QUESTION 1

**Using Built-in Functions in Data Analysis Libraries:**

Many data analysis libraries, such as pandas in Python, provide built-in functions specifically designed to identify missing values. These functions typically scan through the entire dataset and report the number of missing values in each column. They can identify standard representations of missing data like NaN (Not a Number) or None. This method is quick, efficient, and often the first step in exploring missing data in a dataset.

**Visual Inspection through Graphical Representations**:

Visualization techniques can offer intuitive insights into the patterns of missing data. Specialized libraries can create visual representations such as matrices or heatmaps that show where missing values occur across the dataset. These visualizations can reveal patterns in missing data, such as whether certain variables tend to have missing values together, or if missing values are concentrated in particular rows or columns. This approach is particularly useful for large datasets where manual inspection is impractical.

**Custom Functions for Non-Standard Null Representations:**

In some datasets, missing values might be represented by non-standard indicators such as 'N/A', 'NULL', empty strings, or special characters like '?'. A custom function can be created to search for these specific representations across the dataset. This method is crucial when dealing with datasets that use varied or unconventional ways to denote missing information. It allows for a more comprehensive identification of missing values that might be overlooked by standard functions.

**Statistical Approaches for Detecting Potential Missing Values:**

Some missing values might be encoded as extreme values in the dataset. A statistical approach, such as calculating z-scores for numerical variables, can help identify these potential missing values. By setting a threshold (e.g., z-score > 3 or < -3), one can flag values that are unusually high or low, which might indicate missing data encoded as extreme values. This method is particularly useful when missing data might have been replaced with arbitrary large or small numbers during data collection or processing.

**Domain-Specific Value Inspection:**

In certain fields, specific values are commonly used to represent missing data. For instance, in medical datasets, age might be coded as 999 when unknown, or in financial datasets, income might be coded as -1 for missing information. This approach requires domain knowledge to identify these field-specific placeholder values. By checking for the frequency of these known placeholder values in relevant columns, one can identify potential missing data that might not be caught by standard missing value detection methods.

QUESTION 2

1. Prediction Models: You can use machine learning models to predict missing values based on the other features in your dataset. This is especially useful when the missing data has a pattern and can be predicted accurately.
2. Deletion: In cases where the missing data is substantial and can’t be reasonably imputed, you may choose to delete the rows or columns with missing values. However, this should be done cautiously, as it may result in a loss of valuable information.
3. Mean/Median/Mode Imputation: If you have numerical features, you can replace missing values with the mean & median of that column. This is a simple approach and is effective when the missing data is missing completely at random (MCAR) or missing at random (MAR). Easy to implement, preserves variable mean/median. The downside is, it Does not preserve correlations, can distort variable distributions.