**Exploratory Data Analysis (EDA) of Disney Hotstar**

1. Link to Github: <https://github.com/Hriday65/IDAI101-1000257--Hriday-Khubchandani-Summative-Assignment-2>
2. Student Name: Hriday Khubchandani
3. Candidate Registration Number: 1000257
4. CRS Name: Artificial Intelligence
5. Course Name: IDAI1
6. School Name: Ryan Global School, Kharghar

**Introduction**

Entertainment analytics company CineStream Insights has been onboarded by Disney Hotstar to enhance the entertainment platform’s understanding of its content library. The goal of this Exploratory Data Analysis (EDA) is to uncover trends and insights that can help Disney Hotstar in the following areas:

1**.Identifying** Emerging Genres Over: Tempt future demand and place Disney Hotstar in the right spot with the right series at the right time.

2.**Monitoring** Trends in Family-Friendly: Make sure Disney Hotstar caters for all age groups, teenagers and even children.

3. **Analyzing** Content Types Over Time: Max out the observation of the viewers and the industry to alter Disney Hotstar’s content plan.

**Data Preprocessing**

The Disney Hotstar dataset was collected from the link given. The following data preprocessing steps were performed:

1. **Handled Missing Values**: Checked for gaps in the data and made recommendations on how to fix those gaps in order to obtain constant values across the board.

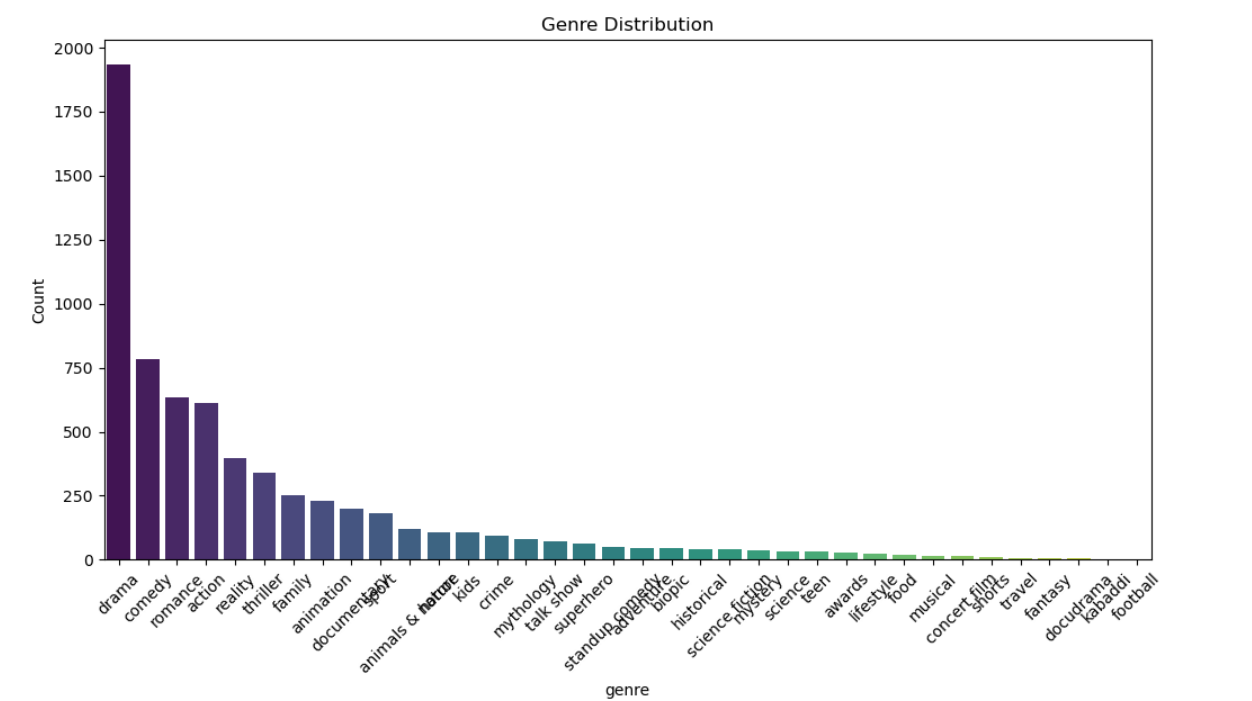
2. **Standardized Formats**: Made sure that all data elements (such as, for example, the release year or the running time) were formatted in a like manner.

