Presented By: Yash Raj Kohli

Email: yashrajkohli5@gmail.com

Table of Contents

About Prime Video

Amazon Prime Video is a premier global streaming entertainment platform that serves as a central component of the broader Amazon Prime membership ecosystem. Originally launched to complement Amazon's e-commerce services, the platform has evolved into a sophisticated digital hub available in over 200 countries and territories. By integrating high-definition streaming with a versatile marketplace, Prime Video allows millions of subscribers to access content through a vast array of internet-connected devices, including smart TVs, mobile phones, tablets, and gaming consoles.

Beyond standard streaming, the platform distinguishes itself through advanced technological features like **X-Ray**, which provides real-time metadata and production insights. Amazon's data-centric infrastructure leverages user viewing patterns to refine its recommendation engine, ensuring a highly personalized interface for every subscriber. Furthermore, the service operates as a comprehensive content aggregator, enabling users to add third-party premium channels and digital rentals to their accounts. This strategic integration of technology and commerce continues to position Amazon Prime Video as a dominant force in shaping the future of on-demand digital media.

Objective

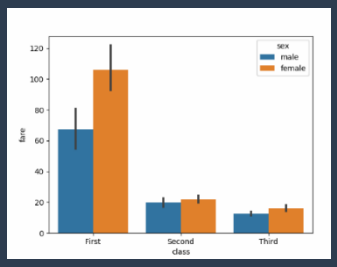
Purpose of the Dataset:

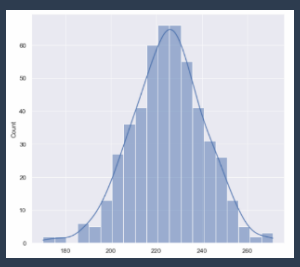
The goal of the **Amazon Prime Video EDA project** is to perform a comprehensive investigation into the content catalogue of one of the world's leading OTT platforms. This involves analysing the **data structure**, ensuring **data quality** by identifying missing entries (such as directors or countries), and using **descriptive statistics** to quantify the diversity of the library.

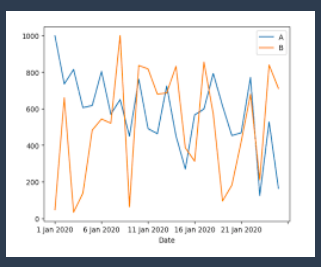
Additionally, the project aims to uncover **content strategy trends**, such as the ratio of Movies vs. TV Shows, and evaluate **audience reach** by analysing maturity ratings and genre distributions. By visualizing these patterns across different release years, the project seeks to derive **data-driven insights** into how Amazon Prime Video curates its global library to stay competitive in the streaming market

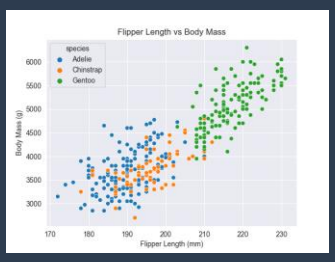
TOOLS

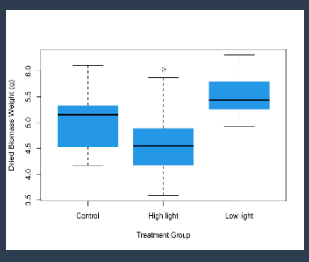
GRAPHS

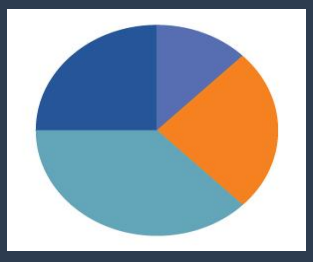












Data Overview

- In this analysis we use Pandas for data manipulation, Numpy for numerical operations, Matplotlib for visualization, Seaborn and Plotly for advanced statistical graphics and datetime for data pre-processing.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import datetime as dt
```

- We use pandas to Read CSV file containing Prime Video data into a DATAFRAME named 'df' for analysis

Overview of dataset:

- **First few rows**

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	The Grand Seduction	Don McKellar	Brendan Gleeson, Taylor Kitsch, Gordon Pinsent	Canada	March 30, 2021	2014	NaN	113 min	Comedy, Drama	A small fishing village must procure a local d...
1	s2	Movie	Take Care Good Night	Girish Joshi	Mahesh Manjrekar, Abhay Mahajan, Sachin Khedekar	India	March 30, 2021	2018	13+	110 min	Drama, International	A Metro Family decides to fight a Cyber Crimin...
2	s3	Movie	Secrets of Deception	Josh Webber	Tom Sizemore, Lorenzo Lamas, Robert LaSardo, R...	United States	March 30, 2021	2017	NaN	74 min	Action, Drama, Suspense	After a man discovers his wife is cheating on ...
3	s4	Movie	Pink: Staying True	Sonia Anderson	Interviews with: Pink, Adele, Beyoncé, Britney...	United States	March 30, 2021	2014	NaN	69 min	Documentary	Pink breaks the mold once again, bringing her ...
4	s5	Movie	Monster Maker	Giles Foster	Harry Dean Stanton, Kieran O'Brien, George Cos...	United Kingdom	March 30, 2021	1989	NaN	45 min	Drama, Fantasy	Teenage Matt Banting wants to work with a famo...

The dataset consists of 9668 rows and 12 columns

Description of columns

- **show_id**: A unique identifier assigned to each movie or TV show in the dataset.
- **type**: Categorizes the entry as either a **Movie** or a **TV Show**.
- **title**: The official name of the movie or television series.
- **director**: The individual(s) responsible for directing the content (contains many null values).
- **cast**: A list of the lead actors and performers featured in the production.
- **country**: The primary country or countries where the content was produced.
- **date_added**: The specific date when the title was first made available on the Prime Video platform.
- **release_year**: The actual year the movie or show was originally released to the public.
- **rating**: The maturity or age-suitability certification (e.g., TV-MA, PG-13, 18+).
- **duration**: The total runtime for movies (in minutes) or the number of seasons for TV shows.
- **listed_in**: The genres or categories assigned to the content (e.g., Comedy, Drama, Horror).

Data Exploration

```
df.shape
```

```
(9668, 12)
```

The dataset consists of 9668 rows and 12 columns

```
df.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9668 entries, 0 to 9667
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       9668 non-null   object
 1   type          9668 non-null   object
 2   title         9668 non-null   object
 3   director      7585 non-null   object
 4   cast          8435 non-null   object
 5   country       672 non-null    object
 6   date_added    155 non-null    object
 7   release_year  9668 non-null   int64 
 8   rating        9331 non-null   object
 9   duration      9668 non-null   object
 10  listed_in     9668 non-null   object
 11  description   9668 non-null   object
dtypes: int64(1), object(11)
memory usage: 906.5+ KB
```

The Amazon Prime Video dataset comprises **9,668 records** and **12 columns**, capturing essential metadata about movies and TV shows available on the platform. The dataset includes **11 categorical features** and **1 numerical feature** (release_year). These variables describe content attributes such as title, director, cast, duration, genres, country of origin, and maturity rating. A few columns, most notably **director**, **cast**, **country**, and **date_added**, contain significant missing values that will be addressed during the data cleaning phase.

Data Cleaning

- Checking for missing values

By Values

```
df.isnull().sum()
✓ 0.0s
```

show_id	0
type	0
title	0
director	2083
cast	1233
country	8996
date_added	9513
release_year	0
rating	337
duration	0
listed_in	0
description	0

dtype: int64

By Percentage

```
df.isnull().sum()/df.shape[0]*100
✓ 0.0s
```

show_id	0.000000
type	0.000000
title	0.000000
director	21.545304
cast	12.753413
country	93.049235
date_added	98.396773
release_year	0.000000
rating	3.485726
duration	0.000000
listed_in	0.000000
description	0.000000

dtype: float64

- **Summary of missing values:**

  The Amazon Prime Video dataset exhibits a significant amount of missing data across several key metadata fields, which will require careful handling during the cleaning process. The **country** column shows the highest number of missing entries (**8,996**), followed by **date_added** (**9,513**) and **director** (**2,083**), indicating highly incomplete geographic and temporal information for much of the catalogue. Additionally, moderate missing values are present in **cast** (**1,233**) and **rating** (**337**), reflecting gaps in the creative and audience classification metadata.

Data Exploration

- **Handling missing values**

```
df['director'] = df['director'].fillna("Unknown")
df["cast"] = df["cast"].fillna("Unknown")
df["country"] = df["country"].fillna("Unknown")
df["rating"] = df["rating"].fillna("Other")
```

Filling ‘director’, ‘cast’, ‘country’ with ‘Unknown’ and ‘rating’ with ‘Other’.

