Datasets & Practice Platforms - for Python Data Scientist, AI, Machine Learning Practitioner

Getting your hands on **datasets** and knowing where to practice is crucial for building skills as a **Python Data Scientist, AI, and Machine Learning Practitioner**. Here's a **curated list of platforms and datasets** (organized by use case):

📊 General Purpose Datasets

- **Kaggle Datasets** – Huge collection across domains; also lets you practice with notebooks & competitions.
- **UCI Machine Learning Repository** – Classic datasets like Iris, Wine, Adult Income, etc.
- **Google Dataset Search** – Search engine for datasets worldwide.
- **Data.gov** – US Government open datasets.
- **Awesome Public Datasets (GitHub)** – Curated list across multiple domains.

🛡️ Healthcare / Medical Datasets

- **PhysioNet** – Physiological & clinical data (ECG, ICU data).
- **MIMIC-III** – ICU patient data (needs credentialing).
- **Kaggle Healthcare Datasets** – Medical images, patient data, genomics.

💰 Finance & Business Datasets

- **Yahoo Finance API** – Stock market data (easy Python integration).
- **Quandl** – Financial, economic, alternative data.
- **World Bank Data** – Global development indicators.
- **FRED** – U.S. Federal Reserve economic data.

📱 NLP (Text & Language) Datasets

- **Hugging Face Datasets** – NLP-focused, ready-to-use with `datasets` library.
- **IMDb Reviews** – Sentiment analysis dataset.

- **20 Newsgroups** – Text classification.
- **Common Crawl** – Massive web-scraped dataset.
- **SQuAD (Stanford Question Answering Dataset)** – For Q&A models.

▓ Computer Vision / Image Datasets

- **MNIST** – Handwritten digits (classic).
- **CIFAR-10 & CIFAR-100** – Small object recognition dataset.
- **ImageNet** – Large-scale image dataset (used in deep learning benchmarks).
- **COCO Dataset** – Object detection, segmentation.
- **Open Images Dataset** – Large-scale annotated images.

🌍 Geospatial & Environmental Datasets

- **Google Earth Engine Datasets** – Satellite imagery & environmental data.
- **NASA Earth Data** – Climate, atmospheric, and remote sensing data.
- **OpenStreetMap** – Geospatial map data.

🏆 Practice Platforms for Projects & Competitions

- **Kaggle** – Competitions, datasets, kernels.
- **DrivenData** – Competitions with a social good impact.
- **Zindi** – Africa-focused AI/ML competitions.
- **Analytics Vidhya Hackathons** – Industry-focused challenges.
- **AIcrowd** – Research-oriented ML challenges.

✅ These platforms give you **end-to-end practice**: from data cleaning → model building → deployment.