

Dear Client,

Thank you for choosing us and for providing us with the three datasets from Sprocket Central Pty Ltd.

Here is some problems in data set and we will work to fix it.

- 1- There is Additional customer\_ids in the 'Transactions table' and 'Customer Address table' but not in 'Customer Master (Customer Demographics)'

NOTE: Please ensure that all tables are from the same period. Only customers in the Customer Master list will be used as a training set for our model.

This indicates that the data received may not be in sync with each other which may skew the analysis results if there are missing data records.

- 2- Various columns, such as the brand of purchase, or job title, have empty values in certain records and you will find me colored the blanks with green color.

NOTE : If only a small number of rows are empty, filter out the record entirely from the training set for prediction. Else, if it is a core field, impute based on the distribution in the training dataset.

- 3- Inconsistent values for the same attribute

(e.g. Victoria being represented as "V", "Vic" and "Victoria" also F for female and M for Male)

NOTE : Use regular expression to replace extended values with abbreviations to ensure consistency across addresses.

Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field. In order to construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value.

Additionally, gender records where 'U' have been replaced based on the distribution from the training dataset.

- 4- Inconsistent data type for the same attribute

(e.g. numeric values for some fields and strings for others and date is text)

NOTE : Convert selected records in characters to numeric. Remove non-numeric characters from a string.

Recommendation: Ensure that fact tables in the given database have constraints on data types. Having different data types for a given field makes it difficult to interpret results at a later stage.

Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

Kindest Regards,

Alaa Radwan