**Netflix Data Cleaning and Insights Generation**

* **SAMARTH JAIN**

The scope of this project was to first clean a Raw Netflix File and generate insights out of it.

So, I started with loading the data using Pandas and cleaning the data.

**Data Cleaning Steps :**

1. **Data Loading**: Used pandas to read netflix\_titles.csv using pd.read\_csv command.
2. **Data Analyzing**: Analyzed the sample rows in the data, number of rows and columns , data types of the every column values in each column using df.shape, df.info() and df.describe() methods.
3. **Duplicate values removal**: Checked for duplicate rows if any and then dropped them using the drop\_duplicates method.
4. **Handled Null values**: Identified the missing and null values in all columns and replaced blanks in director, cast, country *columns* with ‘Unknown’ using fill.na() method and dropped date\_added, and rating columns.
5. **Whitespaces removal**: Removed extra trailing and ending spaces from the text column values to improve accuracy and efficiency using str.strip method.
6. **Text values Capitalization**: Capitalized the intial letter of the text values for better readability using the str.title method.
7. **Date format correction**: Corrected the date format in the date\_added column using the pd.to\_datetime method
8. **Exported the Cleaned Data**: Exported the cleaned Netflix file to /netflix\_titles\_cleaned\_file.csv using the .to\_csv method.

**Insights & Metrics from the Cleaned Data:**

* Identified the most **Common Ratings** from the data: Categories like TV-MA and TV-14 appeared most frequently in the whole Dataset.
* Identified the **Countries with Most Content**: United States dominated the list, India on the second number, followed by United Kingdom and Japan.
* Identify **the directors with the highest number** of Content: TV shows and Movies: Rajiv Chilaka dominated the list, followed by Raúl Campos and Jan Suter with second highest number of Content.

**Regional Popularity -**

* Europe: Comedy and International Movies topped the charts.
* Asia: Drama and Comedy are amongst the most watched genres.
* North America: Documentaries and Stand-Up Comedy are amongst the most watched genres.

**Genre Analysis -**

* Drama and Comedy are the most popular genres followed by international movies.

**Release Year Trends**

* The Highest Number of Content rise took place between the years 2017–2020.
* A Decline in Total number of Content in 2021.

**Country Contributions**

* United States contributed the most Content.
* India is ranked second in terms of Content Contribution.
* UK, Japan, and Canada are also major contributors after United States and India.

**Exploratory Data Analysis (EDA):**

The goal of exploratory data analysis was to identify patterns, trends and outliers from the Netflix cleaned dataset. Importing libraries like Matplotlib , seaborn and plotly to creating visualisations or plotting various charts like : bar charts, line chart, etc.

So, I started with importing matplotlib library as matplotlib.pyplot as plt

and analysed these things:

**Content Growth Over Time-:**

* We need to identify and analyze how the Netflix content (movies or tv shows) has grown over time (year over year). So, I extracted the year column from date\_added column and then grouped the whole content year wise into a separate variable - content\_growth using group by function.
* Then , I plotted a line bar and bar chart to analyze the growth of content over time with ‘Year’ on X axis and ‘No of titles per year’ on Y axis.

**Distribution Analysis:**

1. **Genre Distribution Analysis:** Identify the Top 10 genres in the Dataset.

* Dataset have Genres categorized as (‘International movies’,’Drama’,’Comedy’,etc). Used the .explode() method to convert string or list values into rows with .str.strip() method to generate a genre\_counts column.
* Plotted a horizontal bar chart with the ‘Genres’ column on the Y axis and ‘genre\_counts’ column on the X axis using the plt.barh method.

1. **Rating Distribution Analysis:** Identify the Top Rating Categories.

* Dataset have Ratings categorized as (‘TV-MA’,’TV-14’,’TV-PG’,’PG-13’,etc). Used the .explode() method to convert string or list values into rows with .str.strip() method to generate a ratings counts column.
* Plotted a vertical bar chart with the ‘Ratings column’ on the X axis and ‘ratings\_counts’ column on the Y axis using the plt.bar method.

1. **Content Type Distribution Analysis:** Identify the Content Distribution in the Dataset out of Movies and TV Shows.

* Dataset have Content type categorized as (‘Movies’ and ‘TV Shows’). Generated a content type column and found out the count in each type as stored as type\_counts column.
* Plotted a vertical bar chart with the ‘Content Type’ column on the X axis and ‘type\_counts’ column on the Y axis using the plt.bar method.

**Country-Level Analysis:**

* To identify that which countries produced most content in terms of Movies or TV Shows and contributed the most in terms of content over the whole period.
* So, I generated a ‘country\_counts’ column to give the number of content(Movies or TV Shows) produced by that country.
* And then plotted a bar chart with Countries column on the X axis and the ‘country\_counts’ column on the Y axis.