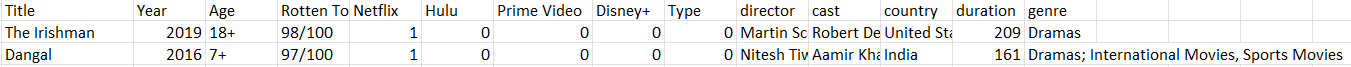
**Data Analysis on OTT Platforms and their Ratings**

**Preliminary Discussion & Data Description**

An over-the-top media service is a media service offered directly to viewers via the Internet and is among top business in the world today. In this report we’ve analyzed the data of these OTT platforms such as Netflix, Amazon, Hulu and Disney+ to understand the content hosted by these platforms. The data for this report is downloaded from Kaggle consisting of 5 data files. Out of these 5 data sets, the main dataset (Movies) comprises of columns such as Title, Release Year, Age group, Rotten Tomatoes Rating and flags indicating the platform on which the movie/show is available. Rest of the 4 datasets have further information such as Director, Cast, Duration(Running time) & Genre of the shows/movies hosted on each platform. These 5 datasets were joined together based on title of the show/movies while retaining the number of rows in the Movies dataset (9515 movies/shows).

Below are 2 rows of the data after joining all 5 datasets:



This micro report involves analyzing this dataset to understand the differences in the content hosted by each platform, for example: What age group does a platform serve more? Or how does a platform select their content. To test these hypotheses, the data needs to be processed to make it ready for statistical analysis. In this process, 4 operations were undertaken to clean the data:

1. The data was checked for NAs, and it was observed that Rotten Tomatoes column had 7 NAs. These rows were removed
2. It was also observed that Title column had movies with a decimal number and such movies didn’t exist. There were 2 such movies and the records were removed
3. Title column had movies starting with # and the has was removed using gsub
4. Duration column contains mins as the unit for duration which was removed using gsub
5. Rotten tomatoes column contained rating out of hundred. This column was split to retain the left part. It was done using separate command of tidyr package

The age column of data is available only in 5,174 records however, the NAs were not removed because this will be loss of significant amount of data. While testing hypothesis regarding age, only 5,174 records will be used.

**Exploratory Analysis**

In the exploratory analysis, it was found that Netflix had the greatest number of data points in the sample followed by Prime. Below is the distribution of Movies across platforms in the dataset:



Multiple platforms group represents movies which were listed across more than 1 platform.

It was also found that ratings vary across these platforms and Hulu had the highest avg rating as compared to others.

Background pattern

Description automatically generated

Based on the above result, the null hypothesis is that platforms and ratings are independent of each other. This hypothesis was tested using Anova test for independence. Anova Test helps determine if there is a statistically significant difference between two or more groups. 253 movies that were hosted on multiple platforms were not considered in the analysis.

: Movie Ratings and Platform are independent.

: Movie Ratings and Platform are not independent.

Text

Description automatically generated

In the above result, p value is below 0.05 therefore, the null hypothesis was rejected and it was established that platforms and ratings are not independent. Tukey HSD test was done to understand how each of the groups are different from each other.

Text

Description automatically generated

Since, p-value of each group is less than 0.05, it is established that all platforms host movies with statistically different ratings. Hulu has movies with 1.7 rating higher than Disney. Disney has movies 4.1 rating higher than Netflix followed by Prime having movies with 4.3 rating lower than Netflix. In below diagram, Rating increases from left to right

Hulu

Disney

Netflix

Prime

Along with this, it was found that out of 5,174 movies that had Age group available, Disney movies had the highest number movies for younger population as compared to other platforms. Text

Description automatically generated with low confidence

This data set serves as a basis to apply Chi-square Test which essentially allows to compare proportions of different groups statistically. We used these methods to test our hypothesis that platform and age rating of movies are independent of each other.

: Age Ratings and Platform are independent.

: Age Ratings and Platform are not independent.

Text

Description automatically generated

Since p-value is less than 0.05, the null hypothesis is rejected, and it is established that Age ratings and platforms are dependent. It was further checked what Platform and Age group is contributing the most to this observation. The expected values for each Platform and age group were compared with their respective actual values and it was found that Disney has very fewer 18+ movies and have very high movies for all age groups than expected.

**Recommendation**

When we did Anova test for Independence, one of the assumptions that the variances of all the groups are equal failed. Given that the sample size of all the group is not equal, it is recommended that we check the box plots and run Barlett’s test.

Chart, box and whisker chart

Description automatically generated

The variance of ratings in each platform can be seen by the length of each box plot. The longer the box, the higher the variance. For example, we can see that the variance is a bit higher for Disney & Netflix as compared to both Hulu & Prime.

**A screenshot of a computer

Description automatically generated with medium confidence**

[Bartlett’s Test](https://www.statology.org/bartletts-test-in-r/) tests the null hypothesis that the samples have equal variances vs. the alternative hypothesis that the samples do not have equal variances. In this case, the p-value of the test is less than the alpha level of 0.05. This suggests that the samples do not have equal variances.

However, if the sample sizes are not the same and this assumption is severely violated, run a [Kruskal-Wallis Test](https://www.statology.org/kruskal-wallis-test/), which is the non-parametric version of the one-way ANOVA.

**References**

Dataset: <https://www.kaggle.com/datasets/titassaha/top-rated-tv-shows>

<https://www.kaggle.com/datasets/shivamb/netflix-shows>

R Markdown & Knit File: https://drive.google.com/drive/folders/1hNWCaXoAuszj7geD1-wBGLRpoo4IjPPL?usp=share\_link