**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.