Disney+

Content Analysis

Group Name: Group Shrestha

Group Members:

First name	Last Name	Student number
Aanal	Patel	C0910376
Bimal Kumar	Shrestha	C0919385
Danilo	Diaz	C0889539
Ernie	Sumoso	C0881591
Jayachandran	Saravanan	C0910392

Submission date: ***March 17, 2024***

Contents

Abstract

The report speaks about the implementation of various data visualization approaches to examine the given dataset of the Disney+ digital streams. The project covers the primary data analysis steps by working on the numeric data to unhide the insights that are useful for marketing and strategic decisions. It also discusses the potential methods of the panda's data frame, NLP for text data processing and visualization tools like matplotlib and Seaborn. Further, the detailed observations for the given challenge questions are recorded with graphs and tables. The analysis is not limited to but includes steaming behaviour across multiple geographic regions, different time periods, and categories like content type, rating certificates and genres. The best inferences are plotted, and the experiments are drafted for future reference. Countries like the US, India, the UK, and Canada contribute more to the streaming content, with TV-MA ratings that are more popular with genres like comedy, romance, international movies, and TV shows. The words (unigram) are analyzed for frequencies, and the results are displayed clearly.

Keywords: Matplotlib, seaborn, Pandas, Disney+, streaming content. Ratings, movies, TV shows, visualization, genre, unigram, Natural Language Processing and data analysis

Introduction

Online media streaming services have revolutionized the way media are being consumed. The flexibility to consume media with any of the devices like smartphones, computers, TVs, etc, has transformed the way people engage in entertainment, allowing them to binge-watch any movie or TV show without being concerned about missing any episodes and watching them on the go. Disney+ is just one of these subscription-based media streaming services providing a wide variety of content ranging from R-rated Action Movies to PG-13 Animations. The objective of

this project is to conduct a complete analysis of given data, particularly unveiling the data impurities and getting more helpful information for processing. In addition, visualization methods are employed to protect the data clearly and interpretably. The following table explains the process and expected outcome.

Process	Explanation
Data loading and validation	The dataset is loaded using a suitable framework, and the quality of the data is checked
Exploratory Analysis	Exploring the characters of the given data and contribution to the analysis
Task interpretation	Complete the study of the given task and identify the action points on the given data
Draft implementation	Basic execution of the task on given data with high-level details
Improvisation of the base analysis	Complete analysis of the data satisfying the required questions/query
Styling and decoration	Experimenting with various charts and styling options to give better picture using available visualization tools

Dataset Description

The provided dataset about Disney+ contained more than 8.8 thousand records of content available in the platform, which were added from the year 2008 to 2021. These records had different information about the contents, which are given below:

- **Show_Id**: Unique identifier for the content.
- **Type**: Type of the content: TV Show or a Movie.
- **Title**: Title of the content.
- **Director**: The ones who directed the TV show or the movie.
- **Cast**: Main characters from the movie or TV show.

- **Country**: Country of origin for the given content. It can have more than one value.
- **Date_Added**: The date when the content was added to Disney+.
- **Release_Year**: The year when the content was released, not necessarily in Disney+.
- **Rating**: Content rating which indicates its character: Family Friendly, Inclusion of violence, nudity, etc
- **Duration**: The timeframe length of the content if it is a movie and the of seasons if it is a TV show
- **Listed_in**: The different genres it belongs to or in which genre it can be found in Disney+
- **Description**: A short description of the movie or the TV show

Data validation and cleansing

Before any data exploration, data validation is a must. This includes checking for absence of fields in the record, unique values and its count in data(column with no variance is not significant for analysis), range of numeric fields and presence of duplicate records.

