Dear [Client],

I hope this email finds you in good health. As per the requirement, I have reviewed the data quality of the given three datasets from your organization, Sprocket Central Pty Ltd. I have identified the data quality issues and have provided strategies to mitigate these issues. First, please let us know if the following figures are aligned with your understanding:

|  |  |
| --- | --- |
| DATASET | No. of Records |
| Customer Demographic | 4000 |
| Transaction | 20000 |
| Customer Address | 4001 |

The data quality issues have been identified and methods to mitigate the issues have been mentioned which are as follows.

* DATA MISSING (Empty values in certain records)
  + Under ‘online\_order’, ‘brand’, ‘product\_line’, ‘product\_class’, ‘product\_size’, ‘standard\_cost’ and ‘product\_first\_sold\_date’ in the Transaction dataset.
  + Under ‘last\_name’, ‘DOB’, ‘job\_title’, ‘default’ and ‘tenure’ in the Customer Demographic dataset.
  + *Mitigation: If only a small number of rows are empty, filter out the record entirely from the given dataset for prediction. Else, if it is a core field, compute and insert the data based on distribution in the training dataset.*
* ADDITIONAL/ MISSING CUSTOMER ID
  + All the customer IDs are between 1 and 4000 under the Customer Demographics dataset. But under the Transaction dataset, there are three transactions (transaction ID: 8708, 16701 and 17469), which has a customer ID - 5034. This ID is not under the Customer Demographics dataset.
  + Also, under the Customer Address dataset, Customers with IDs 3, 10, 22 and 23 are missing with no address or information or a row dedicated to them. Additionally, extra Customer IDs 4001, 4002 and 4003 exist which have no information stored under the Customer Demographics dataset.
  + *Mitigation: Please ensure uniformity of the number of customers and their IDs in all the datasets. This would help us while creating our model.*
* INCONSISTENCY
  + Victoria and VIC are the same states and so are NSW and New South Wales.
  + Gender: ‘Female’ is sometimes written as ‘F’ or ‘Femal’, and, ‘Male’ is at times written as ‘M’ in the Customer Demographics dataset.
  + The Gender ‘U’ is not defined.
  + *Mitigation: There must be consistency maintained in the naming and abbreviations must be understandable or properly defined.*
* MISCELLANEOUS
  + Customer ID: 1787 - Last name of the customer (Godehard.sf) does not seem valid. The excel sheet recognizes it as a link in the Customer Demographics dataset.
  + Customer ID - 34, the customer’s date of birth is ‘1834-12-21’. This implies the customer to be 175 years old, which is definitely an error in the Customer Demographics dataset.
  + Three transactions (transaction ID: 8708, 16701 and 17469) in the Transactions database, the standard costs seem to be an error. The format of a cost value looks invalid. (E.g. Transaction ID 8708 has standard cost 312.7350159 which is an invalid value for a cost)
  + In the Customer Address database, street numbers 0, 00, 000 are invalid. Thus, these addresses are most likely errors.
  + *Mitigation: Small details along with the validity and accuracy of the values and data needs to be rechecked and corrected.*

This the evaluation of the given three datasets. I have mentioned the data quality issues and have provided mitigating strategies as much as possible. Moving forward, the team will continue with the data cleaning, standardisation and transformation process for the purpose of model analysis.

Thank you.

Kind regards,  
[Junior Consultant Name]