**01 DATA PREPARATION AND QUALITY ASSESSMENT**

**General Remarks**

There are attributes in Customer Address and Transactions table that are not in the master table of Customer Demographic. To mitigate this issue, new attributes from other tables must be matched to the master table accordingly, and ensure that all customers are accounted for, i.e. no transactions where customer detail does not exist in master table.

**Value Inconsistency**

* Inconsistent values such as use of ‘M’ or ‘F’ for gender instead of ‘Male’ or ‘Female’. To mitigate this issue, there can be a setup that automatically inputs the complete word when the initials are entered. In addition, the method of data collection for this attribute can be a multiple choice-type.
* Unknown or blank values in gender, occupation and first and last names. This could be an issue when looking up a customer by name because of ambiguity. To mitigate this issue, there is a need to ask the customer more information such as industry or occupation to identify them.
* Error in data input such as birth year of 1843, which falls outside reasonable value; misspelling of ‘Female’ in a particular input. These can be solved by doing a allowable value check and error checks.

**Transactions Dataset**

* Approved transactions with missing values for product line, product class, brand and product size with product id of 0. If these are proven not to be actual transactions, they may be removed from the dataset.
* Missing values for online order attribute.
* Inconsistent product ID number with similar product brand, line, class, size. There is a need to figure what product ID stands for, whether it varies for certain traits such as color to know whether product ID should remain or be corrected accordingly.

**Customer Address**

* Inconsistent values entered for the ‘state’ attribute such as ‘NSW’ and ‘New South Wales’ or ‘VIC’ and ‘Victoria’. Can be solved by setting up automatic input/ correction. Data collection can be done in multiple choices to select to ensure consistency.
* Questionable house/ lot number of ‘0’ in many addresses. Setup of allowable value check.
* Questionable large house/ lot number in many addresses such as ‘93405’ or ‘02663’. Use of automatic address generator with validity check to mitigate this issue.

## 02 DATA ANALYSIS STRATEGY

**2.1 Data Cleansing**

**Variable identification.** Categorising variables into discrete or continuous, descriptive or numerical. This is important in the latter stages in determining the type of analysis to use (qualitative to quantitative etc.)

**Variance, correlation analyses.** To determine the behaviour of each individual variable or two.

**Missing values and outliers treatment.** Need for cleaning data to produce a more accurate model.

**2.2 Feature Engineering**

**Variable transformation.** Ensuring variable types are appropriate for a certain visualisation or analysis. Determining context and purpose of the data to decide on i.e. discrete/ continuous, categorical or numerical issues. Dealing with skewed data by taking averages/medians or other measures when appropriate.

**Variable/ feature creation.** Creating dummy or new variables i.e. converting DOB to age.

**2.3 Model Development**

Creating hypotheses and determining essential variables to test the validity of hypotheses. Stating assumptions, estimations and the like.

**2.4 Interpretation**

Placing raw data into understandable graphs and charts, depending on variable type and purpose of data visualisation.

Help the business make data-driven decisions by pointing out insights from results.

Balance of qualitative to quantitative results in order to create solutions that have a human touch.

Example of data cleaning needed.

Filtering and segmenting demographics and traits.

Creating new variable “age” to better compare purchase patterns by age group.

Variable binning.