TMDb-PROJECT: INVESTIGATE THE DATASET

❖ Table of Contents

❖ Introduction

In the course of the analysis, I'll be analyzing a dataset called The Movie Database (TMDb), which is a dataset that was originally culled from Kaggle. This dataset originally consists of 10,866 rows and 21 columns. Below are the column attributes;

1. id: this is a unique identifier for each movie
2. imdb_id; unique identifier for each movie on the imdb
3. popularity; numeric quantity specifying the movie's popularity
4. budget; tells the movie for each movie
5. revenue; tells how much each movie generated
6. original title; original title before adaptation or translation.
7. Cast; lists the casts in each movie .
8. homepage; is a link to the homepage of the movie.
9. Director; tells the director of each movie.
10. Tagline; the movie's tagline.
11. Keywords; keywords related to the movie found.
12. overview; a summary of the movie
13. runtime; the duration of the movie in minutes
14. genres; the genre of the movie example action, comedy, thriller etc.
15. Production_companies; the production house
16. release_date; the date the movie was released
17. vote_count; number of vote ratings the movie received
18. Vote_average; average of the vote count
19. release_year; the year the movie was released
20. Budget_adj; tells the budget of the associated movie
21. Revenue_adj; tells the revenue of the associated movie

The project is divided into three parts, Data Wrangling, Exploratory Data Analysis, Conclusion and limitation.

❖ Questions

I will be answering the following questions in the course of my analysis,

A: Which Director has produced the highest number of movies?

B: What is the most watched genre?

C: What are the top 3 production companies with the most movies?

D: How has the number of movies produced changed over the years?

E: What are the top 10 most profitable and least profitable movies?

❖ Conclusions

Below are the conclusions I inferred from my explorative analysis of the tmdb dataset;

The first question showed that Steven Spielberg is the director with the highest number of movies produced with a total of 27 movies produced.

The second question revealed drama as the most watched genre with a total count of 243.

The third question revealed the top 3 production companies to be Paramount Pictures with a total of 77 movies, Universal Pictures with a total of 57 movies and Columbia Pictures with a total of 3 movies.

The fourth question sought to answer how the number of movies produced had progressed over the years and findings showed that the number of movies produced where considerably low between 1960 to 1983 with the year 1969 having only 4 movies produced while the number of movies produced spiked up between the year 2012 to 2015 with year 2013 having 179 movies produced which is the highest number.

The fifth question looked at the top 10 most profitable and least profitable movies. This analysis was done using a point plot line and it showed Avatar, Star Wars: The Force Awakens, Titanic, Jurassic World, Furious 7, The Avengers, Harry Potter and the Deathly Hollows: Part 2, Avengers: Age of Ultron, Frozen, The Net as the top 10 most profitable movies respectively while the least 10 profitable movies are The Warrior's Way, The Lone Rangers, The Alamo, Mars Needs Moms, Brother Bear, The 13th Warrior, The Adventures of Pluto Nash, Charlotte's Web, Flushed Away and Australia respectively.

❖ Limitations

During the process of this analysis, I experienced certain limitations considering it was my first time carrying out an analysis solely based on the python tool. Also, the dataset was limited as it contained some missing column and missing entries. I had to clean the dataset, drop some columns, delete duplicates, separate the columns with pipe characters and even create some new columns.

Thanks to the various resources available all over the net, I made use of google, stackover, Github and had to watch some videos to see how this type of analysis is done. I must also mention that my session lead, Tolu Okuwoga was very helpful and fundamental to the completion of this project as she was available to answer my numerous questions and guide me through where I had glitches.