Exploring the Variety of Streaming Services

Sophia Weidner

Bellevue University

DSC 680: Applied Data Science

Professor Catherine Williams

April 2nd, 2023

Table of Contents

Introduction

Business Problem

In today's day and age, the amount of streaming platforms are rapidly increasing. One may dare say that streaming platforms are overtaking cable networks entirely. With each platform comes more exclusivity; for example, Netflix used to have an abundance of movies and television series (including Disney titles). But, that number has dwindled overtime as shows produced by certain companies are shown on their own, or other smaller, streaming services. Disney+ happens to be one of the newer streaming services, but when Disney comes to mind, it's an incredibly niche category of movies: princesses, Pixar, Star Wars, etc… In this project, I want to explore the variety of movies and television shows provided by Disney+ to see if customers are truly getting what they are paying for.

Background/History

As mentioned previously, Disney+ is a relatively new streaming service that launched back in November of 2019. The purpose of Disney+ is to provide a platform to stream any Disney owned television show or movie, and the media provided is unique to only this service (see Appendix A). This service is supposed to directly compete with other established streaming giants, such as Netflix, Hulu, or Amazon Prime Video.

Ethical Assessment

My business problem revolves entirely around consumer opinions, which can be problematic because different people may hold varying views. Furthermore, assessing whether Disney+ is considered 'worth it' is a subjective matter and my personal bias, as an employee of the Walt Disney Company, could affect the analysis. Therefore, I will approach this problem with an open mind and base my conclusions strictly on the data analysis results.

Data Explanation

The data for my analysis was obtained from Kaggle, and includes all titles for any show or movie currently available on the Disney+ service. The dataset provides information such as the director, cast, and the year of release for each media title. Much of the data within this set is categorical, which means it will have to be transformed in order to perform an adequate data analysis.

The first step in preparing my data was to check for any NaN values. When doing this, I found that certain columns had very many of these empty values. These columns included “director”, “country”, “date_added”, “cast”, and “rating”. To combat this issue, I made the following changes:

- I replaced NaN values in “director” with "No Director Listed".
- I removed the “country” column.
- I removed the “date_added” column.
- I replaced NaN values in “rating” with "Unknown".
- I replaced NaN values in “cast” with “No Cast Listed”.

I wanted to keep as many data points as possible, but ultimately decided two columns were worth removing. The “country” column indicated where each movie is made, but the majority of them were made in the United States, so there was not much variety within that datapoint anyways. The “date_added” column had the dates on which the movies or television shows were added, which only really affects the availability of the media, not necessarily the variety of the media intake. The other columns listed were worth filling in the NaN values, as I wanted to use those points in data visualization as well as the analysis portion of the project. After the steps above were taken, I was able to begin creating visualizations for different aspects of my project.

The first visualization created was a simple one: I wanted to show how many movies versus how many television shows are demonstrated within the dataset. To do this, I found it was most effective to use a bar graph (or a histogram could have been utilized, but I ended up using a bar graph). This was the result:

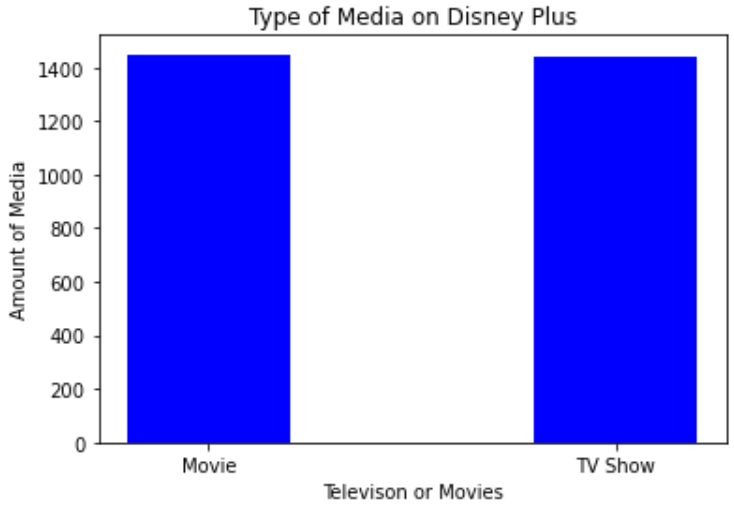

Figure 1: A Bar Graph depicting the amount of movies and television shows on Disney+.

