Movie Data Analysis — Project Summary

By Benjamin Akingbade

1. Project Overview

This project explores a comprehensive dataset of over 4,800 movies using Python to uncover patterns, relationships, and factors that influence commercial success and critical acclaim. The analysis combines exploratory data analysis (EDA), financial correlation studies, time-series trend analysis, and geospatial visualization to understand the dynamics of the film industry. The final insights are curated in Jupyter notebooks and serve as the foundation for a Tableau dashboard.

2. Objective

To conduct an in-depth exploratory and advanced analytical study on movie performance metrics, identify meaningful relationships between budget, genres, and revenue, and visualize global industry trends.

3. Data Source

Dataset: TMDB 5000 Movie Dataset

Source: Kaggle (Open-source)

Description:

The dataset includes metadata on movies released up to 2017. Key variables include:

- **Financials:** budget, revenue, and derived profit & ROI.
- **Metadata:** original_title, runtime, release_date.
- **Categorical:** genres, original_language, status.
- **Cast & Crew:** JSON columns containing cast (actors) and crew (directors, writers).
- **Metrics:** popularity, vote_average, vote_count.
- **Location:** production_countries

4. Data Requirements Check

The dataset meets the criteria for advanced analysis:

Open-source and publicly accessible: Sourced from Kaggle.

Includes continuous variables: Budget, Revenue, Popularity, Runtime, Vote Average.

Includes categorical variables: Genre, Original Language.

Contains ≥ 1,500 rows: 4,803 entries.

Includes geographical information: Production Countries.

Contains relevant historical data: Covers release dates spanning several decades.

5. Data Preparation

Cleaning and preparation steps performed in Python (Notebook 1.1):

- **Data Loading:** Merged tmdb_5000_movies.csv and tmdb_5000_credits.csv on the movie ID.
- **Parsing JSON:** Extracted usable lists from JSON-string columns (genres, cast, crew, production_countries) to isolate Directors, Top Actors, and primary Genres.

- **Handling Missing Values:** Inspected null values in runtime and release_date and handled them appropriately to ensure data integrity.
- **Data Type Conversion:** Converted release_date to datetime objects to extract Year and Month features.
- **Feature Engineering:** Created new financial metrics: Profit (Revenue - Budget) and ROI (Revenue / Budget).

6. Exploratory Analysis

Major visual insights documented across notebooks:

- **Genre Distribution:** Analyzed which genres are most frequently produced and which generate the highest average ROI (Notebook 1.2).
- **Financial Correlations:** Investigated the relationship between budget, revenue, and popularity using scatterplots and heatmaps (Notebook 1.4).
- **Time Trends:** Visualized the explosion in movie production volume and the evolution of average profits over the decades (Notebook 1.3).
- **Word Analysis:** Generated Word Clouds to visualize the most common terms used in movie titles (Notebook 1.6).

7. Hypotheses / Questions Explored

- **Budget Influence:** Does a higher production budget significantly increase box office revenue?
- **Genre Performance:** Which genres consistently perform better in terms of Return on Investment (ROI)?
- **Star/Director Power:** Do movies directed by specific "top-tier" directors consistently generate higher revenue?
- **Geographical Patterns:** Which countries (outside the US) produce movies with the highest average user ratings?

- **Temporal Trends:** Has the average movie runtime changed significantly over the years?

---

8. Advanced Analytics

- **Correlation Analysis (Notebook 1.4):**
    - **Method:** Correlation Matrix & Heatmap.
    - **Findings:** Strong positive correlation (approx. 0.7) found between **Budget** and **Revenue**, confirming that investment drives returns. **Popularity** also showed a strong correlation with **Vote Count**.
- **Time Series Analysis (Notebook 1.3):**
    - **Method:** Aggregation by Year and Line plotting.
    - **Findings:** The data shows a clear upward trend in the number of movies produced per year, with a significant spike in average budgets entering the 21st century.
- **Text Analysis (Notebook 1.6):**
    - **Method:** WordCloud generation from original_title.
    - **Findings:** High-frequency words include "Love", "Man", "Day", and "Life", indicating recurring themes in film naming conventions.

---

9. Geospatial Component:

- **Location-Based Insights (Notebook 1.7):**
    - **Analysis:** Mapped average_rating against production_countries.
    - **Insight:** While the USA produces the highest volume, other nations (e.g., UK, New Zealand, Japan) frequently score higher on average user ratings for their top exports.
    - **Visuals:** Bar charts ranking top countries by average rating.

---

10. Results Summary

- **Budget & Revenue:** Hypothesis supported. There is a definitive strong positive correlation between budget and revenue. However, high budget does not always guarantee high *ROI.*
- **Genre ROI:** Action and Adventure movies earn the most raw revenue, but **Sci-Fi** and **Fantasy** often yield higher ROI due to massive franchise appeal.
- **Director Impact:** A small cluster of directors (Cameron, Nolan, Jackson, Spielberg) account for a disproportionate amount of total industry revenue, supporting the "Star Power" hypothesis for directors.
- **Production Volume:** The industry has shifted from low-volume/moderate-budget in the 20th century to high-volume/high-variance budget in the 21st century.

---

11. Limitations

- **Inflation:** Revenue and budget figures are not adjusted for inflation, which skews financial comparisons between modern movies and older classics (e.g., *Gone with the Wind*).
- **Missing Financials:** Some smaller or older movies in the dataset have budget/revenue listed as 0, requiring filtering for accurate financial analysis.
- **Bias in Ratings:** User ratings (Vote Average) can be subject to recency bias or "review bombing," potentially skewing the perception of quality.

---

12. Ethical Considerations

- **Public Figures:** The dataset contains names of real actors and directors. This data is used strictly for professional performance analysis and is already in the public domain.

- **Data Integrity:** Care was taken not to manipulate the data to force a conclusion (e.g., excluding "flops" to make a specific genre look better).
- **Sourcing:** Data was sourced legitimately from an open-source platform (Kaggle) which scrapes TMDb API.

13. Next Steps

- **Predictive Modeling:** Build a Linear Regression model to predict future box office revenue based on budget and cast.
- **Sentiment Analysis:** Scrape user reviews to perform sentiment analysis and see if text sentiment correlates with the numerical vote_average.
- **Clustering:** Use K-Means clustering to group movies into categories based on performance metrics (e.g., "Blockbusters," "Critical Darlings," "Flops") rather than just genre.