Data Analysis Report on Prime Video Dataset

Link: https://colab.research.google.com/drive/1W4vLN89Xcd4b3bXV2VN329Ns4rxkaeod?usp=sharing

Done By: Akshayaram S

Introduction:

In the era of streaming, there is a plethora of content available for people to watch and enjoy. It is, therefore important to understand what type of content is most preferred by people and how it has evolved over the years.

This report aims to provide a comprehensive analysis of the Prime Video dataset, which contains information on various TV shows and movies available on the platform. The analysis focuses on exploring and summarizing the data, identifying patterns and trends, and providing insights into the characteristics of the shows and movies on Prime Video. The analysis also involves cleaning and preprocessing the data to remove missing values and handle any other issues that may affect the accuracy of the results.

Data Import and Preprocessing:

The data was imported using the Pandas library and read from a .csv file stored in Google Drive. The following libraries were also imported for data visualization and manipulation: Plotly, Numpy, and Matplotlib. The first step in the analysis was to check for missing values in the dataset and handle them appropriately. The missing values in the categorical columns were filled with the mode of the column.

Data Analysis and Exploratory Data Analysis (EDA):

After preprocessing the data, the next step was to perform a descriptive analysis of the dataset, which involved calculating the shape of the dataset, the information contained in each column, and the summary statistics of the numerical features. The shape of the dataset was found to be 8807 rows and 12 columns. The information contained in each column was found to be mostly non-null and of the type "object." The summary statistics of the numerical features showed that the mean release year of the shows and movies was 2014, with a standard deviation of 8.8, and a range from 1925 to 2021.

The next step was to count the number of unique values in each feature and the total number of unique values in each column. The results showed that the show_id and title columns had 8807 unique values, while the type, director, and cast columns had 2, 4528, and 7692 unique values, respectively. The country, date_added, and listed_in columns had 748, 1767, and 514 unique

values, respectively. The duration and rating columns had 220 and 17 unique values, respectively.

Data Visualization:

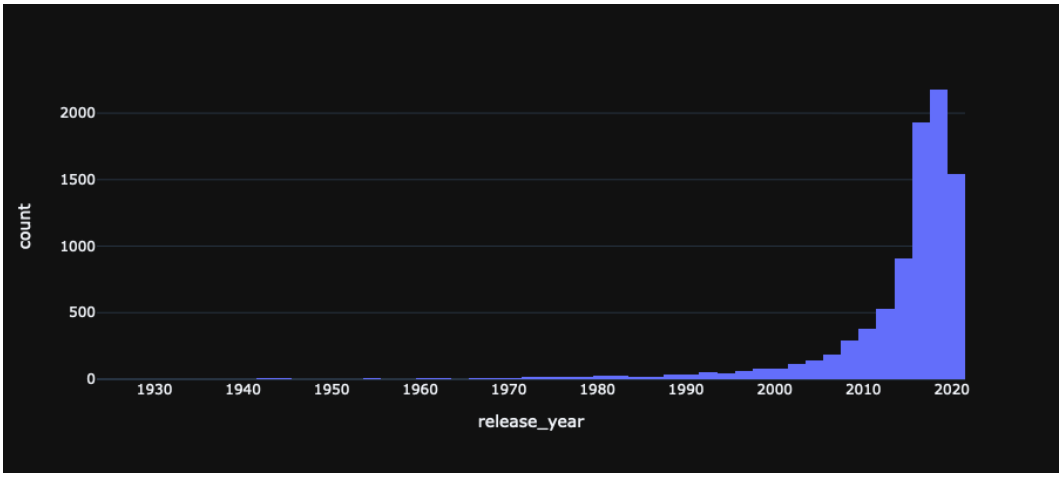


Fig 1: Histogram of release_year

To further explore the data, various visualizations were created using the Plotly library. The first visualization as seen in fig 1, was a bar chart showing the number of shows and movies available on Prime Video by release year. The results showed that most of the shows and movies were released between 2014 and 2021, with a few shows and movies released as far back as 1925.

