technische universität dortmund

Statistical Analysis of Movie Data Across Streaming Platforms

Department of Statistics

TU Dortmund University

\

Github Link: Kambanl (Kamban Loganathan)

Name: Kamban Loganathan

Date: 12/19/2024

Table of Contents

1. Introduction and Motivation

In the era of Industrial Revolution 4.0, Streaming platforms emerged as a dominant force in the television industry, reconceptualising the audiences who consume movies and TV shows. Digital internet has made accessibility offered these platforms have led to a rapid shift away from traditional media consumption methods. Among the numerous streaming services, Netflix and Disney+ stand out as two of the most popular platforms, each catering to distinct audience demographics.

Netflix as a leading streaming platform appealing to a broad range of viewers from childrens to adults. On the other hand, Disney+ a family-friendly platform, prioritizes content suitable for younger audiences. The variations in target audiences and content strategies encourage data scientists to gain a more profound insight into the characteristics of content on these platforms. Are these views sustained by the data, or are they simply assumptions based on branding and market strategies?

This report aims to examine the given data from Kaggle.com - Netflix and Disney+ to explore two key aspects: the age restriction of movies and their Rotten Tomatoes score. Age restrictions provide insights into the platform's target audience, while Rotten Tomatoes scores provide the movie quality perceived by the audiences. By examining these factors, this study seeks to validate or challenge the common perceptions of these streaming services.

To achieve this, the analysis employs statistical techniques to address the following questions:

1. Does Disney+ feature movies with significantly lower age restrictions than Netflix, thereby supporting its family-friendly image?
2. Are the average Rotten Tomatoes scores for movies on Netflix higher than those on Disney+, indicating a focus on quality over quantity?

The given dataset includes information on age restrictions and Rotten Tomatoes scores for movies across multiple platforms like Netflix, Hulu, Disney+, and Prime Video. Descriptive statistics and hypothesis testing are utilized to derive meaningful insights and conclusions.

The statistical analysis may provide valuable information for consumers, directors, entrepreneurs who aspire to touch the streaming industry. For viewers, this report provides suitable platform selection based on their preferences, for eg: parents may find it suitable to stream appropriate content to their kids. For content creators and platform

managers, understanding the patterns and perceptions of their libraries could guide content curation and strategy.

This report is structured as follows: Section 2 provides a detailed description of the problem and dataset, Section 3 outlines the statistical methods employed, Section 4 presents the evaluation results, Section 5 offers a summary of findings, and Sections 6 and 7 include the bibliography.

2. Detailed Description of the Problem

Data Overview

The given dataset used for this analysis includes information for around 9515 movies from various streaming platforms with key variable being:

- **Title:** Name of the movie
- **Year:** Release year
- **Age:** The age restriction for the movie, categorized as "All," "7+," "13+," "16+," and "18+." **This is an ordinal variable.**
- **Platform Availability:** Binary indicators (0 or 1) for availability on Netflix, Hulu, Prime Video, and Disney+.
- **Rotten Tomatoes Score:** A continuous variable representing the movie's rating on Rotten Tomatoes, ranging from 0 to 100.

Initial Observations

- **Variable Types:**
  - Age restrictions are ordinal.
  - Rotten Tomatoes scores are numerical (continuous).
  - Platform indicators are binary (0 or 1).
- **Missing Data:**
  - A significant portion (~40%) of the *Age* column is missing in the dataset, which is essential for our analysis.To address this, we applied **mode-based imputation**.

Research Questions

This report will employ statistical techniques, including hypothesis testing, to answer these questions and provide actionable insights:

1. To determine the distribution of age restrictions for movies on Netflix and Disney+.
2. Does Netflix data reflect a preference for mature audiences, as evidence, preliminary examination shows a substantial proportion of movies rated “18+”?
3. Does Disney+ predominantly feature content suitable for younger audiences?
4. Evaluation of the descriptive statistics (mean, median and standard deviation) of Rotten Tomatoes scores for the movies. Are there any significant differences in average scores?

