TV show analysis

Kiyana Mirbaghestan

December 2024

Contents

1 Introduction

The global rise of streaming platforms has significantly reshaped entertainment consumption, providing diverse content to audiences worldwide. This Analysis investigates the number of viewers on Netflix and Disney+ and measures how age restriction impacts several viewers. Moreover, we can understand which platform is better based on the Rotten Tomatoes score. Here, we start by optimizing performance and strategies.

1.1 Motivation

The age restriction of watching different types of movies is one of the most crucial factors and concerns in parents' opinion. Not all platforms are appropriate for content of all age categories. Moreover, based on the Rotten Tomatoes score, we can see which platform will be our choice if a new movie spreads. Here, we focus on two platforms, Disney and Netflix, more than others.

1.2 Problem-Statement

The Analysis addresses whether the age category influences the type of platforms that different people of different ages are interested in. Then, we want to see which of these two platforms is more pleasant for everyone.

1.3 Approach

To address the questions, we follow these steps for Analysis: 1. Descriptive Analysis: Analysis of datasets using statistical measures (mean, median, standard deviation) and visualize them with (box plots and bar charts) to answer the questions above. 2. Hypothesis testing: Statistics tests to see if there are significant differences in age groups and if there are substantial differences between their scores.

1.4 Main result

The Analysis concludes that there is a significant difference between Disney-age people and Netflix-age people, as well as a dramatic difference in their Rotten Tomatoes score.

1.5 Structure Report

The report contains an organized layout, introduction, motivation, approach, and prominent result. Also, it follows a detailed description, which is the first display of data; methods like data preprocessing and descriptive statistics will be in our report, and Hypothesis testing, visualization, and a summary will be the final section here. In these ways, we can understand how to solve our tasks and even help in the future.

2 Detailed Description of the problem

The task is to investigate the influence of age category on the kind of platform that has been watched. Also, the type of platform has been watched with its score on Rotten Tomatoes. It will be helpful in future, too.

2.1 Research Questions

1. Age category and sorts of the platform: Are older people interested in Netflix and younger people more passionate about Disney? 2. Sorts of the platform and Rotten Tomatoes score: Which platform is more interesting for people?

2.2 Data Overview

There are 9515 rows and columns here:
ID: Unique identifier to each movie
Title: Name of each movie
Year: The year of publishment
Age: The age restriction for each film
Rotten Tomatoes: Score of each movie
Netflix: Availability in Netflix
Hulu: Availability in Hulu
Prime Video: Availability in Prime Video
Disney: Availability in Disney

3 Methods

In this section, I describe statistical methods used to analyze the impact of age categories on watching platforms and recognize which is more exciting based on the Rotten Tomatoes score. This Analysis used R programming language, leveraging libraries like ggplot, dplyr, tidyr, gridExtra, car, and VIM.

3.1 Data preprocessing

Description: Data preprocessing involves cleaning raw data without changing the accuracy and concept of data. So, we can rely on visualization and statistical tests. Here are some steps in preprocessing:
1. Changing character to numeric: To contrast datasets better for condition, we need to convert Rotten Tomatoes and age from character into numeric.
2. Handel missing value: Rotten Tomatoes column has a low missing rate. However, the age column has a high missing rate. Thus, we should solve this issue differently.
●For Rotten tomatoes: we convert the missing part to the mean of this column. However, we did not choose the median because there were few outliers.
●For age: we use KNN methods to fill in the missing data.

Boxplot of Scores
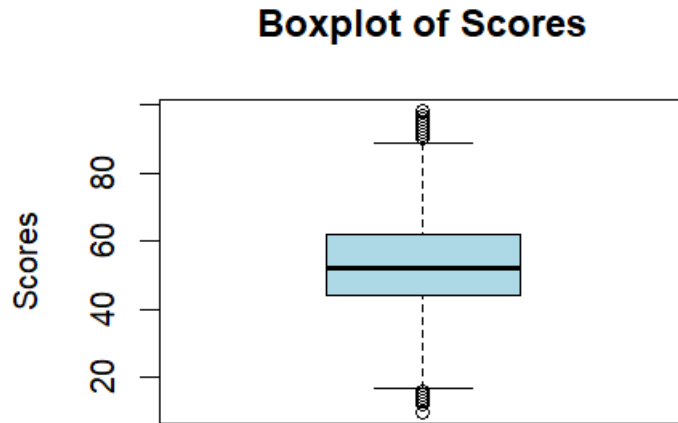

Figure 1: Boxplot

Reference:
o Max Kuhn Kjell Johnson (2013). "Applied Predictive Modeling"

3.2 Descriptive Statistics

Description: Descriptive statistics summarize and organize data to highlight key patterns or features. This includes measures of tendency (mean, median, mode), variability (range, standard deviation) and data visualization (charts, histograms). It provides a snapshot of the dataset without making predictions or inferences.

