

KPMG

Thomas Nguyen

13/07/2021

Dear Mr. A, manager of Analytics, Information & Modelling, Thank you for sending us the three datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the three datasets received. Please let us know if the figures are not aligned with your understanding.

Table name	No. of records	Distinct Customer IDs	Date Data Received
Customer Demographic	4000	4000	2021-07-14
Customer Address	3999	3999	2021-07-14
Transaction Data	20000	3494	2021-07-14

There are some notable data quality issues in the dataset. Detailed information and strategies to mitigate these issues are described below:

1. There are customer ID that appear in Transaction table and Customer Address table but not in Customer Demographic (eg. customer id 5034).
Mitigation: Please make sure all tables are from the same period. Only customer ID that are in Custom Demographic table will be used in further analysis.
2. Various columns contain missing values.
Mitigation: If the number of rows with missing values is small, the rows will be omitted in further analysis. If the records are important, rows will be impute based on the distribution of the data.
3. Inconsistent in recording data for the same attribute.
For example in Customer Demographic table, there are grammar errors and different ways of recording data in gender columns: (Male, Female, Femal, F, M, U). Mitigation: All values will be replaced using standard expression to ensure consistency (Male and M will be replaced with "Male"). Gender record with "U" will be imputed based on the distribution of the data.

Our team will continue with data wrangling, tidying and the transformatin process. Further questions and assumptions will be documented and discussed in future meetings. After we have completed it, both sides should hold a meeting to ensure that all assumptions are aligned with Sprocket Central's understanding.

Yours sincerely,

Thomas Winson.