```
missing_mask = df["date_added"].isnull()
num_missing = missing_mask.sum()

valid_dates = df["date_added"].dropna()

df.loc[missing_mask, 'date_added'] = np.random.choice(valid_dates, size=num_missing)
```

Filling missing values of ‘date_added’ with random sampling.

- **Remove duplicates if any**

```
df.duplicated().sum()
```
✓ 0.0s

```
np.int64(0)
```

- **Correct any inconsistencies or errors in the data.**

```
df.rename(columns = {'listed_in' : 'genres'}, inplace = True)
df['date_added'] = pd.to_datetime(df['date_added'])
df['year_added'] = df['date_added'].dt.year
df['month_added'] = df['date_added'].dt.month
```

Renaming ‘listed_in’ column to ‘genres’ for better understanding and adding columns ‘year_added’ and ‘month_added’ for better analysis.

Data Exploration

- **Summarize basic statistics for the numeric columns.**

	date_added	release_year	year_added	month_added
count	9668	9668.000000	9668.0	9668.000000
mean	2021-03-31 17:00:25.320645376	2008.341849	2021.0	3.062888
min	2021-03-30 00:00:00	1920.000000	2021.0	3.000000
25%	2021-03-30 00:00:00	2007.000000	2021.0	3.000000
50%	2021-03-30 00:00:00	2016.000000	2021.0	3.000000
75%	2021-03-30 00:00:00	2019.000000	2021.0	3.000000
max	2021-10-10 00:00:00	2021.000000	2021.0	10.000000
std	NaN	18.922482	0.0	0.565349

Summary of Descriptive Statistics:

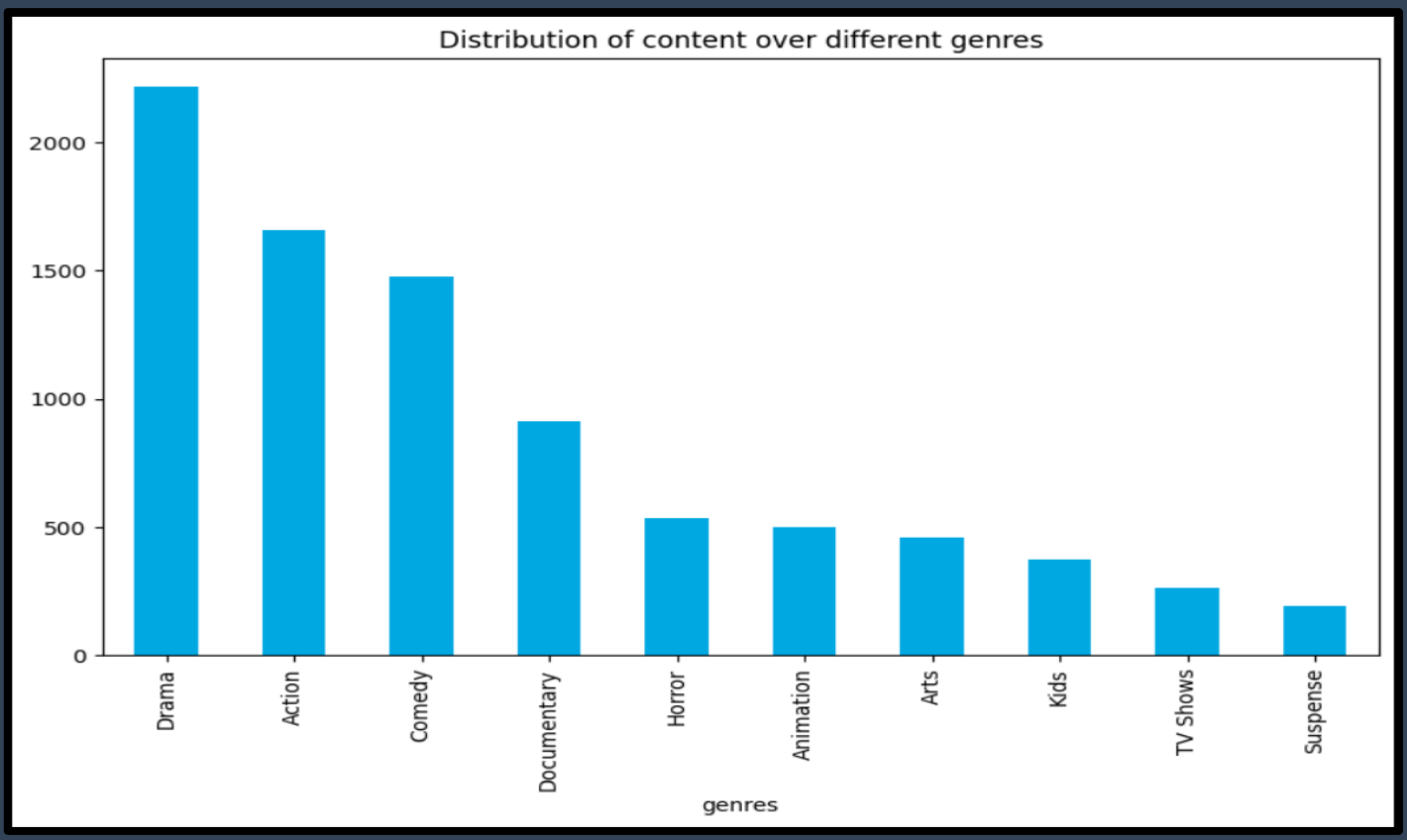
The Amazon Prime Video catalogue is primarily tailored for modern viewers, focusing on fresh content while maintaining historical depth. Although titles in the collection date back as far as 1920, the median release year of 2016 confirms that the library is largely dominated by modern content. In fact, at least 75% of the movies and shows were released from 2007 onwards, highlighting that while the service respects cinematic history, its core strength lies in titles from the last decade.

Overall, the descriptive statistics show that the library underwent a massive expansion in 2021. To ensure a complete and realistic picture of this growth, we used random sampling to fill in missing record dates. This revealed that while there was a significant surge of new additions around March 30, 2021, the catalogue continued to grow steadily through October. This approach provides a much more natural view of how the collection evolves over time, rather than appearing as if it were updated all at once.

Data Visualization

- **Create visualization to represent the distribution of content over top 10 different genres.**

```
top_genres = df['genres'].str.split(',').str.get(0)
plt.figure(figsize = (10,6))
top_genres.value_counts().head(10).plot(kind="bar", color = "#00A8E1")
plt.title("Distribution of content over different genres")
```



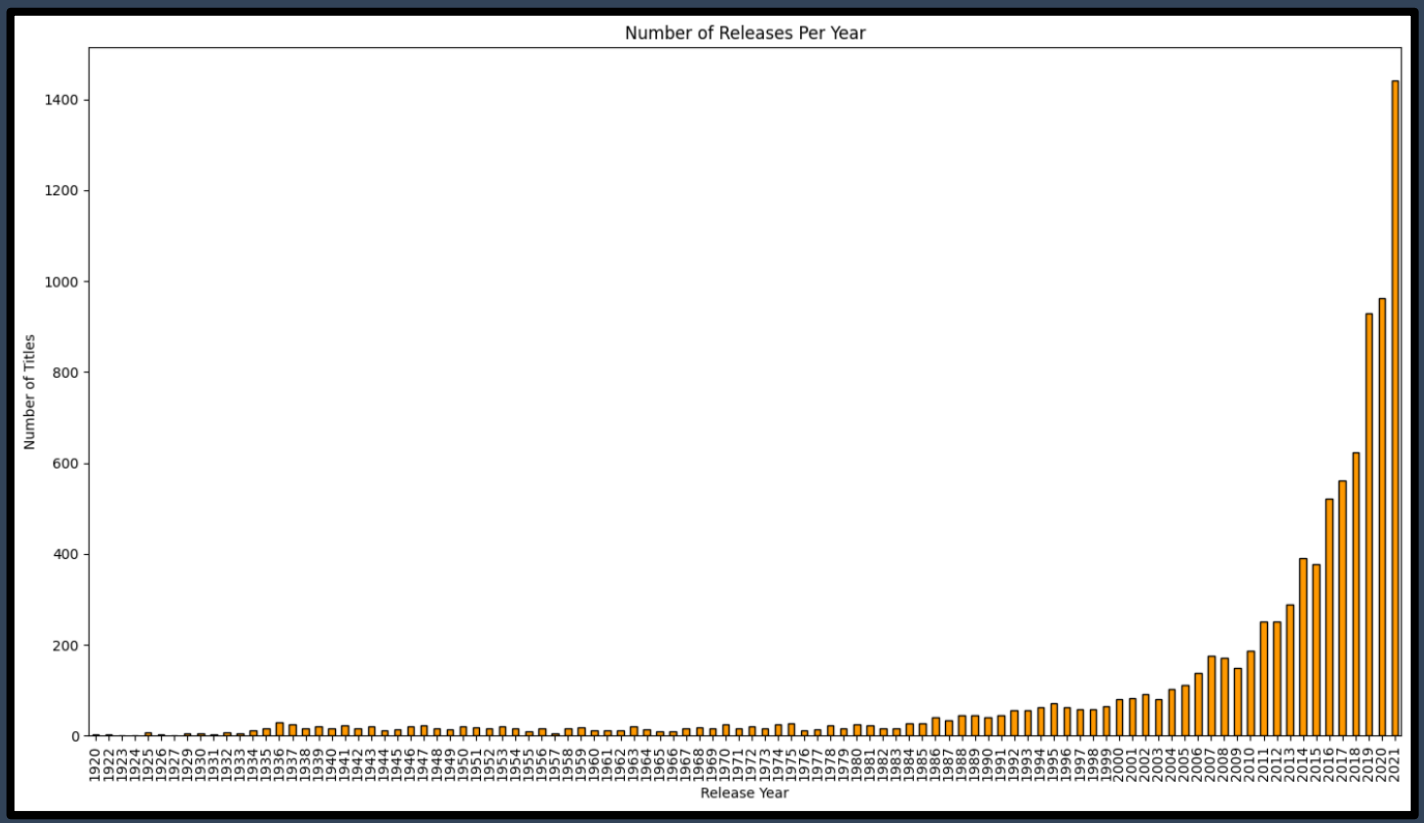
Insights:

The analysis of the top genres shows that a few categories dominate the library. Genres such as **Drama**, **Action**, and **Comedy** appear most frequently, with Drama leading at over **2,200 titles**. This indicates a strong user demand and a content strategy focused on these popular entertainment categories.

Data Visualization

- **Visualize the distribution of content released across the years.**

