Netflix Insights and Metrics

Week 1&2:

Netflix Insights

1. **Content Distribution**
   - The dataset comprises **~8,800 titles**, a combination of both **Movies (~70%)** and **TV Shows (~30%)**.
   - Movies dominate Netflix's catalog, but TV Shows have been increasing in recent years, signaling Netflix's shift towards episodic content.
2. **Temporal Trends**
   - Titles span multiple decades, with older classics alongside recent Netflix Originals.
   - A sharp rise in content is observed post-2015, aligning with Netflix's global expansion strategy.
3. **Genre Representation**
   - A wide variety of genres exist.
   - **Top genres:** Dramas, Documentaries, Comedies.
   - **Emerging genres:** International TV, Stand-up Comedy, and Romantic TV Shows — reflecting user demand.
4. **Geographical Spread**
   - Content originates from over 100 countries, showcasing Netflix's global production and licensing reach.
   - Major contributors**:** United States, India, United Kingdom, Japan, South Korea.
5. **Rating Distribution**
   - Titles are spread across maturity ratings (TV-MA, R, PG-13, TV-14, etc.).
   - A strong presence of mature-rated content (TV-MA, R) indicates a focus on adult audiences, but family-friendly segments (TV-Y, PG) are also well represented.
6. **Missing Data Observations**
   - Director and cast columns had significant missing values, likely due to incomplete metadata.
   - Rating and Duration had gaps, which were filled systematically for consistency.

Netflix Metrics/Scope

1. **Trend Analysis**
   - Evaluate the evolution of Movies vs. TV Shows, genres, and ratings over years.
   - Guide Netflix in shaping its content acquisition and production strategies.
2. **Genre Popularity & Recommendations**
   - Identify top genres globally and regionally.
   - Enable personalized recommendations based on user preferences.

3. **Geographical Expansion Strategy**
    - Assess country-level contributions to Netflix’s catalog.
    - Support regional expansion and localized content production.
4. **Content Duration Insights**
    - Distinguish average movie length vs. average TV Show seasons.
    - Inform viewer engagement and content planning.
5. **Data Quality Improvement**
    - Enhance metadata completeness for directors and casts.
    - Support enriched recommendation systems and talent-based content analysis.

Dataset Loading

The Netflix dataset is sourced from Kaggle and loaded into the workspace for preprocessing and analysis.

Dataset Source: Kaggle — Netflix Movies and TV Shows Dataset.
Dataset Size: ~8,800 titles across multiple years and genres.
Key Columns: type, title, director, cast, country, release_year, rating, duration, listed_in, date_added.

Loading the dataset using pandas:

```
import pandas as pd

df = pd.read_csv("/Volumes/workspace/default/netflix/netflix_titles.csv")

display(df.head())
```

The dataset provides a rich set of features that allow for multi-dimensional analysis of Netflix’s content strategy, such as genre diversity, rating distributions, and country-wise availability.

Data Cleaning Steps Using Pandas

1. Duplicate Removal

- Removed duplicate rows to eliminate redundancy and maintain data integrity.

```
df = df.drop_duplicates()
```

2. Missing Value Handling

- Replaced missing values with default placeholders for consistency.

Column	Handling Strategy	Replacement Value
director	Fill missing values	"Unknown"
cast	Fill missing values	"Not Available"
country	Fill missing values	"Unknown"
date_added	Fill missing values	"Not Available"
rating	Fill missing values	"Not Rated" (later encoded)
duration	Fill missing values	"Unknown" (later normalized)

- Dropped rows missing critical identifiers: title and type.

```
df = df.dropna(subset=['title', 'type'])
```

---

3. Standardization of Text Fields

- Trimmed whitespaces.
- Converted text fields to consistent case formatting.

```
df['type'] = df['type'].str.strip().str.title()

df['country'] = df['country'].str.strip().str.title()

df['rating'] = df['rating'].str.strip()
```

---

4. Data Type Conversion

- Converted release_year → **integer type**.
- Converted date_added → **datetime type** for trend analysis.

```
df['release_year'] = pd.to_numeric(df['release_year'], errors='coerce')

df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
```

Normalization

5. Duration Normalization

- Standardized duration field:
  - For Movies → Extracted minutes (integer).
  - For TV Shows → Extracted number of seasons (integer).
- Missing values filled with 0 for clarity.

Example:

- "90 min" → 90
- "2 Seasons" → 2

---

6. Normalization of Categorical Features

- **Label Encoding** → rating_encoded.
- **One-Hot Encoding** →
  - type → type_Movie, type_Tv Show.
  - listed_in (genres) → multiple genre_* columns.
  - country → multiple country_* columns.

This ensures:

- Single row per title.
- Multi-genre and multi-country support via binary columns.

---

7. Critical Field Validation

- Verified title and type fields are present for all records.
- Ensured no nulls in critical analysis features after cleaning.

```
print(df[['title','type']].isnull().sum())
```