Data Loading and Initial Exploration

- **Library Imports**: The script begins by importing essential libraries — `pandas`, `re`, and `numpy` — for data manipulation and regular expression operations.

- **Dataset Loading**: The Netflix dataset is read from a CSV file named `'netflix_titles.csv'` into a pandas DataFrame named `df`.

- **Initial Inspection**: To get a quick overview of the dataset, `df.head()` is used to display the first few records.

Handling Missing Data

- **Identifying Nulls**: The number of missing values in each column is checked using `df.isnull().sum()`.

- **Filling Missing Values**:

    - Missing entries in the `'director'`, `'cast'`, and `'country'` columns are replaced with `'Unknown'`.

    - Records lacking a `'date_added'` value are removed using `df.dropna()`.

    - For the `'rating'` column, missing values are filled with the most frequently occurring value (mode).

Duration Extraction

- A custom function `extract_duration_minutes` is created to extract numerical duration from the `'duration'` column. It handles both minutes and seasons (by converting seasons into minutes).
- The results are stored in a new column called `'duration_minutes'`.
- Any missing values in this new column are filled with the median duration value.

Date Formatting

- The `'date_added'` column is converted to datetime format using `pd.to_datetime`, accounting for inconsistent formats.
- Any rows where conversion fails (i.e., invalid or missing dates) are dropped using `df.dropna()`.

Data Type Conversion

- The `'show_id'` column is explicitly converted to string type with `astype(str)`.
- The `'release_year'` column is cast to an integer using `astype(int)`.

Duplicate Removal

- Duplicate rows, if any, are removed using `df.drop_duplicates()` to ensure data consistency.

Summary

Overall, the code thoroughly cleans and prepares the Netflix dataset for analysis. It tackles missing data, standardizes formats, extracts useful features like duration, and ensures appropriate data types — resulting in a structured and ready-to-analyze dataset.