3. Methods

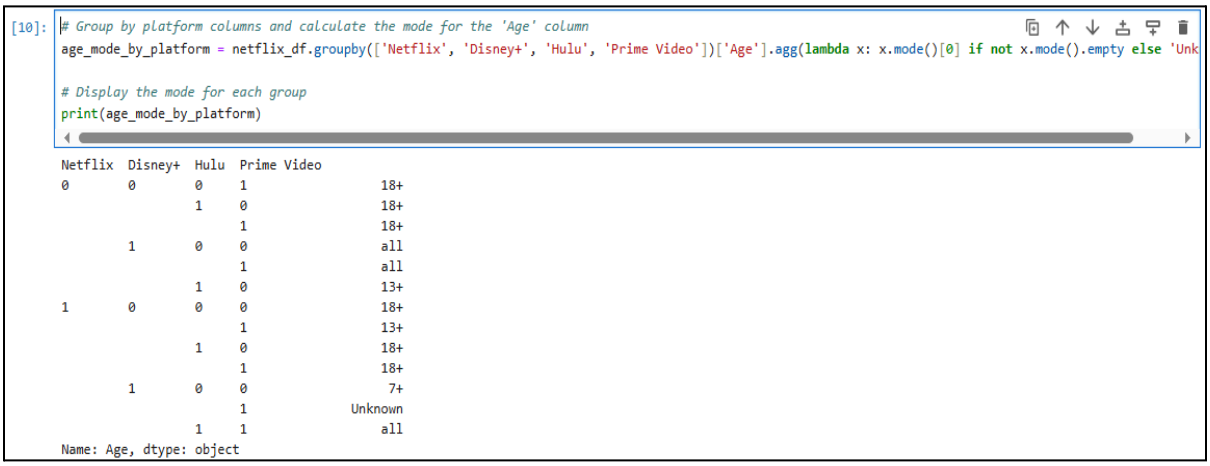
1. Data Cleaning and Preprocessing

The dataset used in this analysis was sourced from Kaggle and required several preprocessing steps to ensure its suitability for statistical analysis. The following data cleaning techniques were employed:

1.1 Handling Missing Values

Missing data were present in columns such as Age and Rotten Tomatoes. These were addressed as follows:

1. Age column: Since our Analysis primarily deals with Age categories and restrictions we cannot drop the 40% of the missing data. Thus we went on with a Mode based Imputation of the missing values.
    a. **Mode Imputation :** The Missing values in the Age column were replaced by the mode of each group. I.e., The Age column was grouped by platform categories (like Netflix, Disney+. Hulu, Prime Video), and if the no mode exists, we simply replace it with Unknown.



```
[10]: # Group by platform columns and calculate the mode for the 'Age' column
age_mode_by_platform = netflix_df.groupby(['Netflix', 'Disney+', 'Hulu', 'Prime Video'])['Age'].agg(lambda x: x.mode()[0] if not x.mode().empty else 'Unk

# Display the mode for each group
print(age_mode_by_platform)
```

```
Netflix  Disney+  Hulu  Prime Video
0        0        0     1                18+
                  1     0                18+
                        1                18+
         1        0     0                all
                        1                all
                  1     0                13+
1        0        0     0                18+
                        1                13+
                  1     0                18+
                        1                18+
         1        0     0                 7+
                        1            Unknown
                  1     1                all
Name: Age, dtype: object
```

Screenshot: Group by platform columns and display the mode for each group

Why mode based imputation and not predictive modelling?

- Predictive modeling for imputing missing values is unfeasible due to several limitation such as:

        - Lack of strong predictors like movie genres or audience demographics
        - Available columns (eg: Rotten tomatoes scores, platform availability) provide insufficient context.
        - ~40% of missing value limits reliable model training - increasing the risk of overfitting.
        - The dataset's structure and lack of relevant variables further complicate feature engineering, validation, and resource efficiency.
2. Rotten Tomatoes: 7 Entries in the column had NAN, thus for maintaining consistency and facilitating analysis, thus these entries were removed from the dataset.

1.2 Data Transformation

To extract Numerical values from the textual data, String Parsing techniques were used.

- The Rotten Tomatoes column originally contained percentage scores as strings such as 98/100, 90/100. The column entries were transformed into numerical values using regular expressions. These scores were then cast as floating point numbers for compatibility with statistical methods.
- Conversion of the Age column to numerical codes (eg: 18+ to 18)

1.3 Removal of Redundant Columns

- Columns such as Unnames: 0, ID, and raw Rotten Tomatoes strings were dropped from the dataset before analysis, since these columns deemed irrelevant to the analysis., This ensures streamlined data for computational efficiency and interpretability.