3.3 Hypothesis Testing

Descriptive statistics summarize and organize data to highlight key patterns or features. This includes measures of tendency (mean, median, mode), variability (range, standard deviation) and data visualization (charts, histograms). It provides a snapshot of the dataset without making predictions or inferences.

3.4 Shapiro Wilk test

Purpose: To determine if data follows a normal distribution. There are four parts for examination: Disney ages, Netflix ages, Disney Rotten score and Netflix Rotten score.
Null Hypothesis (H0): The part of the data we work on comes from a normal distribution.
Alternative Hypothesis (H1): That part of the data does not come from a normal
distribution population.
Mathematical Definition: The Shapiro-wilk test statistic W is calculated as:

$$W = \frac{\left(\sum_{i=1}^{n} a_1 x_{(i)}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

W=1:Perfect agreement with normality.
$W < 1$:Increasing deviation from normality.
References: J. P. Marques de Sa (2003). "Applied Statistics using SPSS, STATISTICA and MATLAB"

3.5 Mann-Whitney test

Purpose: This test determines whether one group tends to have higher or lower values than the other group.
Null Hypothesis(H0): The distributions of the two groups are identical (no difference in central tendency).
Alternative Hypothesis (H1): The distribution of the two groups is different.
Mathematical Definition: Here is its formula:

$$U = n_1 n_2 + \frac{n_1 \left(n_1 + 1\right)}{2} - R_1$$

n1, n2: Sample size of groups 1 and 2.
R1: Sum of ranks for group 1.
For a two-tailed test, reject H_0 if:

$$U \leq U_{\text{lower}} \quad \text{or} \quad U \geq U_{\text{upper}}$$

References: Mwarumba Mwavita, Kamden K. Strunk (2024). "Design and Analysis in Quantitative Educational Research" Davide J. Sheskin (2020). "Handbook of parametric and nonparametric statistical procedures" Davide S. Moore-George P. McCabe-Bruce A. Craig (W. H. Freeman and Company 2009). "Introduction to the practice of statistics 6th ed.

4 Evaluation

4.1 Descriptive Statistics

Both age and Rotten Tomatoes score are categorical. We categorize age as follows: (+1 to all, +7 for kids and more, +13 for young teens and more, +16 for teens and more, and +18 for adults). We convert the Rotten Tomatoes score to numeric values for more meaningful Analysis.

Statistics Comparison

Disney Rotten Score vs. Netflix Rotten Score Disney age vs. Netflix age

Statistics	Disney Rotten Score	Netflix Rotten Score
Median	57.5	53
Mean	58.31	54.45
Standard Deviation	13.95331	13.83454
Minimum	10	10
Maximum	96	98

Table 1: Comparison of Rotten Scores between Disney and Netflix

Statistics	Disney Age	Netflix Age
Mode	1-6 years old	+18 years old

Table 2: Comparison of Age Groups for Disney and Netflix

These statistics indicate that, on average, Disney scores more than Netflix, with a mean of 58.31 for Disney compared to 54.45 for Netflix. The second chart above shows that most Disney movies are appropriate for all ages. However, movies on Netflix are only eligible for adults.

The following charts show that, unfortunately, the proportion of people (with +7) who watch Netflix is more than those who watch Disney. We should restrict them better for these ages. Moreover, people with +18 are not interested in Disney. (Figure2) We have other chats relating to our report, and it's for the Rotten Tomatoes score. I categorize them based on their score in 5 sections. Lower people watched Disney. Thus, it is natural to have a lower score. But here, we consider the proportion of the population, not their count. On both the good and bad sides, Netflix has more proportion. (Figure 3)

