MOVIE ANALYSIS

OPEN DATA 1916-2017

RAYMOND ZHANG

2022-10-20

EXECUTIVE SUMMARY

CONTEXT:

This data contains about 5000 movies from year 1916 to 2017 and was generated from the movie database API by using TMDb API.

There are total of 4803 observations in this dataset and includes at least three continues columns which are popularity, release_date and runtime. Also, it includes at least 3 classification columns like country, status, and language.

CONTENT:

All the data comes from real movie data, and it doesn't contain any columns that need to be anonymized due any ethic concerns in future.

OBJECTIVE:

Would track the most popular movies through the world and what caused those movies are famous based on the profit gained.

Also, find out any correlations between famous movies and unfamous movies.

DATA SOURCE: Movies Dataset | Kaggle

LIMITATIONS AND ETHICS:

- This dataset only contains 5000 movies may cause rank error due to the limitation.
- This dataset was considered as not never update frequency would be caused timeliness problem.

DATA WRANGLIN AND CONSISTENCY CHECKING

- Data wrangling check:
  - ✓ Dropped 8 columns won't be used in analysis and keeps another 14 columns.
  - ✓ Columns names are all easy understanding.
- Data Consistency Check:
  - ✓ Mixed-type columns check, column country, genres, overview, and language contain nan values considered as mixed-type column.
  - ✓ Nan values checked, found 181 nan values in country column and 28 nan values in genres column and 132 nan values in language column.
  - ✓ Didn't do anything with those nan values as they will be ok for my analysis
  - ✓ No duplicates found in this dataset.

DATA PROFILE

Column Name	Description	Data type	Time Variant
budget	The total spending to cast the movie	Quantitative, continuous	No
genres	What kind movie is defined as.	Qualitative, nominal	No
id	Movie unique number	Quantitative, discrete	No
overview	Customers' experiencing about the movie	Qualitative, nominal	No
popularity	The index of how famous bout the movie	Quantitative, continuous	No
country	Stay days in weekend	Quantitative, continuous	No
Release_date	The date that the movie was released.	Quantitative, continuous	Yes
revenue	Total gains from the movie	Quantitative, continuous	No
runtime	The length of the movie	Quantitative, continuous	No
language	Movie's spoken language	Qualitative, nominal	No
status	Whether the movie was released or not	Qualitative, binary	No
title	Movie's name	Qualitative, nominal	No
Vote_average	The average number of movies that was voted	Quantitative, discrete	No
Vote_count	The total number of movies that was voted	Quantitative, discrete	No

QUESTIONS TO DIG OUT:

- How can we define a movie as popularity? Are there any correlation actors among the observations?
- Which region contributes the most revenues and what are the profits? Does the top revenue indicate the top profit?
- Does budget can be connected to a movie length, like high budget means long length?
- Which year contributes the most revenue and which year contain the most popularity movie? Where are these movies come from?