Luckily, we're looking at a very even split: 1,400 movies and 1,400 television shows. I double checked by utilizing the len() function and found this fact to still be true.

Now that we know the type of media available for user consumption, I wanted to start to explore the variety of the titles available. Since many people like to watch certain directors, I figured this should be my first step. Because there are over 2,000 titles with several directors, I decided to limit my data frame to only include movies. Even when doing this, my original visualization was illegible and you weren't able to distinguish the directors and how many movies they released. So, I filtered the information to include the top twenty directors based solely on the amount of movies they have released on Disney+. This is shown in Figure 2 below:

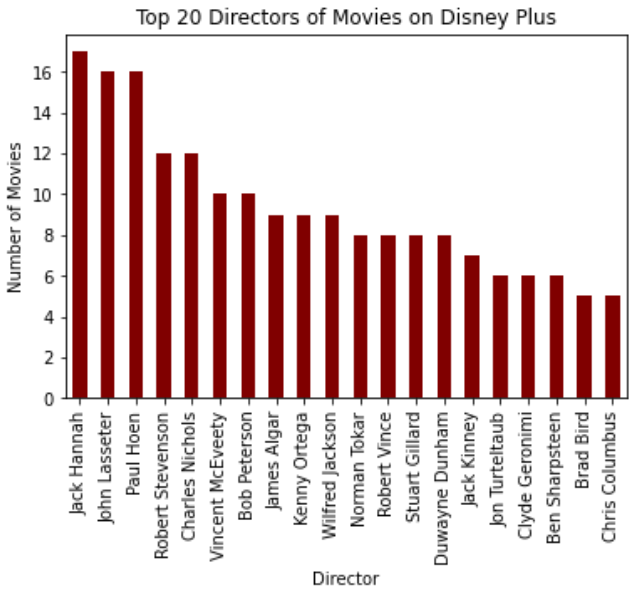

Figure 2. The Top Twenty Movie Directors on Disney+.

The director with the most releases, Jack Hannah, has about 16 movies available on the platform. While this may seem like a lot, that is 16 out of the 1,400 movies available on Disney+, which is roughly 1.1%. As far as Disney+ movies go, it seems that the platform does include a plethora of directors and thus a good variety for consumers seeking that certain criteria.

The next aspect I wanted to analyze included the type of movie (not television show) available to the customer. Unfortunately, this dataset had only one column for the category, which can be found under “listed_in”. This column also had up to three categories available for each title, but not all movies had all three categories filled. I decided it was appropriate to split the “listed_in” column into three separate columns, and those columns were split into the categories using a comma as the delimiter. If the movie didn’t have three categories, then the datapoint was filled in with “None”. I then created three separate pie charts to show the distribution of the categories.

