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**.
- The logic was applied differently for **Movies** and **TV Shows**:
    - **Movies:**
        - ≤ 60 min → Short
        - 61–120 min → Medium
        - 120 min → Long
    - **TV Shows:**
        - 1 Season → Limited
        - 2–4 Seasons → Moderate
        - 4 Seasons → Long-running

Functions Used:

- apply() → to apply the custom function row-wise.

- fillna() → filled missing duration values with “Unknown.”

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 category.
- Many **TV shows** are Limited or Moderate, indicating Netflix’s focus on short-format or limited 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 word “Netflix” in the title name.

Functions Used:

- apply() with a lambda expression.

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

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

Insights:

- All the titles listed are **licensed content**, not Netflix originals.
- This suggests Netflix relies on **acquiring popular shows and movies** in addition to producing originals.

3. Yearly Growth of Netflix Originals

Steps:

- Filtered only “Original” titles and counted their releases per year using: value_counts().sort_index()
- Plotted a **line chart** to show yearly trends.

Functions Used:

- value_counts() → counts per release year.
- sort_index() → ensures chronological order.
- plt.plot() → visualized growth.

Insights:

- Very few Netflix Originals were released in these years.
- Originals appeared sporadically: 2016, 2017, 2020, and 2021

4. Mapping Countries to Regions

Steps:

- Mapped each title’s **country** to a broader **region** (like Asia, Europe, North America, etc.).
- Created a dictionary region_map to assign countries to continents.
- Missing values were filled with “Unknown.”

Functions Used:

- fillna() → handle missing country names.
- apply() → apply get_region() mapping function.

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

Insights:

- The majority of content comes from **North America** (mainly U.S.).
- **Asia**, especially **India** and **South Korea**, shows growing contribution — reflecting Netflix's global content strategy.

5. Extracting Year from "Date Added"

Steps:

- Converted date_added column to proper datetime format.
- Extracted the **year** as a separate column Year.

Functions Used:

- pd.to_datetime() → convert to date format.
- .dt.year → extract the year part.

Output/Process:
Added Year column, enabling analysis of when titles were added to Netflix.

Insights:

- Helped analyze addition trends over time and relate them to release trends.
- Showed that most content was added after **2016**, aligning with Netflix's global expansion.

6. Creating Content Age Group

Steps:

- Converted ratings like PG, TV-MA, R, etc., into simplified audience groups.

- Mapped each rating into **Kids**, **Teens**, **Adults**, or **Unknown** using map().

Functions Used:

- fillna() → filled missing ratings.
- map() → replaced ratings with corresponding age groups.

Output/Process:
Created new column Content_Age_Group to simplify audience segmentation.

Insights:

- The majority of Netflix content is for **Teens** and **Adults**.
- Less content is targeted toward **Kids**, showing Netflix's focus on mature storytelling.

7. Visualizations.

Visuals Created:

- Bar Chart: Distribution of Content Length Categories → Shows most movies are Medium length and TV shows are Limited or Moderate series.
- Bar Chart: Original vs Licensed Titles → Reveals the majority of titles are Licensed, not Netflix Originals.
- Line Chart: Number of Originals per Year → Highlights gradual growth of Netflix Originals, with spikes in 2016, 2017, 2020, and 2021.
- Bar Chart: Audience Category Distribution → Indicates most content targets Teens and Adults, with fewer titles for Kids.
- Horizontal Bar Chart: Number of Titles by Region → Shows North America dominates content, followed by Asia (India, South Korea) and other regions.

Key insights-

- Most titles are licensed, meaning Netflix streams many shows/movies produced by others.

- Content Length: Most movies are Medium length; most TV shows are Limited or Moderate series.

- Audience Target: Majority of content is for Teens and Adults, less for Kids.

- Regional Contribution: Most content comes from North America, followed by Asia (India, South Korea).

- Content Growth: Netflix content added per year has been increasing, especially in recent years, showing platform expansion.