## **Netflix Insights and Metrics**

### **Key Insights**

* **Content Distribution**: The dataset comprises ~8,800 titles, with a mix of both Movies and TV Shows.
* **Temporal Trends**: Titles span multiple decades, highlighting Netflix’s long-term content acquisition and production strategy.
* **Genre Representation**: A diverse set of genres exists, with certain categories (like Dramas, Documentaries, and Comedies) being dominant.
* **Geographical Spread**: Titles originate from numerous countries, reinforcing Netflix’s global reach.
* **Missing Data**: Attributes such as director and cast had significant missing values, which were filled with defaults. For rating and duration, missing values were explicitly filled with 0 for consistency.

### **Potential Applications**

* **Trend Analysis**: Analyze how the mix of Movies vs. TV Shows, genres, and ratings evolved over the years to guide Netflix’s content acquisition.
* **Genre Popularity**: Identify top-performing genres to optimize recommendations and marketing campaigns.
* **Geographical Strategy**: Understand country-level contributions to inform regional expansion and localized content production.
* **Content Recommendations**: Build recommendation systems leveraging attributes like genre, rating, and duration.
* **Data Quality Improvement**: Address data gaps (especially director and cast) to ensure richer analyses.

## **Data Cleaning Steps Using Pandas**

1. **Duplicate Removal**
   * Dropped duplicate rows to ensure uniqueness and remove redundancy.
2. **Missing Value Handling**
   * Filled missing values in:

* director → "Unknown"
* cast → "Not Available"
* country → "Unknown"
* date\_added → "Not Available"
* rating → **0**
* duration → **0**
  + Dropped rows missing critical fields (title, type).

1. **Standardization**
   * Trimmed whitespaces and normalized text formatting.
   * Converted type, rating, and country to consistent casing (e.g., title case).
2. **Data Type Conversion**
   * Converted release\_year to numeric.
   * Parsed date\_added into datetime where applicable.
3. **Duration Normalization**
   * Missing values replaced with **0** (for consistency).
   * Enables clearer differentiation between Movie durations (minutes) and TV Shows (seasons).
4. **Critical Field Validation**
   * Verified that essential columns (title, type) are complete, ensuring dataset integrity.