

# In-Lab assignment: Data Preprocessing and visualization

In this assignment, we shall explore a sample dataset using pandas and matplotlib.

**About this Dataset:** *Netflix is one of the most popular media and video streaming platforms. They have over 8000 movies and tv shows available on their platform, and as of mid-2021, they have over 200M 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, etc.*

## **Information about database:**

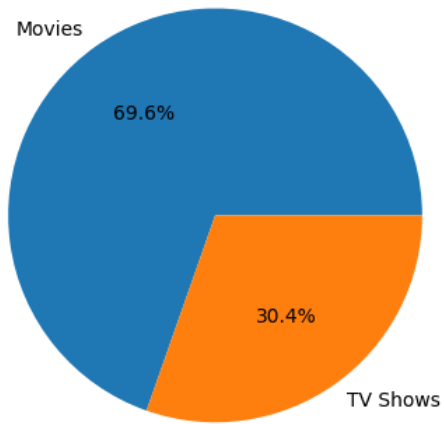
- **Show\_id:** Unique ID for every Movie / Tv Show
- **Type:** Identifier - A Movie or TV Show
- **Title:** Title of the Movie / Tv Show
- **Director:** Director of the Movie
- **Cast:** Actors involved in the movie/show
- **Country:** Country where the movie/show was produced
- **Date\_added:** Date it was added on Netflix
- **Release\_year:** Actual Release year of the movie/show
- **Rating:** TV Rating of the movie/show
- **Duration:** Total Duration - in minutes or number of seasons
- **Listed\_in:** Genre
- **Description:** The summary description

Q1. Using a pie chart, find the percentage share of TV shows and Movies.

Instructions:- Upload pie chart (correctly labeled and the respective percentage values)

Ans:    Movies – 69.6%  
          TV Shows – 30.4%

Percentage: Movies v/s TV Shows



Q2. Plot a bar chart showing top 5 regions having the highest number of shows (both TV Shows and Movies)

Instructions:

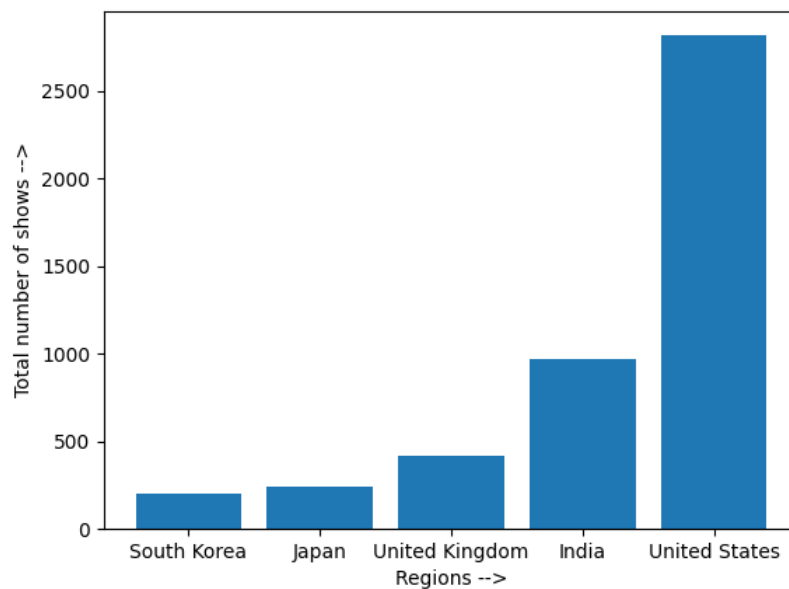
x-axis: Regions.

y-axis: Total number of shows.

The bar chart should be sorted in ascending order and should contain only top 5 countries with proper labels. Upload your bar chart.

