**Task 1 Report**

**Amazon Prime EDA, cleaning and visualization**

**Introduction**

This report presents an analysis of streaming content data obtained from the dataset "prime.csv." Our objective is to understand and visualize various aspects of this dataset, including content genres, content growth over the years, content ratings, duration, and more.

**Data Summary and Cleaning:**

* The analysis begins with a data summary and cleaning process to ensure the dataset is suitable for analysis. We load the data and observe its structure:
* The dataset consists of various columns, including "type," "title," "director," "cast," "country," "date\_added," "release\_year," "rating," "duration," "listed\_in," and "description."
* We perform a basic data summary to view the first few rows, check data types, and identify any missing values.
* Additionally, we calculate summary statistics for numerical variables to gain an understanding of the data's distribution.

**Data Insights and Visualizations**

The analysis includes several data insights and visualizations to provide a comprehensive view of the streaming content dataset.

* Most Popular Genres

We investigate the most popular genres by counting the number of each content type (Movies or TV Shows). The top 10 genres are visualized in a bar chart, allowing us to identify the genres that dominate the platform.

* Top Genres Bar Chart
* Content Growth Over the Years
* We explore the growth of content over the years by examining the distribution of release years. A line plot reveals how content production has evolved throughout time.
* Content Growth Line Plot
* Relationship Between Content Duration and Ratings
* To understand the relationship between content duration and IMDb ratings, we create a scatterplot. This visualization showcases whether there is a correlation between these two factors Duration vs. Rating Scatterplot
* Count of Movies vs. TV Shows
* We differentiate the count of movies and TV shows in the dataset with a countplot. This visualization clearly illustrates the balance between the two content types.
* Count of Movies vs. TV Shows
* Ratings Distribution
* We delve into the distribution of content ratings and visualize it using a countplot. The ratings distribution provides insights into user preferences.
* Ratings Distribution Countplot
* Duration Distribution
* The duration distribution is visualized through a histogram. This provides an overview of content duration patterns within the dataset.
* Duration Distribution Histogram
* Top 10 Countries with the Most Content
* We identify and visualize the top 10 countries with the most content. A barplot highlights the countries that contribute the most to the dataset.
* Top 10 Countries Barplot
* Distribution of Content by Type and Rating
* Lastly, we explore the distribution of content by type and rating using a countplot with hue. This visualization showcases the variety of content in terms of ratings.
* Content Distribution by Type and Rating

**Conclusion**

* The analysis and visualizations presented in this report provide valuable insights into the streaming content dataset. It is evident that content growth has been substantial over the years, with a diverse range of genres and ratings. The relationship between content duration and ratings is also explored, offering a deeper understanding of user preferences.
* These process serves as a foundation for further analysis and decision-making in the context of streaming content.