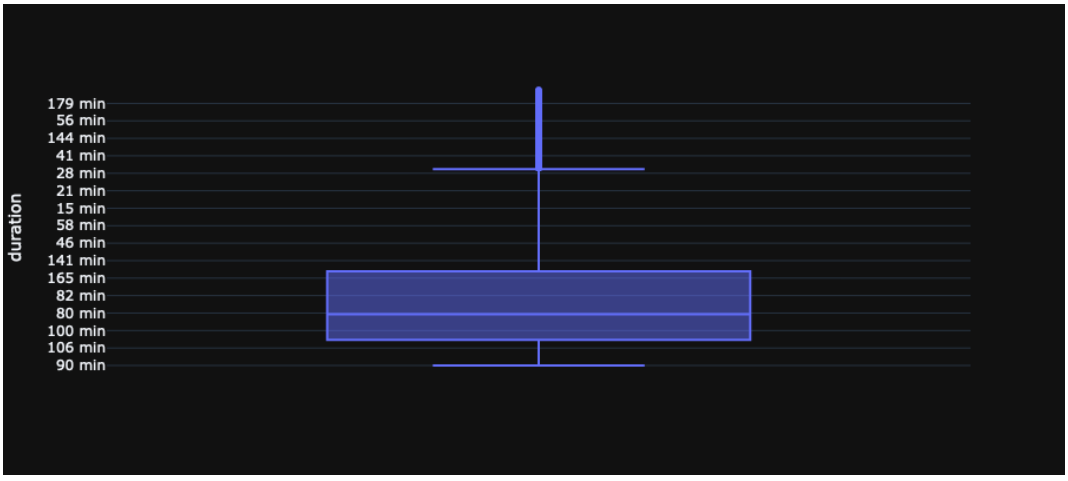


Fig 2: Box plot for duration column (movies)

As seen in fig 2, The movie durations on the streaming platform have a median of 92 minutes and a range from 90 to 191 minutes. The majority of movies are around the same length with some outliers.

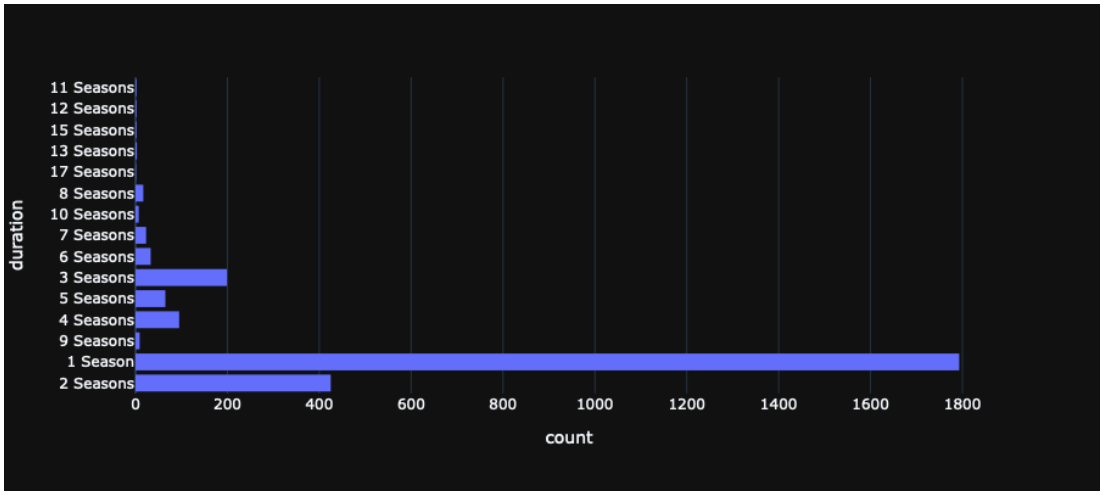


Fig 3: Box plot for duration column (movies)