	Columns	Null Count	Null Ratio	Unique Values
0	show_id	0	0.0	8807
1	type	0	0.0	2
2	title	0	0.0	8807
3	director	2634	0.299	4528
4	cast	825	0.094	7692
5	country	831	0.094	748
6	date_added	10	0.001	1714
7	release_year	0	0.0	74
8	rating	4	0.0	17
9	duration	3	0.0	220
10	listed_in	0	0.0	514
11	description	0	0.0	8775

Fig1. Checking for null values, its ratio and no of unique values

Upon inspection, it was found that the column director had the most null values, followed by cast and country. Similarly, checking the unique values gave us an idea of the variance in data. After

that, unique values for all these columns were skimmed to get information about the values in the data fields. While doing this, some invalid values were seen in the rating column where, instead of ratings like ‘PG-13’, and’ TV-MA’, the duration of the content was seen; so we further explored this in the data to find out the duration for these records were null likely due to incorrect formatting of csv. This was fixed by switching the values.

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	2017-04-04	2017	74 min	NaN	Movies	Louis C.K. muses on religion, eternal love, gi...
s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	2016-09-16	2010	84 min	NaN	Movies	Emmy-winning comedy writer Louis C.K. brings h...
s5814	Movie	Louis C.K.: Live at the Comedy Store	Louis C.K.	Louis C.K.	United States	2016-08-15	2015	66 min	NaN	Movies	The comic puts his trademark hilarious/ thought...

Figure 2. Invalid values in the rating column

Looking for a range in numerical values, date_added and release_year inferred that the oldest contents were from 1925, whereas the contents were added starting in 2008.

	date_added	release_year
count	8797	8807.000000
mean	2019-05-17 05:59:08.436967168	2014.180198
min	2008-01-01 00:00:00	1925.000000
25%	2018-04-06 00:00:00	2013.000000
50%	2019-07-02 00:00:00	2017.000000
75%	2020-08-19 00:00:00	2019.000000
max	2021-09-25 00:00:00	2021.000000
std	NaN	8.819312

Figure 3. Distribution of numerical columns

EDA and Insights

After being done with data wrangling, the next step was to explore the data and find characteristics of the dataset using different statistical measures and with the help of

visualizations using various plots. Distributions of data along different fields were explored first. This gives us an initial idea on what type of data we are working on. The first field explored was **type.** Out of 8807 contents, provided in the dataset, almost 70%(6131) of the contents were Movies whereas the remaining 30% were TV shows.

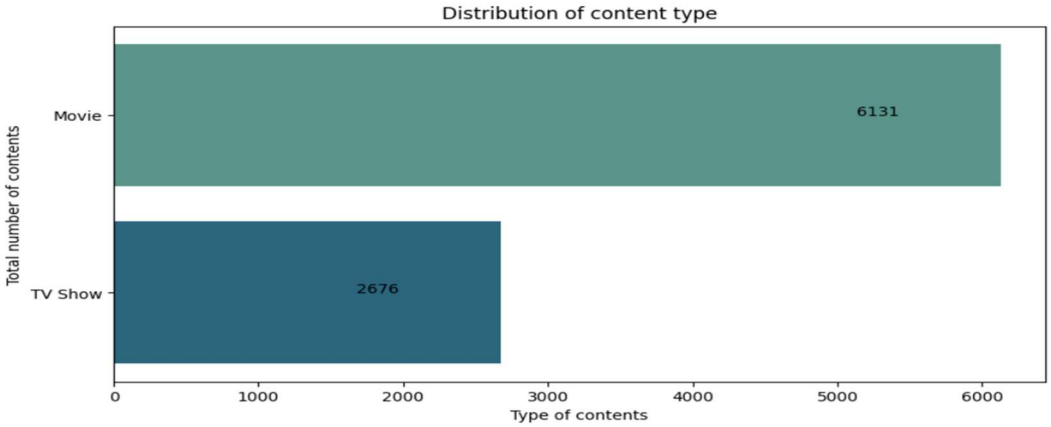
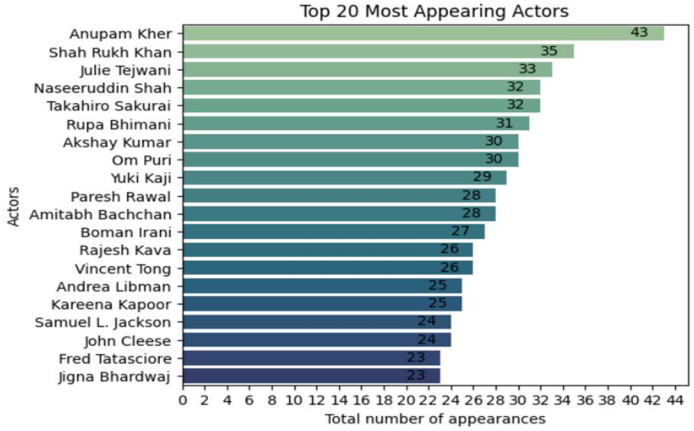


