Dear [Client/point of contact],

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd.
I have gone through all the datasets and created a summary of the datasets. Please refer to the following information and let us know if there are any misaligned numbers.

Dataset	**No. of Records**	**Distinct Customer Ids**	**Date Data Received**
Customer Demographics	4000	4000	2021/07/07
Customer Address	3999	3999	2021/07/07
Transactions	20000	3494	2021/07/07

Data Quality issues have been listed below, we have also provided recommendations to avoid the recurrence of the data quality issues also to improve the accuracy of the data needed to drive decisions.

- Various columns such as job_title, product_class, online_order, standard_cost have missing values
  Mitigation - If a small number of rows are missing then remove those rows, if a large amount of rows data is missing then use imputation based on distribution of data such as mean, mode or median.
  For transactions dataset, less than 1% transactions have missing fields so they will be removed for prediction.

- State column has inconsistent values for NSW and VIC i.e New South Wales and Victoria.
  Mitigation - Use Regex to correctly identify the abbreviations of the data.
  Recommendation - Use dropdown for such fields

- Gender column has inconsistent values such as F, M, Femal.
  Mitigation - Use regex to identify correct values.
  Recommendation - Again use dropdown field rather than a text field.
  We will manipulate Gender value - 'U' with 'Undisclosed' to construct meaning.

- Default column is incorrect/inappropriate.
  Mitigation - Write a script to remove such columns which contain only special characters.

- There are Inconsistent data type for the same attribute (e.g. numeric values for some fields and strings for others)
  Mitigation: Convert selected records in characters to numeric. Remove non-numeric characters from the string.
  Recommendation: Ensure that fact tables in the given database have constraints on data types.Having different data types for a given field make it difficult to interpret results at the

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

We will continue with data cleaning, transformation process for data analysis. We are going to document all the questions and assumptions. After we have completed this we will update you about the same and verify the assumptions.

Warm Regards,
Keyur.