Investigating the Impact of Digital Platforms on the Cinema Industry

Motivation

The emergence and rapid rise of streaming platforms, particularly Netflix, have reshaped the film industry. Traditional cinema now faces competition from digital platforms, affecting production choices, audience behaviors, and revenue structures. This project investigates the extent of that impact by analyzing and comparing key aspects of Netflix

Data Sources

1. **Netflix Dataset**
    - Source: Kaggle
    - Contains top-rated Netflix movies and shows from 2020–2022
    - Attributes: title, release year, duration, genre, production country, IMDb score, and number of votes
2. **Traditional Cinema Dataset**
    - Source: Kaggle
    - Covers over 7,600 movies from 1986–2016
    - Attributes: title, release year, budget, gross revenue, director, genre, country, runtime, and IMDb rating

Both datasets were merged using title and release year as keys, focusing on overlapping movies to enable fair comparisons.

Data Cleaning & Preparation

- Renamed and standardized column names for merging.
- Converted all titles to lowercase for consistency.
- Removed duplicate entries and dropped rows with missing essential values (e.g., IMDb scores, genre, budget).
- Merged datasets on title and year.
- Created IMDb score categories: Low (≤6.0), Medium (6.1–7.5), and High (>7.5) for categorical comparisons.

Exploratory Data Analysis (EDA)

1. IMDb Score Comparison

- **Netflix** movies had an average IMDb score similar to traditional films.
- Boxplots and histograms showed comparable distributions, though Netflix movies skewed slightly higher.

2. Votes Distribution

- Traditional movies typically had more votes, but outliers skewed results. Log-scale boxplots revealed Netflix titles often had fewer but still substantial user engagement.

3. Budget & Gross Analysis (Traditional Only)

- Average budget was high, with a wide variance.
- A weak positive correlation was observed between **budget** and **IMDb rating**.
- Scatterplots showed that high-budget films don't always guarantee high gross returns.

4. Genre Distribution

- Netflix favored genres like **Documentary**, **Drama**, and **Comedy**.
- Traditional cinema had a broader genre spread including **Action**, **Crime**, and **Adventure**.

5. Runtime Distribution

- Netflix and traditional movies had comparable runtimes, with traditional films showing more variability.

6. Yearly Trends

- IMDb ratings over the years remained relatively stable across both datasets with no clear upward or downward trend.

Hypothesis Testing

T-Test: IMDb Score Difference

- **Null Hypothesis (H□):** No difference in average IMDb scores.

- **Result:** *p-value* > *0.05* → Failed to reject H□.
  There is **no significant difference** between the average IMDb scores of Netflix and traditional movies.

Chi-Squared Test: Genre Distribution

- **Null Hypothesis (H□):** Genre and platform are independent.
- **Result:** *p-value* < *0.05* → Rejected H□.
  There is a **significant difference** in genre distribution between Netflix and traditional platforms.

Chi-Squared Test: Genre vs IMDb Rating

- **Null Hypothesis (H□):** Genre is not associated with IMDb score categories.
- **Result:**
  - For Netflix: Significant association found
  - For Traditional: Significant association found
    → IMDb scores vary significantly by genre in both cases.