Handling missing data in **categorical variables** requires different strategies than numerical data because you can't calculate a mean or median. The primary goal is to manage the missingness without creating artificial relationships or losing valuable information.

## Strategies for Handling Missing Categorical Data 🏷️

The most common and generally recommended strategy is to treat "missing" as its own category.

### 1. **"Unknown" Label Imputation (Flagging/Sentinel Value)**

| Theory | Description | Rationale |
| --- | --- | --- |
| **Sentinel Value** | Replace the missing value (NaN) with a new, distinct category label, such as **"Unknown"**, **"Missing"**, or **"Not Disclosed"**. | This method is highly effective because it **preserves all observations** and allows a machine learning model to **learn the predictive power of the missingness itself**. If people who don't report their Education Level behave differently than those who do, the model captures that. |

Example:

In a dataset where Marital Status has categories 'Single', 'Married', and 'Divorced', a missing value is replaced with 'Unknown'.

| Original Marital Status | Imputed Marital Status |
| --- | --- |
| 'Married' | 'Married' |
| NaN | **'Unknown'** |
| 'Single' | 'Single' |

### 2. Mode Imputation (Most Frequent Category)

| Theory | Description | Rationale |
| --- | --- | --- |
| **Mode Replacement** | Replace the missing value with the **mode** (the category that occurs most frequently) of that column. | This is simple and preserves the original distribution of the data somewhat. Best used when the missing data percentage is **very small (<2%)** and you assume the missing values belong to the most popular category. |

Example:

If 'California' is the most frequent value in the State column, all NaN values for State are replaced with 'California'.

**Caution:** If the percentage of missing data is high, mode imputation will artificially **inflate the frequency** of the mode, potentially leading the model to overemphasize that category.

### 3. Advanced Imputation (Predictive Modeling)

For more sophisticated handling of missing categorical data, especially if it is suspected to be Missing At Random (MAR), you can use predictive methods:

* **KNN Imputation:** Replace the missing category with the **mode** of the k nearest neighbors (closest observations) based on other features.
* **Classification Model:** Treat the column with missing data as the **target variable** and train a classification model (e.g., Logistic Regression or Decision Tree) on the complete rows to predict the missing categories.

**The "Unknown" Label strategy (Strategy 1) is generally the safest and most robust method for categorical data in data science projects, as it makes no assumptions about the true value and retains the information about missingness.**