2. Grouping and Aggregation

For further understanding of the dataset, we perform Grouping operations for platform specific patterns. Thus the Data was grouped by Platform categories (Netflix, Disney+, Hulu, Prime Video) to compute aggregated metrics such as:

- Mode: As described earlier, used to impute Missing values
- Counts: For identifying the total numbers of movies associated with each platform.

These operations employed the group-by functionality in Python's Pandas library, leveraging partitioning principles to segment the dataset into mutually exclusive subsets.

3. Statistical and Analytical Techniques

3.1 Descriptive Statistics

Descriptive analysis was performed to summarize key features of the dataset. These include:

- Counts and distributions of content types.
- Measures of central tendency (mean, median) and dispersion (standard deviation) for numerical columns such as the Rotten Tomatoes Scores.

3.2 Advanced Statistical Analysis

While primarily exploratory, the basic inferential analysis and Time-Based analysis were performed:

- **Inferential Techniques**: Hypothesis testing methods may be implied for comparing distributions across platforms.
- **Time-Based Analysis**: Using columns like Data added to analyze trends in content acquisition.

Inferential Analysis

1. **Comparing Age Restrictions:**
    - Method: Mann-Whitney U Test, a non-parametric test, is employed to analyze the ordinal nature of age restrictions.
    - Mathematical Rationale: Let $X1$ and $X2$ represent age restriction distributions for Disney+ and Netflix, respectively. The test evaluates the null hypothesis

$$Ho: Median\,(X1) = Median\,(X2)$$

    - Justification: The Mann-Whitney U Test is suitable for ordinal data that does not meet normality assumptions.

2. **Comparing Rotten Tomatoes Scores:**
    - Method: Two-sample t-test is applied for comparing means of continuous variables.
    - Mathematical Rationale: Let Y1 and Y2 represent Rotten Tomatoes scores for Netflix and Disney+. The test evaluates $Ho: \mu 1 = \mu 2$,where $\mu 1$ and $\mu 2$ are the means of the variable for Netflix and Disney+ respectively.
    - Assumptions: Independent samples, approximate normal distribution of scores, and equal variances. Levene's test checks for variance equality.

3.3 Data Visualization

Several graphical methods were employed to interpret and communicate findings effectively:

- **Bar Charts**: To display the distribution of content genres and country representation across platforms.
- **Pie Charts**: To visualize proportions of content types (e.g., Movies vs. TV Shows).
- **Histograms**: For examining the distribution of numerical variables such as IMDB Ratings.
- **Heatmaps**: To reveal correlations between numerical variables, aiding in identifying relationships or redundancies.

4. Software and Libraries Used

The following tools and libraries were critical to this analysis:

- **Pandas**: For data manipulation, cleaning, and aggregation.
- **NumPy**: For numerical computations and handling arrays.
- **Matplotlib and Seaborn**: For creating visualizations.
- **Regular Expressions (re module)**: For parsing textual data.
- **Jupyter Notebook**: To integrate code, results, and documentation se

4. Evaluation

4.1 Descriptive Analysis

Using descriptive statistics we can describe a chunk of raw data using statistics, graphs and tables.

Let us look into some of the insights into the distribution of age restrictions and Rotten Tomatoes score for movies on Netflix and Disney+. Some of the key observation are as follows:

Age Restrictions

1. **Netflix:**
    - Data reveals a higher count for the Netflix movies in the 18+ and 10 + categories suggesting a focus on content tailor for mature audiences.
    - Noticeable decrease in movies suitable for "All" and "7+" categories, reinforcing the platform's reputation for diverse, adult-focused content.
2. **Disney+:**
    - A dominant portion of the content falls under the "All" and "7+" age categories, reflecting its family-friendly positioning.
    - Very few movies are categorized as "16+" or "18+," supporting the assumption that Disney+ prioritizes younger audiences.

The age distribution confirms the contrasting audience focus of the platforms. Disney+ appears to emphasize content suitable for families and younger viewers, while Netflix targets a broader, more mature audience.

