`Netflix Dataset

Goal of choosing data ?

- the data was fun to explore it and read about the features of it especial it has a good analysis skills to explore them all which are useful for understanding how different product or service categories are perceived. These variables can reveal interesting groupings when clustered

Data discribtion & features :

- Netflix is one of the most popular streaming platforms globally, offering a diverse range of content across genres and ratings. Analyzing this dataset allows us to explore real-world patterns in The dataset includes meaningful categorical variables like Type (e.g., Drama, Comedy) and Rating (e.g., G, PG, R, TV-MA), which are ideal for clustering tasks. These features help us understand how content is structured for different audiences.entertainment media. Since there’s no "target" variable to predict, this dataset is well-suited for unsupervised learning techniques like K-Means and Agglomerative Clustering. It allows for exploration without needing labeled outcomes

- **There are 11 columns in the dataset :**
  - Show_Id: Unique identifier for each show
  - Category: Whether it's a Movie or TV Show
  - Title: Name of the content
  - Director: Director’s name
  - Cast: Main cast
  - Country: Country of production
  - Release_Date: When it was released
  - Rating: Age rating (e.g., PG, R, TV-MA)
  - Duration: Length or number of seasons
  - Type: Genre or type (e.g., Dramas, Action, Horror)
  - Description: Summary of the content

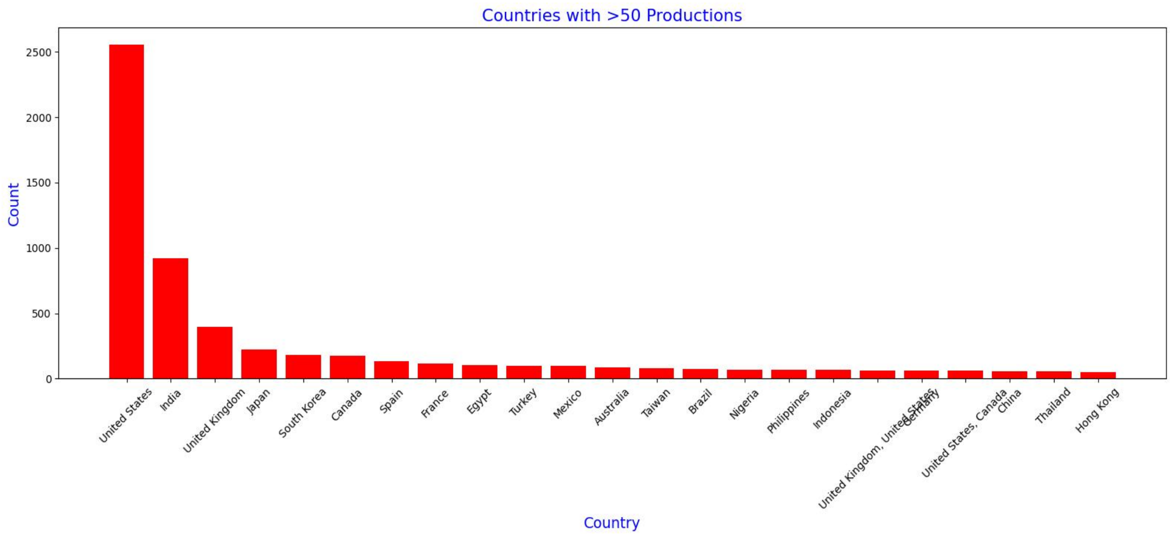
Cleanning phase :

- in cleanning phase we used :

- isnull() : to get know if there any missing values
- fillna() : to fill the missing values with “unknown”
- duplicated() : to checj if there any duplicated value
- unique() : shows the unique values

Visualization phase :

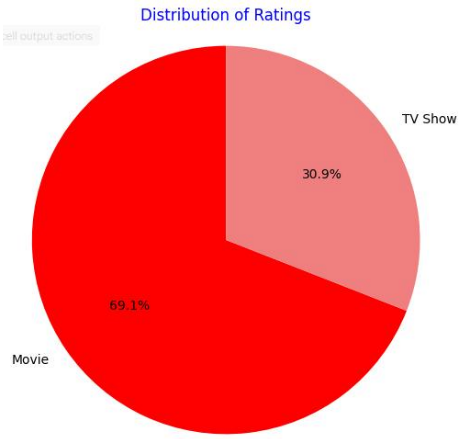
1-

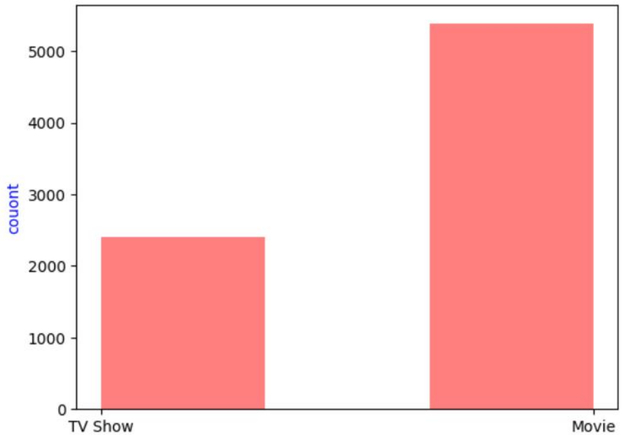


• this graph shows each county count which is bigger than 50 production of movies and TV shows around the world.

• The U.S. and India are the largest content creators, which reflects their strong media industries (Hollywood and Bollywood). Other countries contribute significantly less but still represent important cultural production centers.

