

Transformation cleaning

→ Cleans the dataset to fix quality and consistency issues.

A. Addresses null /missing values

→ 2 ways to do so:

1. Filter out records with missing/null fields
2. Imputation, or replacing the missing values using
 - **Statistical average**: Inserting the mean or median.
 - **Similar records**: Generating values for 'like' customers.
 - **Time series logic**: Using the last valid value in a sequence.

B. Addressing invalid values

→ Caused by inconsistency and ambiguity.

Remedy:

1. _Overwrite with calculated/consistent values.
2. Put it on record but mark it invalid so that two parallel analyses can take place. This helps by telling how much data is skewing your final result and impacting the data.

C. Data standardisation

→ **Edit distance**: Treating similar strings the same i.e (California and California).

→ **Domain knowledge**: Using external logic(area codes/ ZIP codes/ IP addresses) to fix entires like India(North/South).