Datasets

July 31, 2025

0.1 CSV Datasets for Data Preprocessing practices

Below is a table of popular, beginner-friendly datasets widely used to teach data preprocessing and analysis skills. Each entry includes an accessible link for direct download or hands-on exploration.

Dataset	Description	Link
Iris Flower	150 floral samples, 3 species, 4 features. Ideal for learning classification.	Iris (Kaggle)
Titanic Passenger List	Survival prediction; age, gender, class—classic for exploring missing values and feature selection.	Titanic CSV
Wine Quality	Red/white wine attributes with quality ratings; suitable for regression demos.	Wine UCI
Adult Census Income	Predict if income >$50K; covers encoding, scaling, and outlier handling.	Adult Census (Kaggle)
Breast Cancer Wisconsin	Binary class. Cell features for tumor detection. Data cleaning and binary encoding practice.	Breast Cancer (Kaggle)
Boston Housing	Predict Boston home prices; regression and feature engineering.	Boston Housing (Kaggle)
Superstore Sales	Fictional retail sales; groupby, pivot, and visualization skills.	Superstore (Kaggle)
Bank Marketing	Features from real bank campaigns; imbalanced data and categorical encoding.	Bank Marketing (Kaggle)
Online Shoppers Purchasing Intention	E-commerce browsing; practical for classification, handling categorical and missing features.	Online Shoppers (UCI)
Sample People/Products	Small, synthetic customer/product info. Good for first-time practice.	Sample Customers CSV (GitHub)

Dataset	Description	Link
Instacart Orders	Grocery purchase baskets for market basket analysis and recommendations.	Instacart (Kaggle)
Uber NYC Pickups	Taxi trip records; introduction to date-time and geospatial features.	Uber NYC Data (Kaggle)
Citi Bike Data	Public bike trips; great for timelines and location mapping.	Citi Bike NYC (Official)
Household Electric Power	Long-term consumption time series; handle missing values and datetime formats.	Power Consumption (UCI)

All links provide direct access or download to CSV files. These datasets are safe, well-documented, and beginner-appropriate—making them perfect for in-class demos, assignments, or independent experimentation.

Save this formatted text as a `.md` file for easy sharing or use in classroom materials, Jupyter notebooks, or documentation.