

Data Analytics Project

3. Data and Datasets



Aagam Deolasi



WHAT is Data & Dataset?

DATA: The *information collected and shared in various formats across the organizations* is known as Data.

DATASET: A *structured collection of data organized and stored together* for analysis or processing is known as a Dataset.

Datasets can be *Private Datasets* or *Public Datasets*.

Examples of Datasets: Listing of *all real estate sales* in a specific geographic area, data on *regional air quality*, etc.



2

PRIVATE Datasets

A private dataset is **restricted to specific authorized users or organizations with limiting access for confidentiality** or proprietary reasons.

EXAMPLES:

1. Bank Transactions Dataset.
2. Hospital Dataset with Patient Records.
3. Any organization-level research data. etc.



Aagam Deolasi

3

PUBLIC Datasets

A public dataset is **accessible to anyone, often shared openly on the internet** for public reference and use.

It is also known as ***Open Source Datasets***.

EXAMPLES:

1. Kaggle Dataset
2. Data.gov Dataset
3. Github Public Repositories. etc.



Aagam Deolasi

4

METADATA

The data that **provides information about other data, offering insights into the content, structure, and context of a dataset** known as Metadata.

Metadata is one of the most crucial information used by analysts **for understanding the dataset more accurately.**



Aagam Deolasi

Common Data FORMATS

1. Textual Formats:

- a. Comma Separated Values ([CSV](#))
- b. JavaScript Object Notion ([JSON](#))
- c. eXtensible Markup Language ([XML](#))

2. Binary Formats:

- a. Excel ([XLSX](#))
- b. Parquet (For Big Data)

3. Markup Languages:

- a. YAML Ain't Markup Language ([YAML](#)) - It is a *human-friendly data serialization* language.
- b. Hypertext Markup Language ([HTML](#))



THANK YOU!!! FOR YOUR SUPPORT! For now...

Keep Learning, Keep Sharing, & Keep Following
Aagam Deolasi.



Aagam Deolasi

