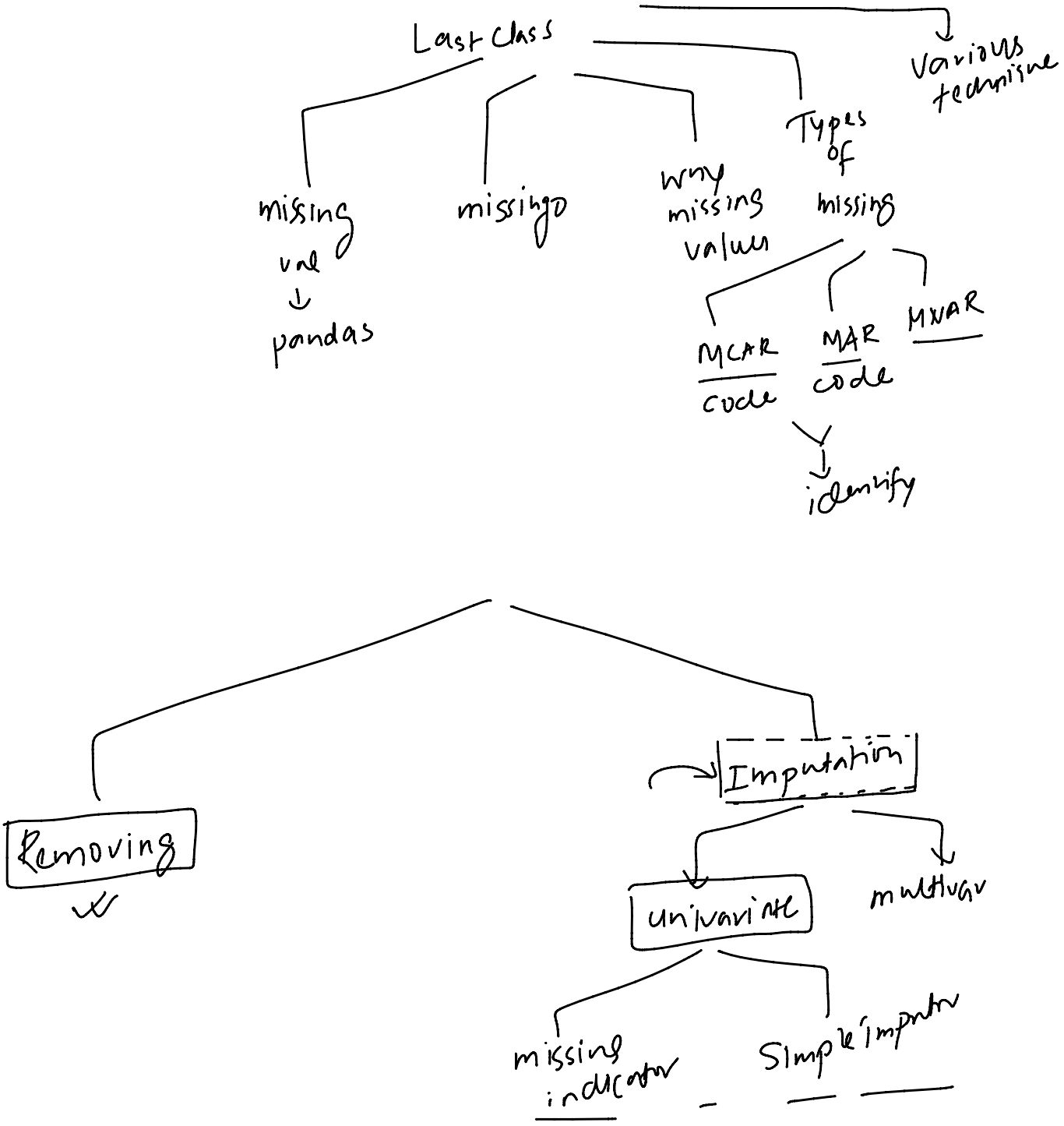
Recap

23 February 2024 09:26



Removing Missing Values

23 February 2024 09:26

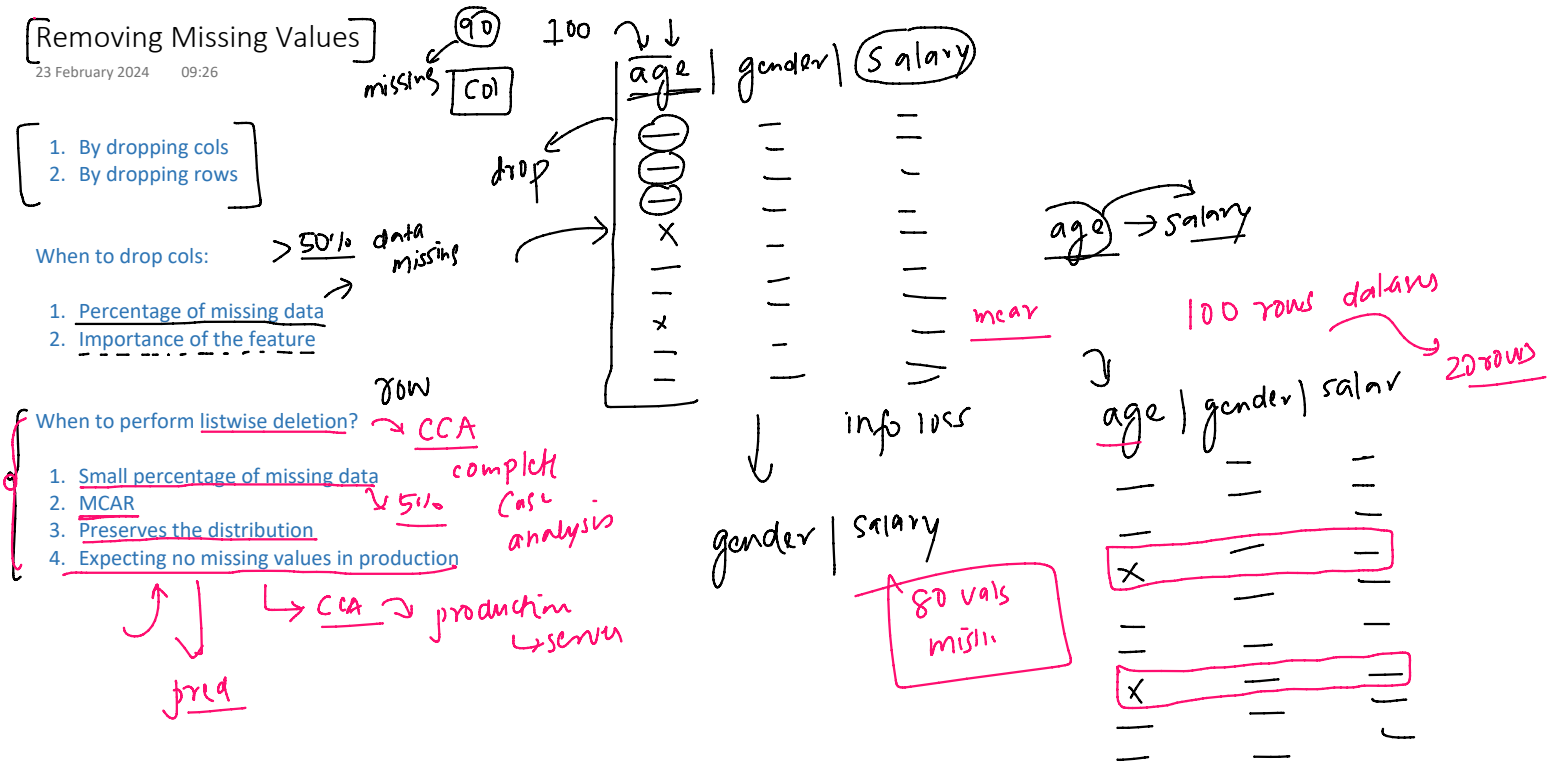
1. By dropping cols
2. By dropping rows

When to drop cols:

1. Percentage of missing data
2. Importance of the feature

When to perform listwise deletion?

