Assignment E: Tidying and Transforming Data in R

In this assignment, you will practice tidying messy datasets and applying tidyverse transformations to prepare data for analysis. You will work with two different real-world datasets:

1. `World Development Bank Data (economic indicators across years)`
2. `Movies Data (with multiple values stored in one cell, and multiple linked files)`

Your task is to identify the messy aspects of each dataset and apply tidy R principles (tidyr, dplyr, stringr).

Part 1: World Development Bank Data Dataset

The dataset contains **~3079 rows and 11 columns.** Each row represents an observation, but the structure is not tidy. Columns include:

- Series ID and Series Name (currently stored as separate columns), which are variables such as death rate, Access to clean fuels and technologies for cooking, etc.
- Years as columns (e.g., 2010, 2011, ...).
- Country Code and Country Name

Values under each year column. At the bottom of the file, you will notice notes and metadata (such as units of measure and explanations).

Tasks

1. Import the dataset into R.
2. Remove metadata rows at the bottom (they are useful as documentation, but not tidy data). Remember, this metadata contains some useful info regarding the unit of measure.
3. Reshape the dataset:
    - Use pivot_longer() to gather year columns into a single Year column.
    - Use pivot_wider() if you want to restructure Series IDs and Series Names.
4. Handle Units of Measure:
    - Some values look unusually small, large, or inconsistent.

- Think about how you could incorporate units into the dataset (e.g., create a Unit column or keep the metadata separately).
- You do not need to fully standardize units, but be ready to explain how you dealt with this problem.

Part 2: Movies Data Dataset (movies.csv)

The dataset includes movie titles with genres stored in one column.

Example:

movieId	title	genres
1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
2	Jumanji (1995)	Adventure|Children|Fantasy

This violates tidy data principles because multiple genres and titles (year) are crammed into one cell.

Tasks

1. Import the dataset into R.
2. Use appropriate tidy functions taught in the class to get the different information in one cell into their own cells.
3. Explore other ways the dataset may need tidying (e.g., column naming, handling missing values).

Extended Analysis with MovieLens Data

To deepen your tidyverse practice, you will also use the related MovieLens files:

1. ratings.csv (user ratings of movies)
2. tags.csv (user-generated tags)
3. links.csv (external identifiers for IMDb and TMDb)

More information about these datasets is provided in README.txt

Suggested Explorations

- **Ratings (ratings.csv)**

1. Compute the average rating and number of ratings per movie.
2. Explore the distribution of ratings across all users.
3. Convert timestamp into readable dates and analyze trends over time.

- **Tags (tags.csv)**

1. Find the most common tags.
   - Explore how tagging activity varies across users and movies.
2. Join with ratings to see if certain tags are linked with higher ratings.

- **Links (links.csv)**

1. Check for missing identifiers (NA in imdbId or tmdbId).
2. Optionally, discuss how these IDs could allow linking to richer external datasets.

Cross-dataset analysis

1. Join ratings.csv and movies.csv to explore genre—rating patterns.
2. Identify top-rated movies (apply a minimum number of ratings filter).
3. Compare user preferences by genre.
4. Explore whether tags like classic or underrated correlate with higher ratings.

Submission

Submit an HTML (.html) or Quarto (.qmd) file showing your code, explanations, and outputs.

Your document should include:

1. A short description of how you identified messy aspects.
2. The tidy transformations you applied (with code).
3. A final cleaned dataset for each case.

(Optional but encouraged) Extra insights you discovered through joining datasets or summarizing with tidyverse.

Advice:

- Always check the bottom of CSV files for metadata that may need to be cleaned out.
- Remember the three rules of tidy data:
  - Each variable forms a column.
  - Each observation forms a row.
  - Each value has its own cell.
- Be creative in handling units—there is no single “right” way, but your approach should be documented.
- When joining datasets, always check for missing values and duplicated IDs.

- You can use AI tools or help from your friends, but remember to cite.