3. **Maintained Consistency**: Compared the reliability of data for different subcategories and period of time.

**Exploratory Data Analysis**

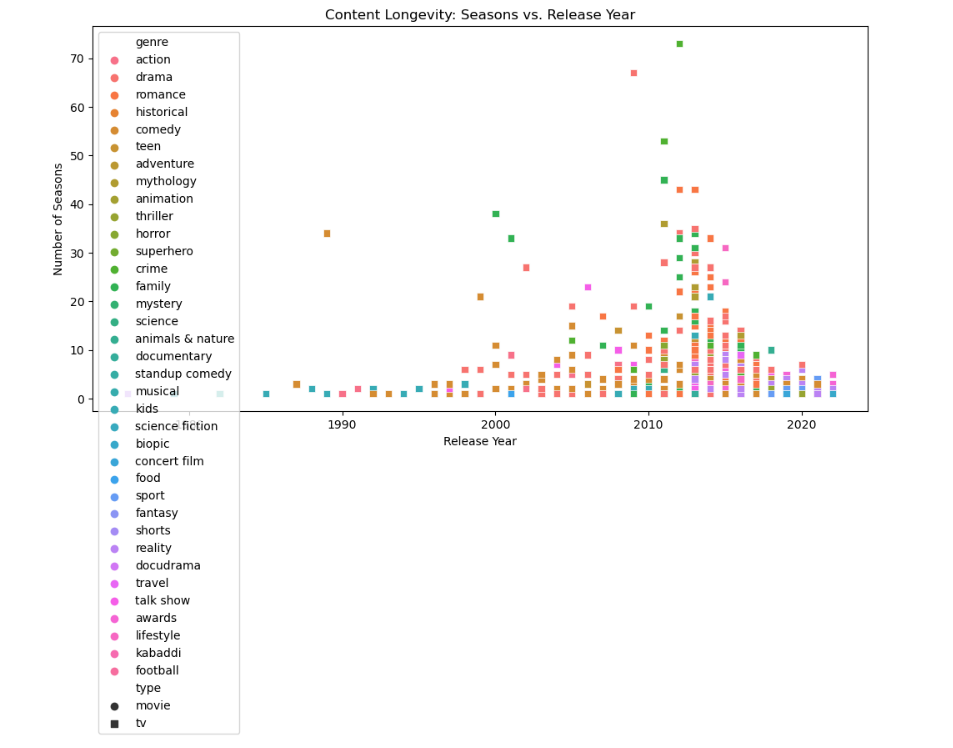
* **Genre Distribution**

The genres distribution in Disney Hotstar dataset was analyzed to find out what are the most common and dominant content types.



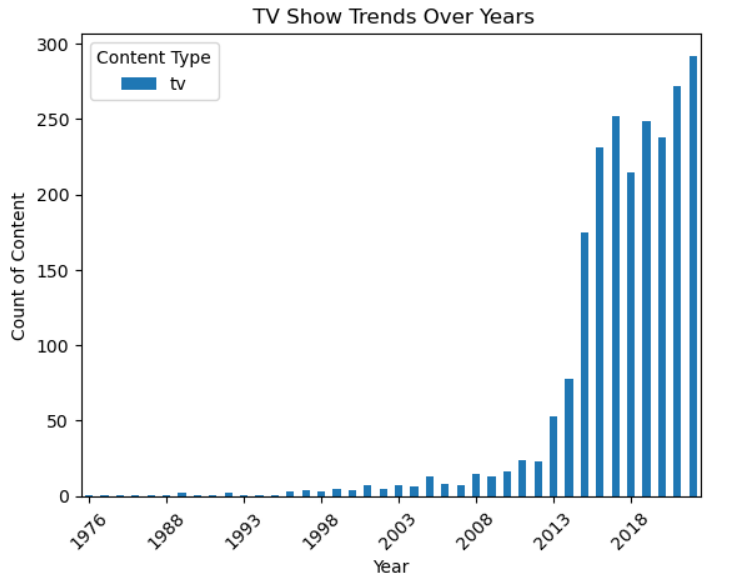
* **Content** **Longevity**

Visualizing the relationship between release year, number of seasons, and episodes provided insights into the longevity of TV shows over time.