1. Small percentage of missing data
2. MCAR
3. Preserves the distribution
4. Expecting no missing values in production

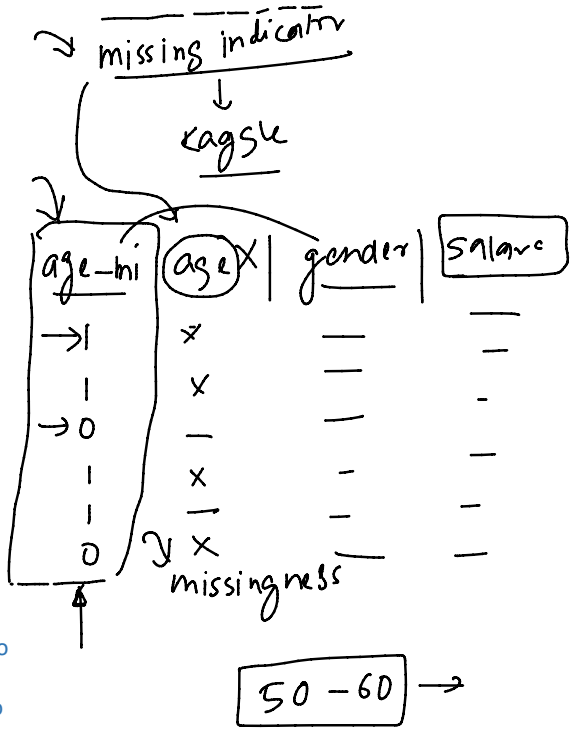


Missing Indicator

23 February 2024 10:22

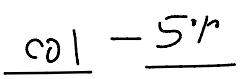
A missing indicator is an additional binary feature (variable) that indicates whether data was missing for a certain observation in another column. Here's how it works:

- For each feature (column) in your dataset with missing values, you create a new feature (column) that will have a binary value (often 1 or 0).
- In this new column, you assign a 1 if the data in the original feature is missing for a particular observation, and a 0 if the data is present.



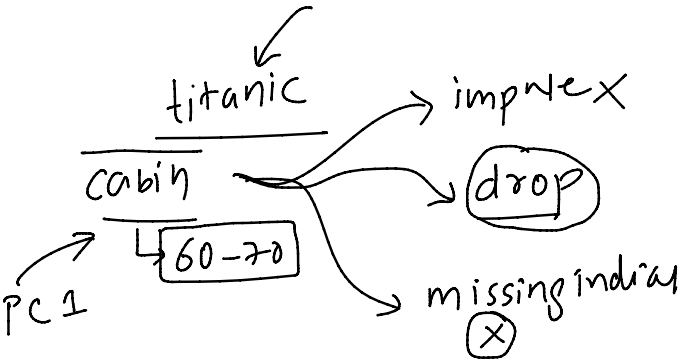
When to use?

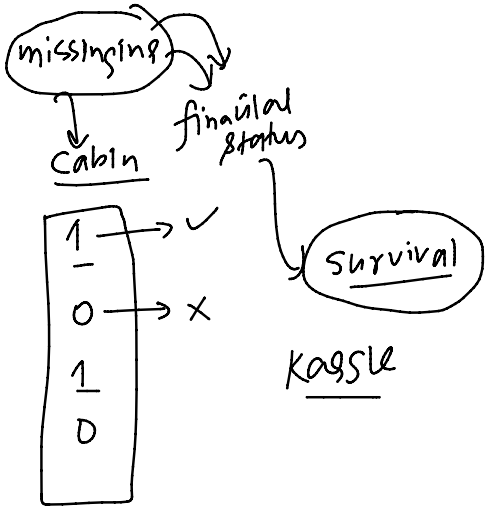
1. Missingness is Informative: If the absence of data itself carries information that might be relevant to your analysis or predictive modeling. For instance, in a medical dataset, the lack of a certain test result could indicate a decision based on the patient's condition, which might be predictive of the outcome of interest.
2. Data is Missing Not at Random (MNAR): When the probability of missing data is related to the missing data itself. For example, people with higher incomes might be less likely to disclose their earnings in surveys. In such cases, a missing indicator can help the model to account for the systematic absence of data.
3. High Prevalence of Missing Data: In scenarios where a significant portion of your data is missing.