```
plt.figure(figsize=(14,8))
df['release_year'].value_counts().sort_index().plot(kind="bar", color = "#FF9900", edgecolor = "black")
plt.title("Number of Releases Per Year")
plt.xlabel("Release Year")
plt.ylabel("Number of Titles")
plt.tight_layout()
plt.show()
```



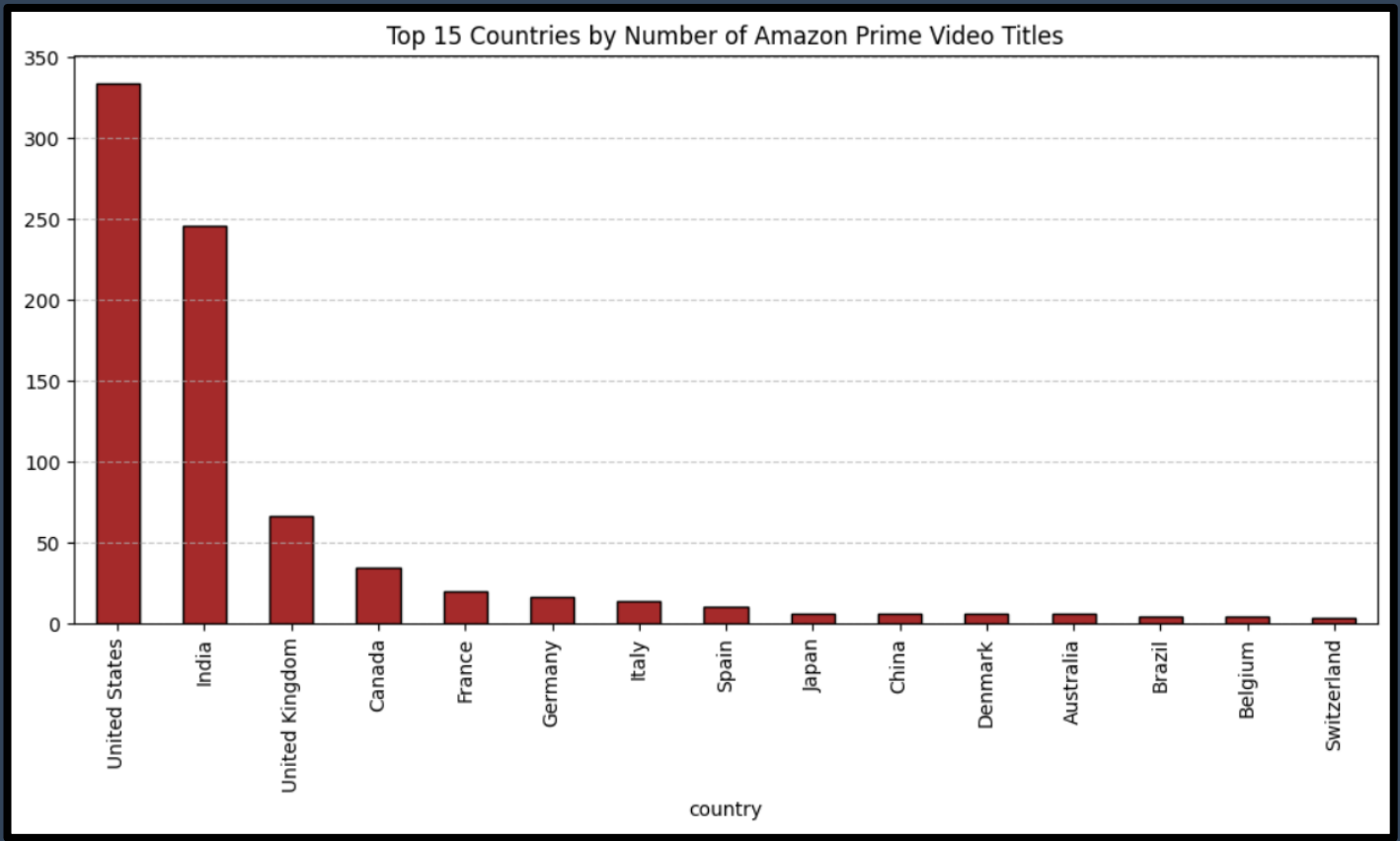
Insights:

The library is defined by a massive concentration of modern titles. While the collection technically stretches back a century, it is now dominated by a recent explosion of content, with half of the entire catalogue released from **2016** onwards. This growth reaches an extreme peak in **2021**, where the number of new releases is at its highest point in history.

Data Visualization

- **Explore the geographical distribution of the content.**

```
country_counts = (df['country'].dropna().str.split(',').explode().str.strip())
country_counts = country_counts[country_counts != "Unknown"].value_counts().head(15)
plt.figure(figsize=(10,6))
country_counts.plot(kind = "bar", color = "brown", edgecolor = "black")
plt.title("Top 15 Countries by Number of Amazon Prime Video Titles")
plt.grid(axis="y", linestyle = '--', alpha = 0.7)
plt.tight_layout()
plt.show()
```



Insights:

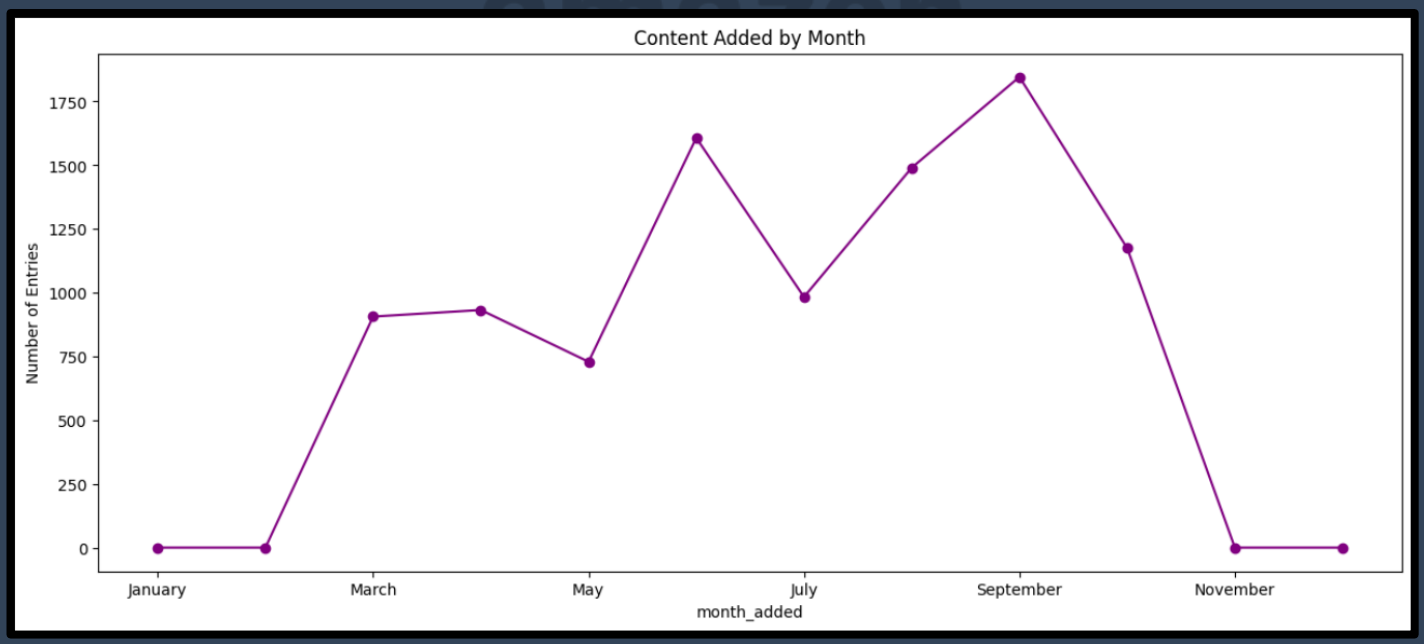
The library is heavily dominated by content from the **United States**, which leads with over **330 titles**. **India** holds a strong second position with nearly **250 titles**, highlighting its status as a critical market for Amazon's international growth. Beyond these two leaders, the library features a long tail of content from the **United Kingdom, Canada, and France**, which round out the top five global contributors.

Data Visualization

- **Time Series Analysis: If there's a temporal component perform time series analysis to identify trends and patterns over time.**