Figure 4. Count Plots for type column

Just for exploration sake, we checked for the most appearing actors or actresses in the contents. Since this column had numerous values per record, this column needed to go under a few transformations like splitting and joining. Similarly we also checked out who had directed the most number of movies. Same proceedings were needed for this columns as some had multiple values as well.



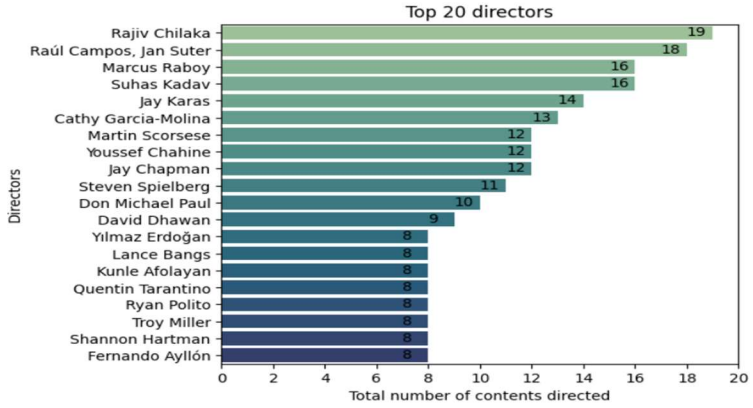


Figure 5. Insights about director and actors

Comprehensive study (task)

This is the crucial step of this project, where the pre-processed data is considered for granular-level examination and processing to reveal hidden insights. Based on the given questions, the process is carried out,

Questions
How does content differ geographically?
Over time, how has the type of content that is being added changed?
Which words recurred among content Titles and Descriptions?
Over a year, what changes have occurred in the overall content being added?
Which is the most popular rating category?

Geographic-wise content distribution

In this task, the steaming is analyzed with respect to the countries listed in the data, particularly the distribution of the content (movie and TV shows). The user-defined function to fetch the country-wise counts is used to get the exact numbers of each type of content. To gather the primary understanding, the choropleth module in the Plotly package gets a more enriched representation of the total count in a thematic map with various values. The count values are passed to the logarithmic function to get a distinct and uniform distribution value, and it also avoids the domination of the large values over the smaller ones.

