Dear Manager,

This is Li Hui from KPMG Data Analytic Team (Virtual Internship). First of all, thank you for considering us and providing us with the three datasets from Sprocket Central Pty Ltd. We have carefully reviewed the datasets provided by your company based on our data quality assessment framework while we discover some issues and errors need to be corrected. Please let us know if the following figures are not aligned with your understanding.

Datasets	No. of Records	Unique customer_id	Date Data Received
Customer Demographic	4000	4000	20 Aug 2021
Customer Address	3999	3999	20 Aug 2021
Transaction	20000	3494	20 Aug 2021

Here are the data quality issues and methods used to mitigate the inconsistencies noted, which will avoid reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

- As all three datasets are relevant to each other, it is advisable to merge them into one table for our analysis

- Additional customer_ids in the 'Transactions table' and 'Customer Address table' but not in 'Customer Master (Customer Demographic)'
  Mitigation: Please ensure that all tables are from the same period.

- Empty values in various columns, such as DOB and job title for the "Customer Demographic" dataset, online order and brand column in the "Transactions" dataset.
  Mitigation: Filter out job title; brand; and online order columns.
  Recommendations: We can take a mode year value for the missing records of customers DOB while eliminate blank orders considering fake orders.

- Inconsistent values for the same attribute (e.g. Victoria being represented as "V", "Vic" and "Victoria")
  Mitigation: Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses.
  Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field.

- Lack of relevance in default column for "Customer Demographic" dataset and order status column in "Transactions" dataset.
  Mitigate: Delete meta-data in default column in "Customer Demographic". Remove the cancelled value from order status column in "Transactions".

The following are the details of error encountered in the dataset.

Customer Demographic (Total records 4000)

FIELD NAME	ERRORS
DOB	01 record 1843 87 records Blanks
last name	125 records Blanks
gender	88 records gender 'U' Values are not consistence M, Male, F, Female, Femal, U
job title	506 records Blanks
job industry	656 records mention 'N/A'
default	3317 records value 'special characters' includes null and Blanks
tenure	87 records Blanks

Transactions (Total records 20000 -past 3months)

FIELD NAME	ERRORS
online order	94 records Blanks
brand	48 records Blanks
product line	48 records Blanks
product class	48 records Blanks
product size	48 records Blanks
standard cost	48 records Blanks
product_first_sold_date	48 records Blanks

CustomerAddress(Total records 4000)

FIELD NAME	ERRORS
state	Combine New South Wales and NSW as NSW Combine Victoria and VIC as VIC

The recommendations and mitigation strategies are effective and easy to implement. The recommendations will improve data quality and make it more effective or efficient. Also these suggestions will improve data analysis output.

It would be great if we can spend some time with your team to ensure that all assumptions are aligned. Kindly contact us if you have any queries.

Best Regards,
Li Hui Han