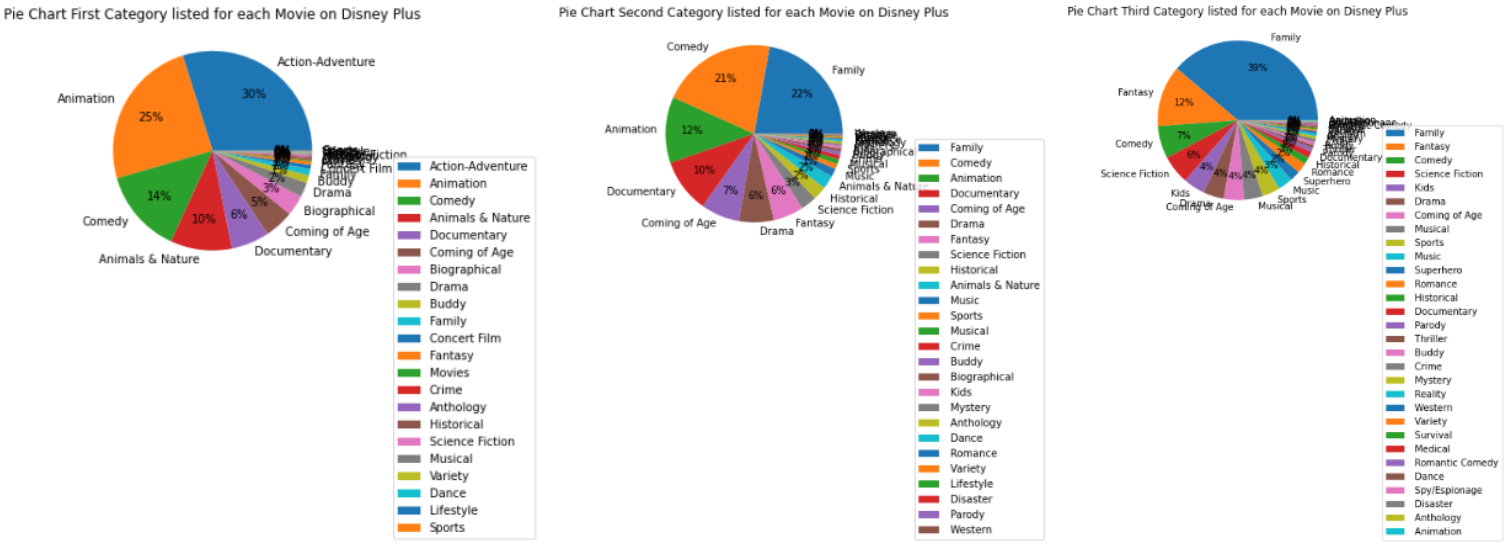

Figure 3. Pie Charts Depicting the Categories Available

Unfortunately, the color scheme across the three graphs did not remain the same for certain categories, like “Family”, and the original dataset did not have a particular order to which the categories were inserted. However, I think the Pie Chart is the best indicator of showing the variety of films available. In the first graph, for example, it’s very clear that “Action-Adventure” and “Animation” are the most popular types of movies and account for over 50% of the films available. In just looking at the first graph alone, there is some variety of movies available, but ideally, the pieces of the pie would be smaller to indicate more categories. As we move from graph to graph, the pieces start to get smaller and smaller, but it’s important to keep in mind that not every movie has a second, or even a third, categorical listing. I believe that the “listed_in” column is not the best datapoint to determine the variety of movies, as it was not well organized and there was not a uniform way to split the available data.

Lastly, I wanted to take a closer look at the release years of both television shows and movies. I split the criteria into media released before and after 1980, since it was a good halfway point in the “release_year” column.

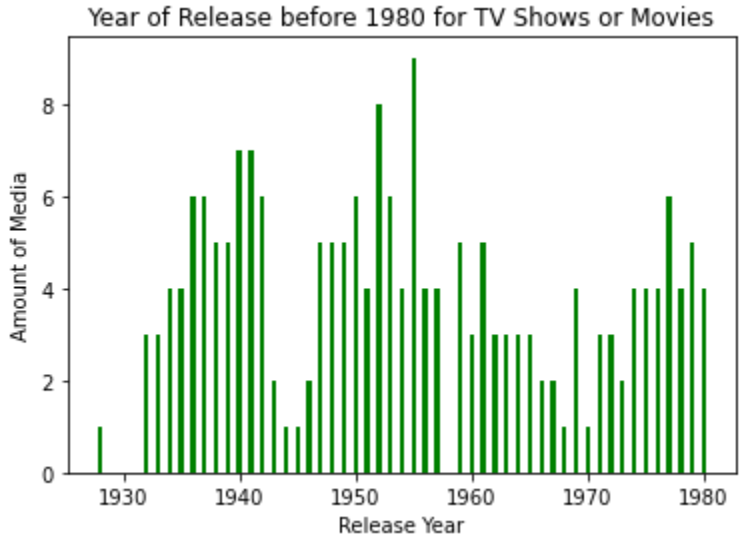

Figure 4a. TV Shows and Movies released before 1980.