Rotten Tomatoes Scores

1. **Netflix:**
    - The Rotten Tomatoes scores range from 0 to 100, with a few outliers scoring near the extremes (e.g., 10 and 90+).
    - The average scores seem widely spread, suggesting Netflix hosts content across varying levels of critical acclaim.
2. **Disney+:**
    - The scores are distributed more narrowly, with fewer movies receiving extremely low or high ratings.
    - This distribution might indicate that Disney+ curates content with consistently moderate to high audience reception.

Conclusion for Rotten Tomatoes Scores:
While both platforms offer movies with varying scores, Disney+ appears to maintain a higher baseline of quality, with fewer extreme outliers. Netflix, on the other hand, accommodates a broader range of content quality, likely to cater to a diverse audience

Visualization: Bar chart shows the distribution of age restrictions for both platforms.

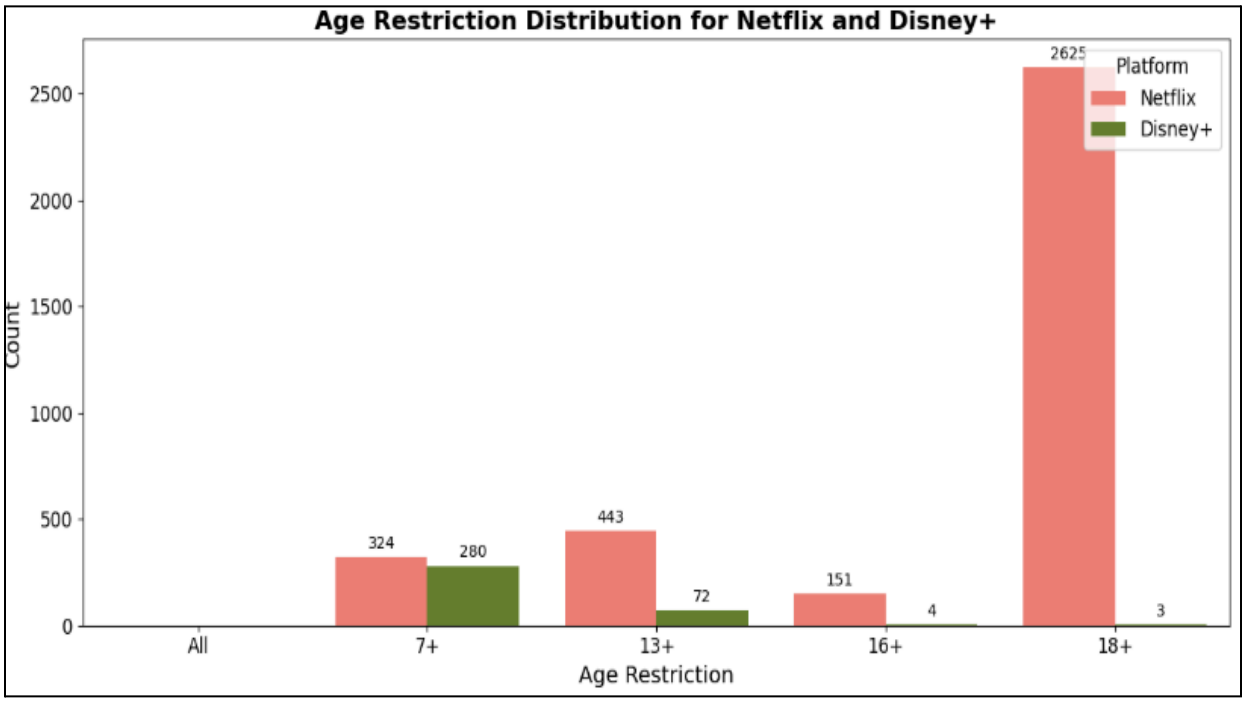


Figure: Distribution of age restrictions for both platforms

Computing basic statistics for Rotten Tomatoes Scores:

- **Netflix:**

Statistical Measure	Value
Count	3688.00
Mean	54.45
Median	53.00
Minimum	10.00
Maximum	98.00
Standard Deviation	13.85

Disney+:

Statistical Measure	Value
Count	922.00
Mean	58.31
Median	57.70
Minimum	10.00
Maximum	96.00
Standard Deviation	13.95

Visualization: Boxplots depict the distribution of Rotten Tomatoes scores.

