General Properties

DataSet chosen for analysis: TMDb Movie Data

The database contains information about movies collected from The Movies Database, including revenue, budget, ratings, and homepage. I displayed 10 rows to get a little more detailed result about the columns, values and structure. I decided to ask questions related financials, popularity and genres.

Questions posed:

1. How has the popularity of Western movies changed over the years?
2. Does budget correlate with popularity? What about the movies with the biggest budget?
3. Which 10 production company profited the most over the years?

The Data Structure

Before working with the data I checked the databese and looked for missing values, inconsistency or inadequate datatype. After getting more information and find out the questions I wanted to pose, I cleaned the database. There were unecessary columns with missing data, inadequate datatypes and rows with 0 value. The columns 'genres' and 'production_companies' contained multiple value that does no meet the requirements of first normal form.

The Cleaning Process

- I removed the columns cast, homepage, tagline, keywords, overview and imdb id to improve database performance.
- The column 'genres' and 'productions_companies' were not in the first normal form which requires that in the table should not have multiple value in the same row of data. I was unable to create a second joined column, so I decided to remove the values after the first '|' sign to get better groupping and cleaner visualization in the further analysis.
- I casted release_date from string to date datatype.
- I converted the columns revenue, budget, revenue_adj and budget_adj from float to int.
- The 0 values would distort the result of forther calculations so I replaced the 0 in revenue, budget, revenue_adj and budget_adj with means.
- I also replaced the Na values with 'Unknown' to improve interpretation.

Exploratory Data Analysis

I recently read an article about the evolution of Western movies. It mentioned that the most prolific era was in the 1930s to the 1960s and the genre almost vanished in the 1980s. Tthere were little sign of resurgence after the 1990s but Western has not got back its popularity yet. As I am a big fan of the genre, I decided to analyze its populatiry over the decades and test the assumptions on the data. However, I use the article only as an interesting starting point - I do not intend to draw far-reaching conclusions.

1. Have changed the popularity of Western movies over the decades?

At first, I created a smaller dataframe which contained movies where in the column 'genre' appeared the word 'Western'. I decided to get every Western influenced or Western styled movie, I did not want to define the conditions too strict to get a bigger dataframe. I also planned to analyze the popularity by decades, not by release_year.

2. Ratings for the Cheapest and Most Expensive Movies

I was curious about if there are any measurable difference between votes of the most and the least expensive movies. I decided to get information about the most and least expensive films and viualize the distribution of their votes in one histogram.

Part 1: Get the most expensive movies

At first, I sorted the movies by budget to get the 200 most expensive movies from the database.

Part 2: get the cheapest movies

I queried the 200 cheapest movies from the database.

Part 3: Compare the results in one diagram

I created a diagram to display the differences between the ratings.

Part 4: Conclusion

As a conclusion, I can say that the most expensive movies generally got better rating that the cheaper ones. We can see on the diagram that the worst rating is 4.5 while the cheapest movies worst rating were lower than 2.

Conclusions

In the first section I examined the popularity of Western movies over the decades. I made my analyzation based on the values of 'released_year' and 'popularity'. I could not find any correlations between the numbers and the assumptions but I found it by taking into account the numbers of released movies.

After that I analyzed the ratings of the most and least expensive movies and I found out that the more expensive movies got higher votes than the cheaper ones.

Limitations

In the first section - although the literature details the phenomenon - I could not find any correlation between 'populary' and 'release year'. It would be good to know more about what is behind the value 'popularity' and what popularity means here. Just to name a few... How was it calculated? Which criterias and values were measured exactly to get these numbers? It could be caculated based on ticket sales? Or based on audience appraisal? However, I found correlation between my assumptions and the number of released western movies but I would not name it causation without a much more detailed further analysis.

In the second section, I made my calculations based on the values of budget adjustment to take the fluctuations into account, I found this really useful. But there were more missing values in the 'budget_adj' column. During the cleaning process I replaced the missing values with the average, but it still can distort the result (for instance, there would be other movies among the most expensive 200 movies).