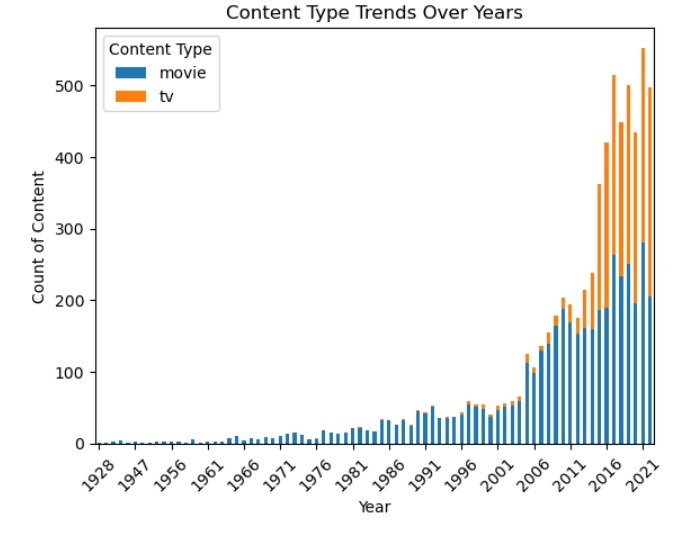
* **Seasonal Trends in TV Show Releases**

Over the years tv shows have got significant rise.



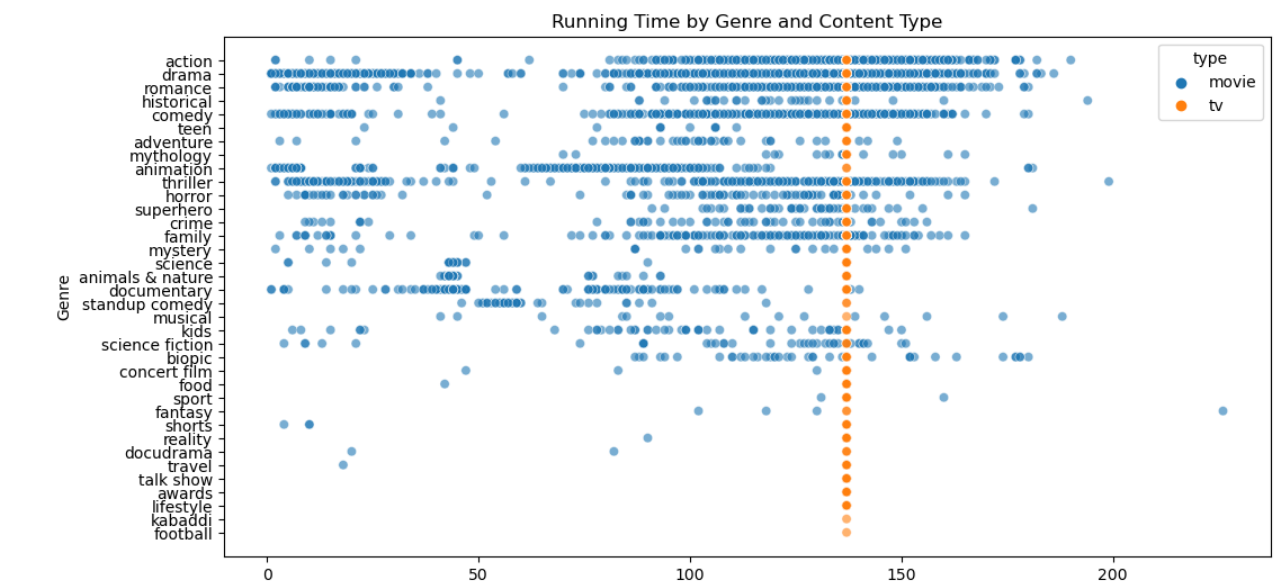
* **Content Types Over Time**

Over the years tv shows have got more priority and views than movies.

