Data Cleaning Practice Datasets (Easy → Hard)

This document contains a curated set of messy datasets designed to practice NumPy, Pandas, and real-world data cleaning skills. The datasets are arranged in increasing order of difficulty.

Dataset 01: Titanic Dataset (Beginner)

Download Link: https://www.kaggle.com/c/titanic/data

Description: Passenger information from the Titanic including survival status, age, gender, and class.

Messiness Level: Low (Beginner)

Problems Present:

- Missing values in Age and Cabin columns

- Categorical columns such as Sex and Embarked

- Irrelevant columns like PassengerId and Ticket

Practical Cleaning Goals:

- Identify and count missing values

- Fill missing ages using mean or median

- Encode categorical columns

- Drop unnecessary columns

- Save a cleaned CSV file

Dataset 02: Students Performance Dataset (Beginner → Medium)

Download Link: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

Description: Student exam scores along with demographic and parental education data.

Messiness Level: Low to Medium

Problems Present:

- Long and inconsistent column names

- Categorical values requiring normalization

- Requires grouping and aggregation

Practical Cleaning Goals:

- Rename columns for clarity

- Standardize categorical values

- Group data by gender and parental education

- Calculate average scores

- Export cleaned dataset

Dataset 03: Netflix Movies and TV Shows (Medium)

Download Link: https://www.kaggle.com/datasets/shivamb/netflix-shows

Description: Netflix catalog containing movies and TV shows with metadata such as cast, director, and release date.

Messiness Level: Medium

Problems Present:

- Missing values in director and cast columns

- Dates stored as strings

- Multiple values stored in single columns

- Duplicate records

Practical Cleaning Goals:

- Convert date strings to datetime format

- Handle missing values logically

- Split multi-value columns

- Remove duplicate rows

- Normalize dataset structure

Dataset 04: US Traffic Accidents Dataset (Medium → Hard)

Download Link: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

Description: Large-scale traffic accident records across multiple US states and years.

Messiness Level: Medium to Hard

Problems Present:

- Very large dataset size

- Datetime columns with inconsistent formats

- Boolean values stored as strings

- Presence of outliers and invalid values

Practical Cleaning Goals:

- Efficiently load large CSV files

- Parse and clean datetime columns

- Filter data by state and year

- Remove invalid records

- Optimize memory usage

Dataset 05: World Bank Indicators (Hard)

Download Link: https://data.worldbank.org

Description: Global economic and social indicators such as GDP, population, and life expectancy.

Messiness Level: High (Real-World Data)

Problems Present:

- Data split across multiple CSV files

- Large blocks of missing values

- Wide-format time series data

- Requires merging and reshaping

Practical Cleaning Goals:

- Merge multiple datasets

- Convert wide format to long format

- Handle missing time-series values

- Produce analysis-ready dataset