Data Cleaning, Normalization Steps, and Metrics:

Data Cleaning Steps

1. Replace 'Unknown' with NaN. Helps mark missing values properly so they can be handled consistently. This prevents misleading data during analysis.
2. Drop Duplicate Rows: Removes repeated records to avoid bias and redundancy in results.
3. Convert 'date_added' to datetime: Ensures dates are in the correct format for filtering and time-based analysis.
4. Strip Whitespace from Titles: Cleans extra spaces for uniformity and prevents mismatches during searches or grouping.
5. Lowercase and Clean 'listed_in': Standardizes category names, making grouping and comparison easier.
6. Normalize Rating (uppercase + strip): Removes inconsistencies in rating labels for accurate frequency counts.
7. Extract Duration Value and Unit: Splits time into numbers and units, making it easier to analyze movies vs. TV shows.
8. Convert Duration Value to Numeric: Ensures durations are numeric so calculations and statistics can be applied.
9. Drop Rows with Missing Critical Info: Removes incomplete data where analysis would not be meaningful.
10. Drop Columns with >50% Nulls: Eliminates columns that have too many missing values to be useful.
11. Fill Missing Director: Assigns a placeholder to preserve rows while marking missing information.
12. Fill Missing Cast: Replaces nulls with “Not Available” to avoid gaps in the dataset.
13. Fill Missing Country: Uses “Unknown” for missing countries to keep rows intact for analysis.
14. Fill Missing Rating: Assigns “Unrated” to ensure all shows/movies have a rating label.
15. Fill Missing Date Added: Keeps timeline structure intact even if exact dates are missing.
16. Fill Missing Duration by Type: Assigns default durations to avoid null values while keeping type-specific logic.
17. Re-extract Duration After Filling: Refreshes numeric and unit fields after filling gaps to maintain consistency.

18. Fill Missing Main Country: Ensures every record has at least one country value for analysis.

Normalization and Encoding Steps

1. Country Encoding: Converts country presence into binary values (1 for known, 0 for unknown).
2. Movie Type Encoding: Assigns numeric labels (1 for movies, 0 for TV shows) for modeling.
3. Duration Unit Encoding: Converts time units into binary values (min = 0, season = 1).
4. Rating Normalization: Calculates the relative frequency of ratings for balanced comparison.
5. Country Label Encoding: Converts each unique country name into a numeric code for machine learning.
6. Director Name Cleaning & Encoding: Assigns numeric IDs to directors, with defaults for missing ones.

Metrics for Data Cleaning and Normalization

1. Percentage of Missing Data per Column: Measures completeness of each feature.
2. Duplicate Row Count and Ratio: Identifies how much repetition exists in the dataset.
3. Data Type Consistency Checks: Ensures each column uses the proper format (e.g., dates, numbers).
4. Outlier Detection (IQR, Z-Score): Flags unusual values that may distort analysis.
5. Unique Value Counts in Categories: Detects anomalies and ensures consistency in categorical data.
6. Distribution of Encoded Variables: Confirms balanced representation after encoding.
7. Normalized Frequency of Ratings: Helps compare ratings fairly by scaling their counts.