****

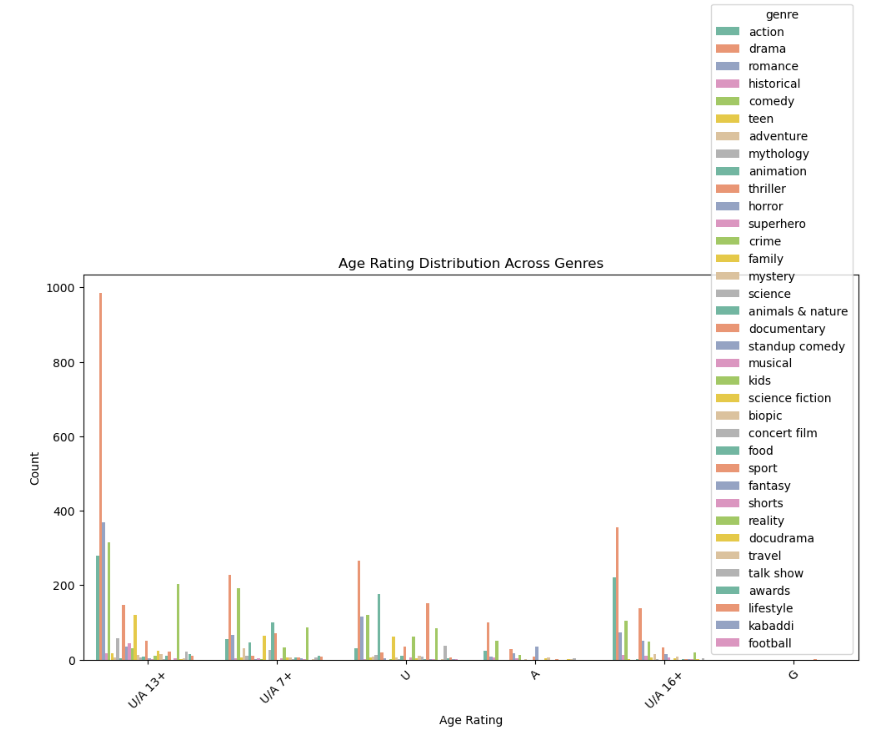
* **Distribution of Content Types by Running Time and Genre**

Drama and thriller shows are most watched.



