Dear Sprocket Central Pty Ltd ,

I hope this message finds you well. As part of our ongoing commitment to providing you with the highest quality data analysis and reporting, we have conducted a thorough assessment of the dataset with the Customer Demographic dataset, Customer Addresses dataset and the Transactions data in the past 3 months

We have identified several data quality issues that may impact the accuracy and reliability of the insights generated from this dataset. In this email, we will outline these issues and propose strategies to mitigate them.

This emails contains the data issues and mitigation strategies for each of the datasets

Data Quality Issues:

Missing Data: One of the most significant issues in this dataset is missing data. There are several records where critical information is absent, such as missing values in the "online_order," "order_status," "product_class" , job titles, industry categories, and wealth segments columns. This can lead to incomplete and unreliable analysis.

Mitigation Strategy:

We recommend a thorough data cleaning process to address missing data. This may involve imputing missing values or, in some cases, removing incomplete records if they cannot be reliably imputed

Data Consistency:

Inconsistent data formats and values were observed in various columns. For example, some columns have inconsistent date formats (e.g., "transaction_date"), and currency symbols are included in the "list_price" and "standard_cost" columns, gender data included both textual and binary entries. Some addresses include full descriptions, while others lack important details. Inconsistencies like these can cause errors in calculations and analysis.

Mitigation strategy:

We suggest standardizing date formats and removing non-numeric characters (e.g., currency symbols) from relevant columns

Data Accuracy:

There are discrepancies in the "product_size" column, where the size values are described as both "medium" and "low", the 'DOB' (Date of Birth) field contained invalid dates, and some 'wealth_segment' entries did not match the corresponding customer profiles. The 'state' and 'country' fields should ideally have consistent and standardized entries. However, we found variations in state abbreviations and country names.

This inconsistency can affect product categorization and related analysis.

Mitigation strategy:

To ensure data accuracy, we recommend validating the columns and other categorical columns for consistency. This involves reviewing and correcting any discrepancies in size descriptions.

Data Integrity:

The dataset appears to contain instances where the "order_status" is marked as both "Approved" and "Cancelled." It's essential to ensure the integrity of the data to avoid incorrect conclusions based on potentially canceled transactions.

Addressing these data quality issues will lead to more accurate reporting, better decision-making, and improved customer service.

We understand that addressing these data quality issues may require time and resources, but the benefits in terms of improved decision-making and customer satisfaction are well worth the investment

We are committed to working closely with you to ensure that these data quality issues are effectively addressed, and you receive reliable insights from your dataset. Please let us know if you have any questions or if you would like to discuss these strategies in more detail.

Thank you for entrusting us with your data analysis needs. We look forward to your feedback and the opportunity to assist you further.

Best Regards,