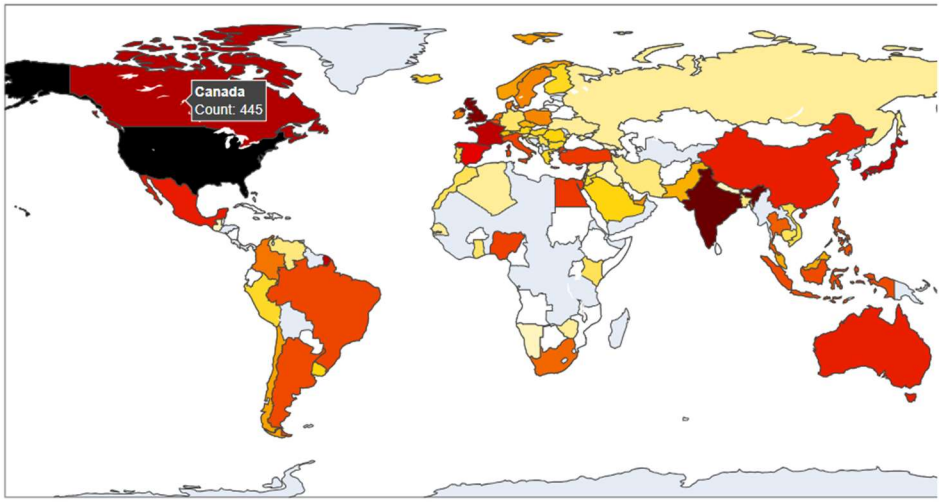


Figure 6. Geographic distributions

Insights:

- 65% of the content is distributed across the top 4 countries: the United States, India, the United Kingdom and Canada, with the US and India accounting for 80% of the contribution to the top layer.
- Countries like Mexico, Germany, South Korea and Spain are producing only less than 2 percentage.

type	country	Movie	TV Show
0	United States	2752	938
1	India	962	84
2	United Kingdom	534	272
3	Canada	319	126
4	France	303	90
5	Japan	119	199
6	Spain	171	61
7	South Korea	61	170
8	Germany	182	44
9	Mexico	111	58

Figure 7. Country-wise content split

- The TV shows are fewer compared to the Movies in most of geographical regions, with the expectation of Japan and South Korea, where series and TV shows are more

The rating-wise exploration of 14 categories is tabulated for all the countries where mature content is preferred with clear distinction, and categories like PG, R, and TV-14 show equal viewership. The pie chart is used to describe rating contents across various countries individually.

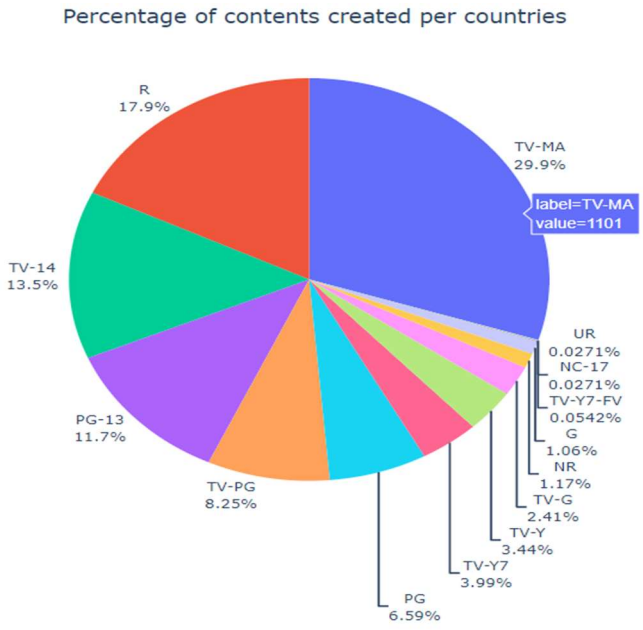


Figure 8. Rating-wise content split

When it comes to genre distribution, the most popular countries are viewed with the help of explode and group by method. The following table and the insights will discuss the details,

new_listed_in country	Comedies	Documentaries	Dramas	International Movies	International TV Shows
Canada	39	21	22	26	19
France	18	24	31	68	32
India	308	19	620	817	65
Japan	0	0	12	58	141
Mexico	17	10	24	49	35
South Korea	6	1	21	38	149
Spain	34	17	37	90	46
Turkey	56	0	27	74	29
United Kingdom	32	84	45	66	112
United States	524	411	591	21	27

Figure 9. Top 10 countries rating categories

The above group table gives an overall understanding of the genre-based contents created in various countries.

