Project report

Data set used:

- TMDb movie data set

Introduction:

This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue. Certain columns, like ‘cast’ and ‘genres’, contain multiple values separated by pipe (|) characters. The final two columns ending with “_adj” show the budget and revenue of the associated movie in terms of 2010 dollars, accounting for inflation over time.

I will explore the following questions:

1. What is the relation between popularity and revenue_adj ?
2. Is there a relation between budget and revenue?
3. What is the average budget for the movies?
4. What is the relationship between movie budget and revenue?
5. How many movies in each genre?
6. What is the most profitable movie?
7. What are the most ten movies have highest revenue?
8. What is the most genres associated with the high revenue movies?
9. Which movie have high popularity overall years?
10. Which genre have high popularity overall years?
11. What are the most ten directors achieve highest revenue from the movies?

Data wrangling:

We want to identify any problems in the quality or structure of the data with general properties:

- Assessing and build intuition.
- Defining shape and data types.
- Display the summary of data including nan-values.
- Take observation about what to need to be changed or removed then try to solve it

Observation:

(1) We have lots of nan-values that we want to remove in:

- homepage
- tagline
- cast
- imdb_id
- keywords
- production_companies

(2) We have unwanted columns for analysis:

- cast

(3) We have lots of zero values in budget, budget_adj, revenue and revenue_adj columns but few in runtime column

(4) We have 44 nan-value for genres and 23 nan-value for director, so we will remove it from data

(5) We have release_date column needs to be transformed to datetime

(6) We have one duplicated row

What will we do?

- We will remove columns with nan-values by drop() method.
- We will remove duplicated rows.
- We will exchange all zero-values with budget, budget_adj, revenue, revenue_adj and runtime columns with the means of each column.
- We will change the release_date column to datetime datatype.
- We will drop all rows that have nan-values in genres and director columns using dropna() function.

Exploring data analysis:

- Answering the questions above using data visualization techniques and libraries like seaborn and matplotlib.

Conclusions:

- We have positive relationship between popularity and revenue
- We have positive relationship between budget and revenue
- Drama is the most genre has number of movies overall years
- Star Wars is the most profitable movie overall years.
- Avatar movie has the highest revenue over all years
- Movies that have high revenues associated with adventure genres the most in all years
- Jurastic world movie has the highest popularity overall years
- Comedy genre has the highest popularity overall years
- Steven Spielberg is the director that achieved the highest revenue from its created movies

Limitations:

- The columns 'budget' and 'revenue' did not display the actual currency for the values.
- I decided to keep all the zero values from the data set because there were too many. That number was too much to drop from the dataset. They were replaced with mean values.
- The dataset does not confirm that every release of every director is listed.
- I decided to remove nan-values from the data that found in each director and genres columns as it is very few to ignore them