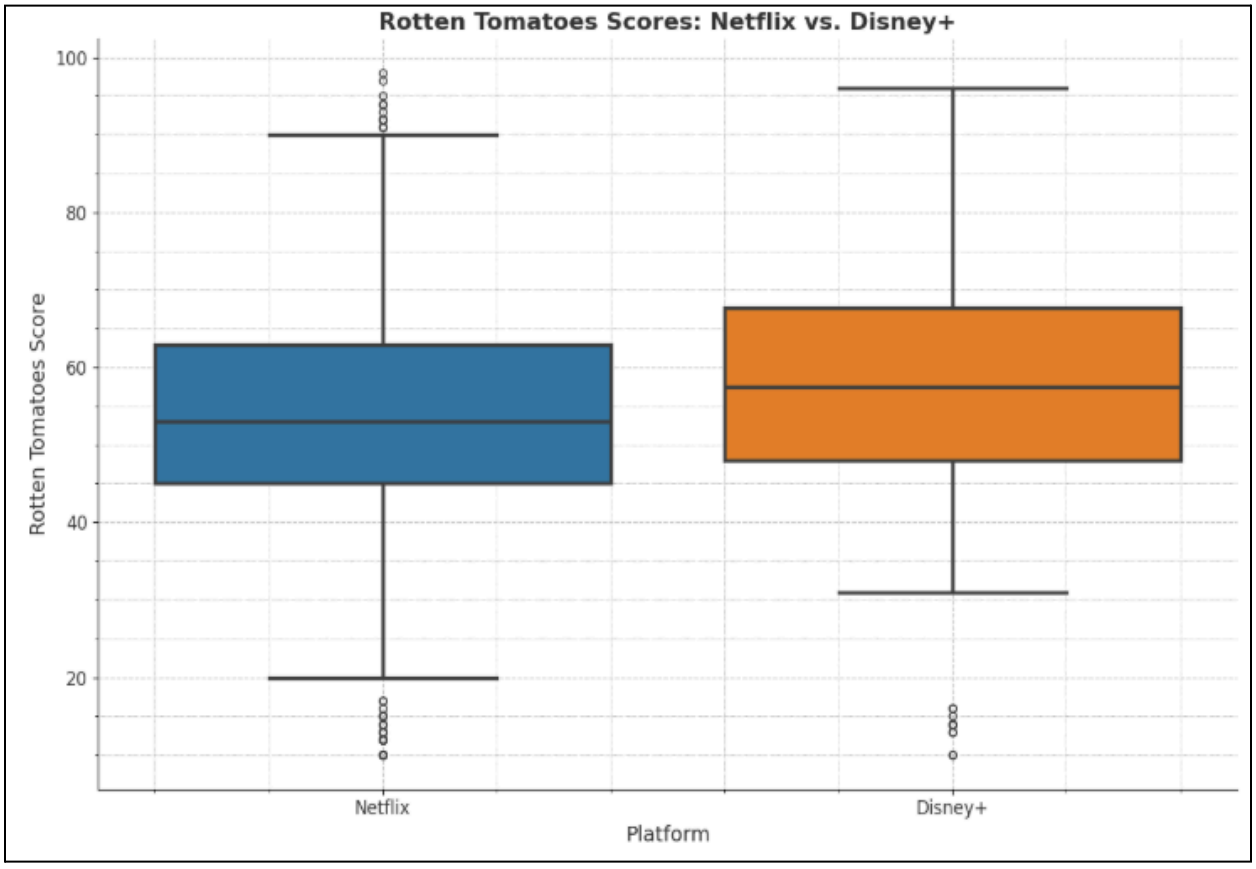


Figure: Distribution of Rotten Tomatoes Score using Box plot

4.2 Inferential Analysis

Code for Mann-Whitney U Test and t-test:

```
# Perform Mann-Whitney U Test for Age Restrictions
age_test_stat, age_p_value = mannwhitneyu(netflix_data['Age_Code'].dropna(), disney_data['Age_Code'].dropna(), alternative='greater')

# Perform Two-Sample T-Test for Rotten Tomatoes Scores
score_test_stat, score_p_value = ttest_ind(rotten_netflix.dropna(), rotten_disney.dropna())

# Results
test_results = {
    "Age Test Statistic": age_test_stat,
    "Age Test P-Value": age_p_value,
    "Rotten Tomatoes Test Statistic": score_test_stat,
    "Rotten Tomatoes Test P-Value": score_p_value
}
print(test_results)
```

```
{'Age Test Statistic': np.float64(3208869.5), 'Age Test P-Value': np.float64(0.0), 'Rotten Tomatoes Test Statistic': np.float64(-7.567567531710881), 'Rot
ten Tomatoes Test P-Value': np.float64(4.5645960458960904e-14)}
```

```
test_results
```

```
{'Age Test Statistic': np.float64(3208869.5),
 'Age Test P-Value': np.float64(0.0),
 'Rotten Tomatoes Test Statistic': np.float64(-7.567567531710881),
 'Rotten Tomatoes Test P-Value': np.float64(4.5645960458960904e-14)}
```

Figure: Python Code for Mann-Whitney U Test and t-test

Age Restriction Comparison

- **Result:** The Mann-Whitney U Test yielded a test statistic of 3208869.5 and a p-value of 0.0, leading to rejecting the null hypothesis. Thus, Disney+ has significantly lower age restrictions than Netflix.

Rotten Tomatoes Score Comparison

- **Result:** The two-sample t-test yielded a test statistic of -7.567567531710881 and a p-value of approximately , leading to rejecting the null hypothesis. Hence, the average Rotten Tomatoes scores for movies on Netflix and Disney+ are significantly different.

5. Summary

The Netflix analysis conducted using both descriptive and inferential statistical evaluations to compare age restrictions and movie rating between the two top streaming platforms Netflix and Disney+:

1. **Descriptive Analysis:** It was evident from the analysis that Disney+ emphasis on family-friendly content as their majority of its movies were classified as ‘All’ or ‘7+’. On the other hand, a higher percentage of Netflix platforms has a greater proportion of content rated “16+" and "18+". Rotten Tomatoes scores vary significantly among platforms, as seen by boxplots and summary statistics.
