Data Quality Issues And Strategies

Customer Demographic Dataset

- 125 Missing values in the Last name column
- Gender column contains spelling mistakes: F,Female,Female, M,Male,U
- 87 Missing values in the DOB column
- A cell record says '1843-12-21' which is in text format and not Date. Also, this date means the person is 175 years old which is impossible.
- 83 missing values in the DOB column
- 481 missing values in Job_title column
- 631 entries marked n/a in the job_industry_category column
- Default column contains many dirty entries
- 83 blank cells in tenure column

Customer Address Dataset

- State column contains both full names and its abbreviations e.g. New South Wales and NSW.
- There is no need for the Country columns as all the states belongs to Australia
- In the Address Column, the initial digits are unclear, they are either representing the street no. or flat no. They are not really important and can simply be removed

Transactions Dataset

- Online Order Column contains 360 blank entries
- Brand,product_line,product_class,product_size,standard_cost and product_first_sold_date also contains blank entries.

Possible Strategies

- **Dropping** observations that have missing values if we can replace it on the basis of other observations
- The missing values can also be impute by mean, median or mode values in the column.
- Keep the Date column type as date and not custom.
- Missing **numeric data** can be filled in with say, 0, but has these zeros must be ignored when calculating any statistical value or plotting the distribution
- While **categorical data** can be filled in with say, "Missing": A new category which tells that this piece of data is missing.
- There should be a uniformity in the data column either use abbreviations or full form but not both.