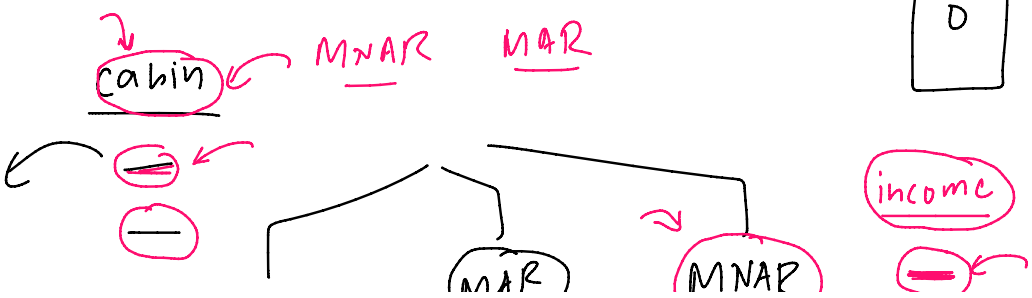


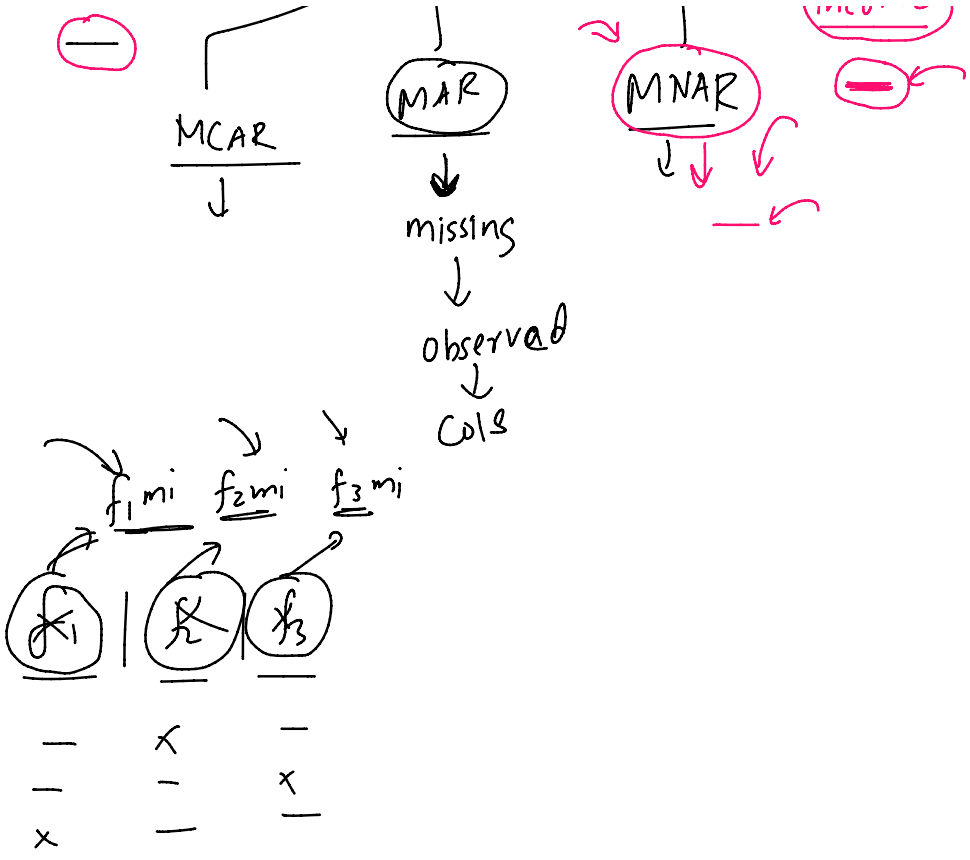
When Not to Use:

