A Comprehensive Analysis Of Streaming Services

Crystal Diaz, Maximillian Gutierrez, Juan Rivera

College of Engineering, California State University, Long Beach

CECS 450 Data Visualization

Professor Anthony Giacalone

12/8/2023 *(for student papers)*

1. Introduction

In the midst of numerous streaming services, the task of choosing the right one can be overwhelming. To assist users in making informed decisions about movies, TV shows, or both, we delved into a comprehensive dataset. The dataset we used had over 14,000 entries of data for our research and modeling purposes. Leveraging R Studio, we meticulously cleaned the data and built models to compare various categories. Because there are numerous factors that can affect why or why or not a person might want to subscribe to our evaluation, the study employed a point-based system considering factors such as title variety, age groups, ratings, and pricing. However, it's noteworthy that each service excelled in specific aspects, emphasizing the diversity of strengths across streaming platforms.

2. Dataset and Features

The dataset we used for this project, “Movies on Netflix, Prime Video, Hulu and Disney+” (1), was obtained from Kaggle, which is a well-known data science competition platform as well as online community for data scientists. Our dataset was initially two separate datasets, with one containing only movies and the other containing only shows. The resulting dataset is a csv file containing approximately 14,500 instances of data gathered from four major streaming services: Netflix, Hulu, Amazon Prime, and Disney +. The dataset contains 10 different categories:

1. ID: This is a number used only to classify the number, in order, of the data entry.
2. Title: This is the title of the show/movie.
3. Age: This is the age group that the respective show/movie has. The range is from 7+ to NR (Not Rated).

4. RottenTomatoes: This is the show/movie's rating on the website Rotten Tomatoes, which is a review aggregation website.
5. Netflix, Hulu, Prime, Disney: Either have a 1 or a 0. This means that the respective service either has the show/title or does not have it respectively.
6. Type: This is how it is determined whether a title is either a show or a movie. A 1 means the title is a show and a 0 means the title is a movie.

This was the layout of our dataset after extensive cleaning. There were some issues with the dataset that needed fixing before it was able to be used.

Cleaning and Preprocessing

Our dataset was originally two separate datasets, one that had only shows and the other which had only movies. We wanted a more comprehensive analysis of the streaming service overall, so we merged the two. The columns for both were identical, so merging was not a very complicated process. Once we merged the data, there were some issues. We did not have a way to determine whether a title was a movie or a show. We did not think about this issue prior to merging, but the solution we came up with was quite simple. We created an additional column in both datasets called *Type* discussed earlier. The addition of this column solved our issue and we then merged the dataset together once again. The *RottenTomatoes* column contained some null values which did not allow us to use the column for much at first without encountering errors. We replaced every null value with a value of zero so that the column was usable. Fortunately, the inclusion of zeros did not affect the average scores of the column severely. Another issue we faced came from the *RottenTomatoes* column in the form of unusable values yet again. The scores are displayed as a ratio such as x/y. The scores were unable to be used or converted into numerical values, so our

solution was to change the respective scores into the decimal value equivalent of the ratios. This change was done in Rstudio and is not reflected on the dataset itself.

3. Methods

Hypothesis

We did not have a concrete hypothesis for this dataset initially, but after cleaning and preprocessing we decided it best to create one so that our models would have something to prove or disprove. As stated previously, there are four streaming services in this dataset. Netflix was the first major streaming service to debut, and, amongst ourselves, was the first streaming service we ever used. According to a study from Staista, Netflix continues to be the most popular streaming service of most modern choices. We decided our hypothesis should be that Netflix would be the “best” streaming service i.e. it will be awarded the most points.

Visualization

After preprocessing and cleaning the data, we created numerous visualizations and plots to model our data so that it can be easier understood. Our goal was to identify relationships between attributes in our dataset that could help prove or disprove our hypothesis. It should be noted that two of the plots are based on values not included in the dataset which are the prices of a monthly subscription to the services, with and without ads.

Streaming Service Average Score with Type

Approach

We wanted to plot and see each streaming service average Rotten Tomatoes score for the movie and show separately. We decided that using a boxplot is the most appropriate graph to show the

average score and show extra information like the 1st quartile, 3rd quartile , minimum, and maximum.

Problem

The dataset included all of the streaming services.

