

Dear [Client],

We appreciate you granting us access to the three Sprocket Central Pvt. Ltd. datasets. The key quality issues that we discovered across three data sets are listed in the Summary table that follows. Please get in touch with us if you have any inquiries concerning the problems mentioned.

Summary Table:

	Accuracy	Completeness	Consistency	Validity
Customer Demographic	- DOB : Inaccurate - Age : Missing	- Job Titles: Blanks - Customer ID: Incomplete	- Gender : Inconsistency	
Customer Address		- Customer id : Incomplete	- States - Inconsistency	
Transactions	Profit : Missing	- Customer id: Incomplete -Online order: Blanks - Brand: Blanks		- List Price : Format - Product Sold Date: Format

The difficulties with data quality that were found and the mitigation techniques that were implemented are described in more detail below. To prevent more problems with data quality in the future, recommendations and justifications have also been added. The data Sprocket Central Pvt. Ltd. will utilize to inform future business decisions will be more accurate as a result of the recommendations made.

Accuracy

- **The "Customer Demographic" part is missing an age_column, contains an inaccurate DOB, and lacks a profit_column for the "Transaction."**

Mitigation: Remove the outlier from DOB by eliminating the data.

Recommendation: To make data easier to interpret and to make it easier to check for issues, add an age_column. Add a profit_column to the "Transaction" table in order to confirm that the sales are accurate.

Prediction: The age and profit columns will make it easier to identify errors. For upcoming financial analysis, the profit_column will be helpful.

Completeness

- **Additional customer_ids in the fields "Customer Demographic," "Customer Address," and "Transaction" were inconsistent.**

Mitigation : Ensure that all tables are current.

Recommendation: We advise against using the data since it contains missing numbers that can skew the findings. Only the first 3500 customer ids will be used for our model because they contain all the data. Cross-checking spreadsheets and synchronizing data are advised to stop future incidents.

- **In some records, many columns, including brand, only_order, and job title, have empty values.**

Mitigation : If only a few records are empty, filter out the data; however, if the field is a core field, impute the missing values.

Recommendation: To make the data easier to understand, add dropdown menus for online_order, brand, and job_title.

Consistency

- **varying values for the same characteristic**

Mitigation: To ensure uniformity, replace values into abbreviations using regular expression.

Recommendation: It is advised to make a dropdown list with all state and gender abbreviations.

Prediction: Dropdown menus reduce manual entry and human error, as predicted. allows for increased terminology consistency

Relevancy

- **Lack of significance in Default_column**

Mitigation: Remove the metadata from default_column.

Recommendation: Verify the metadata for ambiguity and remove or alter it as necessary to make it comprehensible.

Validity

- **Heterogeneous data type**

Mitigation: Format product_sale_date to a short date and list price to a currency. Data in number columns should be changed from general/Text to numeric type.

Recommendation: Ensure that fact tables in a particular database have data type restrictions. To guarantee consistent data type for the given data, appropriate data transformations should be applied.

For the goal of conducting a model analysis, the team will keep working on cleaning, standardizing, and transforming the data. Along the journey, queries and presumptions will be recorded.

If you have any questions or encounter any data quality issues, kindly contact us.

Kind Regards,
Rutuja Sawant.
[Junior Consultant]