Table of Contents

Week 2 Project: Movie Ratings Analysis

Objective

The aim of this project was to analyze historical movie performance and audience ratings to identify patterns and insights. Using a dataset containing movie box office data (budget, domestic/foreign earnings) and audience ratings (IMDb scores), we performed data cleaning, feature engineering, and visualization to explore relationships between movie production metrics and audience reception.

Data Cleaning and Preparation

- **Duplicate removal:** Ensured all entries were unique.
- **Missing values:** Identified zeros and nulls in numeric columns ($Domestic, $Foreign, Domestic %, Foreign %, Rating, Vote_Count) and replaced them with the median of the respective columns.
- **String cleaning:** Converted rating strings like "6.126/10" to numeric float values.
- **Categorical columns:** Filled missing values in Genres and Production_Countries with the most frequent category.
- **Grades:** Created a new column Grade based on ratings:
    - A (Rating ≥ 8), B (Rating ≥ 6), C (Rating ≥ 4), D (Rating < 4).

Feature Engineering

- **Profit Margin:** Calculated as (Worldwide Gross - Domestic Budget) / Domestic Budget to understand financial success relative to production costs.
- **Grade:** Mapped numeric ratings to letter grades to categorize audience reception.

Exploratory Data Analysis and Visualization

Several Seaborn visualizations were created to explore patterns and distributions:

1. **Domestic Budget vs Worldwide Gross (Scatter Plot)**
    a. Colored by Grade, this plot shows the relationship between production budget and global earnings. Higher-budget movies generally earned more worldwide, with top-rated movies concentrated in higher revenue ranges.
2. **Rating Distribution by Grade (Box Plot)**

    a. Box plots revealed the spread of ratings for each grade category, highlighting consistency in high-rated movies.

3. Profit Margin Distribution (Histogram)
    a. Visualizing profit margins showed most movies had moderate returns, with a few highly profitable outliers.
4. Average Rating by Grade (Bar Plot)
    a. Confirmed that movies in higher grade categories received higher average audience ratings.
5. Correlation Heatmap
    a. Correlations between numeric variables were calculated and visualized, showing strong relationships between domestic and worldwide earnings, and moderate correlations between ratings and vote counts.

Insights

- Movies with higher domestic budgets tend to earn more worldwide, but audience ratings are not solely dependent on budget.
- Profit margins vary widely; some lower-budget films achieve higher relative returns.
- Genre and production country play a role in audience reception, as seen in median ratings across categories.
- Cleaning and transforming data (handling zeros, missing values, and string formats) is crucial for meaningful analysis.

Conclusion

This exercise demonstrates how integrating multiple metrics from box office performance and audience ratings can provide actionable insights into movie success. Through data cleaning, feature engineering, and visualization, we can identify patterns in financial performance, audience reception, and genre-specific trends. This project highlights essential data analysis skills, including Pandas manipulation, NumPy calculations, and Seaborn-based visualization for comparative insights.