TASK

Exploratory Data Analysis on the Movies Data Set

Visit our website

Introduction

Summary of the data set

Data Cleaning

During the data cleansing phase, several steps were taken to prepare the dataset for analysis.

1. Column Removal:

Redundant or unnecessary columns were identified and dropped from the dataset. These included 'homepage', 'keywords', 'original_language', 'original_title', 'overview', 'production_companies', 'status', and 'tagline'.

2. Remove Duplicate Rows:
Duplicate rows were removed to maintain data integrity and ensure that each entry in the dataset is unique.

3. Remove Rows with Missing Data:
Rows with missing or zero values in crucial columns, such as budget or revenue, were filtered out. This step helps in focusing the analysis on complete and meaningful data.

4. Change Data Types:
- Data types were adjusted to facilitate easier manipulation.
- 'release_date' was converted to the DateTime format for better date handling.
- The release year was extracted from each release date.
- budget' and 'revenue' columns were converted to the integer data type (int64) using NumPy.

5. Flatten JSON Columns:
- JSON-formatted columns ('genres', 'production_countries', 'spoken_languages') were flattened, making the data more accessible for analysis.

Data Stories and visualisations

Data Exploration:

Identify Relationships Between Variables:
In the exploration phase, the goal is to identify interesting relationships or patterns within the dataset. Specific questions or relationships can be defined for further investigation.

Top 5 Most Expensive Movies:

To understand the budget distribution, the top 5 most expensive movies were identified, providing insights into high-budget productions.

Title	Budget
Pirates of the Caribbean: On Stranger Tides	380,000,000
Pirates of the Caribbean: At World's End	300,000,000
Avengers: Age of Ultron	280,000,000
Superman Returns	270,000,000
John Carter	260,000,000

Top 5 Most Profitable Movies:

Profitability, calculated as the difference between revenue and budget, was used to determine the top 5 most profitable movies. This sheds light on successful ventures in terms of financial gains.

Title	Profit
Avatar	2,550,965,087
Titanic	1,645,034,188
Jurassic World	1,363,528,810
Furious 7	1,316,249,360

Title	Profit
The Avengers	1,299,557,910

Top 5 Most Popular Movies:

Sorting the dataset based on the 'popular' column revealed the top 5 most popular movies, giving an indication of audience interest and engagement.

Title	Popularity
Minions	875.581305
Interstellar	724.247784
Deadpool	514.569956
Guardians of the Galaxy	481.098624
Mad Max: Fury Road	434.278564

Movies Rated Above 7:

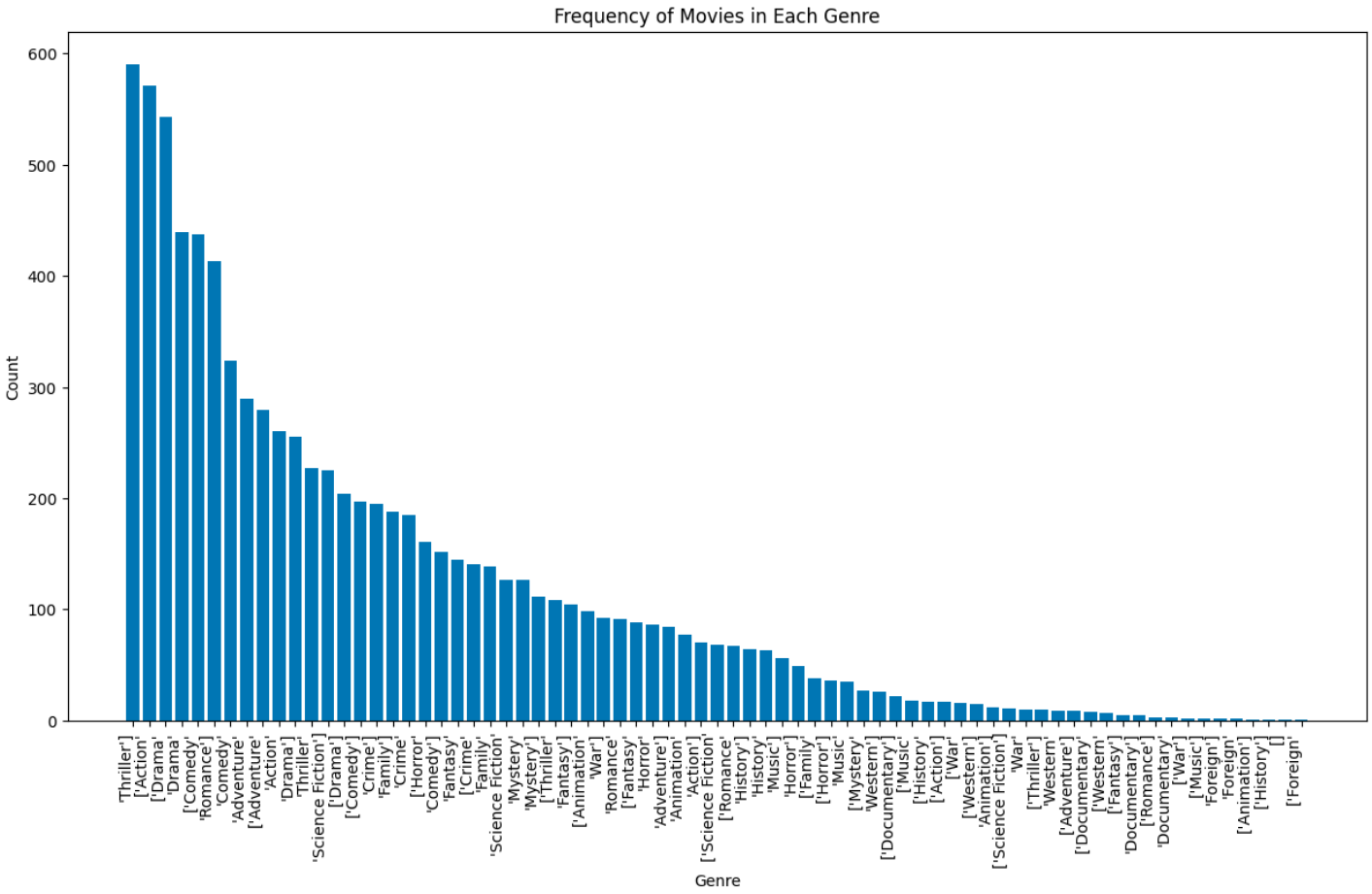
Movies with a rating above 7 were singled out, providing insights into critically acclaimed films based on user ratings.

