INVESTIGATE A DATASET

I analyzed TMDB Movie Dataset. This dataset contains around 10K movies information. These movies are released between 1960 and 2015. After checking the dataset.

My research questions are;

- Does movies with homepages add any value to the revenue in a Bayesian approach?

- How genres matter in terms of revenue and popularity at all?

- Does specific companies attract more revenue? If yes, what do these companies specialize in terms of genre?

- How is the popular cast in high revenue movies?

- How is the popular directors in high revenue movies?

LIMITATIONS:

While dataset has over 10K samples, my analysis was limited to 3855 samples since %70 of movies have no budget and revenue information. My analysis showed that there was little correlation between budget and revenue of movies, but, this analysis is limited by the lack of available budget and revenue data

Another limitation is about collecting the data. While Western movies have lack of representations (14 movies), this is surprising since earliest release date of movies in dataset is in 1960. I would expect more samples from Western genre in high revenue movies.

Movie dataset includes up to 2015 release year, so, this analysis doesn’t give intuition about latest movie industry

Analysis:

I checked the summary statistics and data types of the dataset. I removed to rows that have no budget and revenue, since my main argument was based on the high revenue/budget movies

For every research questions, I checked the counts of the feature of the dataset and compared with population features that have no filter of revenue and budget. This gave me an idea how revenue and budget played a role provided that variables are solely dependent on revenue and budget. Here is an example of high-revenue movies have genres;

A close up of a fence

Description automatically generated

This plot shows most 10 sought-after actors/actresses in these movies;

A picture containing object, comb, fence, drawing

Description automatically generated

This analysis does not show concrete evidence since there may be many underlying reasons for these aggregated results.

In analyzing movies with home pages, I used mean statistics of the movies to compare vote counts, and average budget. I also used pandas series value counts, min, max, mean and matplotlib functions to plot my findings. Though, I split the list-like column to multiple columns, it was meant only for the sake of data manipulation practice and was not used further to analyze the data, since my research questions have nothing about apriority and effect of group of casts or genres in a movie.