```
import calendar

df['month_added'] = df['month_added'].apply(lambda x: calendar.month_name[int(x)])
month_order = list(calendar.month_name)[1:]
df['month_added'] = pd.Categorical(df['month_added'], categories=month_order, ordered=True)
plt.figure(figsize=(15,6))
df['month_added'].value_counts().sort_index().plot(kind='line', marker='o', color='purple')
plt.title("Content Added by Month")
plt.ylabel("Number of Entries")
```



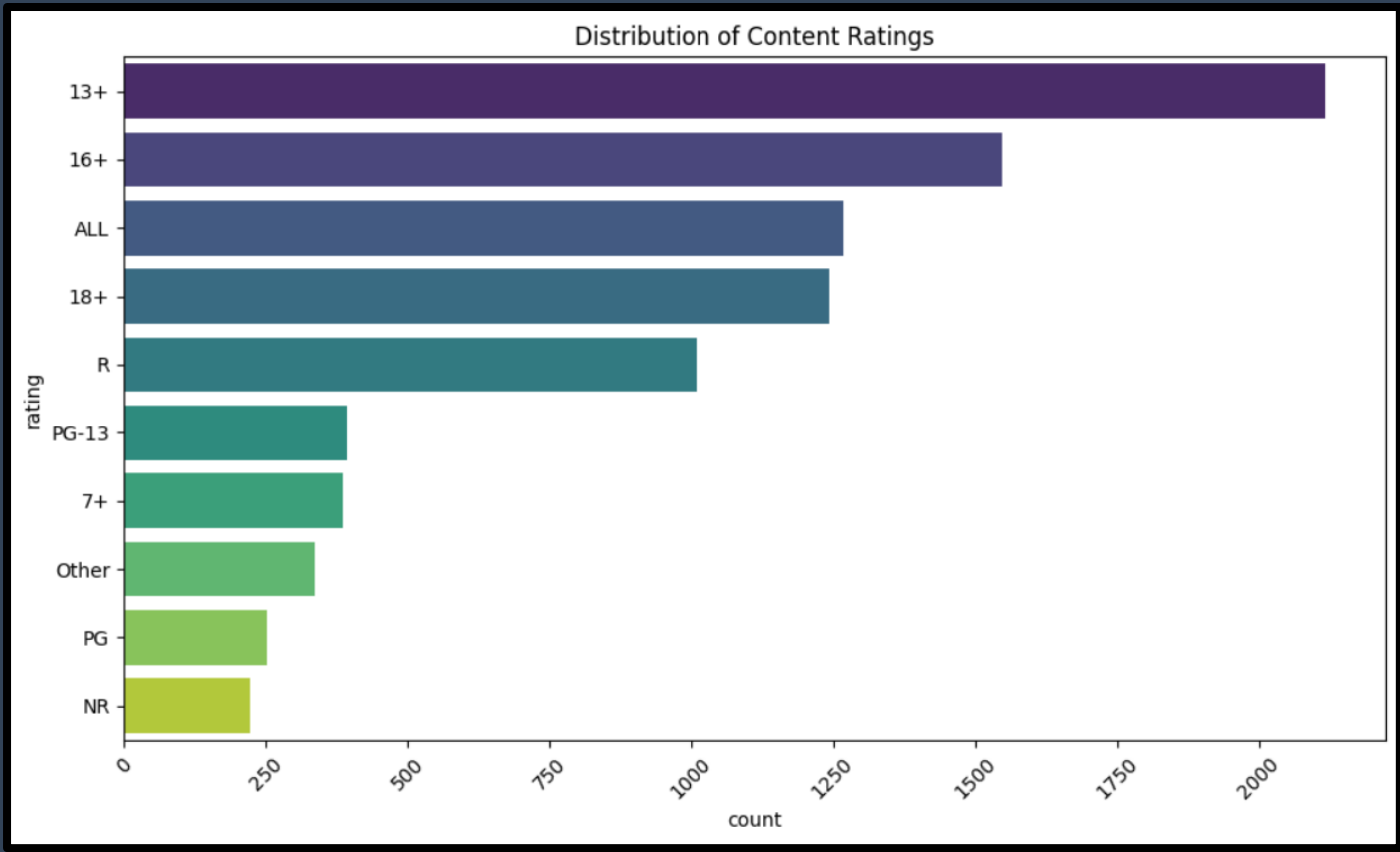
Insights:

The Amazon Prime Video catalogue follows a highly strategic and seasonal pattern in how it expands its library throughout the year. Data shows a massive concentration of activity during the mid-year months, characterized by a **primary peak in September** where content additions reach a record-high of over **1,750 entries**. A **secondary surge** occurs in **June** with approximately **1,600 new titles**, contributing to a consistent eight-month window from March to October where the platform maintains a steady pace of at least **700 new entries per month**.

Data Visualization

- **Analyse the distribution of content ratings.**