Ans:

Top 5 regions having the highest number of shows



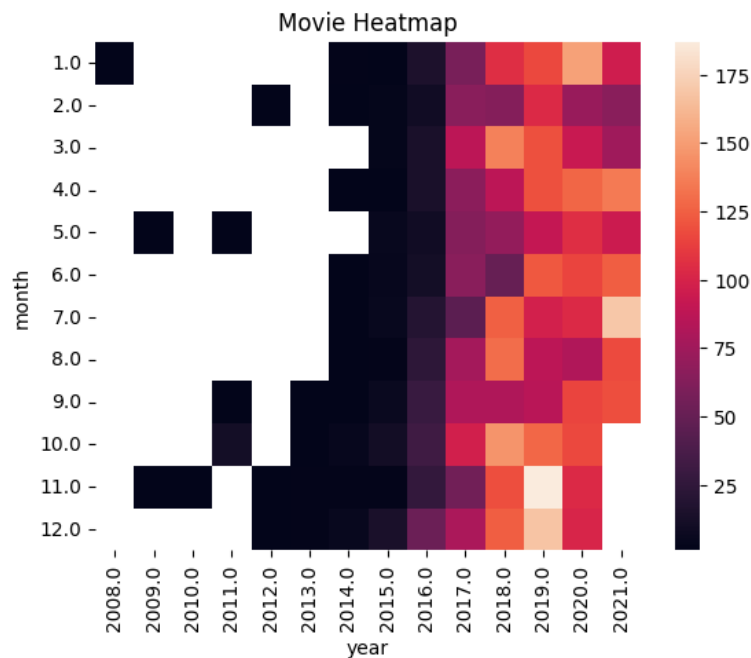
Q3. Using a heatmap answer the following questions.

Do for both Movies and TV shows separately:

- After which year do the shows start appearing on Netflix?
- In 2021, in which month were most shows added?

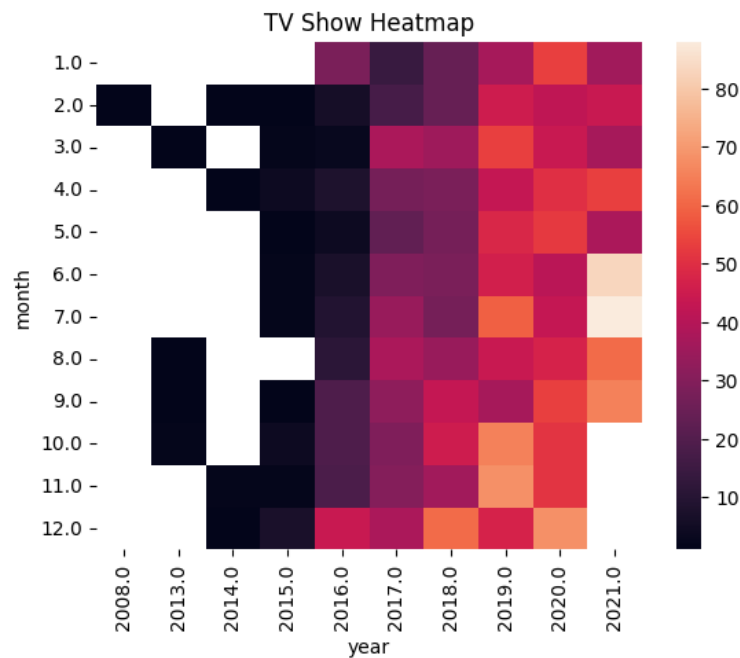
Note:- Make two separate heatmaps, one for TV shows and one for Movies and answer the two questions in the context of both the types  
Heatmap should have the y-axis as months and the x-axis as years.

Ans: Movies



- The movies start appearing in 2008. However, there is a significant increase after 2015, i.e., 2016. There is an increase of 450% in the number of movies.
- In 2021, most movies were added in the month of July (no. 7).

