Missing Value Imputation Method	Assumptions	Advantages	Disadvantages	Observations
Mean / Median imputation	Values are missing completely at random	• Easy to implement • Fast way of obtaining complete datasets	làm sai lệch • Distortion of original variance • Distortion of covariance / correlation with other variables within the dataset	Mean / Median imputation + adding a variable to indicate “missingness” is widely used in data science competitions. Mean should be used for Gaussian distributions and median otherwise. Although in practice most people replace by the mean regardless of the variable distribution.
Random sample imputation	Values are missing completely at random	• Easy to implement • Fast way of obtaining complete datasets • Preserves the variance of the original variable	• Randomness • Distortion of covariance / correlation with other variables within the dataset	Random sample imputation consist of taking a random sample of the variable where observations are available and using those to fill the NA. Not so widely used in data competitions, but it is used by businesses. Need to control randomness when scoring customers. Customers with same conditions should receive same treatment.
Adding a missing indicator	Missing data is predictive	• Easy to implement • Captures importance of “missingness” if there is one	• Increases feature space • May lead to similar or highly correlated added missing indicators	Mean / Median / Mode imputation + adding a missing indicator are widely used in data science competitions and in organisations.
End of tail / distribution imputation	Values are not missing at random	• Easy to implement • Captures importance of “missingness” if there is one	• Distorts the original distribution of the variable • If “missingness” is not important, it may mask the predictive power of the original variable • If the number of NA is big, it will mask true outliers in the distribution • If the number of NA is small, the replaced NA may be considered outliers and pre-processed in a subsequent step of feature engineering	Used by companies, who do not want to attribute to missing values the most common occurrence of the variable (mean / median). The rationale is that if the value is missing, it is for a reason, therefore, NA would not be replaced by the mean which makes them look like the majority of the observations. Instead, NA are flagged as different by assigning a value at the tail of the distribution, where observations are rarely represented in the population.
Arbitrary value imputation	Values are not missing at random	• Easy to implement • Captures importance of “missingness” if there is one	• Distorts the original distribution of the variable • If “missingnes”s is not important, it may mask the predictive power of the original variable by distorting its distribution • Hard to decide which arbitrary value to use • If the value is outside the distribution it may mask or create outliers	When variables are captured by third parties, like credit agencies, they place arbitrary numbers already to signal this “missingness”. So if not common practice in data competitions, it is common practice in real life data collections. Typical arbitrary numbers are 9999, -9999.
Frequent category imputation	Values are missing at random	• Easy to implement • Fast way of obtaining complete datasets	• Distortion the relation of the most frequent label with other variables within the dataset • May lead to an over-representation of the most frequent label if there is a big number of NA	This is the equivalent of mode imputation and it is used only for categorical variables (mode imputation is not normally used for numerical variables).
In categorical variables: treating NA as an additional category (‘Missing’)	None	• Easy to implement • Captures importance of “missingness” if there is one	• If the number of NA is small, creating an additional category may lead to an additional rare label	Method of choice, as it treats missing values as a separate category, without making any assumption on their “missingness”.