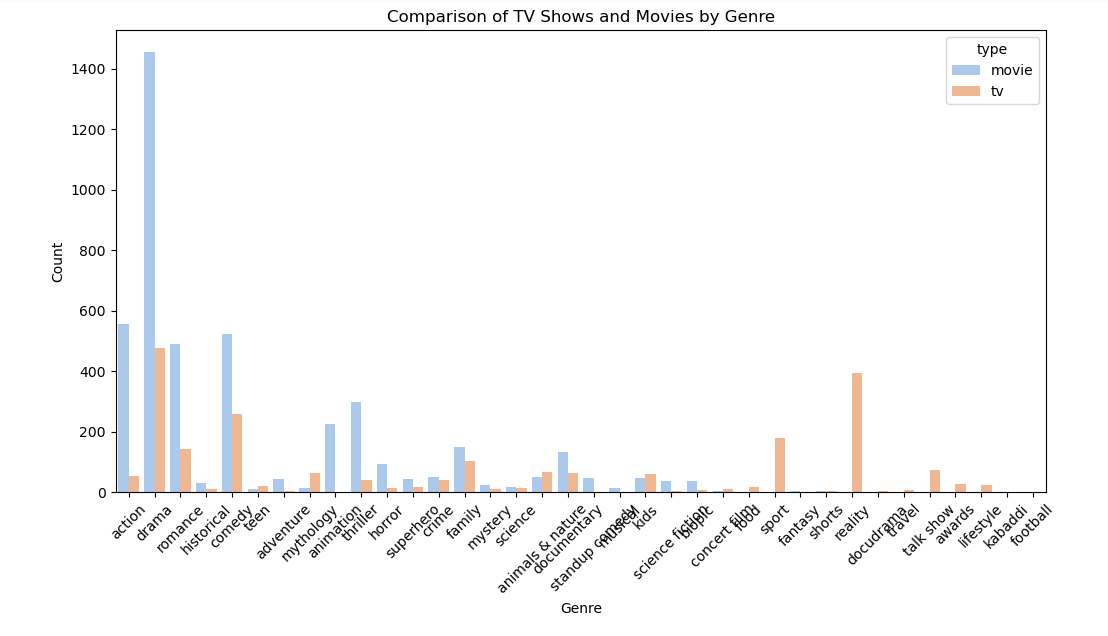
* **Age-Rating Distribution**

Biopic shows are most watched in UA 13+ and UA 16+

****

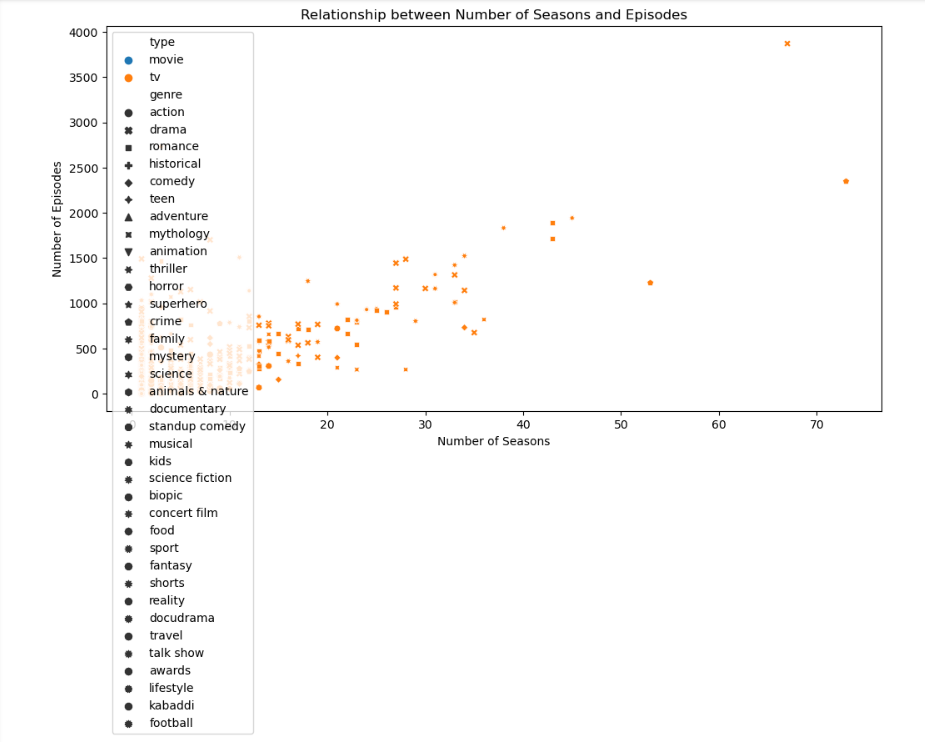
* **Comparing TV Shows and Movies**

Drama, action and historical movies are most watched while reality and drama tv shows were watched the most.

