Hello,

Thank you for sending the datasets from Sprocket Central Pty Ltd. I have reviewed the datasets and have emphasized the summary statistics of it in the table below.

Table Name	No. of Records	Distinct Customer IDs	Date Data Received
Transactions	20,000	3494	12th June 2023
Customer Demographic	4000	4000	12th June 2023
Customer Address	3999	3999	12th June 2023

Please let us know if these figures match your records and any other discrepancies in this.

Furthermore, I have explained my findings on the data quality with recommendations to improve it for phase two analysis.

- Empty Cells under different columns

  Under transactions, 197 of them are missing information on the brand, product line, product class, product size, and standard cost. In addition to this, the order type is missing in 360 transactions. Under customer demographic, data on date of birth and tenure is missing for 87 customers.
  Mitigation: *If the number of records with missing data is insignificant, it is better to remove them from the dataset. If the missing data is significant, compute and assign the value using the distribution in the training dataset.*
  Recommendation: *Fill out the other empty cells with 'n/a' or any other recognizable string rather than leaving it empty. If any data is missing under significant records, please refer to the client and fill in the gaps.*

- Missing units and currency

  Under the list price column, the currency of the transactions has not been mentioned.
  Mitigation: *Make sure all the prices mentioned have the right currency.*
  Recommendation: *It is assumed to be US Dollars (USD) as the standard cost is in terms of USD, however, it is better to confirm with the client and include it.*

- Inconsistent values for the same attribute

  Under the gender column, both male and female are represented by 'Male' or 'M' and 'Female' or 'F', respectively. In some cases, it is recorded as 'U'. Under the state column, the same state has been mentioned by its full name and an abbreviation at different places (Eg: New South Wales is represented as "NSW" and "New South Wales")
  Mitigation: *Replacing all the values with their abbreviations would make them consistent. In some cases, the gender has been represented as 'U'. Replace these values by using the distribution in the training dataset.*
  Recommendation: *Implementing a drop-down menu for fields with limited values will mitigate this issue.*

- Inconsistent data type for the same attribute

  The address field contains both strings and numeric values.
  Mitigation: *Convert all the characters to numeric to follow a consistent data type in a field.*
  Recommendation: *Ensuring all the fields have limitations on the data types will make it consistent and easier for further analysis.*

- Clarifications from the client
  - The difference in the number of records between the 'Customer Demographic' (master) and the other two sheets is pretty significant. As we would only use the master sheet for analysis, it would be nice to bridge any gap between these datasets.
  - The data under the 'product_first_sold_date' is very confusing as it does not follow any date format.
  - The data from the 'default' column does not have any relevant information about the customers. This should be clarified with the client to check we are not missing any important information.
  - It would be useful to clarify what exact information the 'property_valuation' column offers.
  - For one of the customers (Jephthah Bachmann), the date of birth given seems to be an error as it mentions the year 1843, and that is quite impossible.

Eventually, the team can continue cleaning the data with the above recommendations and further processes to begin model analysis. Any other questions or assumptions can be resolved along the way.

Thank you.

Kind Regards,

Jaswant Jayacumaar