Handling Categorical Missing Data.

In the context of categorical data, we primarily focus on two techniques. First, we can replace missing values with the most frequent value. Second, we can create a new category named "missing" to represent the missing values. We apply the first technique by finding the mode of the column (e.g., Cities) and replacing the missing values with the mode value. This approach is suitable when dealing with Missing Completely at Random (MCAR) data.

Another technique we have is called missing category imputation. Let's consider a scenario where the dataset has a column named "city" with values A, B, and C, but a significant portion (e.g., 10% or more) of values are missing. In this case, simply replacing missing values with the mode may not yield accurate results. Instead, we create a new category named "missing" to represent the missing values. This approach ensures that our dataset has four categories. We employ this technique when instructing our machine learning algorithm on how to handle missing values in case they are provided as inputs.