Handling Missing Categorical Data | SimpleImputer

When your dataset contains **missing values in categorical columns** (like "Gender", "Country", "Category"), machine learning models can't work with them directly. We must fill (impute) these missing entries.

➢ Techniques Overview

Technique	Description	When to Use
Most Frequent	Replaces missing values with the **most common category** in that column.	When a category dominates (e.g., 80% of values are "Male")
Missing Category	Fills missing values with a **placeholder like "Missing" or "Unknown"**	When you want the model to learn from the fact that it was missing
Constant Imputation	Similar to missing category, just uses a **fixed custom value**	For consistency or model interpretability

➢ Example

Original Column:

Gender
Male
NaN
Female
Male
NaN

▶ **After Most Frequent Imputation:**

Gender
Male
Male
Female
Male
Male

▶ **After Missing Category:**

Gender
Male

Missing
Female
Male
Missing

Tip:

Always treat categorical imputations **before encoding** (e.g., OneHot or Label Encoding). Imputing after encoding can corrupt the feature structure.