- The close observation gives us the insights that the US is having an equal amount of genre content, which shows the diversity.
- India, on the other hand, gives more focus on Dramas, International movies
- Japan and South Korea focus on international TV shows (series)
- Canada is the having moderate genre contents
- UK has a higher document ratio compared with other genres than any other countries

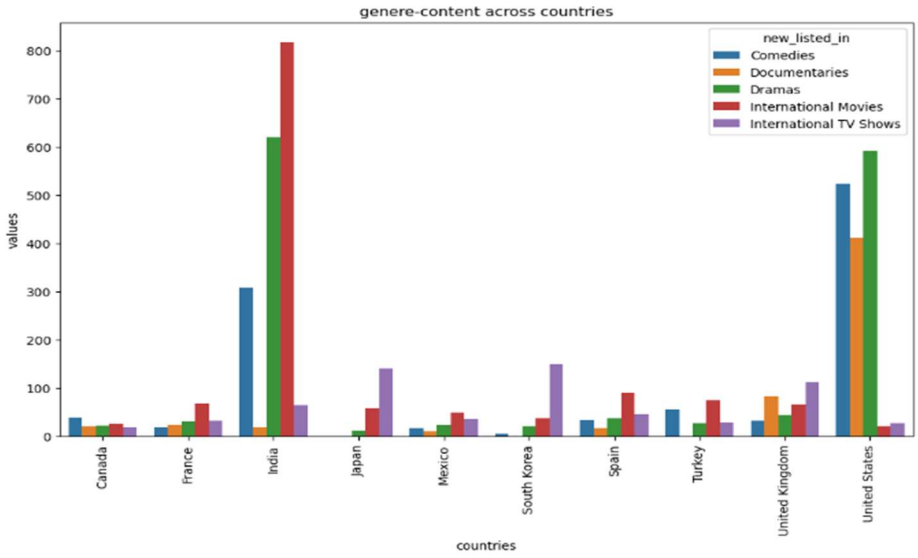


Figure 10. comparison plot

Addition of content over time

The addition of content over time in the Disney+ platform depicts the slow growth in the initial six years, and the transition of exponent growth is seen in the span of 4 years between 2014 and 2018. There was a decrement starting in 2020, which was seen with a high margin in movie-related content. 200% increment took place in the calendar year 2016 with the emergence of the internet.

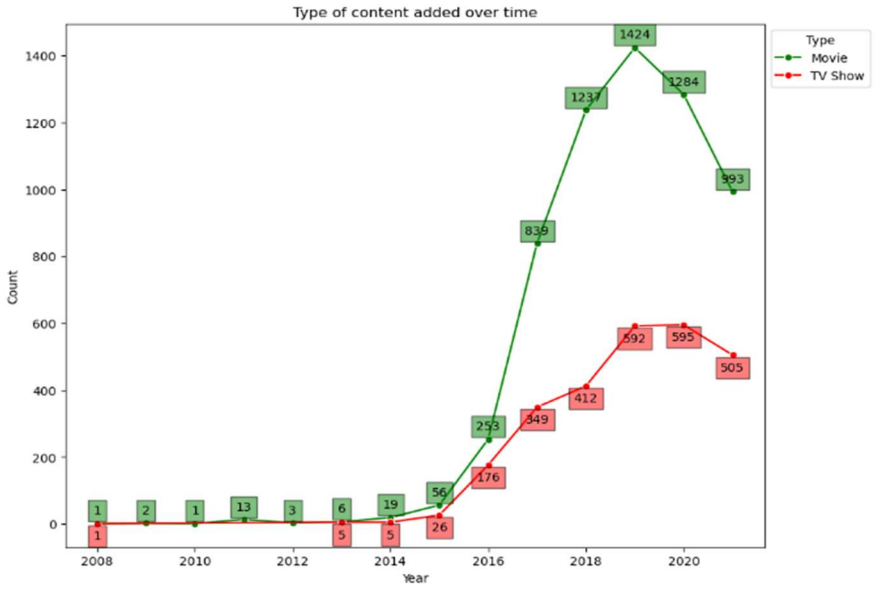


Figure 11. over time content insertion

Frequency of words in Titles and Descriptions

Usage of the words plays a vital role in any streaming content. Analyzing the title and description needs the utilization of Natural Language Processing (NLP) to get rid of the stop words and the punctuation marks. Further, the unigram and bigram-based approach will give more collocation of words in the title and descriptions. The country-wise title and description of the movie with more frequencies are view through the word cloud.

