Dear Client,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The table below 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	4,000 *13 cols	4,000	2023/06/20
Customer Address	3,999 * 6 cols	3,999	2023/06/20
Transaction Data	20,000 * 13 cols	3,494	2023/06/20

Notable data quality issues that were encountered and the methods used to mitigate the indentified data inconsistencies are as follows. Recommendations and explanations have also been included to avoid further data quality issues in the future. Following recommendations will improve the accuracy of data used to influence business decisions of Sprocket Central Pty Ltd in the future.

- **Inaccurate data**
  (e.g. Customer Demographic table: DOB column was inaccurate, missing age column and profit column)

  Mitigation: Filtered out outliers in DOB. Add an age column in Customer Demographic table and a profit column in Transactions table

  *Recommendation: Create an **age column**, allowing for more comprehensible data and easier to check for errors. Create a **profit column** in **"Transactions" table** to check accuracy of sales.* Creating additional columns for age and profit will allow for easier identification of errors. The **profit column** will assist in future analysis.

- **Inconsistent data values for the same attribute**
  (e.g. NSW being represented as "New South Wales", "F/Femal" for "Female" and "M" for "Male")

  Mitigation: Use regular expression to replaced extended values into abbreviations to ensure consistency across address.

  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.

- **Columns have empty values in certain records**
  (e.g. product brand and details in Transactions table, job title in Customer Demographic table)

  Mitigation: If only a small number of rows are empty, filter out the record entirely from the dataset for prediction. Else, if it is a core field, impute based on distribution in the dataset. For key datasets such as Transactions, less than 1% of transactions have missing fields. This records have been removed from the dataset.

- **Inaccurate data type**
  (e.g. product_first_sold_date should be date type)

  Mitigation: change list_price and standard_cost into currency type, product_first_sold_date into short date type.

Recommendation: Ensure that fact tables in the given database have constraints of 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.

- **Data Relevancy**
  (e.g. default column in Customer Demographic)

  Mitigation: deleted meta-data in default column.

  Recommendation: Exclude unnecessary data would make data to be interpreted more easily. If this column is meaningful, data type need to be reviewed.

That summarises all data quality issues discovered through the first stage of the data quality analysis. The mitigation strategies suggested are simple and effective ways of improving data quality for future analysis. They will not only improve the analysis output that one can perform within the company but will increase the level of analysis that can be performed by KPMG and other hired analysis teams.

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 Sporcket Central's understanding.

Kind regards,
Stephy Zhu