2. **Inferential Analysis:**
    - **Age Restrictions:** The Mann-Whitney U Test (test statistic: 3208869.5, p-value: 0.0) confirmed that Disney+ features significantly lower age restrictions compared to Netflix, reinforcing its reputation as a family-oriented platform, catering to audiences seeking wholesome entertainment among all the age groups.
    - **Rotten Tomatoes Scores:** The two-sample t-test (test statistic: -7.567567531710881, p-value: ~0.0) demonstrated a significant difference in the average Rotten Tomatoes scores, with Netflix showing higher variability and potentially higher mean scores than Disney+. Netflix's diverse content library, including mature themes, resonates with its broader target audience.

Open Questions and Future Research

Future research could explore additional factors, such as:

- The impact of genres in shaping audiences preferences on these platforms
- The impact of movie popularity (e.g., viewer ratings, box office revenue) on platform strategies.
- A deeper dive into the distribution and trends of original versus licensed content.

A comprehensive Understanding of these factors will enable Netflix and Disney+ to better serve their customers and adjust to changing consumer needs.

6. Bibliography

1. Python Software Foundation. Python Language Reference, version 3.8. Available at Python for Data science
2. Kaggle website: Netflix and Disney+ Movie Data
3. Seaborn Documentation:Seaborn - Statistical Data Visualization
4. Scipy Documentation: SciPy - Open Source Scientific Tools
5. Python for Data Analysis by Wes McKinney (O'Reilly Media, 2017)
6. Matplotlib Documentation:Matplotlib: Visualization with Python
7. Pandas Documentation: Pandas with Python
8. NumPy Documentation:Numpy: Scientific Computing with Python
9. Levene's Test Overview: Brown, M. B., & Forsythe, A. B. (1974). Robust Tests for Equality of Variances. Journal of the American Statistical Association, 69(346), 364-367.
10. Mann-Whitney U Test Reference: Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics, 18(1), 50-60.
11. Two-Sample T-Test Overview: Student. (1908), Lehmann, E.L. (1992). Introduction to Student (1908) The Probable Error of a Mean. In: Kotz, S., Johnson, N.L. (eds) Breakthroughs in Statistics. Springer Series in Statistics. Springer, New York, NY. The Probable Error of a Mean