The distribution between good and bad sides is approximately the same. On both sides, the positive side is more than the negative. As we saw, their mean and median in the descriptive part were roughly the same. However, we must be sure of the tests. Positive side: perfect, great, good Negative side: You need more effort, so don't worry next time. (Figures 4 and 5)
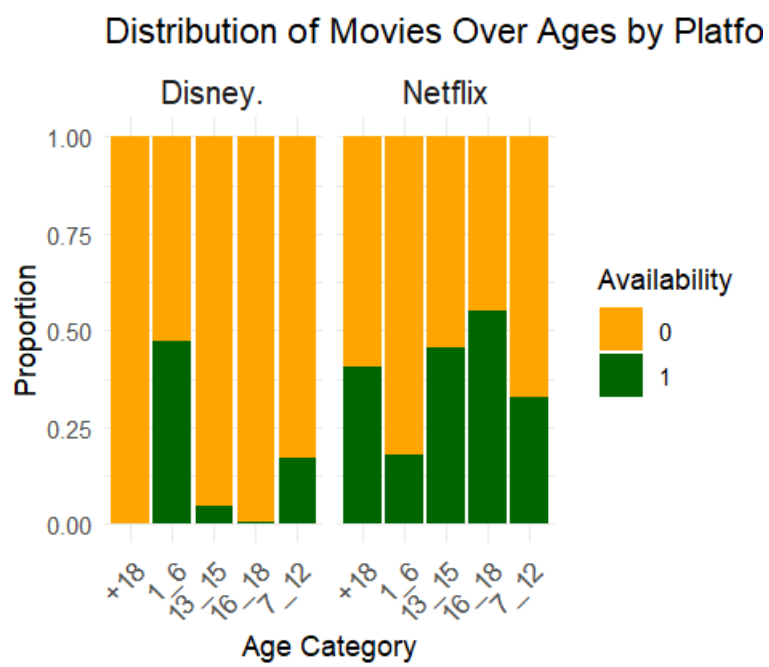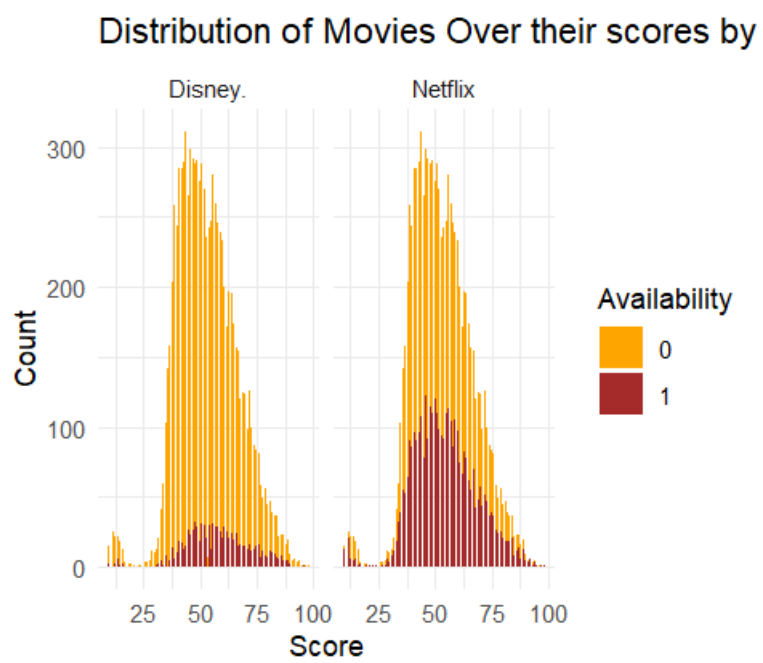
Figure 2: Bartcharts of Ages
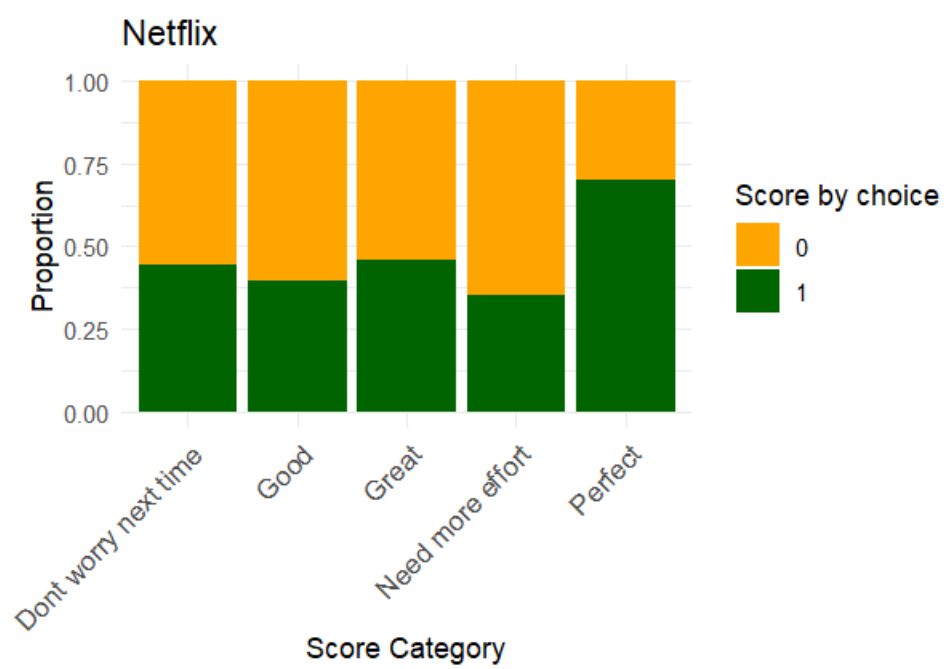
Figure 3: Distribution