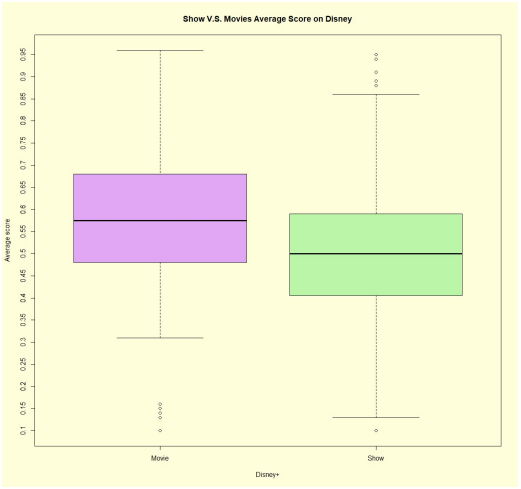
Solution

We created subsets to filter the dataset, each subset belonged to a streaming service.

This is the code we used:

```
net <- dplyr::filter(dt,Netflix == 1)
dis <- dplyr::filter(dt,Disney == 1)
hulu <- dplyr::filter(dt,Hulu == 1)
prime <- dplyr::filter(dt,PrimeVideo == 1)
```

Afterward, we graphed the subsets and looked like this:




```
boxplot(unlist(RottenTomatoes) ~ Type, data = net,
        names = c("Movie","Show"),
        col = c("#E1A7F5", "#BBF5A7"),
        main = "Show V.S. Movies Average Score on
Netflix",
        xlab = "Netflix",
        ylab = "Average score"
)
axis(2, at = seq(0.05, 1, by = 0.05), labels = seq(0.05, 1, by = 0.05))
```

<u>Average score (Movies and Shows) Plotting Process</u>

Approach

We decided to use the Rotten Tomatoes scores and calculate the average score for each streaming service. Since we are showing average scores, we decided to keep using the boxplot.

Problems

Since we had our dataset combined into one, we just needed to find a way to calculate the Rotten Tomatoes score. A problem we had was plotting the average score of each streaming service in one graph. The problem was that each streaming service had a column with the values 1 or 0, the

value 1 means the film/show is in the streaming service and the value 0 means the film/show is not in the streaming service. Since shows/films can be streamed in multiple streaming services, this would cause the average not to be calculated correctly since the title would count in one streaming service and not the other. We also could not filter the dataset correctly because trying to filter the dataset would do a or filter such as Netflix == 1| Hulu == 1 | Disney ==1| Prime ==1.

Solution

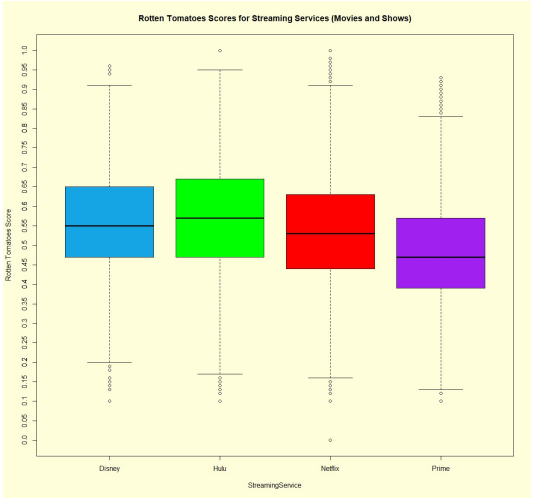
Using the same subsets from average score types, we combined the subsets into one where it had two columns: score and streaming service. We entered the scores and service parallel to each other from the subsets so we have a correct dataset.

```
#Get the length of titles in each streaming platform
n <- as.numeric(length(unlist(net$RottenTomatoes))) #netflix
h <- as.numeric(length(unlist(hulu$RottenTomatoes))) #hulu
d <- as.numeric(length(unlist(dis$RottenTomatoes))) #disney
p <- as.numeric(length(unlist(prime$RottenTomatoes))) #prime video
#Combines the subset score and name parallel
combined_data <- data.frame(
  RottenTomatoes = c(unlist(net$RottenTomatoes), unlist(hulu$RottenTomatoes),
                     unlist(dis$RottenTomatoes), unlist(prime$RottenTomatoes)),
#multiples the Service name with the length of titles the service has
  StreamingService = rep(c("Netflix", "Hulu", "Disney", "Prime"),
                         times = c(n,h,d,p)))
```

The graph:

```
boxplot(RottenTomatoes ~ StreamingService, data =
combined_data, col = c("#16A5E4", "green", "red",
"purple"), main = "Rotten Tomatoes Scores for
Streaming Services (Movies and Shows)", ylab =
"Rotten Tomatoes Scores)

#Adds labels on y axis
axis(2, at = seq(0.05, 0.95, by = 0.05), labels = seq(0.05,
0.95, by = 0.05))
```



We proceeded to use the same process to graph average score for shows only, movies only, and for the age

groups. We kept reusing the 4 first subsets and would filter what we needed and then combine the data. Finally, we would plot the data using boxplot since it was the most appropriate graph to use for our data and what we wanted to show.

