6.1 Sourcing open data.
Tanvi Gadage.

Data Analysis on the Netflix Datasets

Data Source :

Netflix is one of the most popular media and video streaming platforms. They have over 8k+ movies or tv shows available on their platform, they have over 282 million Subscribers globally. This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration and more.

Data collection :

This dataset consists of TV shows and movies available on Netflix as of 2021. The dataset is collected from kaggle.
Netflix Movies and TV Shows | Kaggle

We got the dataset from Kaggle and we are going to utilize data that to understand the trend of movies and TV shows released on the platform.
This dataset consists of CSV file contain(netflix_titles.csv) .
We will utilize the following columns to understand what movies and TV shows were released in specific year, what genres they were, date when they were released and the rating the audience gave and so on.
From the column names, we could observe that there are twelve columns: show_id, type, title, director,
cast, country, date_added, release_year, rating, duration, listed_in, description.
There are 8807 rows of data.

Objectives of the Project :

- To understand what content is available in different countries.
- No of movies/shows released based on the year.
- The most rated content based on the genre.
- Network analysis of actors/directors.
- Is Netflix has increasingly focusing on TV shows rather than movies in recent years.
- The most observed rating by category in TV shows and movies.
- How many movies or shows it has released year differ from its year added.

SCOPE OF THE PROJECT:

- Understanding what content is available in different countries
- Identifying similar content by matching text-based features
- Is Netflix has increasingly focusing on TV rather than movies in recent years ?

Netflix in the past 5-10 years has captured a large populate of viewers. With more viewers, there most likely an increase of show variety. The advancement in technology and streaming platforms have made it possible for us to watch movies from all over the world. Due to the different taste and preferences and according to current situation the number of movies released in a year may vary. So, Let’s see the frequency of the TV shows / movies releasing in a year.

Data Cleaning and Preprocessing

- Before diving into the analysis, we need to clean and preprocess the data. This includes handling missing values, converting data types, and extracting useful information from existing columns.

```
show_id          0
type             0
title            0
director         2634
cast             825
country          831
date_added       10
release_year     0
rating           4
duration         3
listed_in        0
description      0
```

There are some missing values in column director, cast country, date_added, rating . this values filled with 0.

- No duplicates were found in dataset.