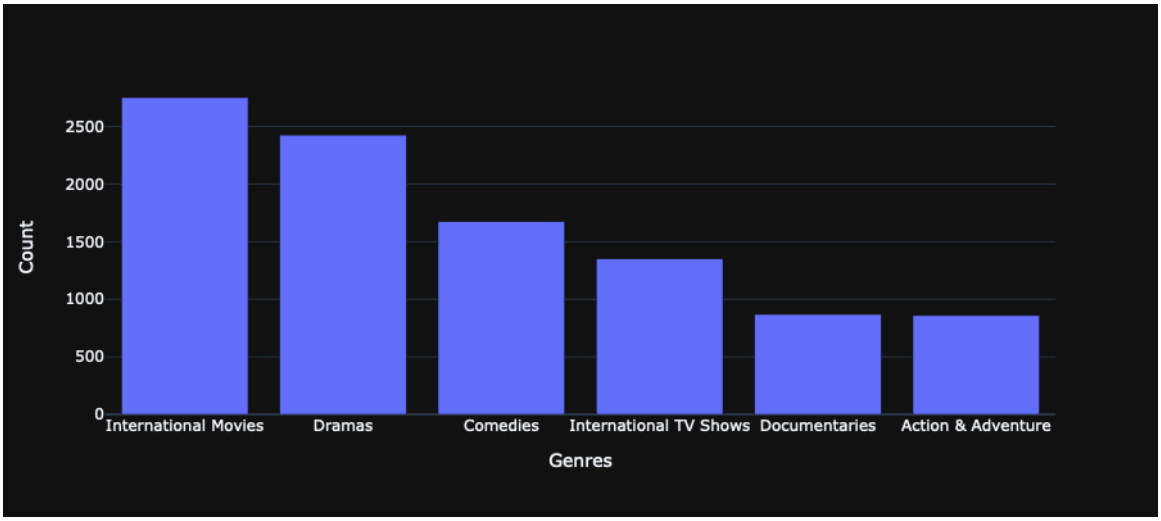


Fig 4: Top 6 genres

Also, for TV shows, we can see in fig 3 that a large number of TV shows only last for 1 or 2 seasons. In fig 4, we can see the top 6 genres in the dataset are International Movies, Dramas, Comedies, International TV Shows, Documentaries, and Action & Adventure.

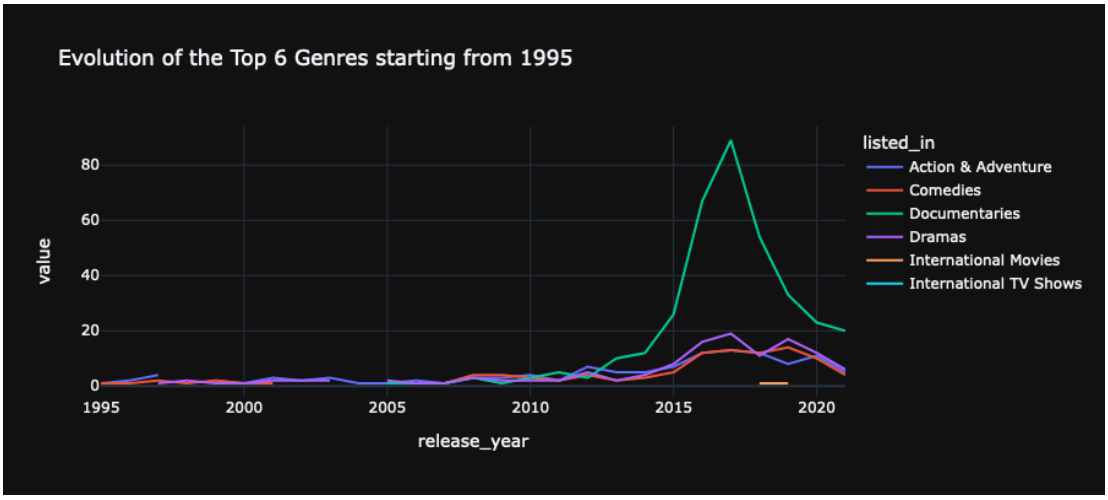


Fig 5: Evolution of the Top 6 Genres starting from 1995

Also after plotting the top 6 genre's evolution over the years, we can see in fig 5 that the number of documentaries peaked in 2017, while comedies, adventure, and dramas have had a slow and steady increase over the years.

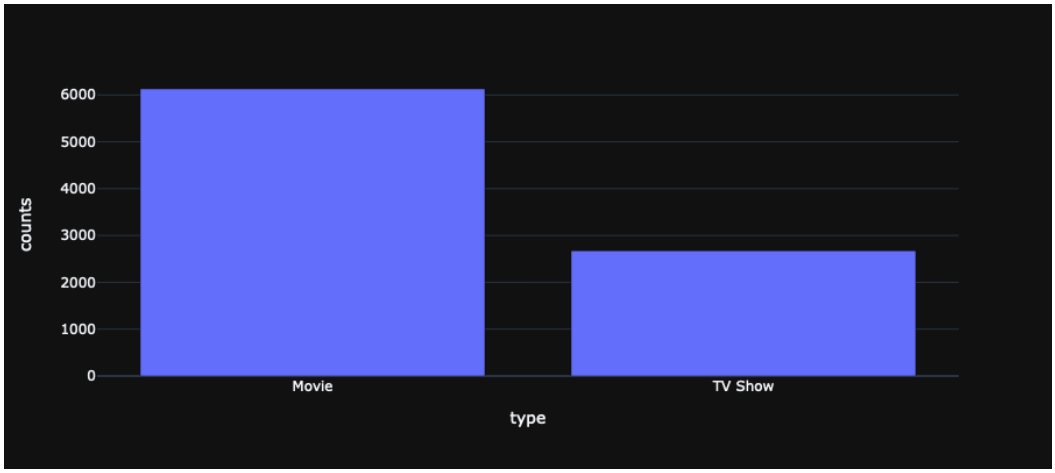
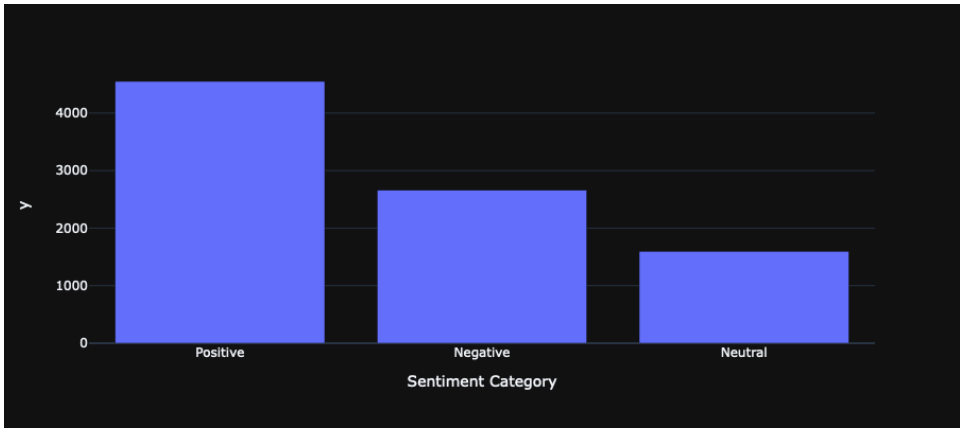