- Data is Missing Completely at Random (MCAR): If the missing data has no relationship with other data or the missing data itself, the addition of a missing indicator may not provide additional information to the model and could unnecessarily complicate the model.
- Minimal Missing Data: If the amount of missing data is negligible, the complexity added by introducing missing indicators might not be justified.
- Risk of Overfitting: Introducing many missing indicators in a dataset with lots of missing values across many variables can significantly increase the feature space, which might lead to overfitting, especially in smaller datasets.








MCAR
MAR
MNAR
missing
observed
cols
f1 mi
f2 mi
f3 mi

Simple Imputer - Mean & Median

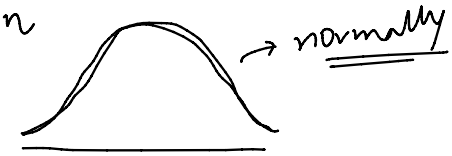
23 February 2024 10:22

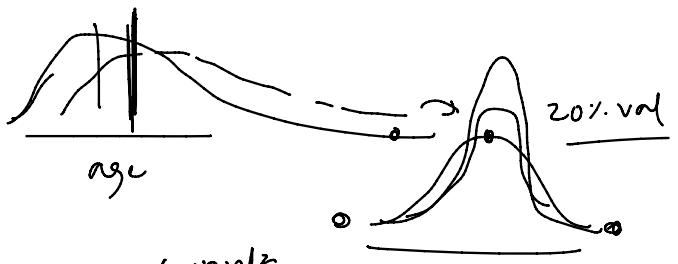
SimpleImputer

age

outlier → mean → median

non-mis → mean





When to use:

1. When missing values are MCAR.
2. Data is approximately normally distributed.
3. The proportion of missing data is small. → % → 20% missing → mean impute

When not to use:

1. Missing data is not MCAR - it might induce bias
2. Skewed data or outliers(use median instead) → median
3. If variance is reduced a lot → mean impu → variance
4. Introduces outliers sometimes(use median instead,) → outlier

numerical cols

mcar → mean/median

#data < 5%

outlier / skewed → mean X median

→ prediction → impute

Simple Imputer - Most Frequent

23 February 2024 09:29

num → mean / median

cat → mode → most_frequent category

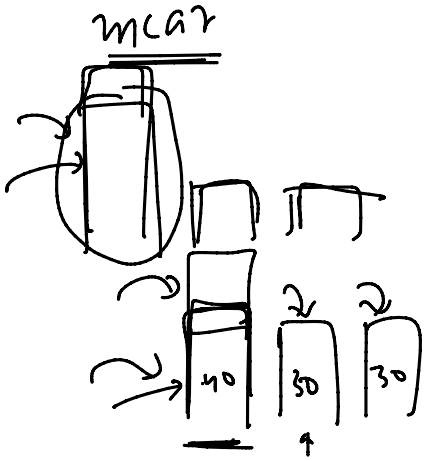
city → Mumbai → 40%, pune → 20%, kolkata → 20%