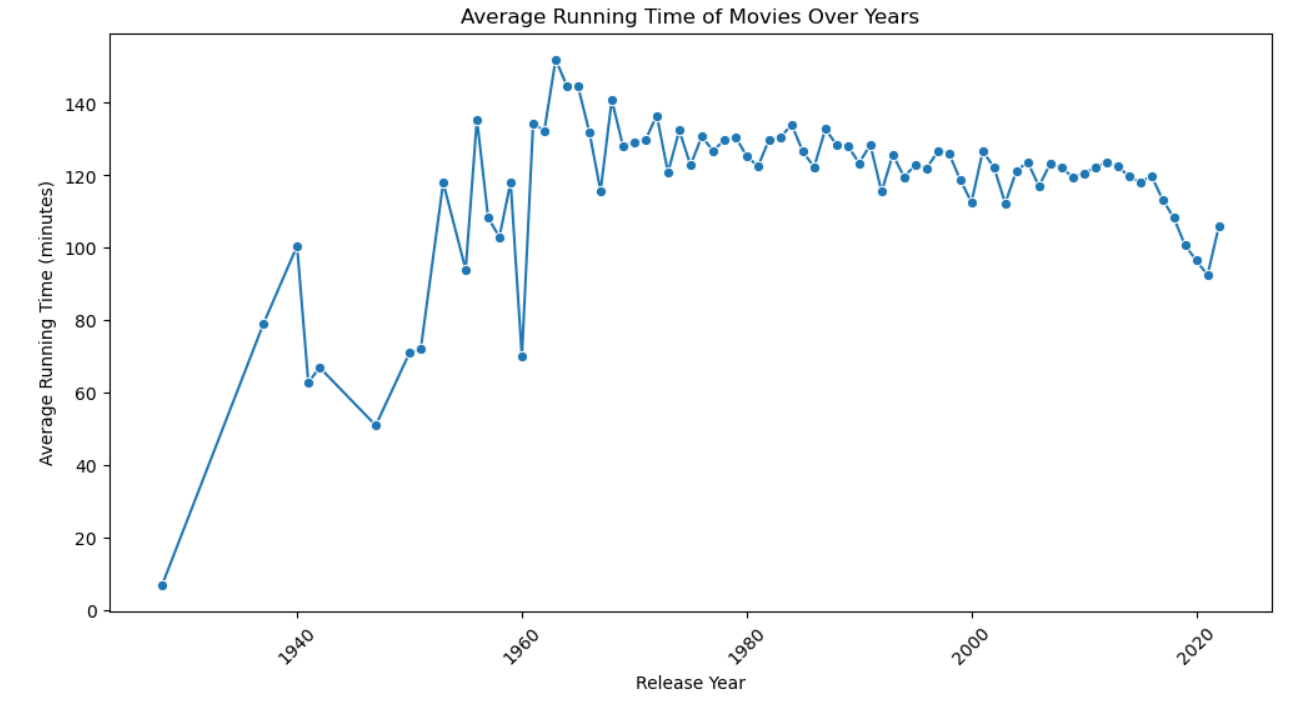


* **Seasons** and **Episode** **Count**

Mostly there are 1000-1500 episodes under 20 seasons.

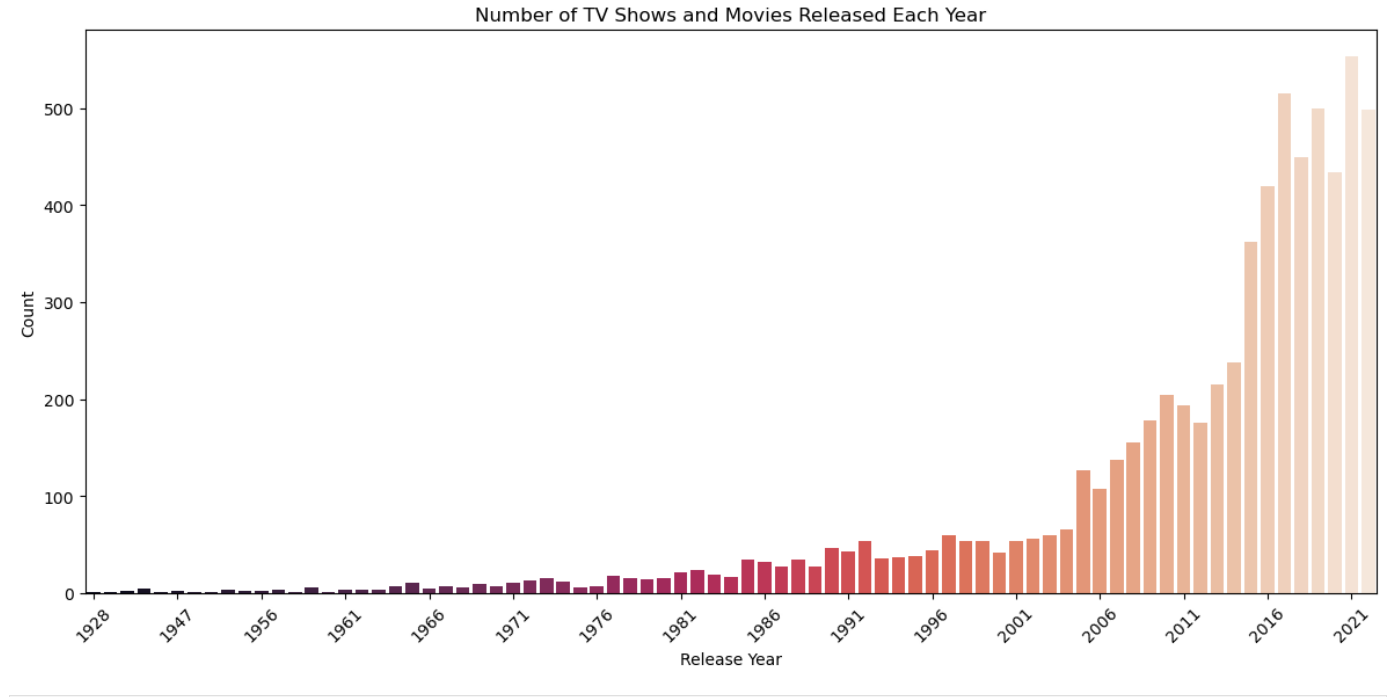


* **Content Length Trend**

Movies were most watched in between 1960-1980

* **Release Year Trends**

Number of tv shows and movies have increased with years.



