# Dataset Platforms & Types of Data

In any DS or ML project, the first thing we need is a good dataset, there are many online platforms that provide datasets for free, and each platform offers different types of data depending on the purpose.
In this documentation, I have explained the commonly used dataset sources, the different types of data we usually work with, and where each type of dataset can be found.

**Platforms to Find Datasets:**

Below I have listed some of the most reliable and widely used platforms for datasets. These are commonly used by students, researchers, and data scientists.

- Kaggle >
  Kaggle is probably the most popular platform for datasets. It has thousands of datasets, Most datasets are available in CSV format and are easy to understand.
- UCI Machine Learning Repository >
  This is one of the oldest and most trusted dataset sources. Many academic papers and research studies use data from UCI. The datasets here are usually simple, clean, and great for beginners.
- Google Dataset Search >
  Google provides a dedicated search engine for datasets. Instead of hosting datasets itself, it shows datasets from different websites, used when looking for specific dataset
- Government Open Data Portals >
  Government portals provide officially recorded datasets from different sectors.
  These datasets are real-world and trusted since they come from government departments.
- GitHub >
  Many users upload datasets directly into their project repositories.

**Other Data Collection Methods:**

| SOURCE | DESCRIPTION |
|---|---|
| **1. SENSORS / IOT DEVICES** | Real-time data from devices (e.g., temperature, pressure, ECG sensors). |
| **2. 2.WEB SCRAPING** | Extracting structured/unstructured data from websites using scripts or tools. |
| **3. APIs** | Accessing external data via services (e.g. Twitter API, Google Maps API). |
| **4. DATABASES** | Structured data from SQL, or NOSQL databases. |
| **5. PUBLIC DATASETS** | Datasets made available by research or academic institutions (e.g., MNIST, CIFAR-10, ImageNet). |
| **6. MANUAL COLLECTION** | Surveys, questionnaires, human labeling, and experiments. |

**Types of Data:**

Understanding the types of data is important because different datasets need different methods and models.

- **Structured Data:** This Data is organized in rows and columns, this is the easiest type of data to work with.
  > Examples: Sales data, customer data, student records
  > Where to find: Kaggle, UCI, government portals
- **Unstructured Data:** This type of data does not have a clear format. Images, videos, and free-text documents fall under this category.
  > Examples: Photos, emails, audio recordings
  > Where to find: Kaggle, Google Open Images, GitHub projects
- **Semi-structured data:** this data is partially organized. It does not fit into rows and columns, but it still contains tags or structure.
  > Examples: JSON files, XML, log files
  > Where to find: GitHub, API outputs
- **Time-Series Data:** This is data collected over time with timestamps. It is used for forecasting and trend analysis.
  > Examples: Stock prices, weather data, sensor data
  > Where to find: Sensors, IOT Devices
- **Big data**: this refers to datasets that are extremely large and cannot be processed on a normal computer.
  > Examples: Satellite imagery, social media streams

**Conclusion:**

Having access to the right dataset is very important for any project in data science. Different platforms provide different types of datasets, and each type of data requires a suitable approach for cleaning, processing, and modelling. Understanding where to find datasets and the different categories of data will help in choosing the right dataset for any future project or assignment.