Subject: Data Quality Issues and Mitigation Strategies

Dear [Name of the Client],

I hope this email finds you well. Our team has been working with your data and we have identified some data quality issues that need to be addressed in order to ensure the accuracy and reliability of the insights we will be providing. Below are the details of the issues and our recommendations to mitigate them.

1. **Completeness**
   Issue: We have noticed that some columns such as online order, last name, date of birth (DOB), job title, and job industry category have missing values. These missing values can potentially affect the accuracy of our analysis and insights.

   Recommendation: To handle the missing values, we recommend using imputation methods such as mean, median, or mode imputation. In some cases, additional data may be needed to fill in the missing values. Alternatively, we can consider removing rows with missing values if the percentage of missing values is relatively low compared to the entire dataset. However, we need to carefully assess the potential impact of removing these rows on the analysis and insights we will be providing.

2. **Consistency**
   Issue: We have identified inconsistencies in the data across the Transactions, Customer Demographic, and Customer Address tables. For instance, the Transactions table has more customer ID entries than the other two tables. Moreover, there are typos, differences in data types or formats, and inconsistencies in fields such as gender, address, state, and default. These issues can lead to inaccurate analysis and conclusions.

   Recommendation: We recommend using customer ID as a unique identifier to ensure consistency across all three customer tables, and syncing data using this identifier. Standardisation methods such as using drop-down menus for fields like gender and product category can also help. Additionally, automated validation rules can be put in place to detect and correct errors in real-time. These steps can help improve the accuracy and reliability of our analysis and insights.

3. **Validity**:
   Issue: We have also found that the State column in the Customer Address table contains duplicate values such as "VIC" and "Victoria" and "NSW" and "New South Wales". Furthermore, the Gender column in the Customer Demographic table has inconsistencies. Additionally, some of the data entered in the system is not valid, such as invalid DOB. This can lead to incorrect analysis and insights.

Recommendation: To ensure data validity, we recommend implementing data validation checks to ensure that all data entered in the system is valid. It would be best to use state abbreviations instead of full names for all records to ensure validity across addresses. This can include the use of data validation rules for fields like postal codes, dates, and product codes (id, line, class, and size).

We believe that implementing these strategies will help us ensure the accuracy and reliability of our analysis and insights. Please address the above-mentioned quality issues along with the recommended changes to ensure consistent quality of the dataset across all tables. If all the suggestions are implemented, we can proceed with further analysis of the data to find suitable insights for the company.

Please let us know if you have any questions or concerns. We look forward to continuing our work with you.

Best regards,

[redacted]