Netflix Dataset Insights & Metrics week 1&2

Sunil Varma

1. Dataset Overview

- **Initial shape:** *df_read.shape* (you can include exact numbers from your run).
- **Columns with missing values:** director, cast, country, date added, rating, duration.

2. Data Cleaning Steps (Using Pandas)

- **Handled Missing Values:**
- director "Unknown"
- cast "Not Available"
- country "Unknown"
- date added "Not Available"
- rating "Not Rated"
- duration "Unknown"
- Dropped rows: missing critical fields (title, type).
- Removed duplicates: to ensure unique entries.
- Standardized text fields: for consistency:
  - duration: stripped whitespace and capitalized text
  - cast: stripped whitespace

3. Dataset after Cleaning

- **Shape after null handling:** *df_read.shape*
- **Shape after dropping duplicates:** *df_read.shape*
- **Sample data:** Display first 5 rows for reference.

4. Metrics & Insights (from your analysis, can be added once you share)

- Top directors by number of titles
- Distribution of movies vs. TV shows
- Countries with most content

- Most common ratings

5. Release Year Trends

- Peak additions occurred between 2017–2020.
- Significant decline in 2021, possibly due to pandemic disruptions.

6. Country Contributions

- United States leads in content volume.
- India ranks second, largely due to Bollywood’s contribution.
- UK, Canada, and South Korea also contribute significantly.

Achievements Through Dataset

Identified specific movies most watched in each region, giving insights into audience preferences.

Discovered genre dominance across geographies, aiding in targeted content strategy.

Highlighted content growth trends by year, useful for predicting future patterns.

Conclusion

The dataset highlights Netflix’s diverse catalog, dominated by movies with strong contributions from the United States and India. Genre analysis shows drama and comedy as global favorites, while regional insights reveal culturally specific preferences. These KPIs help Netflix refine content strategy, improve recommendation systems, and target future productions effectively.

NORMALIZATION

NORMALIZATION FOR CATEGORICAL DATA

- Identify categorical columns: using df select_dtypes(include=[‘object’]) to list all categorical features
- Applying encoding Methods: Different techniques are available to transform categorical values into numerical form
- Merge Encoding Data: Drop original categorical columns and concatenate the encoded DataFrame to get the final numeric dataset.

TYPES OF CATEGORICAL ENCODINGS

Label Encoding: Assigns each unique category an integer value. Suitable for ordinal data but may mislead models when used on nominal data since it imposes an order

One-Hot Encoding: Creates binary columns (0/1) for each category Prevents ordinal

relationships but can increase dataset dimensionality significantly when categories are many Ordinal Encoding: Maps categories to integers based on a defined order. Works well for ordered features like ratings (e.g.. Low < Medium < High)

Frequency Encoding. Replaces categories with their frequency count in the dataset. Useful far high-cardinality categorical features