Dear [Client point-of-contact],

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. 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	14/04/2023
Customer Address	3999	3999	14/04/2023
Transaction Data	20000	3494	14/04/2023

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 the Customer Master (Customer Demographic) table
    - **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.
      This indicates that the data received may not be in sync with each other which may skew the analysis results if there are missing data records.
    - **Recommendation**: Ensure that all tables are from the same period to avoid missing data records. Currently, the data reflects 12 months, not 3 months as per business requirements.
- Various columns have empty values. This can be observed with the ‘brand’ column in the transaction table, and ‘job_title’ in the Customer Demographic table.
    - **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 the distribution in the training dataset.
    - **Recommendation:** For key datasets, such as transactions, less than 1% of transactions (totalling less than 0.1% of revenue) have missing fields. These records should have been removed from the training dataset.

- Inconsistent values for the same data. This can be observed Customer demographic table with the ‘gender’ column, ['F' 'Male' 'Female' 'U' 'Femal' 'M']. And the ‘state’ column in the Customer Address table as ['New South Wales' 'QLD' 'VIC' 'NSW' 'Victoria'].

    - **Mitigation**: Use regular expressions to replace extended values into abbreviations to ensure consistency.

    - **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.

- Incostitant datatype for the same attribute and corrupt values. This can be observed the ‘default’ column in the Customer Demographic table. Some values are numbers, some are strings, and some appear to be corrupted text.

    - **Mitigation**: Convert selected records in characters to numeric. Remove non-numeric characters from a string. Also, Investigate the corrupted text to identify and correct any errors in the data.

    - **Recommendation:** Ensure that fact tables in the given database have constraints on data types.
      Having different data types for a given field makes it difficult to interpret results at a later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

- Investigate and correct the incorrect entries and missing values. This can be observed in the ‘product_first_sold_date’ column in the transaction column, as it shows the first product was sold 25 years ago, which is highly unlikely. As well as the ‘product_id’ 0 in the same table, where most items have empty attributes.

    - **Mitigation**: The data entry could have been due to human error or system error, therefore, the data needs to be checked for inconsistencies and corrected.

    - **Recommendation:** Verify the data source and the data entry processes. If the data is accurate, it may be worth removing the data record from the dataset. If the data is inaccurate, it may be worth correcting the data by using historical data. Additionally, it may be worth conducting more in-depth checks of the data quality to identify any other inaccuracies or inconsistencies that may have been introduced during data entry or processing.

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,
[Consultant Name]