```
plt.figure(figsize=(10,6))
sns.countplot(data=df, y='rating',
order = df['rating'].value_counts().index[:10], palette="viridis")
plt.title("Distribution of Content Ratings")
plt.xticks(rotation = 45)
plt.tight_layout()
plt.show()
```

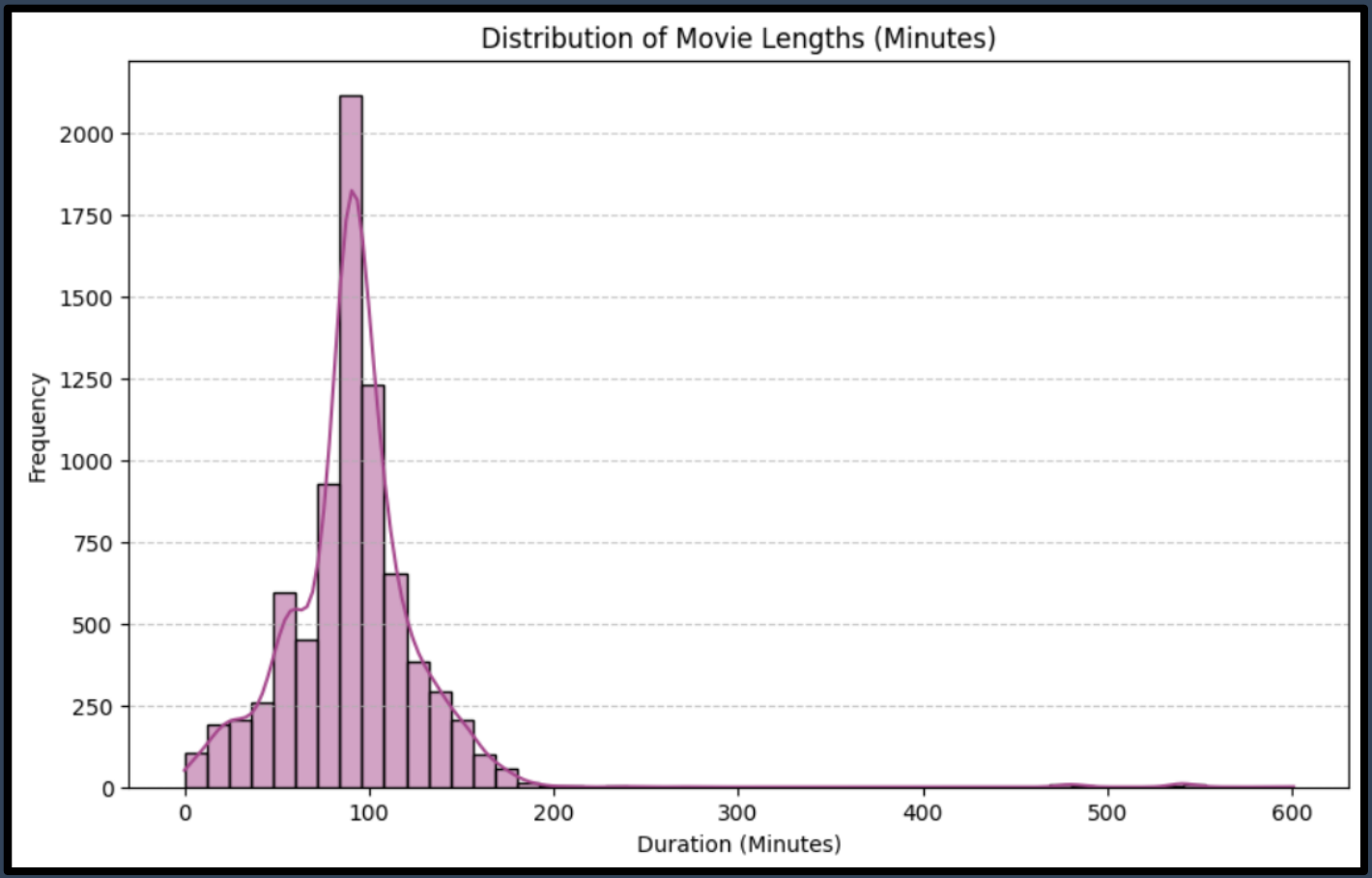


Insights:

- **Mature Audience Focus**: The library is dominated by **13+** and **16+** categories, with the 13+ rating alone surpassing **2,100 titles**.
- **General Entertainment Priority**: Mature content (**18+** and **R**) forms a substantial portion of the catalogue, while family-friendly segments like **ALL** and **PG** represent smaller, middle-tier volumes.
- **Limited Niche Content**: Ratings such as **7+** and **NR** are among the least common, confirming a library built for general entertainment rather than specialized children's programming.

Data Visualization

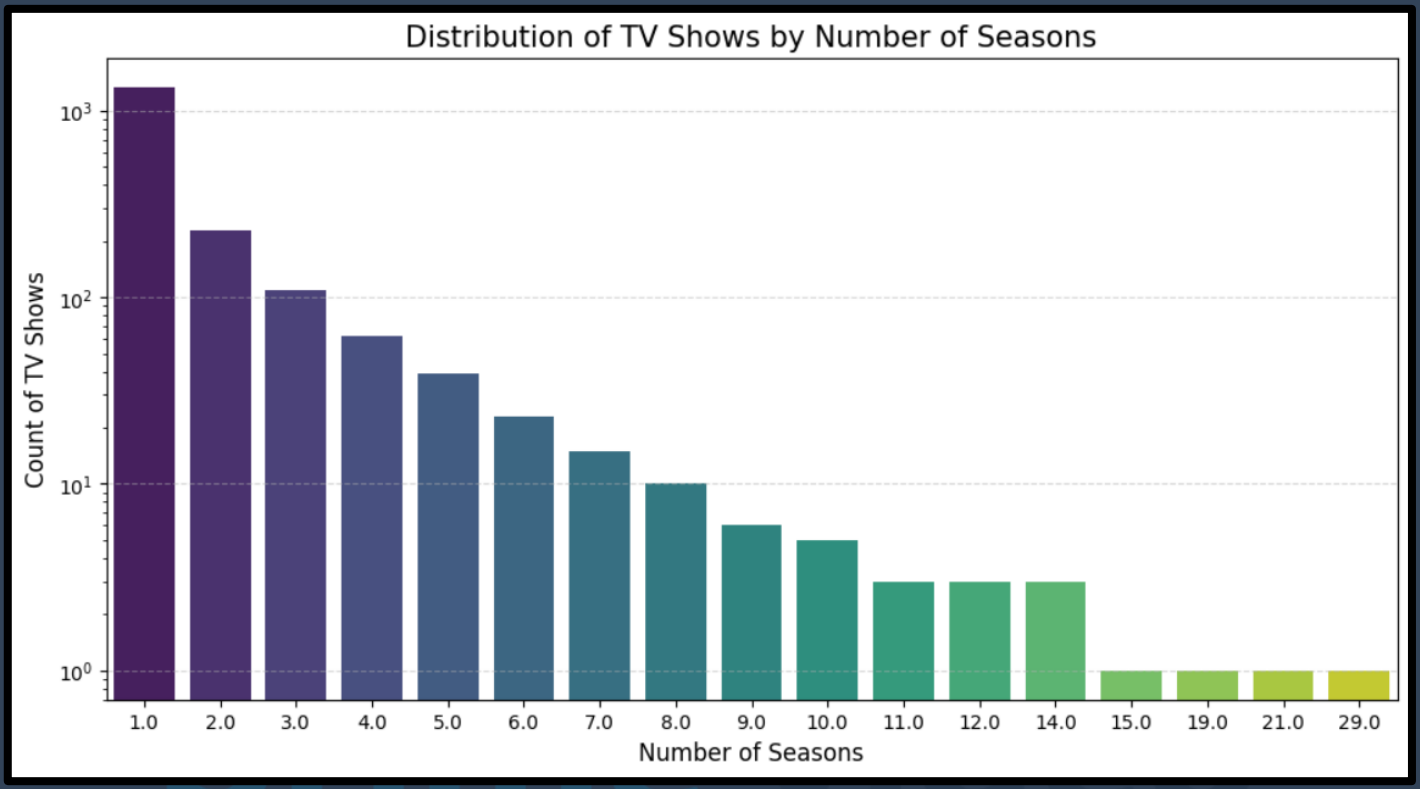
- **Explore the length of movies or episodes and identify any trends.**



Insights:

The distribution of movie lengths on Amazon Prime Video reveals a significant concentration around the **90-minute mark**, which serves as the primary peak of the catalogue. While the majority of content is clustered between **75 and 125 minutes**, there is a notable "long tail" featuring extreme outliers that extend up to **600 minutes**. This pattern confirms that the platform prioritizes standard feature-length films while also maintaining a small selection of exceptionally long specialized content.

Data Visualization



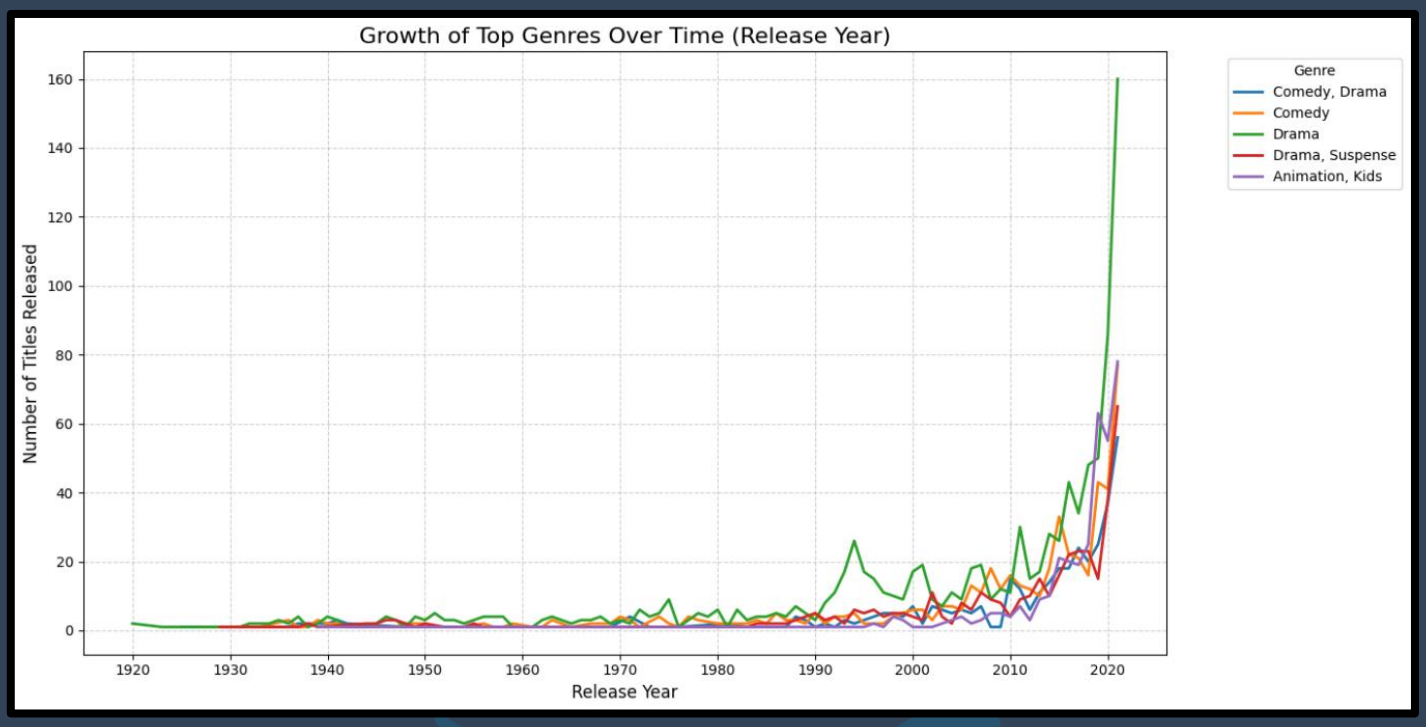
Insights:

The TV show catalogue is heavily front-loaded with single-season titles, indicating a high volume of new releases or limited series. While the number of series drops sharply as season counts increase, a significant "long tail" persists, featuring veteran franchises with up to **29 seasons**. This structure reveals a strategy that balances a constant stream of fresh, one-season content with a small but deep archive of binge-heavy legacy hits.

Data Visualization

Genre Trends:

- **Analyse the trends in the popularity of different genres overtime.**



Insights:

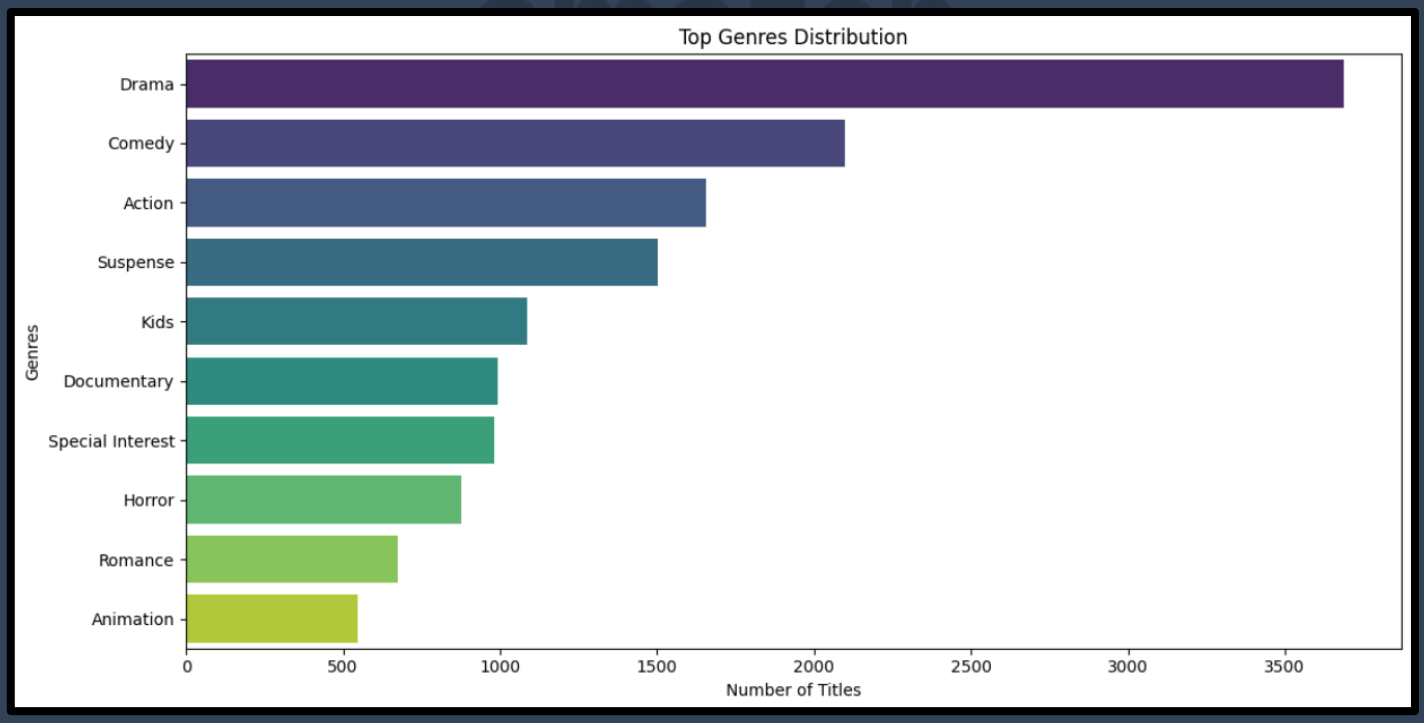
- **Massive Growth After 2015**: The line graph shows a sharp, near-vertical increase in the number of movies and shows added to the platform starting around 2015.
- **Expansion of Content Types**: While the chart shows a few dominant categories early on, the lines begin to branch out significantly after 2010 as **Documentaries** and **Special Interest** programs gain volume.
- **2016–2019 Surge**: The data points for these four years show a significant and rapid jump in total new releases compared to any previous period.
- **Leading Categories**: The lines representing **Drama** and **Comedy** consistently stay at the top of the y-axis, maintaining the highest volume of titles throughout the timeline.
- **2020 Growth and Rebound**: The graph indicates that content additions continued through 2020, followed by a very steep "catch-up" spike in the data immediately after the pandemic period.

Data Visualization

- **Evaluate the diversity of content by analysing unique genres.**