Value based on Ratings for Price Plotting Process

Approach

To better gauge the returns on streaming service, we decided to create a variable that wasn't in the dataset by combining ratings with the pricing. This way, we could illustrate the relationship between the two factors and help consumers who are concerned with both prices and quality.

Problem

We found out we had not yet created an average score variable for each of the subsets that we could actually manipulate.

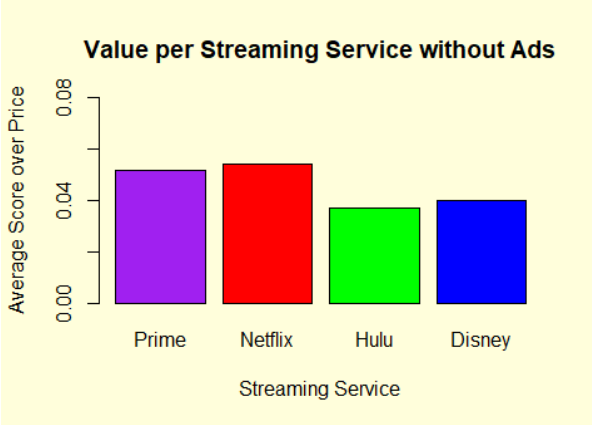
Solution

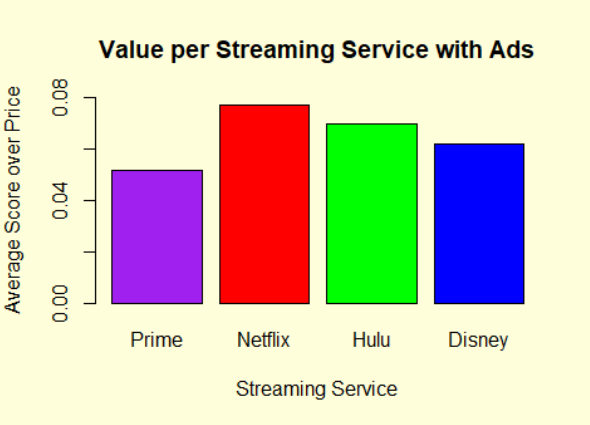
```
 #Average Rotten Tomato Score per Streaming Service
prime_avg <- mean(unlist(prime$RottenTomatoes))
net_avg <- mean(unlist(net$RottenTomatoes))
hulu_avg <- mean(unlist(hulu$RottenTomatoes))
dis_avg <- mean(unlist(dis$RottenTomatoes))
```

From here, we could make our dependent variable use the equation:
`prime_value<- (prime_avg/prime_price)`

Because we were using our dependent variables to derive our y-axis, we chose to make our x-axis categorical. To that end, we decided to keep using a bar chart for easy comparison

between streaming services. The outputs for the graphs were produced below:





Results

Using our points-system, the streaming services finished as follows: 1st: Netflix, 2nd: Hulu, 3rd: Disney+, and in last place, Amazon Prime. Neflix won in the following categories: 1. Best-scoring non-rated series, 2. Best-scoring 16+ rated series, 3. Best Cost with Ads, 4. Best Ratings for Price with Ads, and 5. Best Ratings for Price without Ads. Hulu grabbed the following: 1. Best Rotten Tomato Scores (Film), 2. Best Rotten Tomato Scores (Shows), 3. Best Rotten Tomato Scores (Cumulative), and 4. Best-scoring 18+ rated series. Disney+ scored a little less: 1. Best-scoring 7+ rated series, 2. Best-scoring 13+ rated series, and 3. Best-scoring E-rated series. Finally, Prime won only two categories: 1. Most Titles and 2. Best Cost without Ads.

Conclusion

The results of the study proved the initial hypothesis that Netflix would place best and that Netflix is best for consumers who value a holistic approach. However, the gap between Netflix and other services such as Hulu and Disney was not considerable, as each service has their own strengths. For example, Hulu would appeal to customers who value quality over pricing and quantity while Disney might draw in family-driven consumers. Prime would likely be the preferred choice for those who want the greatest amount of titles for the most affordable price.

References

Bhatia, Ruchi. “Movies on Netflix, Prime Video, Hulu and Disney+.” *Kaggle*, 16 Dec. 2021, www.kaggle.com/datasets/ruchi798/movies-on-netflix-prime-video-hulu-and-disney.