Dear **Sprocket Central Pty Ltd,**
Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the three datasets received.

Table Name	No. of records	Distinct Customer İDs	Date Data received
Demograohic	4000	3415	10.10.21
Customer Address	3999	3999	10.10.21
Transaction Data	20000	19445	10.10.21

Notable data quality issues that were encountered and the methods used to mitigate the identified data
inconsistencies are as follows. Furthermore, recommendations have been provided to avoid the re-occurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

●
Additional customer_ids in the 'Transactions table' and 'Customer Address table'
but not in 'Customer Master (Customer Demographic)'
Mitigation: 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.

●
Various columns, such as the brand of a purchase, or job title, have empty values in
certain records
Mitigation: 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 distribution in the training dataset .
For key datasets, such as transactions, less than 1% of transactions (totalling less than 0.1% of revenue) have missing fields. These records have been removed from the training dataset.

●
Inconsistent values for the same attribute
(e.g. Victoria being represented as "V", "Vic" and "Victoria")(e.g Female gender being represented as F ,Femal,M gender being Male)
Mitigation: Use regular expression to replaced extended values into 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.

●
Appropriate data transformations are made to ensure consistent data types for a given field.Moving forward, the team will continue with the data cleaning, standardisation and transformation process
for the purpose of model analysis. Questions will be raised along the way and assumptions documented.
After we have completed this, it would be great to spend some time with your data SME to ensure that all
assumptions are aligned with Sprocket Central's understanding.

Kind regards,
Junior Consultant Nazrin Jafarova