**Key Insights:**

1. **Genre Distribution:** The Genre Distribution graph shows how many times each genre appears in the dataset. This makes it possible to discover which genres are the most and least represented on Hotstar platform. For a similar reason, it may be possible that genres such as Action, Drama and Comedy are more frequently populated than others.
2. **Content Type Trends Over Years:** The Content Type Trends Over Years graph then shows how the production of TV shows has changed over the years. When the data has already been sifted for TV shows, one can see shifts like the annual production of more or fewer TV shows.
3. **Content Longevity:** The Content Longevity: Seasons vs. Release Year scatter plot gives a clue about how long television series are likely to exist. Plotting the data is as simple as taking the number of seasons and the release year and then projecting it to understand if older shows have more seasons than the new ones.
4. **Missing Data:** From the first check it is clear that there are null values in the running\_time, seasons, episodes columns. This tells us that not all the entries have all the information which may be an issue in future analysis and therefore missing data should be dealt with properly.
5. **Standardization of Data:** There are steps that are followed in EDA process, for example converting the genre and type columns to something standardized. This step is also fundamental for correct analysis since problems originating in word differences (e.g., Action vs action) can arise.
6. **Age Rating Distribution:** Analyzing the age\_rating column gives us some insights into who is the target audience of the Hotstar’s content. A good example is a high occurrence rate of U/A 13+ ratings may mean such a focus on family content.
7. **Running Time Analysis:** When we looked at the running\_times column, we realize that the average movie or TV show on Hotstar is around that duration. It can even help you identify bigger trends — like newer pieces longer or shorter in duration.
8. **Season and Episode Count:** The seasons and episodes columns give clues into the structure of a tv show. These columns can be analyzed to know whether most TV shows on Hotstar have more than just one season and episode or are limited series.
9. **Yearly Content Production:** Now we can find out the number of movies and TV shows produced in a year. By executing this, we’ll be able to find out whether we’ve seen increased content production trends in the past or if in fact a lot of new releases occurred in the last few years.
10. **Content Type Proportions**: From the type column we can see what percentage of the movies offered on Hotstar or TV shows. It can also offer you insights into the kind of content the platform has chosen to deliver —movies or TV shows.

**Conclusion:**

The Exploratory Data Analysis (EDA) of the Disney Hotstar dataset has demonstrated that it is possible to apply these revelations to Disney Hotstar's efforts to optimize a range of issues such as recommended content, content procurement, viewer retention and targeted marketing.

They include catching up on emerging genres, monitoring family friendly content, and the change in content type preferences. These insights will help Disney Hotstar strategize their content offerings, their content strategies and correspond better to the changing viewers’ necessities.

With this being performed, Disney Hotstar will be able to predict the industry trends better, satisfy divergent audience segments, and an edge in importing data using pandas as pd.

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Load the Disney Hotstar dataset

url = 'C:\\Users\\Hriday\\Downloads\\hotstar.csv' # Replace with the actual dataset link

data = pd.read\_csv(url)

# Display the first few rows of the dataset

print(data.head())

# Check for missing values

print(data.isnull().sum())

# Ensure consistency across different categories

data['genre'] = data['genre'].str.lower() # Standardizing genre names

data['type'] = data['type'].str.lower() # Standardizing content types

data['running\_time'] = data['running\_time']

# Display cleaned data

print(data.info())

display(data)

## Exploratory Data Analysis (EDA)

# 1. Genre Distribution

plt.figure(figsize=(12, 6))

sns.countplot(data=data, x='genre', order=data['genre'].value\_counts().index, palette='viridis')

plt.title('Genre Distribution')

plt.xticks(rotation=45)

plt.ylabel('Count')

plt.show()

# 2. Content Longevity

plt.figure(figsize=(12, 6))

sns.scatterplot(data=data, x='year', y='seasons', hue='genre' , style='type', markers=["o", "s"])

plt.title('Content Longevity: Seasons vs. Release Year')

plt.xlabel('Release Year')

plt.ylabel('Number of Seasons')

plt.legend(loc='upper left')

plt.show()

# 3. Seasonal Trends in TV Show Releases

import pandas as pd

import matplotlib.pyplot as plt

from matplotlib.ticker import MultipleLocator

df = pd.read\_csv('hotstar.csv') # Read the CSV file

# Filter the DataFrame to include only 'tv' type

df\_tv = df[df['type'] == 'tv']

# Group the data by year and type, count occurrences, and reshape

content\_type\_trends = df\_tv.groupby(['year', 'type']).size().unstack()