```
all_genres = genre_df['genre_list'].explode().str.strip()
genre_counts = all_genres.value_counts().head(15)

plt.figure(figsize=(12,6))
sns.barplot(x=genre_counts.values, y=genre_counts.index, palette='viridis')
plt.title("Top Genres Distribution")
plt.xlabel("Number of Titles")
plt.ylabel("Genres")
plt.tight_layout()
plt.show()
```



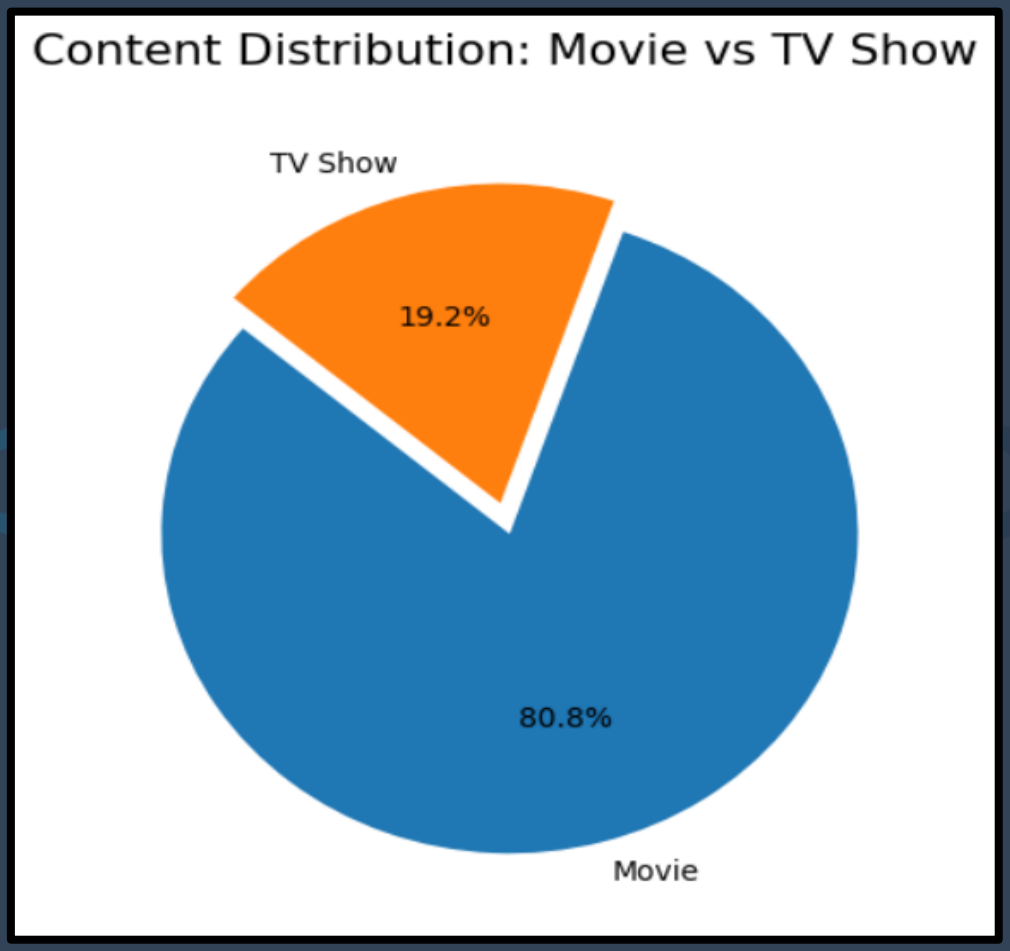
Insights:

- Genres such as Dramas, Comedy and Action appears most frequently indicating strong user interest.
- Niche genres like Romance, Animation are present but less dominant.
- The presence of multiple genres represents high cultural diversity, showing that platform serves a global audience.

Data Visualization

- **Evaluate the diversity of content by category.**

