The Data Analysis Report: Amazon Prime Data

Task given : Amazon Prime dataset was given to perform exploratory data analysis and visualize the findings to draw conclusions.

Critical Analysis Plan Components

Dataset Description :

The dataset consisted of 12 columns and 8807 rows with a total of 108654 data points. It consisted the information about vast catalog of movies broadly classified in two categories TV Shows and Movies. Each of them has a description, listings, released year, cast, director and duration. The dataset also has the rating given according to who can watch the movies and shows.

Purpose :

Performing exploratory data analysis on the given data enables us to identify the most watched category of the movies that helps us to identify the interests of people across the globe

Dataset Cleaning :

The data consisted of many missing values for the director, country and cast. The missing values in the country column was replaced by United States, being the most common country. The missing values in other columns were either filled with "No data" or dropped. The values of Rating column is abbreviated for for better understanding.

Critical Analysis Plan Components

Exploratory data analysis :

- **Content in Amazon Prime**
  Total content in amazon prime is 8794 out of which 70% are Movies and 30% are TV shows.

- **Content contributed country-wise**
  United State has the most content in the platform, which almost 49% of the total available content, followed by India and UK.

  Also the available content in India is highly skewed towards the movies compared to other countries whereas the South Korea has the highest percent of TV shows

- **Yearly trend of movies & TV shows**
  There was no much content uploaded till the year 2014. The most content was uploaded between the years 2018 and 2019 and there was a gradual decrease in the content being uploaded on the platform. This might be due to the effects of COVID19. Also Disney plus was launched in 2019, which could be a reason for the decrease in contents.

- **Contents based on Rating**
  Most of the movies and TV shows that are available in the amazon prime is suitable for adults and teens to watch. The content for kids is very low.

- **Most preferred genre**
  The listed_in column in the dataset described the appearance of each content is relevant categories. To identify the genre based preference of people, the listed_in column was striped to find that international movies, dramas and comedies are the most liked genres that people watch in amazon prime.

- **Relation between genres**
  Dramas and documentaries are negatively correlated, hence there is no similarity between them. There are many dramas for independent and international films as there is good correlation observed between them.

- **Relation between duration and release years**
  The average duration (mean) for the movies in the amazon prime is 100 minutes. It is observed the the duration of the movies is normally distributed. The interquartile range in the box plot is fairly small with many outliers.
  The outliers are dispersed from the scatter plot denoting the movies released between 1960s and 1980s. Hence there is a correlation between the duration and release years.

- **Duration of TV Shows**
  It is observed that most of the TV shows in amazon prime has one season, followed by 2 or 3 seasons.

- **Most used words in title**
  Love, girl, life, christmas, world are the most frequently used words in the titles of the movie/TV shows.

- **Directors provided content**
  The words David, Michael, John are found to appear frequently in the data set with respect to directors. The movie directors with most content in India are David Dhawan, Anurag Kashyap, Dibakar.

- **Cast provided content**
  The words David, Paul, Lee, Michael are found to appear frequently in the data set with respect to actors. Anupam kher, Shah Rukh Khan, Naseeruddin Shah, Akshay kumar are the actors who provided more content from India.

KEY FINDINGS

- If children are using amazon prime for more time, they are highly likely to watch restricted content.
- Although United states contributes the most contents to amazon prime, South Korea is the major source of TV shows and India being the major source of movies.
- India has contributed a lot of movies provided that most content is from the Bollywood industry.
- Movies are often listed in 3 genres and TV Shows in 2 genres.
- The most uploaded content was between the years 2018-2019.
- More than 90% of the films in amazon prime are approximately 100 mins or more than that.
- People are more interested in international films, dramas and comedies.

CONCLUSION

This report summarizes the most preferred genres, relation between various fields like duration and release year, insights about the actors & directors, each countries contribution to the content, etc.. These results will be beneficial in improving both sales and customer satisfaction.