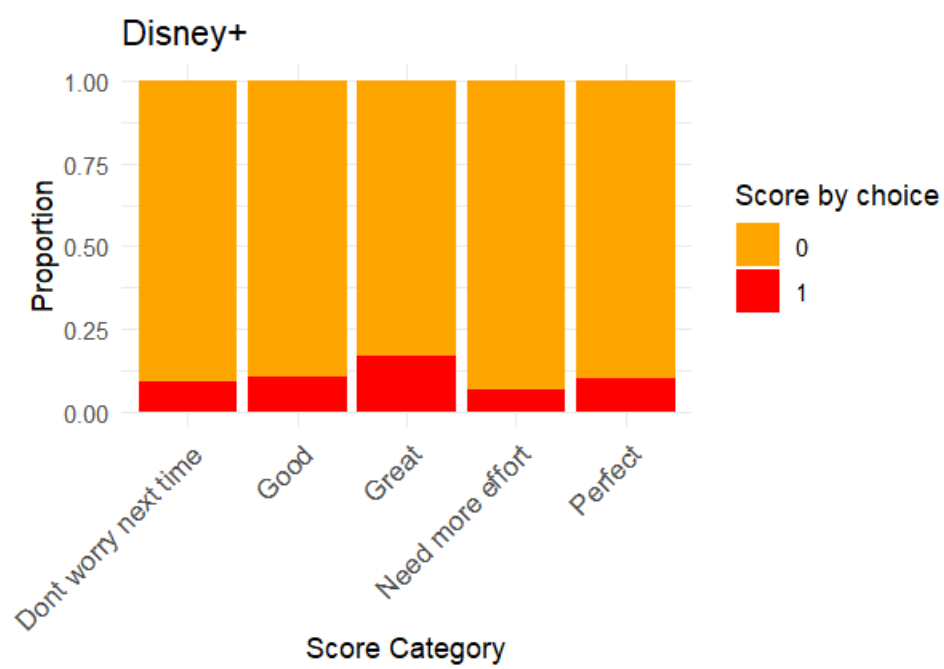Figure 4: Bartcharts of scores for Netflix

Figure 5: Barcharts of score for Disney

4.2   Hypothesis testing

About ages: To formally answer research questions, I used appropriate tests. Question 1: Do our considering samples have a normal distribution? Disney and Netflix ages are not typical. W for Disney's age in the Shapiro test is 0.75207, and for Netflix, age is 0.7969. Question 2: Is their age different? I used the Mann-Whitney test to understand it because I wanted to see if I should use tests for standard or non-normal samples. H0: Do children watch Disney more than Netflix? H1: children watch Netflix more. We reject Ho because the p-value is less than 0.05 and w=356011. From large w, we can understand a significant difference between them. Surprisingly, Disney has more limitations than Netflix movies. That's what charts told us before. About scores: Question 1: Are Disney and Netflix scores come from normal distribution? Disney and Netflix scores come from normal distribution. W for Disney and Netflix are 0.98304 and 0.98445 simultaneously. However, their p-value is tiny, and the normality assumption is rejected. Question 2: Are Disney and Netflix different in medians? Their p-value is 1. Therefore, we understand no significant difference between median and total distributions. The w=1988799

5   Summary

The project investigated about two platforms, Disney and Netflix. We address two questions to find out about a total of the research: 1. Do the limitations impact the availability of platforms, especially Netflix? 2. Which of Netflix and Disney are better based on Rotten Tomatoes?

5.1   Key Findings

1. Descriptive Analysis: Age: It does not influence the sort of platforms that people have watched by people. However, young people of some unique age, like +7, are more interested in Netflix. The mode for ages for Disney is 1-6, and for Netflix, it is +18. Score: There is no significant difference in score. Both sides have a mean and median between 50 and 60. Visualization: box-plots indicate the same result. In the age section, the restrictions do not affect viewers, especially children, and the distribution of scores is approximately the same. 2. Hypothesis Testing: Shapiro-Wilk: All four societies have very low p-values. Thus, they all have a normal distribution. (w=0.75207 w=0.7969 for ages),(w=0.98304 w=0.98445) Man-Whitney: The limitation for Netflix doesn't impact a young number of viewers. (w=356011 p-value= very low, for age) The scores that both platforms get are not that much different (w=1988799 p-value=1, for scores)

5.2 Discussion

The Analysis revealed that despite some limitations of Netflix, it has more viewers and availability in children with +7 and these limitations just influences on children under age 7. Also, the scores that the two platforms have gotten are almost the same based on the statistical tests and the visualizations which had been done on the previous pages.