Cleaning & Exporting Data

## Types of Dirty Data:

# Dirty Data is simply customer or business information that is corrupt, missing, duplicated, or inaccurate. The following are the most common types of dirty data:

* Duplicate Data.
* Outdated Data.
* Insecure Data.
* Incomplete Data.
* Incorrect/Inaccurate Data.
* Inconsistent Data.
* Too Much Data.

## Cleaning steps for SuperFoodsMax Dataset:

# **Validity:** 1633 duplicate records found and deleted using *pandas.DataFrame.drop\_duplicates()* method.

**Accuracy:** Dropped the 19 inaccurate entries using pandas methods from the dataset whose customer\_id and basket\_id are 7 digits against the usual 5-digit entries with an extra trailing ‘00’.

**Completeness:** Customers over 19 are grouped into different age bands but the dataset doesn’t cater to 65-year-old. ‘Greater than 65’ is renamed to ’65 and above’ to bring completeness.

**Consistency:** Transaction dates have been reset to the start of the month to make the data consistent for analysis & forecast.

**Uniformity:** A non-uniformity was observed with the 19-24 age band (with 6 years range) against the other age bands (with 10 years range) don’t seem to impact analysis.

I exported the cleaned data as dataset\_2017\_2020\_clean and used it for the project.