

SUMMARY CHANGES

1. Standardized Column Names

- Converted all column names to lowercase and replaced spaces with underscores for consistency.

2. Handling Missing Values

- Filled missing categorical values with meaningful placeholders:
 - director → "Unknown"
 - cast → "Unknown"
 - country → "Unknown"
 - rating → "Not Rated"
- Dropped rows where date_added or duration were missing, as these are critical for trend and time-based analysis.

3. Parsing the duration Column

- Created a custom parse_duration() function to split the original duration column into two separate numeric columns:
 - **duration_minutes** → for Movies (e.g., "90 min" → 90).
 - **num_seasons** → for TV Shows (e.g., "2 Seasons" → 2).
- Ensured Movies have num_seasons = 0 and TV Shows have duration_minutes = 0.

4. Handling Missing and Inconsistent Duration Values

- Filled missing movie durations with the **median movie duration** for better accuracy.
- Any remaining missing duration_minutes for TV Shows were replaced with 0.

5. Date Formatting & Feature Engineering

- Converted date_added to proper datetime format.
- Extracted **year_added** and **month_added** from the date_added column for trend-based analysis.

6. Data Type Standardization

- Converted key columns to proper numeric data types:
 - release_year, num_seasons, year_added, and month_added → int.

7. Duplicates Check

- Verified that there were **no duplicate rows** in the dataset.

8. Final Cleaned Dataset

- Exported the final clean dataset as **netflix_cleaned.csv**, ensuring no critical missing values remain.
- The dataset is now fully standardized, structured, and ready for visualization or machine learning tasks.