Dear Sprocket Central Pty Ltd.,

I hope this email finds you well. My name is Pitiya Jansiri, and I am an intern at KPMG in the Analytics, Information & Modelling team. I wanted to provide you with a summary of the statistics obtained from the three datasets we received. The information is presented in the table below to align with our collective understanding

Table Name	No. of records	Distinct Customer IDs	Date Data Received
Customer Demographic	4,000	4,000	07/07/2023
Customer Address	3,999	3,999	07/07/2023
Transaction Data	20,000	3,494	07/07/2023

Upon conducting a comprehensive analysis of the dataset you provided, I wanted to provide you with an update on our progress in identifying data quality issues and share some important findings and recommendations.

- REDUNDANT OULINERS

Issue: In the customer DOB records from Customer Demographic table, Jephthah Bachmann was born in 1843, meaning that they are 175 years old.

Recommendation: If the extent of incorrect data is minimal, we can proceed with targeted measures to mitigate inaccuracies, limiting their impact on the analysis outcomes.

- MISSING VALUES

Issue: In Customer Demographic table found the blank and n/a cell in “job_industry_category” column, “job_title” column, “last_name” column and, “DOB” column. Also in Transactions sheet found the blank and n/a cell in “online_order” column, “brand” column, “product_line” column, “product_class” column, “product_size” column, “standard_cost” column, and “product_first_sold_date” column

Recommendation: If the percentage of missing values is minimal as compared to the whole dataset we can ignore or remove them. And also the correlated data, techniques can be used to approximate or infer missing values, improving dataset completeness for analysis.

- INCONSISTENT ENTRIES ACROSS THE DATASETS

Issue: Some customer IDs in the Demographic dataset don't have corresponding Transitions and Address data.

Recommendation: If these values don't significantly impact the analysis or require high resolution, they can be ignored. And please refer to the 'data_outliers.xlsx' file for a list of outliers between the datasets.

- UPDATE TIMELINESS OF DATASET

Issue: The transitions dataset you provided is based on information from 2017, but we are currently in 2023.

Recommendation: Updating the data to reflect the most current information.

- MULTIPLE DATATYPES AND WRONG DATATYPES

Issue: In Transitions Table, some record in “Standard Cost” column have special string characters and some record is integer. Also “product_first_sold_date” column are stored as integers instead of the expected date datatype.

Recommendation: Remove the special characters and convert all to true datatypes.

- INCONSISTENT VALUES FOR THE SAME MEANING

Issue: In the “State” Column of the Customer Address Table were found “VIC” and “Victoria” for Victoria, and “NSW” and “New South Wales” for New South Wales. Also in Customer Demographic Table were found "F", "Femal," and "Female" for females, and "M" and "Male" for males.

Recommendation: Choose a standardized meaning for data values.

Best regards,

Pitiya Jansiri