Dear Sprocket Central Pty Ltd,

I hope this email finds you well. Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. I am writing to discuss some data quality issues we have identified in our recent analysis of your data. We believe that addressing these issues is crucial for ensuring accurate insights and better decision-making.

After reviewing the data, we have identified the following quality issues:

Table Name	No. of records	Distinct Customer IDs	Date Data Received
Customer Demographic	4001	4001	23.03.2023
Customer Address	4000	4000	23.03.2023
Transaction Data	20001	3495	23.03.2023

During our analysis of the data, we have identified several notable data quality issues that require attention. We have utilized various methods to mitigate these issues, including data validation, correction of inconsistencies, and removal of duplicates.

- **Additional Customer IDs in the Transaction table but not in Customer Demographic**
  we need to insure that all tables are from the same period and to indicate if the data received may not be in sync with each other.
- **Various columns have empty values in certain records in the tables**
  It is important to address missing values as they can significantly affect the accuracy of data analysis, resulting in poor business decisions.
  Recommendation: If the number of empty rows is small, we recommend removing the entire record from the training set for prediction. However, if it is a critical field, it is recommended to impute the missing values based on the distribution in the training dataset.
- **Inconsistent data type for the same attribute**
  Inconsistencies in data types can arise due to various reasons, If left unaddressed, inconsistent data types can cause confusion and errors in data analysis, leading to poor decision-making.
  Recommendation: ensure that the data types for each column in the table are consistent and appropriate to the data they represent.
- **Inconsistent values for the same attribute (e.g. Female or male represented as “F” or “M”, and Victoria as “Vic”, “V”**

To ensure consistency across addresses, we recommend using regular expressions to replace extended values with abbreviations.
Recommendation: to avoid such data quality issues in the future is to enforce a drop-down list for the user entering the data rather than a free text field. This would allow us to control the input and ensure that the data is entered in a consistent format.

We understand that data quality issues can be complex and time-consuming to address, but we believe that the benefits of accurate data are essential for informed decision-making. We are committed to working with you to address these issues and improve the quality of your data.

Please let us know if you have any questions or concerns. We look forward to working with you to achieve your goals.

Best regards,

Sara Novak