Netflix Dataset Cleaning & Preprocessing Report

Objective

To clean and prepare a raw Netflix dataset by handling null values, fixing inconsistent formatting, ensuring correct data types, and generating new useful columns for downstream analysis or visualization.

Tools Used

- **Python** with **Pandas** for data wrangling and transformation
- Jupyter Notebook for step-by-step development and testing

Key Cleaning & Preprocessing Steps

1. Missing Value Handling

- director: Filled with 'Unknown'
- cast: Filled with 'Not Available'
- country: Filled with **mode** (most common value)
- rating: Filled with 'Not Rated'
- Dropped rows where date_added or duration was missing

2. Duplicate Check

- No duplicate rows were found in the dataset.

3. Column Renaming for Consistency

- All column names were converted to:
    - lowercase
    - stripped of whitespace
    - spaces replaced with underscores
      (e.g., "Date Added" → "date_added")

4. Text Standardization

- type, country: converted to **title case**

- rating: converted to **uppercase**
- listed_in, director, cast, description: stripped extra whitespaces

5. Date Handling

- date_added converted to **datetime** format using pd.to_datetime()
- Created a new column date_added_ddmmyyyy with readable date format (dd-mm-yyyy)

6. Data Type Fixes

- release_year: converted to int
- duration: ensured it is a string format

7. Feature Extraction from duration

- Created:
    - duration_num: Extracted numeric part (e.g., 90 from 90 min)
    - duration_type: Extracted time unit (e.g., min or season)

📌 **Final Dataset Summary**

- Cleaned and ready for analysis or visualization
- All missing values addressed appropriately
- Text formats are consistent
- No duplicates present
- New features (duration_num, duration_type, formatted date) added

---

Conclusion

The dataset has been successfully cleaned and standardized. It is now consistent, reliable, and ready for meaningful data exploration, visualization, or modeling.