Dear Manager,

I have reviewed the three datasets provided by Sprocket Central Pty Ltd, namely the Customer Demographic, Customer Addresses, and Transaction data. Based on the data quality framework and guidelines you shared, I have identified several issues and concerns that need to be addressed before proceeding with our analysis in phase two. Below is a summary of my findings:

**Missing Values:** Some columns, such as brand names, Product\_Line, Product\_Size, Product\_class, Online\_order, Standard\_cost, and Job title, have empty values. These missing values could impact our analysis and need to be handled appropriately.

|  |  |
| --- | --- |
| Column Name | % of Data Empty |
| Brand | 1% |
| Product\_Line, Product\_class, Product\_Size | 1% |
| Standard\_cost | 1% |
| Job\_Title | 12% |
| Tenure | 2% |
| Product\_first\_order\_date | 1% |

**Inconsistent Values:** Inconsistent representations exist for the same attribute.

For example, the gender "Female" is represented as both 'F' and 'Femal'. Similarly, the state "New South Wales" is represented as both "NSW" and "New South Wales," and "Victoria" is represented as both "VIC" and "Victoria". This inconsistency hampers data integrity and should be resolved for accurate analysis.

**Inaccurate Customer Demographic Data:**

The Date of Birth (DOB) information in the Customer Demographic dataset is not accurate. This could be due to data entry errors or inconsistencies in recording birth dates. We need to verify and correct this information to ensure the reliability of demographic analysis.

**Missing Column Headings**: Some columns in the datasets are missing their corresponding headings. This lack of clarity makes it difficult to interpret the data correctly. We should address this issue by adding appropriate column headings to improve data understandability.

**Inaccurate Data Types:** The data type of certain columns is not accurate.

For example, the DOB column is currently marked as "Text" instead of "Date." This incorrect data type classification may hinder date-based calculations and analysis. It is essential to rectify the data types to ensure proper analysis.

**Garbage Value:** There is a column named "Default" present in the dataset that contains garbage values. This column needs to be addressed to maintain data quality and avoid any misleading interpretations.

In terms of standard data quality, these findings reflect various aspects that may affect the reliability and usefulness of the data. Addressing these issues will ensure the data's accuracy, consistency, completeness, and appropriate representation. It is crucial to communicate these concerns to the client and recommend specific actions to mitigate the current data quality concerns.

**Recommendations:**

* Data Cleaning: Perform data cleaning operations to handle missing values and inconsistent representations.
* Standardization: Standardize attribute values, such as gender and state, to ensure consistent and accurate analysis.
* Data Validation: Validate the accuracy of the DOB column in the Customer Demographic dataset and rectify any inconsistencies found.
* Column Headings: Add appropriate column headings to enhance data interpretability.
* Data Type Correction: Correct the data type of the DOB column from "Text" to "Date" for improved date-based analysis.
* Garbage Value Removal: Address the "Default" column containing garbage values by either eliminating it or cleaning the data within it.
* I hope this analysis and the recommended actions will help improve the data quality for our phase two analysis. Please let me know if you have any further questions or if there are additional instructions from the client.

Thank you for your guidance.

Kind regards,

Rohan Kumar