Netflix Insights

Data Cleaning Steps (using Pandas)

1. Loaded the raw Netflix dataset into a Pandas DataFrame.

2. Removed duplicate records to ensure uniqueness of shows.

3. Handled missing values by either filling or removing them depending on the column.

4. Dropped unnecessary columns (e.g., 'Unnamed: 0' index column).

5. Standardized date formats in the 'date_added' column.

6. Converted 'release_year' to integer type for proper analysis.

7. Ensured 'duration' field consistency (minutes for Movies, seasons for TV Shows).

8. Cleaned and standardized text fields such as 'title', 'director', 'cast', and 'country'.

Metrics and Insights

- Movies vs TV Shows distribution:
  - Movie: 5185
  - TV Show: 147
- Top 5 most common ratings:
  - TV-MA: 1822
  - TV-14: 1214
  - R: 778
  - PG-13: 470
  - TV-PG: 431
- Top 5 countries producing content:
  - United States: 1846
  - India: 875
  - United Kingdom: 183
  - Canada: 107
  - Spain: 91
- Content release trend (last 5 years):
  - 2017: 657

- 2018: 648
- 2019: 519
- 2020: 442
- 2021: 161

• Top 5 most popular genres:

- International Movies: 2369
- Dramas: 2293
- Comedies: 1553
- Action & Adventure: 806
- Independent Movies: 740

SOME FUNCTIONS I HAVE USED TO CLEAN THE DATA

1) print("\nmissing values per coloumn:")

print(df_read.isnull().sum())

df_cleaned = df_read.dropna()

dropna() removes **all rows** that contain **any missing values**, regardless of the column.

2) import pandas as pd

df_read = pd.read_csv("/Volumes/workspace/default/infoysys/INFOYSYS.csv")

display(df_read)

print("\nmissing values per coloumn:")

print(df_read.isnull().sum())

The dataset contains incomplete records, especially in descriptive fields like director and cast.

Further cleaning is required before analysis to ensure accuracy and consistency.

3) df_read['type'] = df_Read['type'].str.strip()

df_read['title'] = df_Read['rating'].str.strip()

Removed leading and trailing whitespaces from key text columns (type, title, rating) using .str.strip().

4) display(df_read)

df_read.country.value_counts()

Counts how many times each unique country appears in the country column.

5) df_read['duration_minutes'] = df_read['duration'].str.extract(r'(\d+)').astype(float)

Extract only the numeric part of the duration (e.g., "90" from "90 min" or "1" from "1 Season").

6) df_read.type.value_counts()

Found that the dataset contains 6131 Movies and 2676 TV Shows.
This shows that Movies make up the majority of the content in the dataset.

NORMALIZED DATA STEPS

Summary of Data Cleaning and Normalization Steps

1. **Data Loading:**
   - Loaded the raw Netflix dataset (INFOYSYS.csv) into a Pandas DataFrame.
2. **Missing Values Check:**
   - Identified columns with missing/null values such as director, cast, country, date_added, and rating.
3. **Data Cleaning:**
   - Removed rows with missing values (dropna) to ensure complete records for analysis.
   - Stripped leading and trailing whitespaces from important text columns like type, title, and rating for consistency.
4. **Verification:**
   - Checked for remaining missing values to confirm all nulls were removed.
   - Verified the shape of the cleaned dataset (reduced number of rows after dropping nulls).
5. **Saving Cleaned Data:**
   - Saved the cleaned and normalized dataset as INFOYSYS_NORMALIZED.csv for future use.

1. One-Hot Encoding

- **What it does:**
  Converts each category in a feature into a new binary (0 or 1) column. For example, a type column with categories Movie and TV Show becomes two columns: type_Movie

and type_TV Show where each row has a 1 in the column corresponding to its category and 0 elsewhere.

- df_encoded = pd.get_dummies(df_cleaned, columns=['type'])

2) Label Encoding

- **What it does:**
  Assigns each unique category a unique integer label. For example, rating categories like PG, R, TV-MA might be encoded as 0, 1, 2, etc.

  ```
  le = LabelEncoder()

  df_cleaned['rating_encoded'] = le.fit_transform(df_cleaned['rating'])
  ```