Recurrence of Words in Title

Figure 11. Overall Title word cloud

Recurrence of words in description

Figure 12. Overall, description word cloud

```
[('life', 774),
 ('young', 728),
 ('new', 699),
 ('family', 570),
 ('love', 497),
 ('two', 495),
 ('man', 491),
 ('world', 491),
 ('friends', 466),
 ('woman', 452)]
```

```
[('love', 170),
 ('2', 129),
 ('man', 81),
 ('christmas', 78),
 ('life', 76),
 ('story', 75),
 ('movie', 73),
 ('world', 70),
 ('little', 64),
 ('one', 63)]
```

```
: tile,des=word_analysis(df[df['country']=='Canada'],'title')

: tile

: [('trailer', 10),
   ('boys', 10),
   ('true', 8),
   ('christmas', 8),
   ('monsters', 5),
   ('cop', 5),
   ('day', 4),
   ('', 4),
   ('la', 4),
   ('bad', 4)]

: des

: [('', 20),
   ('two', 13),
   ('friends', 12),
   ('world', 11),
   ('series', 11),
   ('school', 11),
   ('young', 10),
   ('three', 10),
   ('documentary', 9),
   ('true', 9)]
```

Figure 13. Country wise analysis

Insights

- Title: Most of the shows or content has "Love and "2" --> signifies the part of various shows (a trend in the decade)
- Description: in the description as well, it has more words like "love," "young," "life," "world," "and friends," creating a more sense of the movie's plot
- The US is obsessed with the words Christmas and American in the titles, and the plots are described through words like new, life and world
- India uses their original languages in the title with love. Similar words with different synonyms are present, and Mumbai is used more often. Their plots are biased with preference to the male actors, so the word man is kept repeating.
- In Canada, Trailer, Christmas and monsters are repeated with friends to describe the plots

Changes in content creation over a year

This task required a clear understanding of the content in various aspects and brought more parameters into consideration. The steps and analysis followed are described clearly below

1. Overall cumulative count of the last two years

   The analysis shows a positive increment in the contents added overall, considering the past values. The saturation period is not seen anywhere, suggesting each there is an addition of the contents that happened

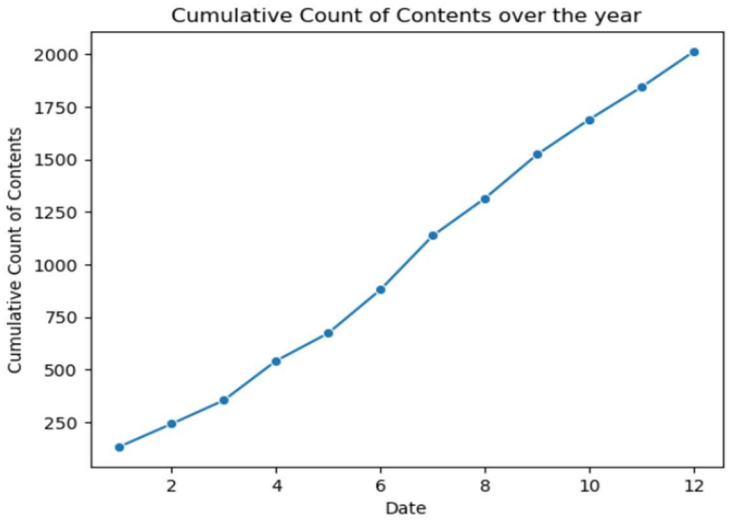


Figure 13. Cumulative content count

2. Year-wise cumulative counts are plotted

Each year's content distributions are plotted with a clear distinction of the months present in the year. Years greater than 2016 show a smooth increment, whereas the years 2015 and 2014 have some stagnated values in a few months. cumulatively the count rate has seen improvement irrespective of the content and regions