# Create a plot

plt.figure(figsize=(12, 6))

content\_type\_trends.plot(kind='bar', stacked=True)

# Customize the plot

plt.title(' TV Show Trends Over Years')

plt.xlabel('Year')

plt.ylabel('Count of Content')

plt.xticks(rotation=45)

plt.legend(title='Content Type')

# Set x-axis major ticks to 5-year intervals

ax = plt.gca() # Get the current axis

ax.xaxis.set\_major\_locator(MultipleLocator(5)) # Set major ticks to every 5 years

# Display the plot

plt.show()

# 4. Content Types Over Time

from matplotlib.ticker import MultipleLocator

# Group the data by year and type, count occurrences, and reshape

content\_type\_trends = data.groupby(['year', 'type']).size().unstack()

# Create a plot

plt.figure(figsize=(12, 6))

content\_type\_trends.plot(kind='bar', stacked=True)

# Customize the plot

plt.title('Content Type Trends Over Years')

plt.xlabel('Year')

plt.ylabel('Count of Content')

plt.xticks(rotation=45)

plt.legend(title='Content Type')

# Set x-axis major ticks to 5-year intervals

ax = plt.gca() # Get the current axis

ax.xaxis.set\_major\_locator(MultipleLocator(5)) # Set major ticks to every 50 years

# Display the plot

plt.show()

# 5. Distribution of Content Types by Running Time and Genre

plt.figure(figsize=(12, 6))

sns.scatterplot(data=data, x='running\_time', y='genre', hue='type', alpha=0.6)

plt.title('Running Time by Genre and Content Type')

plt.xlabel('Running Time (minutes)')

plt.ylabel('Genre')

plt.show()

# 6. Age-Rating Distribution

plt.figure(figsize=(12, 6))

sns.countplot(data=data, x='age\_rating', hue='genre', palette='Set2')

plt.title('Age Rating Distribution Across Genres')

plt.xlabel('Age Rating')

plt.ylabel('Count')

plt.xticks(rotation=45)

plt.show()

# 7. Comparing TV Shows and Movies

plt.figure(figsize=(12, 6))

sns.countplot(data=data, x='genre', hue='type', palette='pastel')

plt.title('Comparison of TV Shows and Movies by Genre')

plt.xlabel('Genre')

plt.ylabel('Count')

plt.xticks(rotation=45)

plt.show()

# 8. Seasons and Episode Count

plt.figure(figsize=(12, 6))

sns.scatterplot(data=data, x='seasons', y='episodes', hue='type', style='genre')

plt.title('Relationship between Number of Seasons and Episodes')

plt.xlabel('Number of Seasons')

plt.ylabel('Number of Episodes')

plt.legend(loc='upper left')

plt.show()

# 9. Content Length Trends

plt.figure(figsize=(12, 6))

sns.lineplot(data=data.groupby('year')['running\_time'].mean(), marker='o')

plt.title('Average Running Time of Movies Over Years')

plt.xlabel('Release Year')

plt.ylabel('Average Running Time (minutes)')

plt.xticks(rotation=45)

plt.show()

# 10. Release Year Trends

plt.figure(figsize=(12, 6))

sns.countplot(data=data, x='year', palette='rocket')

plt.title('Number of TV Shows and Movies Released Each Year')

plt.xlabel('Release Year')

plt.ylabel('Count')

plt.xticks(rotation=45)

# Set x-axis major ticks to intervals of 5 years

ax = plt.gca() # Get current axis

ax.xaxis.set\_major\_locator(MultipleLocator(5)) # Set major ticks every 5 years

plt.tight\_layout()

plt.show()

This code implements the key aspects of the Exploratory Data Analysis (EDA) outlined in the previous response. It covers the following:

1. **Data Preprocessing**:
   * Handling missing values
   * Standardizing data formats
   * Ensuring data consistency
2. **Exploratory Data Analysis**:
   * Genre distribution
   * Content longevity
   * Seasonal trends in TV show releases
   * Content types over time
   * Distribution of content types by running time and genre
   * Age-rating distribution
   * Comparing TV shows and movies
   * Seasons and episode count
   * Content length trends
   * Release year trends
3. The code uses popular Python libraries like Pandas, Matplotlib, and Seaborn to load the dataset, perform data manipulation, and create visualizations to uncover the key insights outlined in the previous response