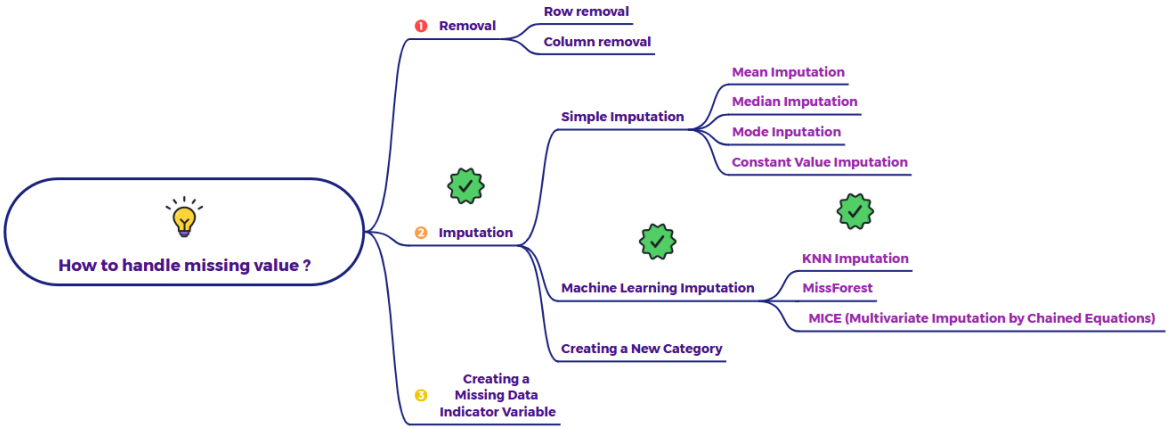
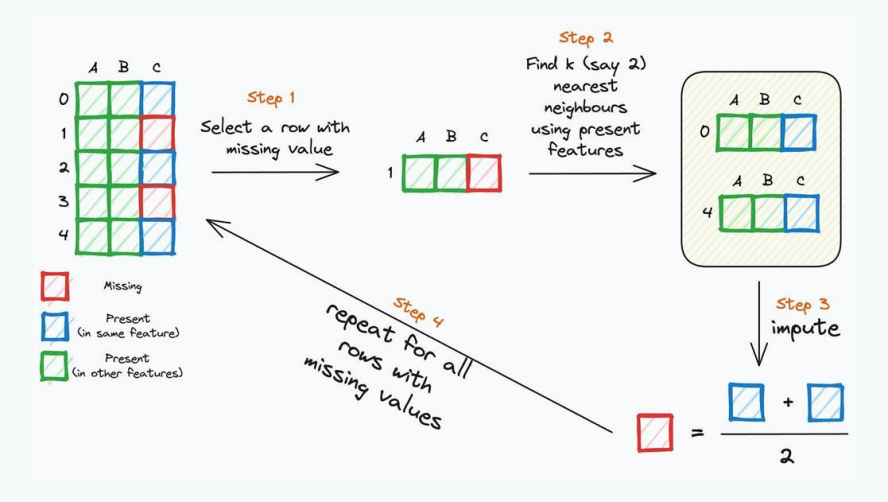
# Explain KNN Imputation with an example




## What is KNN Imputation?

KNN imputation is a machine learning-based method for handling missing values. It imputes the missing values in a feature for a particular data point by looking at the 'k' most similar data points (its nearest neighbors) in the feature space and then using the values of those neighbors for imputation. The similarity is typically determined by a distance metric (like Euclidean distance) calculated based on the other *present* features in the dataset.




## How it Works:

**Step 1: Select a row with a missing value.**

- Identify a data point (row) where one or more features have missing values (represented by red boxes in the image).

**Step 2: Find k (say 2) nearest neighbors using present features.**

- For the row with the missing value, find the 'k' (in the image, k=2) most similar rows in the dataset based on the *other* features that have present (non-missing) values (represented by green and blue boxes).
- The similarity is calculated using a distance metric across these present features. Rows with more similar values in the present features are considered closer neighbors.

