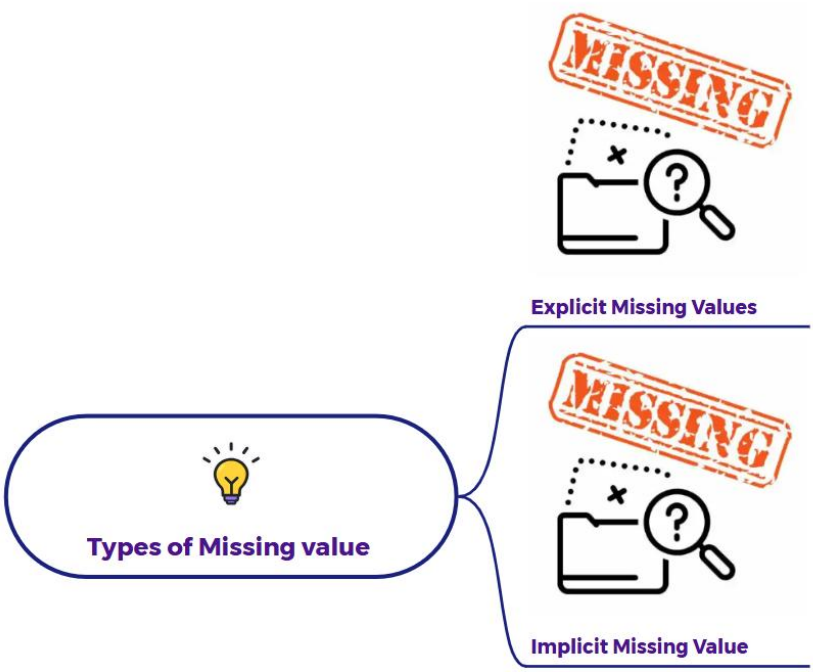
What are different types of Missing values?



We can categorize missing values into different types based on how they are represented in the dataset. Here are the common types of missing values you might encounter:

- **Explicit Missing Values:** These are values that are clearly and intentionally represented as missing using a specific placeholder. Common representations include:
    - **NaN (Not a Number):** This is a standard floating-point representation often used in programming languages like Python (with libraries like NumPy and Pandas) to indicate missing numerical data.
    - **None:** In Python, None is often used to represent the absence of a value, and Pandas can treat it as a missing value (converting to NaN in many cases).
    - **Specific Placeholder Strings:** Sometimes, missing values are represented by specific strings like "NA", "N/A", "Missing", or even empty strings (""). These need to be identified and treated as missing.

- **Implicit Missing Values:** These are values that aren't explicitly labeled as missing but should be considered as such based on the context or domain knowledge. They might appear as seemingly valid entries but actually signify missing information. Examples include:
    - **Sentinel Values:** Specific numerical values used to indicate missingness, such as -99, 0, or very large/small numbers that are clearly outside the expected range for that variable. For instance, if age is expected to be positive, a value of -1 might represent a missing entry.
    - **Empty Strings (""):** While sometimes a valid entry for a string variable, an empty string can also represent missing information, especially if the variable is expected to have some text content.
    - **Whitespace-Only Strings:** Strings containing only spaces, tabs, or newlines might be intended as missing values.
    - **Domain-Specific Codes:** Certain codes or abbreviations might be used within a specific domain to indicate missing information. For example, in a medical dataset, a code like "UNK" (Unknown) might represent a missing diagnosis.

Distinction based on the Nature of Missingness:

While not a "type" in terms of representation, it's crucial to remember the mechanisms of missingness **(MCAR, MAR, MNAR)** as they significantly influence how you should handle the missing values. Understanding *why* the data is missing is often more important than just *how* it's represented.

In summary, when dealing with missing data, you need to:

- Identify the representation: Determine how missing values are encoded in your dataset (e.g., NaN, "NA", -99, empty strings).
- Understand the context: Use domain knowledge to identify implicit missing values that might not be obvious.
- Investigate the mechanism: Try to understand why the data is missing (MCAR, MAR, or MNAR) as this will guide your handling strategies.

By recognizing these different types and understanding the reasons behind missing data, you can make more informed decisions about how to clean and prepare your data for analysis and modeling.