**Cleaning tasks (questions) — grouped by concept**

**A. Structural & schema tasks**

1. Which columns have the wrong data type (e.g., time and releasedate as strings) and which should be converted?
2. Are there any rows or columns that are entirely empty or nearly empty — should any be dropped?
3. Are column names consistent (lowercase, underscores)? Should you rename releasedate → release\_date, stars → rating\_text, etc.?

**B. Parsing / splitting fields**

1. How can you remove the prefixes Writtenby: and Narratedby: from author and narrator and store only the names?
2. Can you split name into parts (e.g., series name vs. book numbers, or title vs subtitle) where appropriate?
3. How can you parse time into consistent numeric fields like duration\_minutes (or duration\_hours + duration\_minutes)?
4. How can you parse stars to extract a numeric rating (e.g., 4.5) and the number of ratings (e.g., 34)? What about when spacing or words are missing?

**C. Dates & timezones**

1. Are releasedate values in a consistent format (mm/dd/yyyy vs dd/mm/yyyy)? How can you programmatically detect the format and convert to datetime64[ns]?
2. After converting, can you extract useful features like release\_year, release\_month, release\_day, or release\_weekday?
3. Are there impossible or invalid dates (e.g., month > 12 if format misinterpreted)? How will you detect and fix them?

**D. Missing values and special null indicators**

1. Which values should be considered "missing" (e.g., "", "NA", "None", "-", "n/a") and how will you standardize them to np.nan?
2. For each column, how many missing values exist and are there patterns (missing by author, by year, etc.)?
3. For columns with missing values, which imputation strategy is appropriate (leave as missing, fill with mode/median, forward/backfill, or domain default)?

**E. Text cleaning & normalization**

1. Are there leading/trailing spaces or extra internal whitespace in any text fields? How will you trim them?
2. Does casing matter? Should you standardize language, author, and narrator to title case or lowercase?
3. Are there punctuation or special characters to remove (e.g., stray colons, HTML entities)?
4. Do you need to correct common misspellings or inconsistent variants (e.g., Englis, english, ENGLISH)?

**F. Numeric parsing & cleaning**

1. Is price stored as numeric or string? If string, how will you strip currency symbols and thousands separators then convert to float?
2. Are there non-numeric characters in numeric fields (e.g., ,, $, %)? How will you remove them safely?
3. After extracting rating (numeric) and rating\_count (integer), are there any non-numeric or malformed values to handle?

**G. Consistency & canonicalization**

1. Should author and narrator names be normalized to a consistent format (e.g., First Last)? How to handle initials, middle names, etc.?

**H. Duplicates & near-duplicates**

1. Are there exact duplicate rows — how many and which ones should be removed?
2. Are there near-duplicates (same title + author but slightly different punctuation or whitespace)? How will you detect and consolidate them?