

Dear Client,

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

Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows.

- Inconsistent values for the attribute (e.g. New South Wales being represented as “NSW” and “New South Wales”, Gender records for “Female” being represented as “Femal”, “F” and “Female”)

*Mitigation: Ensure consistency across addresses by using regular expressions. Enforce a drop-down list for the user entering the data rather than a free text field.*

- Various columns, such as the brand of a purchase, standard cost, product size 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 or fill out the missing values based on distribution in the training dataset.*

- Additional customer\_ids in the ‘Transactions table’ but not in other tables.

*Mitigation: Please ensure that all tables have the same set of customer\_ids and there are no missing records between tables. Only customers in the Customer Master list will be used as a training set for our model.*

In the next step, the team will continue with the data cleaning and analysis.

Kind regards,  
Sreejith S.