```
type_counts = df['type'].value_counts()
plt.pie(type_counts, labels=type_counts.index, autopct='%1.1f%%', startangle=140, explode=[0.05] * len(type_counts))
plt.title('Content Distribution: Movie vs TV Show', fontsize=16, pad=20)
plt.tight_layout()
plt.show()
```

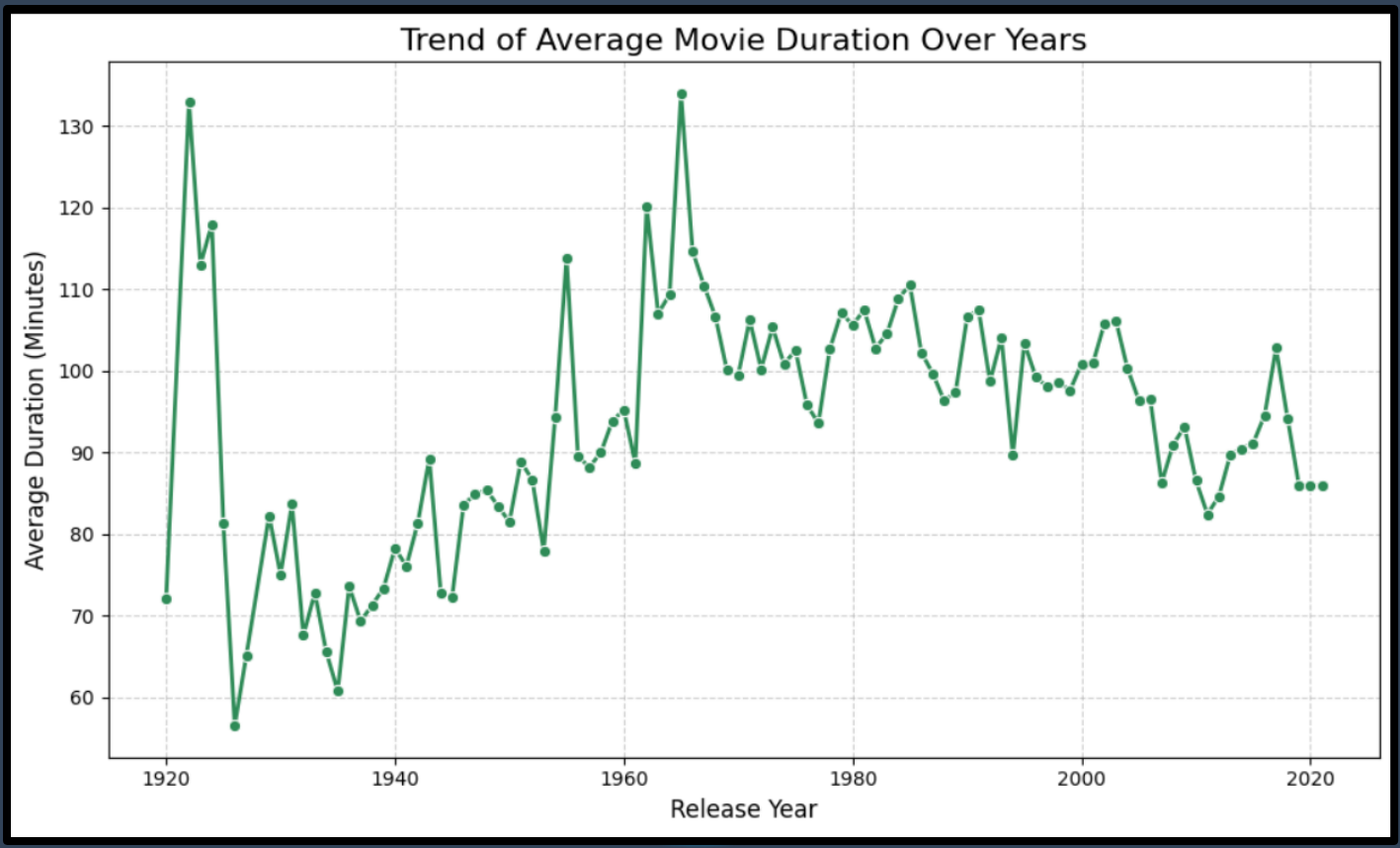


Insights:

- The catalogue is overwhelmingly movie-centric, with feature films outnumbering television series by more than four to one.

Data Visualization

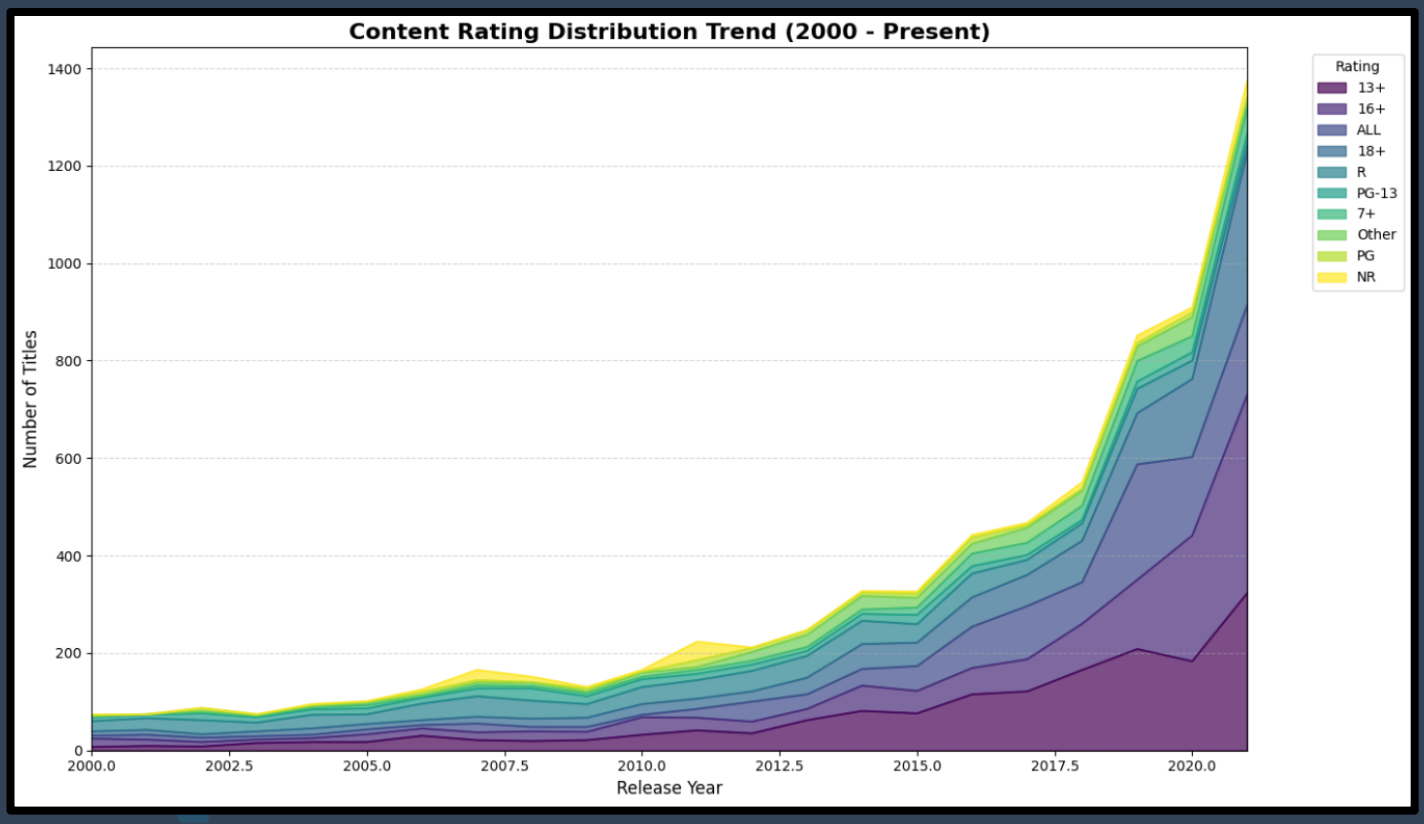
- **Explore how the characteristics of content (e.g., duration, ratings) have evolved over the years.**



Insights:

This graph shows that movie lengths on Amazon Prime Video have been on a wild ride over the last century. In the early days, the timing was like a roller coaster, frequently jumping between being very short and over **130 minutes** long. After 1970, movies stayed more consistent, usually lasting between **90 and 110 minutes**. However, since **2010**, movies have started getting shorter again, with the average length dropping down toward **85 minutes**.

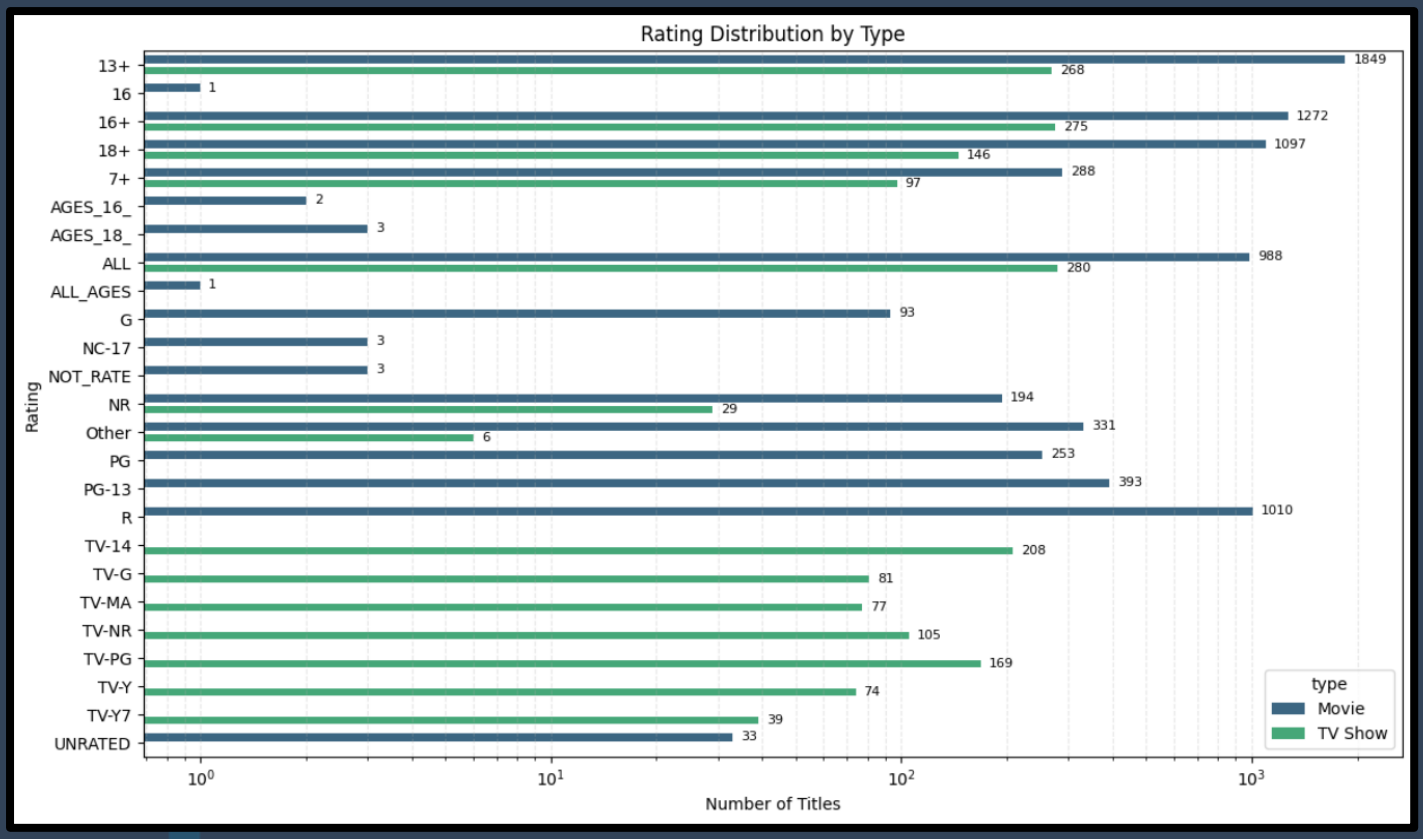
Data Visualization



Insights:

The data shows that the number of movies and shows on Amazon Prime Video stayed very small for a long time, but suddenly exploded after **2015**. Most of this new content is made for older audiences, with ratings like **13+**, **16+**, and **18+** making up the biggest chunks of the library today. While there are still shows for everyone (rated **ALL**), the amount of grown-up content has grown much faster than kids' movies. Overall, the platform has gone from having just a few hundred titles to over **1,300 new releases** in a single year by 2021.

Data Visualization



Insights:

The data shows that movies and TV shows for older audiences are the most popular on Amazon Prime Video. Ratings like **13+**, **16+**, and **18+** have the highest number of titles, with **13+ movies** leading the way at over **1,800** items. While there are plenty of movies for adults, there are also many shows for kids and families, though they aren't as common as the grown-up options. Interestingly, most of the "R" and "13+" rated content consists of movies rather than TV shows.

Conclusion

Key Findings

The Amazon Prime Video catalogue is a massive, movie-centric library that has undergone an explosive period of growth in recent years. While the collection includes titles dating back to 1920, over 75% of the content was released from 2007 onwards, with a record-breaking surge of new additions occurring in 2021. Geographically, the United States and India emerge as the primary content contributors, collectively accounting for over 580 titles and highlighting these regions as critical markets for the platform's global strategy.

User Engagement Patterns

Content strategy is heavily geared toward mature audiences and popular mainstream genres.

- **Genre Preference:** Drama, Action, and Comedy are the most dominant categories, with Drama alone leading the library with over 2,200 titles.
- **Maturity Ratings:** The platform prioritizes "general entertainment" for adults, as ratings like 13+, 16+, and 18+ make up the largest portion of the library, while specialized children's programming remains a smaller niche.
- **Viewing Habits:** Most users engage with standard feature-length films (averaging 90–110 minutes) and single-season TV shows, though a "long tail" of veteran franchises with up to 29 seasons provides depth for binge-watchers.

Conclusion

The analysis concludes that Amazon Prime Video has successfully transitioned from a standard content aggregator into a dominant global streaming hub by focusing on high-volume, modern releases. By maintaining a library that is 80.8% movies, the service differentiates itself from competitors that may lean more heavily into serialized TV content. The strategic seasonal release pattern, peaking in September, ensures that the platform remains relevant during high-engagement periods throughout the year.

Recommendations

To maintain its competitive edge, the platform should consider the following:

- **Diversify TV Content:** Increase the production or acquisition of multi-season TV series to improve long-term subscriber retention beyond the current "one-season" dominant trend.
- **Targeted Regional Expansion:** Leverage the success in India and the U.S. to build similar deep-content hubs in the "long tail" markets like the UK, Canada, and France.
- **Balance Duration Trends:** While movie lengths are trending shorter (dropping toward 85 minutes), maintaining a small selection of "epic" specialized content can serve as a prestige draw for serious cinema enthusiasts.

Enhance Personalization

Amazon should further leverage its data-centric infrastructure to refine its recommendation engine. By using X-Ray metadata and specific user viewing patterns, the platform can create a more "bespoke" interface that highlights niche genres like Animation and Romance to the specific users who enjoy them, ensuring these smaller categories are not overshadowed by the dominant Drama and Action titles.

Visit my Github for repository and code:
yashrajkohli5/Amazon-Prime-Video-EDA

Thanks For Reading