

## Business Rules for Data Cleaning

### Customer Master Dataset

#### 1. Remove Duplicates:

- Drop exact duplicate rows based on all columns.

#### 2. Handle Missing Values:

- Drop rows with missing cust\_id, name, or address if essential.
- Impute missing city, state, and zipcode using mode or external reference.

#### 3. Zipcode Normalization:

- Ensure all zipcodes are 5 digits (U.S. format).
- Remove or correct zipcodes with 4 or 6 digits.

#### 4. Text Cleanup:

- Strip extra spaces in name, address, city, and state.
- Title case or proper case the text fields.

### Sales Dataset

#### 1. Remove Duplicates:

- Drop duplicate sales entries.

#### 2. Fix Missing Dates:

- If viewed\_datetime is missing, drop the row.
- If cart\_added\_datetime or purchased\_datetime is missing, impute using average time gap from viewed\_datetime.

#### 3. Fix Negative Values:

- Remove or replace negative quantity or unit\_price.
- Replace with median/mean or flag for further analysis.

#### 4. Category & Item Cleanup:

- Remove rows with null item or category.
- Standardize spelling or naming conventions.

#### 5. Datetime Validation:

- Ensure the sequence: viewed\_datetime < cart\_added\_datetime < purchased\_datetime.