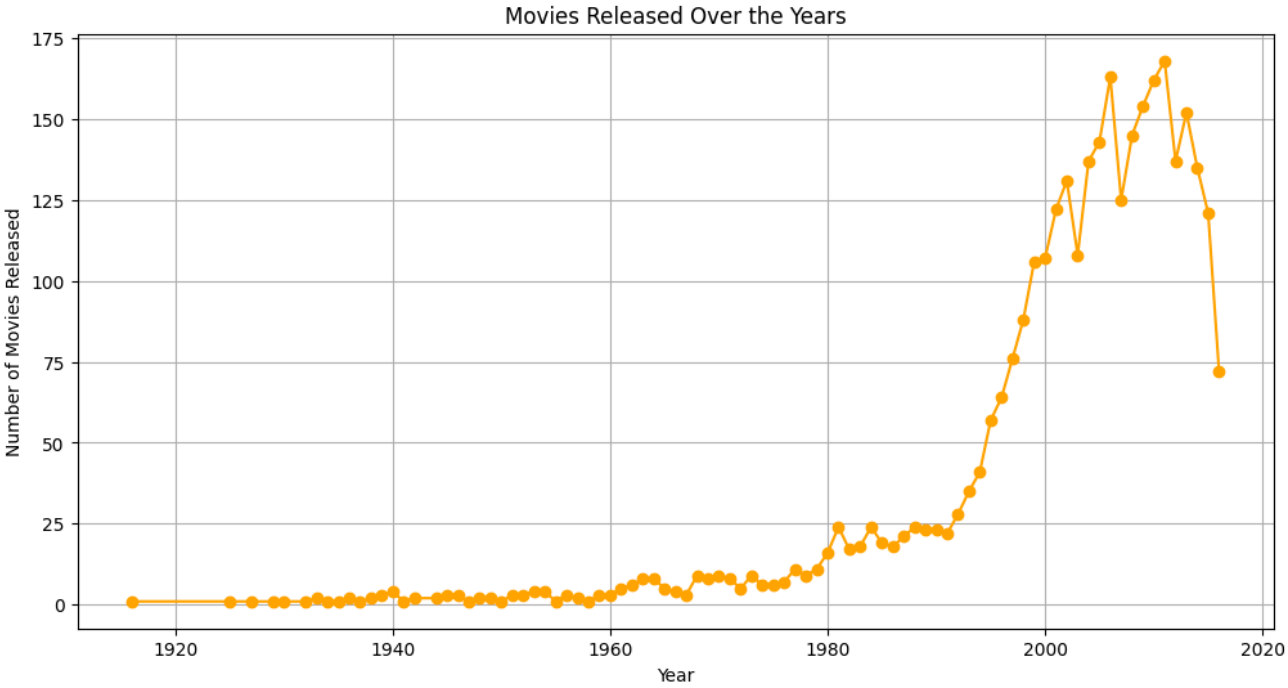
Title	Vote Average
Avatar	7.2
The Dark Knight Rises	7.6
Tangled	7.4
Avengers: Age of Ultron	7.3
Harry Potter and the Half-Blood Prince	7.4
Roger & Me	7.4
Eraserhead	7.5

Title	Vote Average
Pi	7.1
Clerks	7.4
Cure	7.4

Most Successful Genres:

A bar plot was created to visualise the frequency of movies in each genre. This analysis helps understand the distribution of genres in the dataset, potentially revealing trends or preferences.





This line graph depicts two notable trends: first, the upward trajectory in the production of movie releases over time, and second, a distinct decline in popularity. This decline could be attributed to either the impact of the 2020 pandemic or the surge in popularity of digital home movie platforms, which are less reliant on physical audience attendance.

This comprehensive data cleansing and exploration process set the foundation for further in-depth analysis and storytelling. The cleaned dataset now provides a solid basis for extracting meaningful insights from the world of movies.

THIS REPORT WAS WRITTEN BY : Elsy Theledi