

Dear [Name of the Client],

I hope this email finds you well. In reference to your voicemail, we have conducted a comprehensive analysis of the datasets that has revealed several quality concerns that necessitate attention. Below, you'll find a summary statistics table for the dataset you provided. Please review it and let us know if it aligns with your understanding.

| Name of the Table | Number of records | Distinct Customer_IDs |
|----------------------|-------------------|-----------------------|
| Transaction | 20,000 | 3494 |
| Customer Address | 3999 | 3999 |
| Customer Demographic | 4000 | 4000 |

These proposed recommendations have been formulated to effectively address the identified inconsistencies, elevate the data quality and bolster the precision of the data to drive business decisions.

- **Validity:** Additional customer_ids in the Transaction and Customer Address Tables.
Recommendation: Ensure data consistency by using only customers from the Customer Demographic Table for the analysis.
- **Validity:** In the Customer Demographic Table, the customer named Jephthah Bachmann appears to be born in 1843, clearly his DOB presents an error.
Recommendation: Replacing this date with the DOB mode, so it does not affect the distribution of the dataset.
- **Missing values:** The three datasets present multiple columns with missing values.
Recommendation: Evaluating the relevancy of each column to decide whether to drop or replace those values. Since the percentage of missing values is low compared to the whole dataset, we suggest replacing them to potentially maintain the dataset's representativeness and statistical power while resolving the gaps.

For example, we recommend complete DOB entries with the DOB mode and for the 'tenure' column, we recommend replacing the missing values with the tenure mean. However, further investigation may be required depending on the specific circumstances.

- **Inconsistent values:** The 'default' column from the Customer Democratic Table presents several inconsistencies and do not provide relevant information for the data set.
Recommendation: Dropping the 'default' column. The same issue occurs with the five 'Unnamed' columns in the Customer Demographic table. We suggest dropping those columns.
- **Uniqueness:** There are duplicate values in the 'Gender' column in the Customer Demographic Table and in the 'State' column of the Customer Address Table.
Recommendation: Implement a dropdown menu for data entry instead of a free-text field. Standardize Gender Values employing consistent gender like 'Male,' 'Female,' and 'Unidentified' across all records. The same for State values like 'Victoria', 'New South Wales'.

- **Validity:** Incorrect data type across values. i.e. DOB and 'product_first_sold_date' should be datetime data type, but they are not.
- **Recommendations:** Changing the data type to the correct one, so it shows an accurate representation of the data.

Please conduct an examination of the quality concerns above-mentioned, including the proposed modifications, to ensure uniform data quality throughout all tables. Should you need additional clarification or support, please do not hesitate to contact us.

Kind regards,

Nicolle Chuquilin