Dear Marketing team manager,

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	February 21, 2023
Customer Address	3999	3999	February 21, 2023
Transaction Data	20000	3494	February 21, 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.

Table: Transactions:

- In the column “online_order“, “brand“, “product_line“, “product_class “, “product_size “, “standard_cost “, “product_first_sold_date “, there are some missing values.
  The values that may impact the results of analysis will be deleted or filtered out.
- In the column “list_price“, numbers after decimals are not consistent.
- In the column “standard_cost“, numbers after decimals are not consistent.
- In the column “product_first_sold_date “, dates are machine dates.

Table: CustomerDemographic:

- On the column “DOB”, the date value of client “Jephthah Bachmann” is not accurate
- On the “gender” column there are cells that have “U” letter instead of the gender type
- On the column “DOB” there are 86 date values missing.
- Missing data in the column “last_name”.
- Missing data in the column “job_title”.
- Missing data in the column “job_industry_category” indicated as ‘n/a’.

- Missing data in the column “tenure”.
- Missing data in the column “default”.
- two values refer to clients who have passed away in the column “deceased_indicator”
- Invalid characters in the column “default”.
- One duplicate row was found.

Table: CustomerAddress:

- Some addresses have the number “0” before them in the column “address”.
- Two customer ID values are missing in the column “customer_id”.
- In the column “state”, Some states are written with full names and some are written with initial letters.

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.

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.

Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses.

Convert selected records in characters to numeric. Remove non-numeric characters from string.

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.

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.

Moving forward, the team will continue with the data cleaning, standardization 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,

Ayman Shehab