and conflicts of interest. As such, the framework should account for the differences in reviewer perception and understanding.





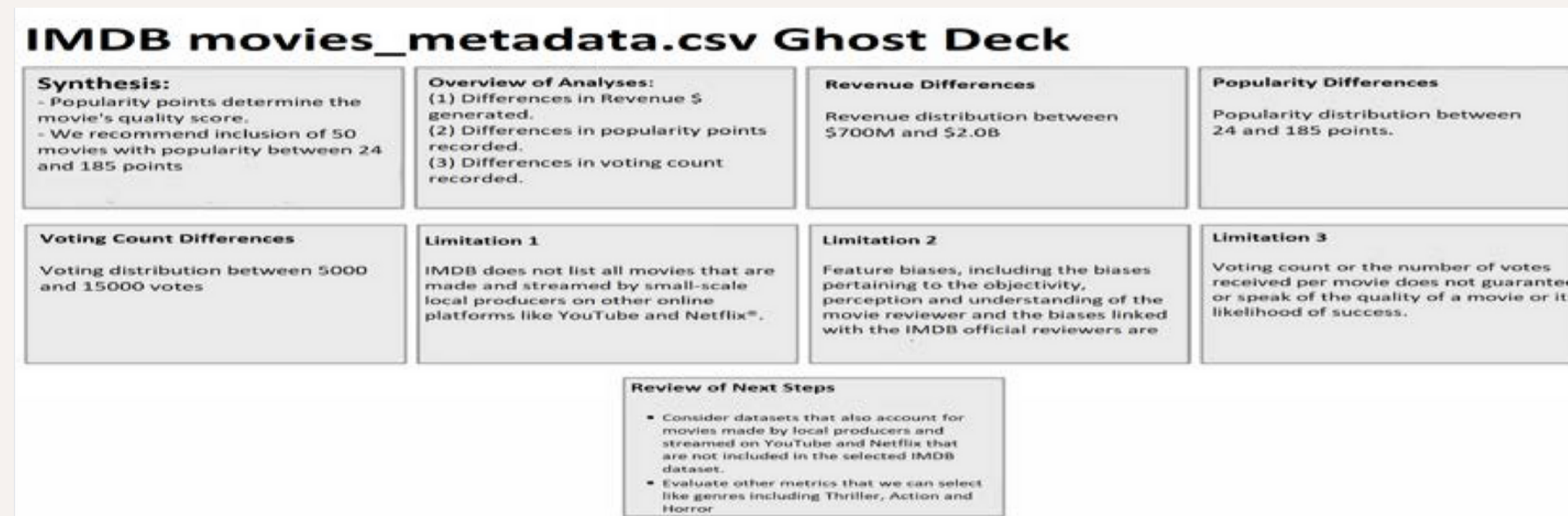
Review of Next Steps

Consider datasets that also account for movies made by local producers and streamed on YouTube and Netflix that are not included in the selected IMDB dataset.

Evaluate other metrics that we can select and test in order to ascertain the top 50 movies to stream (for example; specific movie genres like Science-Fiction, Thriller, Action and Horror)

The Ghost Deck

The IMDB movies_metadata.csv problem's Ghost Deck is represented in the chart shared below;



Conclusion

We have seen above that the popularity genres will affect on the revenue so we should choose the popularity genres in our future productions, also the released time of movies will affect on the revenue so we should choose the appropriate time to release our future movies and also should determine or prepare for the budget of the movie to be appropriate with movies revenue.