## How to ascertain the success of missing value imputation

It's crucial to evaluate how well an imputation method performs to ensure that the imputed data doesn't distort your analysis or model. Here's a breakdown of criteria to consider before and after imputation:

### 1. For Numerical Variables

- **Original Distribution:**
  - **Before Imputation:** Examine the shape of the distribution (e.g., normal, skewed), the mean, variance, and any outliers.
  - **After Imputation:** The imputed data should ideally maintain a distribution similar to the original. Large deviations suggest the imputation method might be introducing bias.
- **Example:**
  - Suppose you have a dataset of customer ages with some missing values.
  - If the original age distribution is right-skewed with a mean of 35, your imputation method shouldn't produce a symmetrical distribution with a mean of 25.
- **Outliers**
  - **Before Imputation**: Identify potential outliers.
  - **After Imputation**: Check if the imputation method introduces new, artificial outliers or excessively smooths out existing ones.
- **Relationships with Other Variables:**
  - **Before Imputation:** Analyze how the variable with missing values correlates with other variables.
  - **After Imputation:** The imputation should preserve these relationships. For instance, if age was initially positively correlated with income, that correlation should still be present after imputation.

## 2. For Categorical Variables

- **Category Frequency Distribution:**
    - **Before Imputation:** Look at the proportion of each category.
    - **After Imputation:** The imputed category frequencies should be reasonably close to the original distribution. Avoid methods that over-represent certain categories.
- **Example:**
    - Imagine a "Product Category" column where, originally, 40% are "Electronics," 30% are "Clothing," and 30% are "Home Goods."
    - A good imputation method shouldn't change this to, say, 70% "Electronics," 10% "Clothing," and 20% "Home Goods."
- **Preservation of Relationships:**
    - **Before Imputation:** Examine how the categorical variable is related to other variables (e.g., using chi-squared tests for independence).
    - **After Imputation:** The imputation method should aim to maintain these relationships.
- **Example:**
    - If, before imputation, "Product Category" was related to "Customer Region" (e.g., more "Electronics" purchases in urban areas), this association should still be evident after imputation.

## Additional Considerations

- **Visualize:** Always visualize the data before and after imputation using histograms, box plots, or bar charts to compare distributions.
- **Statistical Tests:** Use statistical tests (e.g., t-tests for comparing means, chi-squared tests for comparing category frequencies) to formally assess differences before and after imputation.
- **Domain Knowledge:** Consider whether the imputed values make sense in the real-world context.

- **Model Performance:** If the imputed data is used in a machine learning model, the ultimate test is whether the model's performance improves or remains consistent after imputation.