pune → pune

kolkata → mumbai

When to use

1. Less number of missing values
2. One dominant category
3. MCAR

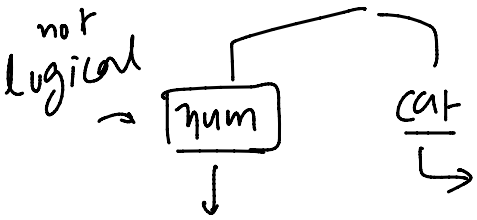


10%

10%

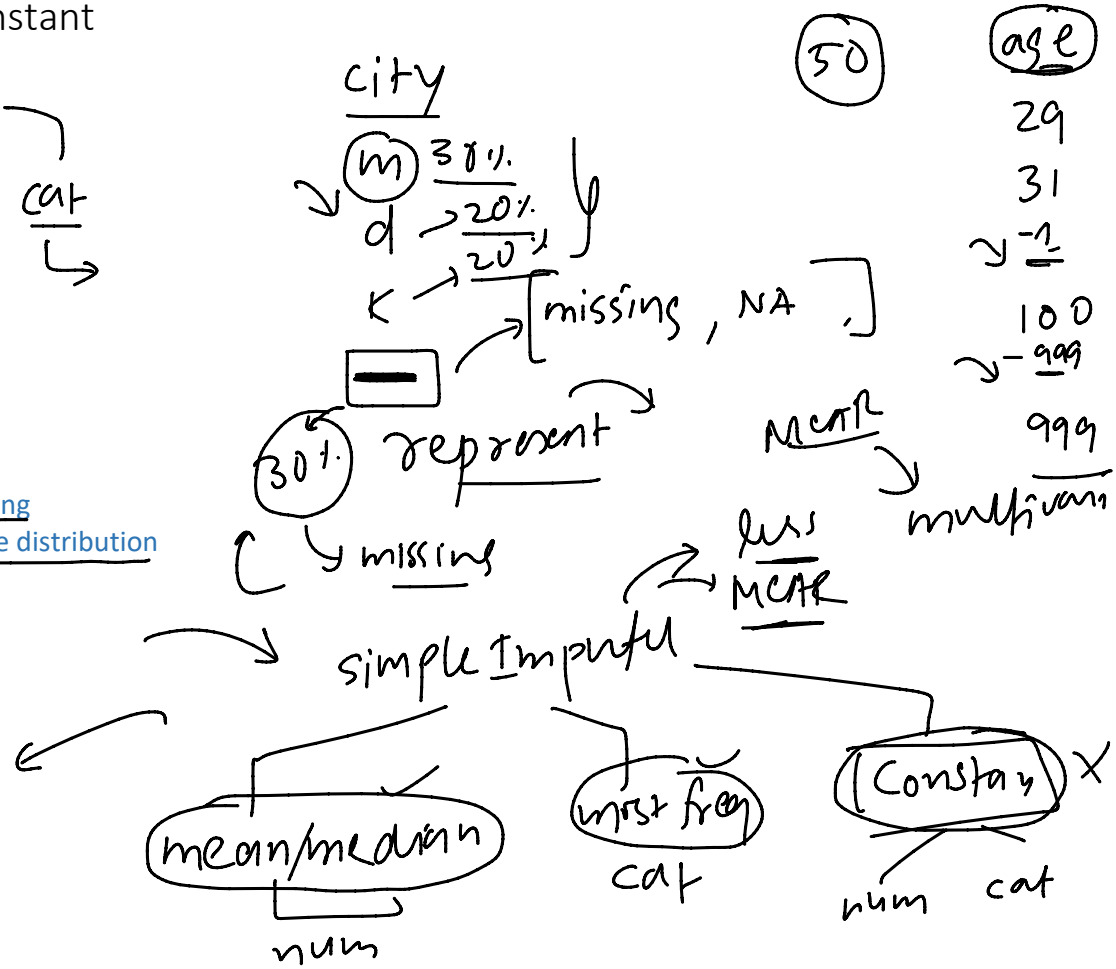
Simple Imputer - Constant

23 February 2024 09:29



When to use:

1. MNAR
2. Missingness has a meaning
3. You want to preserve the distribution



How to select the best

23 February 2024 10:23