## TV Shows



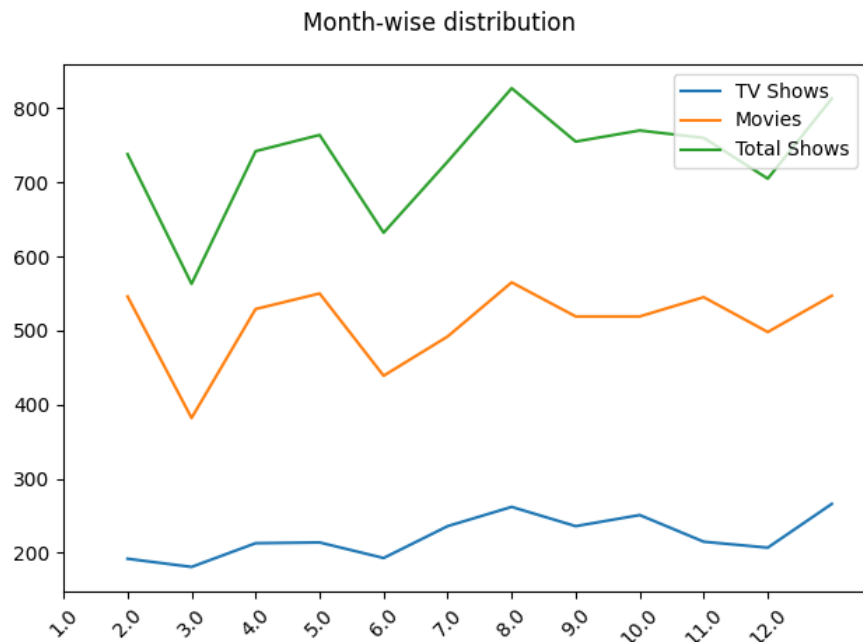
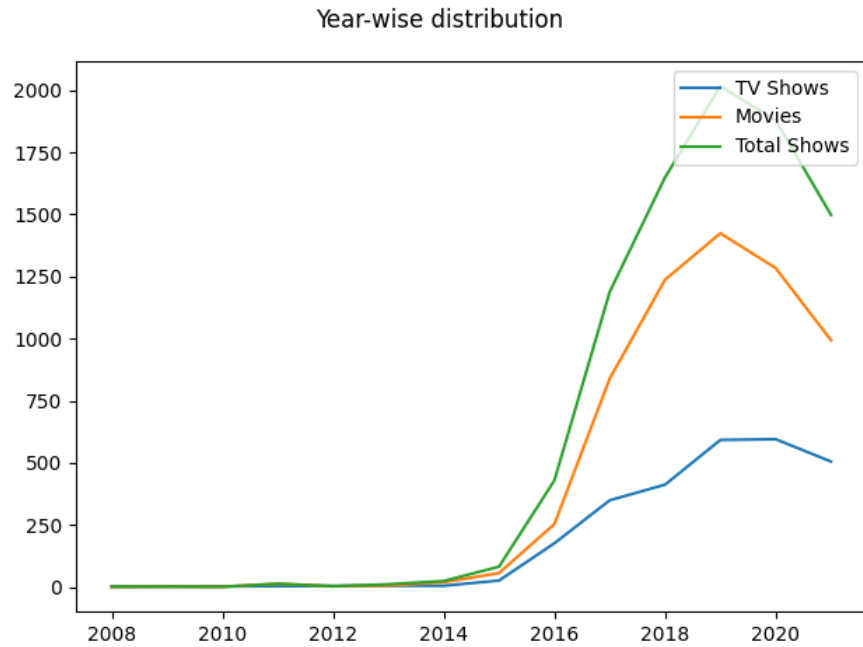
- The TV Shows start appearing in 2008, however there is a brief silence for 5 years until 2013. There is a significant increase in the number of shows in the year 2016.
- In 2021, most TV Shows were added in the month of July (no. 7).

Q4. Using a line graph, tell which year and which month has the highest:

- TV Shows
- Movies
- Total shows (TV shows + Movies)

Note:- 2 separate plots 1 for year-wise distribution and one for monthly distribution (three different data in the same plot:- Movies, TV shows and Total; each with a different color)

Ans:



Q5. Does Netflix have more focus on TV Shows than movies in recent years? Backup your answer using some plot/graph and upload it.

Ans: No, Netflix does not have more focus on TV Shows than movies in the recent years. This can be inferred from the graph 'Year-wise distribution'. Infact, after 2016, the number of movies shot up, and has been much more than the number of TV Shows.

