Sources of data

Data by data types

There are two types of variables you’ll find in your data- numerical and categorical. Numerical data has two subtypes: continuous or discrete data types. Categorical data also has two subtypes: nominal and ordinal

Numerical data type

Numerical data is data presented as numbers, not words or text(string). There are two types of numerical data. Continuous data is data that can take on any value between two numbers and thus can be meaningfully divided into finer parts (such as glucose readings). Discrete data can only take on certain values and cannot be divided into smaller parts (such as the number of students in a class).

Categorical data type

This is any data that isn’t a number, which can mean a string of text. These variables can be broken down into nominal and ordinal values, though you wont often see this done.

Ordinal values are values that have an inherent order to them. Examples or ordinal values include having a priority on a support ticket such as “critical” of “regular”. Nominal values are the opposite of ordinal values, and they represent values with no set order to them. Nominal value examples include variables such as “country” or “marital status”

In addition to ordinal and nominal values, there is a special type of categorical data called binary. Binary data types only have two values: yes or no. this can be represented in different way such as “True” and “False” or 1 and 0. Binary data is used heavily for classification models. Examples of binary variables can include whether a person has stopped their subscription service or not, or if a person bought a car or not.

Data by Structure

Unstructured data includes content such as video, email, images, podcasts, social media posts, and any free-form texts. It has no internal identifier to let search functions recognize it.

Structured data exists in a format that facilitates the capturing, storage, and subsequent data organization and analysis. Due to its inherent structure, structured data is capable of dealing with high volumes of information

In a data analyst and data scientist’s work, however, structured data and unstructured data coexist, and both play a role. They complement and enrich each other, and you often find insights in your unstructured data sets. Furthermore, structured data records can hold unstructured data within them. Consider a form that offers a question with a list of answers available in a drop-down menu, but also allows users to add free-form comments. The answers generated from the toggled list is structured data, but the comments field yields unstructured data.

Most data is a hybrid to some degree. For that reason, you may also see the term semi-structured data, which is a loosely defined subset of structured data. This format includes the capability to add tags, keywords, and metadata to data types that were once considered unstructured data. Adding descriptive elements to images, email, and word-processing files are examples of semi-structured data. Markup languages such as XML: are often used to manage semi-structured data

Sources of structured data

The two primary examples of where structured data is generated are databases and search algorithms.

The term structured data is often associated with RDBMSs, which was proposed by Edgar Codd in 1970 at IBMs San Jose Research Laboratory. Codd’s model organizes data into one or more tables (also known as relations) of columns and row. A few years later, fellow IBMers Donald D. Chamberlin and Raymond Boyce designed the SQL, which is used with the vast majority of relational databases.

In addition to relational databases, spreadsheets are also a common source of structured data. Whether it’s a complex SQL database or an Excel spreadsheet (because structured data depends on you creating a data model), you must plan for how you will capture, store, and access data. For example, will you be storing numeric, monetary, or alphabetical data?

Conversely, structured data is also generated by people to feed databases and spreadsheets. This type of structured data is created by humans who interact with computers and other devices. Examples include form data generated through interaction with online surveys, kiosks, games, and so on.

Sources of unstructured data

Unstructured data is created in two ways: the first is machine-generated data by devices or sensors without human intervention. Our devices that are connected to the internet generate a continuous stream of unstructured data, such as data generated by sensors (for example, GPSs, RFID tags, medical devices, data from networks and weblogs, and retail and e-commerce data). Unstructured data sources deal with data such as email messages, word-processing documents, audio or video files, collaboration software, or instant messages.