**Step 3: Impute.**

- Once the 'k' nearest neighbors are identified, the missing value in the selected row is imputed based on the values of the same feature in those neighbors.
- **For numerical features:** The missing value is typically imputed by the average (mean) or median of the values of that feature in the 'k' neighbors. The image shows an example of averaging the values from the two neighbors.
- **For categorical features:** The missing value is typically imputed by the mode (most frequent category) of that feature among the 'k' neighbors.

**Step 4: Repeat for all rows with missing values.**

- This process is repeated for every data point in the dataset that has missing values.

**Example:**

Let's say you have a dataset of houses with features like 'Area (sq ft)', 'Number of Bedrooms', and 'Price ($)'. Some houses have a missing 'Price'. We can use KNN imputation to estimate these missing prices.

**Original Data:**

| House ID | Area (sq ft) | Bedrooms | Price ($) |
|---|---|---|---|
| 1 | 1500 | 3 | 250000 |
| 2 | 2000 | 4 | 350000 |
| 3 | 1800 | 3 | NaN |
| 4 | 1600 | 2 | 280000 |
| 5 | 2200 | 4 | NaN |
| 6 | 1700 | 3 | 300000 |

Let's say we want to impute the missing 'Price' for House ID 3 (Area=1800, Bedrooms=3) using k=2 neighbors.

1. **Select row with missing value:** Row 3 has a missing 'Price'.
2. **Find k=2 nearest neighbors using present features ('Area' and 'Bedrooms'):** We calculate the distance between House 3 and all other houses based on 'Area' and 'Bedrooms'. Let's assume (for simplicity, without showing the distance calculations) that the 2 nearest neighbors to House 3 based on 'Area' and 'Bedrooms' are House 1 (Area=1500, Bedrooms=3) and House 6 (Area=1700, Bedrooms=3).
3. **Impute 'Price':** Since 'Price' is a numerical feature, we impute the missing price for House 3 by taking the average of the 'Price' of its 2 nearest neighbors:

   Imputed Price (House 3) = (Price of House 1 + Price of House 6) / 2

   = (250000 + 300000) / 2

   = 550000 / 2

   = 275000

Now, let's impute the missing 'Price' for House ID 5 (Area=2200, Bedrooms=4) using k=2 neighbors. Let's assume its 2 nearest neighbors based on 'Area' and 'Bedrooms' are House 2 (Area=2000, Bedrooms=4) and (hypothetically) another house with similar area and bedrooms (not shown in the initial data, let's say Area=2100, Bedrooms=4, Price=380000).

Imputed Price (House 5) = (Price of House 2 + Price of Hypothetical Neighbor) / 2

= (350000 + 380000) / 2

= 730000 / 2

= 365000

**Data After KNN Imputation (k=2, hypothetical neighbors):**

| House ID | Area (sq ft) | Bedrooms | Price ($) |
|---|---|---|---|
| 1 | 1500 | 3 | 250000 |
| 2 | 2000 | 4 | 350000 |
| 3 | 1800 | 3 | 275000 |
| 4 | 1600 | 2 | 280000 |
| 5 | 2200 | 4 | 365000 |
| 6 | 1700 | 3 | 300000 |

### Key Considerations for KNN Imputation:

- **Choice of 'k':** The number of neighbors to consider is crucial. A small 'k' can be sensitive to noise, while a large 'k' might smooth out local variations.
- **Distance Metric:** The choice of distance metric (e.g., Euclidean, Manhattan) affects which neighbors are considered closest. Scaling features is important when using distance-based metrics.
- **Handling Mixed Data Types:** For datasets with both numerical and categorical features, appropriate distance metrics or separate handling for each type might be needed.
- **Computational Cost:** Finding the nearest neighbors can be computationally expensive for large datasets.
- **Assumption:** It assumes that similar data points will have similar values for the missing feature.

KNN imputation is a more sophisticated method than simple mean/median/mode imputation and can often provide more accurate estimates by leveraging the relationships between features.