**Pre\_processing:**

1-drop("date\_added") column

2-Replace both '&' and '-' with a comma to have a consistent delimiter in 'Listed\_in' column

3- Explode the categories after splitting by the new consistent delimiter (comma) in a new column called 'Category' column and drop 'Listed\_in' column

4- Trim extra whitespace around the categories after splitting

5- Remove numeric entries like years (e.g., 2016, 2020) and durations (e.g., 93 min) from 'Category' column

6- Define a UDF to check if a value is a country in 'Category' column

7- Create a new column 'country\_detected' for detected countries from the 'Category' column

8- Remove countries from 'listed\_in' column by filtering out country names

9- Create a mapping dictionary for Categories

10- Map the categories using a user-defined function (UDF)

11- drop("Category") column

12-col("Total\_views").cast(IntegerType())), col("user\_reviews").cast(IntegerType())) & col("user\_rating").cast(FloatType()))

13-Replace Nulls in directors, cast and description with Unknown

14- Filter out rows where 'director' and 'cast' columns that are fully numeric (integers)

15- col("release\_year").cast(IntegerType())

16- Classify movies based on duration directly

17- Replace Nulls in country and duration with Unknown

**Final Standardized Categories:**

['Romance', 'Young Adult Audience', 'Thriller', 'Entertainment', 'Art', 'Adventure', 'Sports', 'Drama', 'Military and War', 'Documentary', 'Health', 'Culture', 'Fantasy', 'LGBTQ', 'Reality', 'Other', 'History', 'Animation', 'Music', 'Special Interest', 'Television', 'Horror', 'Western', 'Comedy', 'Children', 'Action', 'Sci-Fi', 'Talk Show', 'Spiritual', 'Crime', 'Science', 'Family', 'Movies', 'Mystery', 'Musical', 'Nature', 'Cult', 'Indie', 'Classic', 'Food', 'Teen', 'Game Show', 'Technology', 'Lifestyle', 'Black Stories', 'News', 'Survival', 'Travel', 'Disaster', 'Series', 'Variety', 'Screen Dance', 'Animals', 'Romance Comedy', 'Biography', 'Medical']