**REPORT – TASK 1**

Amazon prime is a popular streaming service that offers a vast catalog of movies, TV shows, and original content. The data consists of contents added to Prime from 2008 to 2021. The oldest content is as old as 1925 and the newest as 2021.

**Dataset Description:**

The dataset contained attributes such as show id, type, title, director, country, cast, date added, release year, rating, duration, listed and description of the movies and TV shows available in amazon prime.

**On Handling Missing Values:**

On performing initial preprocessing of dataset, it consists of missing values in certain attributes such as director, cast, country, date added, rating and duration.

The rows which contained the missing values have been removed since the other ways of replacing missing values in an amazon prime dataset which contains the descriptive explanations mostly. So replacing them would give false narratives.

**Correlation Matrix:**

A correlation coefficient of 1 for the "release year" variable in the correlation matrix means that there is a perfect positive linear relationship between "release year" and itself. In other words, it indicates that as the "release year" increases, the value of "release year" also increases in a perfectly linear manner. This is not surprising, as "release year" is a continuous variable that represents the year in which a movie was released, and, naturally, it will have a perfect positive correlation.

Finally, **positive correlation – release year.**

**Visualizations**

Various visualizations are performed to understand the data more precisely as a part of EDA.

* Bar Chart
* Line Chart
* Histogram
* Word Count
* Box plot

A graph of a number of movies and tv shows

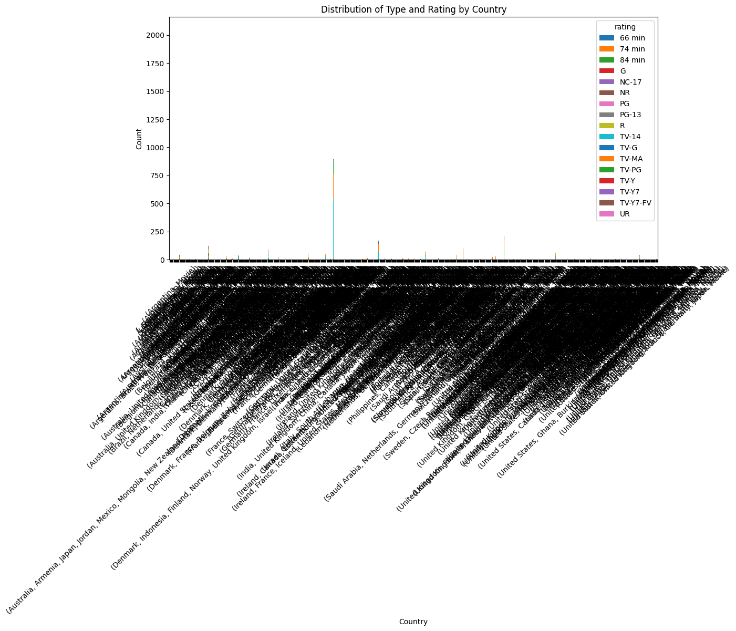
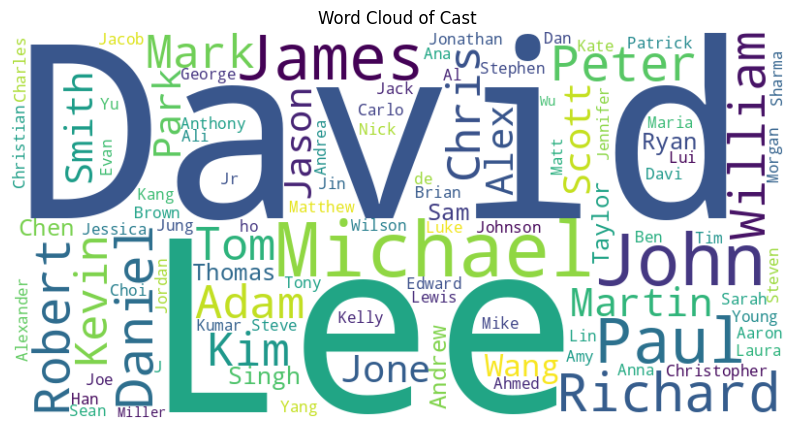
Description automatically generatedA graph showing a movie release

Description automatically generated

A graph of different colored bars

Description automatically generatedA diagram of a movie duration

Description automatically generated



**INSIGHTS:**

* On comparison between movies and TV shows, the count of number of movies released is greater than TV shows.
* In the line chart, the number of movies and TV shows released during the year 1940 has been consistently increasing and has fluctuation in both movies and TV shows till 1980 and culminating in a decrease in both categories by 2020.
* In 1750, the duration of the movies was quite longer than usual but then year by year movies maintained same duration mostly ranging from approximately 90 to 120 minutes.
* In word cloud of cast, the famous actors have been depicted who have fame throughout the year 1940 to 2020. The names are David, Lee, Michael, John, etc.,
* The Distribution, rating and type has all countries which shows that people are interested in both TV shows and movies with support for rating whether it is good or bad.