Task 1:

Amazon prime is a popular streaming service that offers a vast catalog of movies, TV shows, and original contents. The data consist of contents added to Prime from 2008 to 2021. The oldest content is as old as 1925 and the newest as 2021 (see attachment prime.csv). The tasks that you need to perform are

- Undertake an exploratory data analysis and produce a 2 page report on your findings/insights
- Clean the data and produce good visualisation for story telling
- Submit your code and report via Github repo

Result:

The libraries which are needed for the EDA process are imported and the dataset also has been read into a variable 'prime'. The preliminary process is conducted to understand the dataset better. Information such as the type of the variables are revealed. It helps deciding whether the variable is categorical or numerical. For, the amazon prime dataset all the variables except release_year is of type object, while release_year is of type integer.

Null value treatment is done as the next step after the preliminary exploration of the dataset. The variables such as director, cast, country, date_added, rating, duration. The rows with the null values in the variables date_added, rating and duration are dropped. The null values in the variables director, cast and country as too large to be dropped. So, a constant value is used to replace the null values in the dataset.

The Univariate Analysis is performed on the dataset.

- Movie Type – This attribute contains only two unique values. They are Movie and TV show. Plotting the values in a bar plot revealed that the number of movies (69.7%) in the amazon prime platform exceeds the number of tv shows (30.3%).
- Release Year – Plotting this attribute in a bar plot revealed that most of the content were released in the years 2018, 2017, 2019, 2020 and 2016.
- Rating – The bar plot of the attribute revealed the most popular rating of the content in the platform. They are TV-MA, TV-14, TV-PG, R, TV-Y7.
- Director – The popular directors are Rajiv Chilaka, Raul Campos, Jan Suter, Suhas Kadav and Marcus Raboy.
- Cast – There are multiple values in a single row of this attribute. The multiple values are divided into a separate row and then they are plotted. The top 5 actors who have the most content in this platform are Anupam Kher, Shah Rukh Khan, Julie Tejwani, Naseeruddin Shah, Takahiro Sakurai.
- Date_added – The values contain both the month, date and the year. The date and year are separated into different columns in a new variable. From the date attribute we can infer that most of the content were added at the beginning of the month. And the year attribute reveals that the years 2019,2020,2018,2021 and 2017 are the ones when the content were added the most.
- Country – Most of the content were release from the countries such as United States, India, United Kingdom, Japan and South Korea.
- Duration – The popular tv shows are usually had 1,2 or 3 seasons. And the popular movies had duration from 90 – 100 minutes.

- Listed_in – There are multiple values in a single row of this attribute. The values are divided into separate rows. The top 5 genres are International Movies, Dramas, Comedies, International TV Shows, Documentaries.
- Description – The values are made into a word cloud. Some of the popular words which are used to describe the content available in the platform are Young, Documentary, Teen, Friends, Comedian etc. From this we can infer that the movies which are watched are focused on young adult, documentary series, comedy content etc.

Bivariate Analysis

- Popular content by Rating – TV-MA rating has more movies than TV shows. TV-14 rating has also more movies than TV shows. TV-PG has more movie content than tv shows. TV- Y has more TV shows than movies. PG, PG-13 has only movie content. The content type of R rating is mostly movies.
- Popular content by Country – Data containing only the popular countries are selected and a count plot is plotted with the x axis as country and its hue as the movie type. United States and India has more movies than tv shows. United Kingdom has almost equal amount of movie and tv shows and South Korea has more TV shows than movies.
- Ratings of Popular Genres – Documentaries mostly has the rating of TV -MA. Comedies and Dramas mostly has the rating R. International Movies and TV Shows mostly has the rating TV-PG.
- Content by Release Year – Content has been largely released in the years 2015 – 2020 and all those years has movies released the most.
- Popular Genres by Countries – The popular genre in United States is Documentaries followed by comedies and then Dramas. In UK, the popular genre is Documentaries, followed by Comedies, Dramas and International Movies. In India, the popular genres are Dramas and International TV Shows.