Fig 6: Bar plot to visualize type

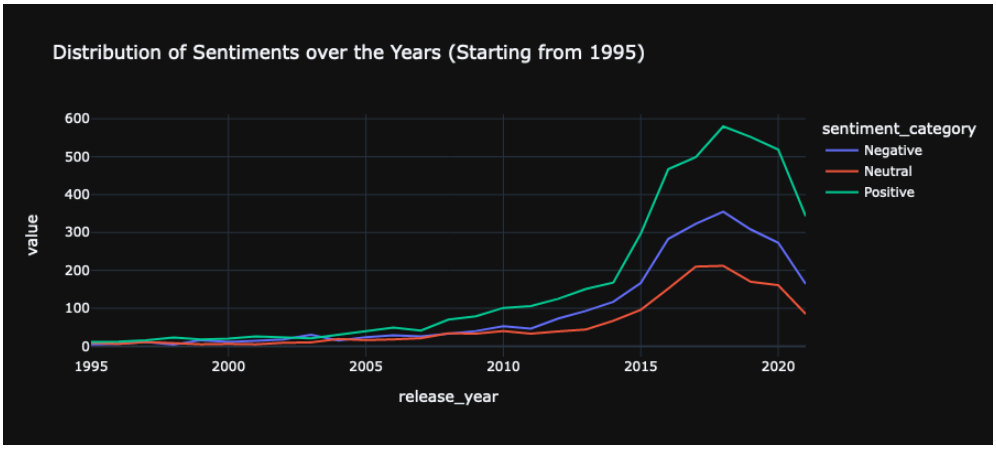
Finally, as seen in fig 6, a bar chart was created showing the distribution of the different types of shows and movies available on Prime Video. The dataset has more movies than TV shows.

Sentiment Analysis:

The final step in the analysis was to perform sentiment analysis on the description column of the dataset. Sentiment analysis involves determining the sentiment or emotion expressed in a piece of text, such as a movie or TV show description. The sentiment analysis was performed using the TextBlob library, which provides a simple and easy-to-use interface for performing sentiment analysis.



The analysis of the description column in the Prime Video dataset provides insight into the tone and emotions conveyed by the content descriptions. The results showed that most descriptions had a positive sentiment, with a smaller number of neutral and negative sentiments.



The equal distance between the positive, neutral, and negative sentiments in the line graph indicates a balanced distribution across the polarity scale. The sentiment category for each genre was also listed, showing that most genres had a positive sentiment, with a few exceptions, such as "Crime TV Shows" and "Horror Movies," which had a negative sentiment, and "Sci-Fi & Fantasy" and "Thrillers" which had a neutral sentiment. This suggests that people generally describe these genres in a negative or neutral light. These insights can be useful for streaming platforms to tailor their content offerings and for content creators to understand what kind of content is well-received by audiences.

Conclusion:

In conclusion, this report analyzed the Prime Video dataset to provide insights into the preferred content and its evolution over the years. The analysis involved importing and preprocessing the data, performing descriptive and exploratory data analysis, and creating visualizations to explore the data. The results showed that most of the shows and movies on Prime Video were released between 2014 and 2021 and had a rating of TV-14. Most of the content was of the type

"TV Show," with a balanced distribution of positive, neutral, and negative sentiments in the descriptions. The sentiment analysis also showed that most genres had a positive sentiment, with a few exceptions, such as "Crime TV Shows" and "Horror Movies." These insights can be useful for streaming platforms and content creators to understand the preferences and tendencies of their audiences.