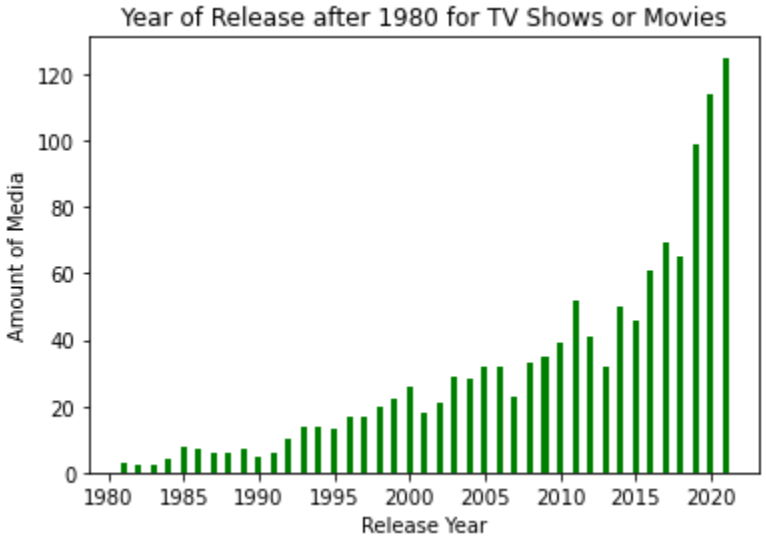

Figure 4b. TV Shows and Movies released after 1980.

In Figure 4a, 1955 was the year with the most releases, topping out at around 10. In Figure 4b, 2021 was the year with the most releases at over 120. From the graphs above, there is a lot less

content released from before 1980; so, for those who wanted older television, there is some content available. However, Disney+ streaming caters more towards modern day releases. One reason for this could be because Disney+ is actively creating their own television series and movies only available on the platform. In summary: if you're looking for older television, Disney+ may not provide the type of variety you are seeking. If you want modern releases, Disney+ is a great service.

Methods

To evaluate variety, there were two methods I wanted to choose from: Regression Analysis or Clustering Analysis. Linear Regression Analysis is used to find correlations between data points, predict outcomes, and to evaluate trends (Amazon Web Services, n.d.). Clustering analysis focuses on grouping similar data together (Displayr, n.d.). I decided against using a method such as a Decision Tree Classifier as I want to evaluate variety, not necessarily the decisions the consumer may make.

I originally tried using Clustering analysis, but I found it challenging to get accurate results with my dataset. Because the data is largely categorical, I had to create dummy variables, and this caused clustering the data to be difficult. I wanted to see variety rather than patterns amongst the data, so I decided it was best to go with a linear regression model. This way, I could determine a relationship between the data and predict what a consumer may want to watch based on their input.

Analysis

To start, I split my data into independent and dependent variables based on the release year. This was the cleanest datapoint when creating my visualizations, and was one of the columns that did not have any NaN values at the beginning. Furthermore, the age of a film or

show can be a factor when a customer chooses a certain media to watch, so I felt it was appropriate to split the data on this factor. Afterwards, I split the data into training and testing sets randomly.

When fitting the Linear Regression Model to my data, I got an RMSE of 12.72. The RMSE indicates that any predicted values deviate from actual values by about 12. In the context of the data, the RMSE value is saying that based on other factors, like the movie category, the predicted year of release for the movie is within 12 years. This is not terrible given the circumstances, but a decade gives plenty of time for technological improvements. For example, think of a movie made in the 70s versus one in the 80s; there are differences in the quality of the films.

Another value I generated was the R squared coefficient, which was 0.65. So, roughly 65% of the variance in other variables can be determined by the release year of the media. Ideally, the R squared value should fall within the 0.5-0.99 range to make a model valid; which, in this case, 0.65 falls into (Soytaş & Sari, 2020). I would like this number to be a bit higher, which would require the data to be split amongst another independent variable. However, because the dataset consists of so many dummy variables, the year of the release is still the best variable that could have been used for this analysis.

Conclusion

Overall, Disney+ offers a wide variety of television and movies for users to consume. There is a range of categories, directors, and release years to please the customer for the monthly price that is paid. This is shown mostly through the data visualizations provided as well as the data analysis that was performed.

Limitations and Challenges

The business problem I sought to analyze set me up for a difficult time. Data analysis is difficult to perform when wanting to assess variety. Thus, I had to shift my goals and see if there was a relationship between the year of release and other factors. Furthermore, the dataset I worked with was rather small (the titles provided were only those listed through 2021) and only included 2,800 datapoints. Some of these data points, such as "listed_in", included movie categories but were not organized alphabetically nor all had the same amount of criteria. At times, it was difficult to extract the data to create visualizations. It was also hard to find a variable to split the data for the Linear Regression Analysis because dummy variables had to be created.

