Dear Client

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The summary table below highlights key quality issues that we discovered within the three datasets. Please let us know if you have any queries surrounding the issues presented.

Summary Table

Table Name	**Accuracy**	**Completeness**	**Consistency**	**Currency**	**Relevancy**	**Validity**
Customer Demographic	DOB: Inaccurate Age: Missing	Job Title: Blanks Customer id: Incomplete	Gender: Inconsistent	Deceased customers: Filter out	Default column: Delete	
Customer Address		Customer id: Incomplete	States: Inconsistent			
Transaction	Profit: Missing	Customer id: Incomplete Online order: Blanks Brand: Blanks			Cancelled status order: Filter out	List price: Format Product sold date: format

Below are more in-depth descriptions of Data quality issues discovered and methods of mitigation used are as follows:

Accuracy

- **Outlier in attribute date of birth from Customer Demographic Table and missing an age column; missing a profit column in Transaction Table**
  Mitigation: we can simply remove it from the dataset as there is no chance to have a customer of age 121.
  Recommendation: Create age column in Customer Demographic Table to check errors and Profit column for Transaction Table to check accuracy.

Completeness

- **Additional customer IDs in the Transaction Table and Customer Address Table**
  Mitigation: It seems like all tables are not from same period.
  Recommendation: Only customer ids from 1 to 3500 will be used as they have complete data.
  This indicate that the data received may not be in sync with each other which can affect the training set modelling.

- **There are various columns with empty records. Some of important attributes are job title in Customer Demographic, online order and brand in Transactions.**

Mitigation: filter out blanks for Job title, online order and brand
Recommendation: there are some other attributes having missing records which we will handle based on the distribution of data and percentage of missing values.

List of columns with empty records in **Transaction Table:**

Column Name	**Empty records**	**Empty records in %**
Online order	360	1.8
brand	197	0.985
Product line	197	0.985
Product class	197	0.985
Product size	197	0.985
Standard cost	197	0.985
Product first sold date	197	0.985

List of columns with empty records in **Customer Demographic Table**:

Column Name	**Empty records**	**Empty records in %**
Last name	125	3.125
DOB	87	2.175
Job title	506	12.65
Job Industry Category	656	16.4
tenure	87	2.175

Consistency

- **Inconsistent values in gender such as F & femal for female and in states VIC & V for Victoria.**
  Mitigation: filter all ‘M’, ‘F’ and ‘Femal’ under attribute gender for Customer Demographic and filter all ‘VIC’ and ‘New south wales’ under states for Customer Address
  Recommendation: Enforce a drop-down list for the users entering the data rather than a free text field
  In order to obtain meaningful insights from models, we have cleaned the data to avoid multiple representations for a single category. Also, we have replaced category U in gender from Customer Demographic Table based on the distribution of dataset.
- **Inconsistent data type for the same attribute like numerical values for some fields and string for others**
  Mitigation: convert selected records to numeric and remove non-numeric characters from string.
  Having different data types for a given field make it difficult to interpret the result, therefore data transformation is done to ensure consistent data type for a given field

Currency

- **People that are ‘Y’ in deceased indicator are not current customers for Customer Demographic Table.**
  Mitigation: filter out customers with ‘Y’ in deceased indicator.
  Recommendation: Once this information is received one should update data accordingly.

Relevancy

- **Lack of relevancy in default column for Customer Demographic and order status for Transactions.**
  Mitigation: Deleted metadata in default column and Filter out cancelled order status
  Cancelled order status is irrelevant information for future analysis, as it can skew data.

Validity

- **Format of list price, product sale date for Transactions Table.**
  Mitigation: format product sale date to short date, list price to currency.
  Recommendation: Set up columns so that formats of prices and decimals will be in place when entering new data.

Moving forward, the team will continue with data cleaning, standardisation and transformation process for the purpose of model analysis. After we have completed this, it would be great to spend time with your data SME to ensure that all the assumptions are aligned with Sprocket Central’s understanding.

Kindly regards,
Junior Consultant