DS – Netflix Data Cleaning & Aggregation

The Netflix dataset, consisting of movies and TV shows, presents a rich source of information for data analysis. However, to extract meaningful insights, the dataset must first undergo thorough cleaning and transformation. The raw data contains inconsistencies, missing values, and unstructured information, making it unsuitable for immediate analysis. This project focuses on transforming the dataset into a clean, structured format that is easy to work with for further analysis, reporting, and dashboard creation.

**Objective:**

The primary objective of this project is to apply data cleaning and transformation techniques using Python libraries such as pandas and NumPy. By executing these tasks, the dataset will be prepared for advanced analytical processes and visual representation. The final output will be a refined dataset, ready for reporting, generating actionable insights, and creating dashboards that showcase trends and patterns within the Netflix content.

**Problem Statement:**

The current state of the Netflix dataset, with its missing values, redundant columns, and unstructured data, makes it challenging to perform in-depth analysis. Without proper data cleaning and transformation, generating accurate insights and building a reliable dashboard is not feasible. Therefore, this project aims to address these issues by cleaning and structuring the dataset, ensuring that it is ready for in-depth analysis and reporting.

Download dataset: [Netflix Show Open Data (kaggle.com)](https://www.kaggle.com/datasets/shivamb/netflix-shows)

Clone notebook-script from GitHub (Use the following on your command line):

git clone <https://github.com/nbaloyi4/ds-assgnment-repo.git>

# References:

1. Finding text patterns in Pandas with regular expressions : <https://www.youtube.com/watch?v=hiJDlvKDy0o>
2. Real World Data Cleaning in Python Pandas: <https://www.youtube.com/watch?v=iaZQF8SLHJs>
3. Pandas datetime manipulation: <https://www.youtube.com/watch?v=ZjMTZIxgkYo>

# Questions

The following sections contain code-based questions that focus on cleaning, manipulating, and analyzing the dataset. You are expected to provide code without additional explanations.

**Section A: Data Cleaning and Transformation**

In this section, the focus is on cleaning, manipulating, and transforming data using pandas to ensure consistency and usability.

1. Remove unnecessary columns like ***show\_id***, ***description***, and ***director***.
2. Eliminate rows with missing values from the dataset.
3. Convert the ***date\_added*** and ***release\_year*** columns into datetime format, and extract ***year***, ***month***, and ***day*** from ***date\_added***, along with extracting the year from ***release\_year***.
4. Calculate the total number of cast members for each show or movie and determine how to extract the lead actor’s name from the cast column, assuming the first actor listed is the lead.
5. Standardize the duration column by converting it into an integer format.
6. Clean and capitalize the country names in the country column, ensuring no extra spaces.
7. Calculate how many genres each show or movie is listed under (i.e., column to be added is ***genre count***).
8. Transform columns like ***listed\_in*** -> ***genre***, ***country***, and ***cast*** -> ***actor*** into list format.
9. Explode the genre, country, and actor columns into individual rows, and remove any null values.
10. Filter out rows with empty country fields and drop unnecessary columns like ***date\_added***, ***cast***, and ***listed\_in***.
11. Preview the first few rows of the cleaned and transformed dataset to verify the changes.

**Section B: Grouping and Aggregation**

This section involves using groupby and aggregation methods to derive insights from the dataset based on specific criteria. Select and complete 7 of the following questions.

1. What is the most popular genre based on the number of movies or TV shows listed?
2. Which country has the highest number of movies or TV shows listed?
3. Which movie or TV show has the most listings?
4. Which movie or TV show has the largest cast?
5. Which movie or TV show has been broadcast the most times?
6. Which movie or TV show is listed in the most unique countries?
7. Which actor has had the most lead roles?
8. Which actor has appeared in the most movies or TV shows?
9. What is the most common rating assigned to movies or TV shows?
10. Are there more TV shows or movies listed on Netflix?
11. On which day of the week are the most movies or TV shows added to Netflix?
12. In which year were the most movies or TV shows added to Netflix?
13. Which month sees the highest number of movies or TV shows added to Netflix?
14. Which release year had the most movies or TV shows?

# Submission Mode

You are expected to submit your assignment a week after it was given, with this, Submission starts from Friday 7am Nigerian time to Saturday 3pm Nigerian time.

This submission will be done either on Twitter or LinkedIn, you can choose either of the two to submit your assignment or you can submit on both platforms. You would take a screenshot of your work Answers, you can add a write up if you want.

1. **Primary submission:** *GitHub – If you would need assistance configuring git, please post on the DS group; Thus a session will be conducted for demonstration.*
2. For twitter Submission, you would tag:
3. The TDI Official page @TDataInitiative
4. Annie @ DabereNnamnai
5. Jacob @Jacob\_Ajala
6. Use the #TDIDataScience
7. For LinkedIn Submission, you would tag:
8. TDI page @TheData Initiative
9. Annie @ Anne Nnamani