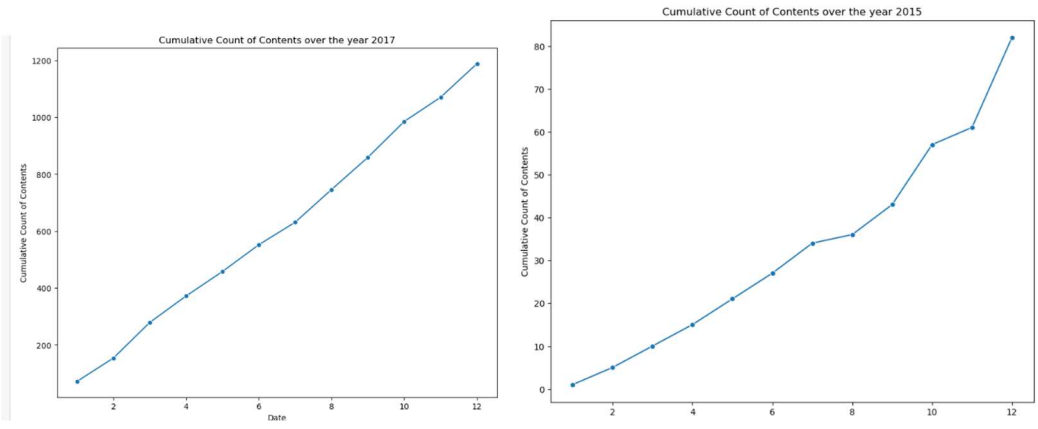


Figure 14. Year-wise Cumulative content count

3. Month-wise analysis with ratings content

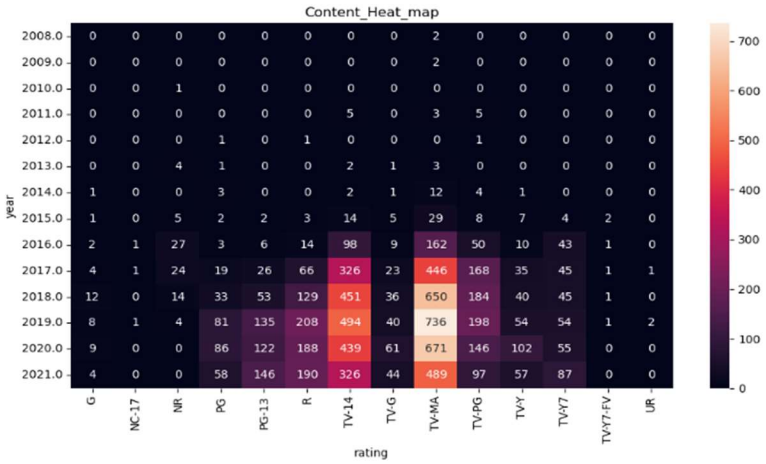


Figure 15. Hear map of rating contents year wise

4. Content distribution

- The months November, December, and January (the holiday season) have the highest content being added in all the years
- surprisingly, the years 2017 and 2018 follow different patterns, paving summer vacation time to have more content (particularly TV shows)

