



Project Summary – Data Cleaning and Preprocessing

Dataset Used: [Netflix Movies and TV Shows \(Kaggle\)](#)

Objective: Clean, preprocess, and prepare the dataset for analysis and visualization to extract meaningful insights.

Steps Taken:

1. Handled Missing Values

- Replaced missing text values with "Unknown" for readability.
- Filled numeric columns (e.g., release year) where necessary.

2. Removed Duplicates

- Although there were no duplicates.

3. Standardized Text Columns

- Converted text to lowercase and stripped whitespace (director, cast, country, rating, listed_in).
- Converted show_id to uppercase.

4. Data Type Conversion

- Converted release_year to integer.
- Extracted numeric duration from the duration column.
- Ensured all text columns use the pandas string dtype.

5. Feature Engineering

- Extracted release_year from release_date (or kept year directly).

6. Exporting Clean Data

- Saved the cleaned dataset to CSV for further analysis:

Outcome:

- A cleaned, well-typed, and analysis-ready dataset
- Ready for exploratory data analysis, visualization, or machine learning projects

[GitHub Repository link](#)

[LinkedIn profile](#)