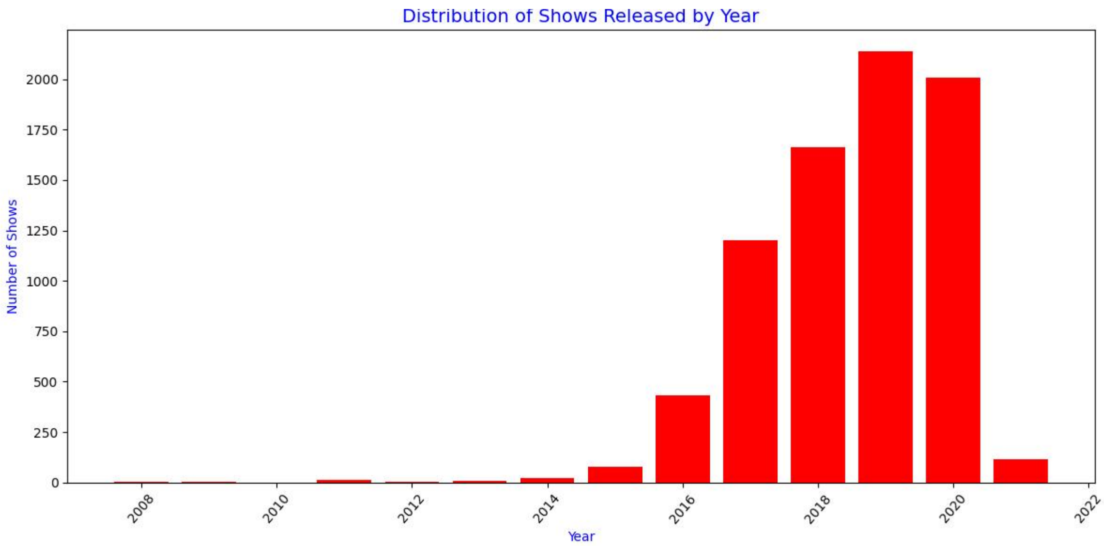
2-





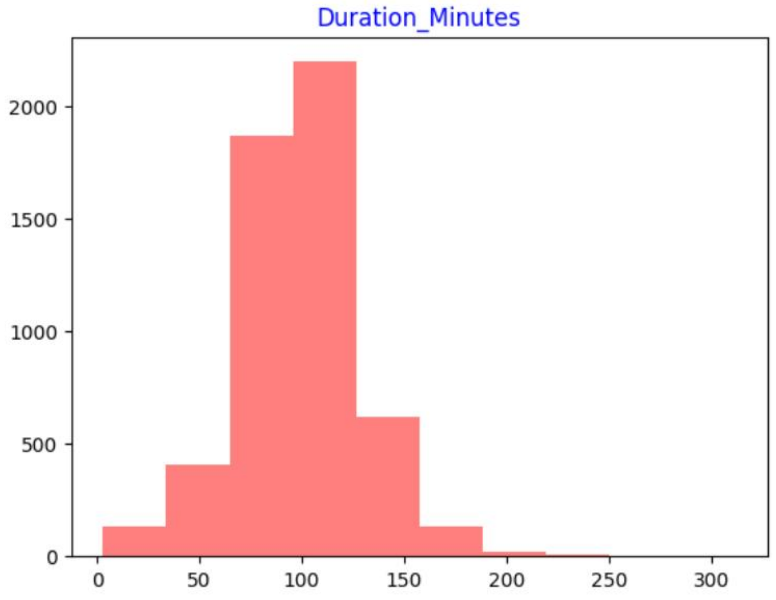
- **pie chart** shows the percentage of how many movies and TV shows are watched
- **histogram** shows more details about how many movies or TV shows are watched
- movies are more to watchable than TV shows.

3-



- this graph shows the number of TV shows and movies are watched over years and the count of it
- the industry of making show began to be more highly productive after the year 2016 and the range between 2018 and 2020 are more highly productive

4-



• The data seems to be right-skewed (positively skewed), meaning most durations are on the lower end, with fewer longer durations.

• The most frequent durations lie between 80 to 110 minutes, peaking around 100 minutes.

• There's a sharp drop in frequency after 120 minutes, and very few instances above 200 minutes

Modeling phase :

Why K-Means Clustering?

• It's a simple and efficient algorithm for partitioning data into a predefined number of clusters (k).

• It works best when you expect distinct, non-overlapping groups.

• It's fast and scales well for larger datasets.

How it works:

• Converts Type and Rating into numeric values (using Label Encoding).

• Groups content by minimizing the distance between points and their cluster centers (centroids).

Why Agglomerative Clustering?

• It’s a type of hierarchical clustering that builds a tree of clusters.

• It doesn’t require you to define the number of clusters beforehand (although you can cut the tree at any point).

• It's useful when the relationships between clusters are hierarchical or nested, such as "TV Shows for Teens" inside a larger "TV Shows" cluster.

How it works:

• Starts with each content item as its own cluster.

• Merges the closest pairs of clusters step-by-step based on similarity (distance) until only one cluster remains or a stopping point is defined.

Result :

The analysis focuses on identifying trends in **viewer preferences, content popularity, and the influence of attributes** (genre, release year, content type, and ratings) on engagement. Key findings include:

> **Content Distribution:** Movies dominate the catalog, shaping acquisition and production decisions.
>
> **Trends Over Time:** A significant rise in content additions, especially originals, highlights Netflix’s strategic shift.
>
> **Viewer Preferences:** Genres and ratings vary widely, strongly influencing engagement levels.
>
> **Feature Relationships:** Limited correlations in categorical data; however, release year and duration show meaningful patterns.
>
> **Data Preparation:** Encoding and feature extraction (e.g., first genre) were vital for effective analysis and modeling.

Purpose: These insights help optimize content strategy, enhance user experience, and improve recommendation systems