Future Uses/Additional Applications/Recommendations

The methods I used for this project could be used to analyze another streaming service, such as Netflix or Hulu. Ideally, this tool would be best used for a dataset that has more numerical data rather than categorical. Also, it would be best to replace any NaN values with some sort of numerical data point, like the mean or median, rather than a string (like I did during my data cleansing steps).

Resources

Displayr. (n.d.). What is k-means cluster analysis? Retrieved April 2, 2023, from https://www.displayr.com/what-is-k-means-cluster-analysis/

Amazon Web Services. (n.d.). Linear Regression. Retrieved April 2, 2023, from https://aws.amazon.com/what-is/linear-regression/#:~:text=Linear%20regression%20is%20a%20data,variable%20as%20a%20linear%20equation.

Soytas, U., & Sari, R. (2020). An analysis of causality between energy consumption and economic growth: Evidence from African countries. Munich Personal RePEc Archive. Retrieved April 2, 2023, from https://mpra.ub.uni-muenchen.de/115769/1/MPRA_paper_115769.pdf

Appendix A: Disney Plus Information

Disney Plus is a subscription-based streaming service that offers a wide range of movies, TV shows, and original content from various Disney-owned franchises. The following resources provide information on Disney Plus:

Disney+. (n.d.). About Disney+. Retrieved April 2, 2023, from https://www.disneyplus.com/welcome/about

Spangler, T. (2022, March 10). Disney Plus guide: What to know about price, shows, movies, bundles and more. Variety. Retrieved April 2, 2023, from https://variety.com/guide/disney-plus-streaming-service-everything-to-know/

D'Alessandro, A. (2021, December 13). Disney+ now boasts 145 million global subscribers, Company aims for 230M-260M by 2024. Deadline. Retrieved April 2, 2023, from https://deadline.com/2021/12/disney-plus-global-subscribers-2021-230-million-target-2024-1234894155/

Note: This appendix provides information related to Disney Plus, including an overview of the service, a guide to its price, shows, movies, and bundles, and the current number of global subscribers.

10 Questions from the Audience (Q

1. What types of content are available on Disney+?
    a. Television Shows and Movies are available on this streaming platform.
2. How much does a Disney+ subscription cost?
    a. There are different packages available, but a base Disney+ Duo Basic (with Ads) subscription costs $9.99 USD (source)
3. What price would justify the variety Disney+ has to offer?
    a. This is largely based on the consumer. Typically, a consumer would not want to pay more for a service with less to offer. I would advise this customer to compare Disney+ to a streaming service that they do enjoy and feel as if they get their money's worth from.
4. How much do other subscriptions for other companies, such as Netflix, cost in comparison?
    a. Netflix: basic (with ads): $6.99 USD (source)
5. Are there any exclusive shows or movies on Disney+?
    a. Yes, there are several exclusive shows and movies on this platform. Movies such as *Cruella* were only released on the platform and **not** in theaters.
6. How does Disney+ compare to other streaming services like Netflix or Hulu?
    a. Disney+ only has titles related to the Walt Disney brand and the brands that the company has obtained. Netflix and Hulu do not have any of these Disney titles, and even share some of the same television shows such as *Grey's Anatomy*.
7. What is the most popular content Disney+?
    a. This link offers some popular content that Disney+ has on their streaming service. A lot of these titles are Disney exclusive brands, such as *Star Wars* and *Marvel*.
8. What criteria offers the most variety on Disney+?
    a. After performing the data analysis, I believe that Disney+ is a great platform to see content from different directors.
9. Are the titles listed in the dataset available in other languages? Do they provide closed captioning?
    a. Yes, the titles listed are available in multiple languages and do include closed captioning.
10. How is Disney+ contributing to the growth of the Walt Disney Company as a whole?
    a. "Disney+ made a revenue of $7.423 Billion in 2022. That is 2.13 billion more than past year. In 2021, Disney Plus registered a revenue of $5.293 billion." (source)