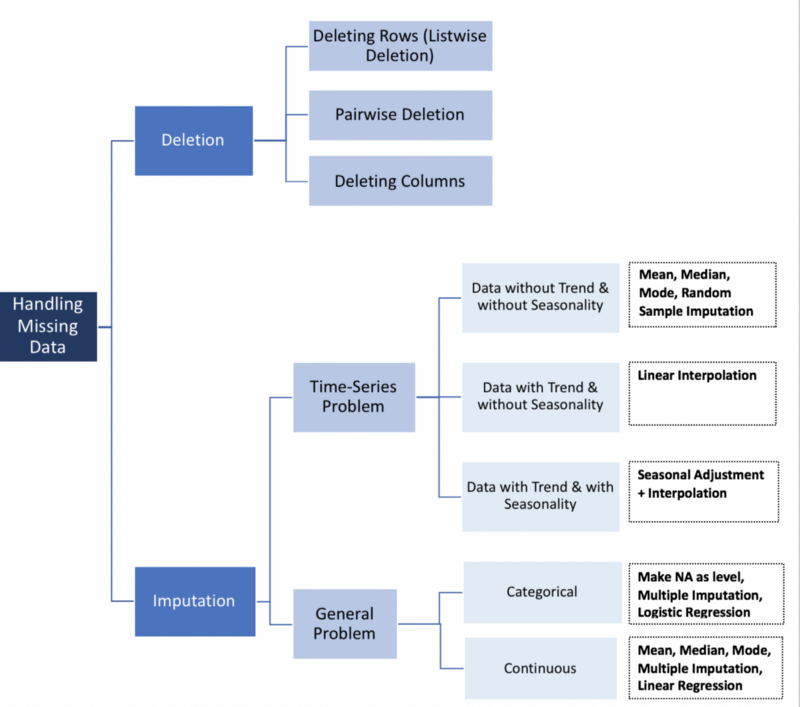
**Imputation Techniques**

Different solutions for data imputation depending on the kind of problem — Time series Analysis, ML, Regression etc.

**Why data goes missing”:**

1. **Missing at Random** :  For example: respondents of data collection process decide that they will declare their earning after tossing a fair coin.
2. **Missing completely at Random** : This is a case when variable is missing at random and missing ratio varies for different values / level of other input variables. For example: We are collecting data for age and female has higher missing value compare to male.
3. **Missing not a Random** :
   * 1. Missing that depends on unobserved predictors: This is a case when the missing values are not random and are related to the unobserved input variable. For example: In a medical study, if a particular diagnostic causes discomfort, then there is higher chance of drop out from the study. This missing value is not at random unless we have included “discomfort” as an input variable for all patients.
     2. Missing that depends on the missing value itself: This is a case when the probability of missing value is directly correlated with missing value itself. For example: People with higher or lower income are likely to provide non-response to their earning.

In the first two cases, it is safe to remove the data with missing values depending upon their occurrences. You can drop variables if the data is missing for more than 60% observations but only if that variable is insignificant.



1. **Time-Series Specific Methods**
   1. **Last Observation Carried Forward (LOCF) & Next Observation Carried Backward (NOCB)**
   2. **Linear Interpolation**  
       This method works well for a time series with some trend but is not suitable for seasonal data
   3. **Seasonal Adjustment + Linear Interpolation**  
      This method works well for data with both trend and seasonality

#### Mean, Median and Mode

#### ****Linear Regression****

several predictors of the variable with missing values are identified using a correlation matrix; Cases with complete data for the predictor variables are used to generate the regression equation

1. **Multiple Imputation**
2. **Imputation**: Impute the missing entries of the incomplete data sets *m* times (*m*=3 in the figure). Note that imputed values are drawn from a distribution. Simulating random draws doesn’t include uncertainty in model parameters. Better approach is to use Markov Chain Monte Carlo (MCMC) simulation. This step results in m complete data sets.
3. **Analysis**: Analyze each of the *m* completed data sets.
4. **Pooling**: Integrate the *m* analysis results into a final result

#### Imputation of Categorical Variables

* Missing values can be treated as a separate category by itself
* Prediction models
* Multiple Imputations

#### 5.1 KNN (K Nearest Neighbors)

The distance metric varies according to the type of data:  
1. Continuous Data: The commonly used distance metrics for continuous data are Euclidean, Manhattan and Cosine  
2. Categorical Data: Hamming distance is generally used in this case. It takes all the categorical attributes and for each, count one if the value is not the same between two points. The Hamming distance is then equal to the number of attributes for which the value was different

*Among all the methods discussed above, multiple imputation and KNN are widely used, and multiple imputation being simpler is generally preferred*