Netflix Content Analysis

Phase 4: Feature Engineering

Name: Gagan Dhanapune
Gmail: gagandhanapune@gmail.com

Objective Of This Milestone

- Analyze Netflix content growth over time.
- Visualize the distribution of genres, ratings, and content type.
- Identify country-level content contributions.
- Create derived features such as “Content Length Category” and “Original vs. Licensed.”

Feature Engineering -

The goal of this milestone is to perform feature engineering on the cleaned Netflix Titles dataset.

Feature engineering means creating new, meaningful features (columns) from existing data to make analysis and modeling more powerful and insightful. This milestone focuses on transforming text-based and categorical information into structured, analyzable formats.

1. Creating Content Length Category

Steps:

- A custom function get_length_category() was created to classify each title based on its duration_num and duration_type.
- The logic was applied differently for **Movies** and **TV Shows**:
    - **Movies:**
        - ≤ 60 min → **Short**
        - 61–120 min → **Medium**
        - > 120 min → **Long**

Functions Used:

- **apply()** → to apply the custom function row-wise across the DataFrame.
- **fillna()** → used in pre-processing to handle any missing duration values, ensuring the categorization function works smoothly.

Output/Process:
A new column **Content_Length_Category** was added that categorizes every title by its duration type.

Insights:

- Most movies fall into the **Medium** length category.
- A significant number of TV shows are categorized as **Limited** or **Moderate**, indicating Netflix's focus on short-format and limited-run series.

2. Identifying Netflix Originals

Steps:

- Created a new feature **is_original** to label whether a title is a **Netflix Original** or **Licensed**.
- The logic checked for the presence of the word "Netflix" within the title of each entry.

Functions Used:

- **apply()** with a **lambda** expression to efficiently check each title.

Output/Process:
A new column **is_original** was created with two possible values:

- **Original** → if “Netflix” is present in the title.
- **Licensed** → otherwise.

Insights:

- The analysis shows that the vast majority of titles listed are **licensed content**, not Netflix originals.
- This suggests that a core part of Netflix's strategy relies on **acquiring popular existing shows and movies** to supplement its original productions.

3. Yearly Growth of Netflix Content

Steps:

- Filtered titles by their type (Movie or TV Show) and counted their releases per year using groupby() and value_counts().
- Plotted a **line chart** to visualize the growth trends for both content types over the years.

Functions Used:

- **pd.to_datetime()** → to convert the date_added column into a proper datetime format for time-series analysis.
- **.dt.year** → to extract the year part from the datetime objects.
- **groupby()** & **value_counts()** → to aggregate and count content by year and type.
- **plt.plot()** → to visualize the growth trend.

Insights:

- Netflix has seen a dramatic increase in the amount of content added per year, especially in more recent years, highlighting the platform's rapid expansion.

- The trend shows a significant uptick in content acquisition after **2016**, aligning with Netflix's global expansion efforts.

4. Mapping Countries to Regions

Steps:

- Mapped each title's country to a broader geographical **region** (e.g., Asia, Europe, North America).
- Created a dictionary region_map to assign countries to their respective continents.
- Missing country values were filled with "Unknown."

Functions Used:

- **fillna()** → to handle missing country names before mapping.
- **apply()** → to apply the get_region() mapping function to each title's country data.

Output/Process:
A new column **Region** was created representing the continent of origin for each title.

Insights:

- The majority of content on Netflix comes from **North America** (primarily the U.S.).
- **Asia**, especially **India** and **South Korea**, shows a strong and growing contribution, which reflects Netflix's global content strategy and focus on key international markets.

5. Creating Content Age Group

Steps:

- Converted the detailed MPAA and TV ratings (e.g., PG, TV-MA, R) into simplified, broader audience groups.
- Mapped each rating into one of four categories: **Kids**, **Teens**, **Adults**, or **Unknown** using the map() function.

Functions Used:

- **fillna()** → to handle any missing rating values.
- **map()** → to replace the original ratings with their corresponding simplified age group.

Output/Process:
Created a new column **Content_Age_Group** to simplify audience segmentation for analysis and visualization.

Insights:

- The vast majority of Netflix content is targeted toward **Teens** and **Adults**.
- A smaller percentage of content is aimed at **Kids**, indicating Netflix's primary focus on mature storytelling and programming for older audiences.