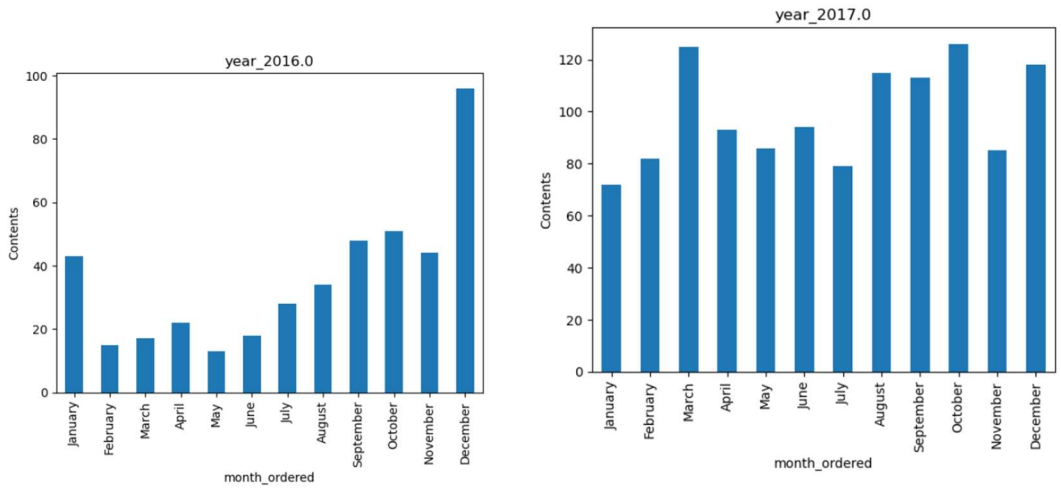
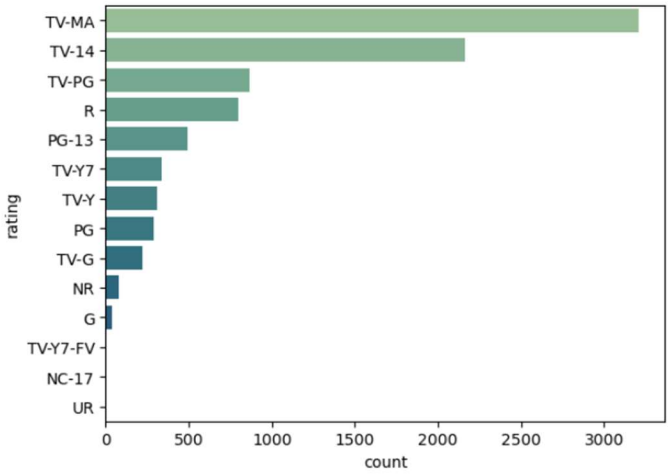


Figure 16. Month-wise content changes

Popular rating category

The popularity of the content based on the rating types is computed in descending order with the value count method.

- TV Mature Audience is the most popular one
- No one under 17 is the lowest known rating category in the Disney+
- there are 3 UR which is not rated (might be experimental shows)



	rating
TV-MA	3207
TV-14	2160
TV-PG	863
R	799
PG-13	490
TV-Y7	334
TV-Y	307
PG	287
TV-G	220
NR	80
G	41
TV-Y7-FV	6
NC-17	3
UR	3

Figure 17. Popular ratings

Conclusion

The complete analysis of the data given is completed, and the 360-degree analysis of the data is completed in this activity. All the inferences are attached, with room for a few more improvements on correlating the hidden information as the future scope.

Reference:

Stream the greatest Movies, Series, Originals and more | Disney+ Canada. (n.d.). Disney+. https://www.disneyplus.com/en-ca

DataFrame — pandas 2.2.2 documentation. (n.d.). https://pandas.pydata.org/docs/reference/frame.html

Python API reference for plotly — 5.21.0 documentation. (n.d.). https://plotly.com/python-api-reference/

seaborn: statistical data visualization — seaborn 0.13.2 documentation. (n.d.). https://seaborn.pydata.org/

Matplotlib documentation — Matplotlib 3.8.4 documentation. (n.d.). https://matplotlib.org/stable/index.html

Git and OneDrive links

Whole working folder - https://mylambton-my.sharepoint.com/:f:/g/personal/c0910392_mylambton_ca/EkXo2J-rOutFro1vMVPyzBoBLGY75bo1E25Dn6eTmHpN9A?e=PvI7XB

Git hub - https://github.com/svjai/project_data_viz