Dear [Client point-of-contact],

Thank you for providing us the dataset from Sprocket Pty Ltd. I have gone through the data quality evaluation process. Here is a summary of the dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | No. of Records | Distinct IDs | Received Date |
| Transactions | 20000 | 20000 | 2021-05-03 |
| New CustomerList | 1000 | - | 2021-05-03 |
| CustomerDemographic | 4000 | 4000 | 2021-05-03 |
| CustomerAddress | 3999 | 3999 | 2021-05-03 |

Attached are some issues that are worth attention, as well as suggestions for the methods that can be used to mitigate these issues:

**Missing values in the dataset**:

Transactions:

* 197 rows with product id 0 do not have record for “brand”, “product line”, “product class”, “product size”, “standard cost” and “product first sold date”.

New CustomerList:

* 360 values are missing in “online order” column.
* 17 values are missing in “DOB” column.
* 29 values are missing in “Last Name” column.
* 106 values are missing in “job title” column.
* 655 values are “n/a” in “job industry category” column.

CustomerDemographic:

* 87 values are missing in “DOB” column and “tenure” column.
* 125 values are missing in “Last Name” column.
* 506 values are missing in “job title” column.
* 656 values are “n/a” in “job industry category” column.

Solution: If the number of missing values is small, remove the data entry. If the data is important for further analysis on the whole dataset, fill the missing cell with average value.

**Inconsistent naming or data type for data entries**:

Transactions:

* Inconsistent data type in “Standard Cost”. Some cells are set to general type while most of the cells are set to currency type.

New CustomerList:

* Inconsistent naming in “Gender” column. There is a typo with customer ID 54’s gender where “Femal” should be corrected to “Female”. Customer ID 1’s gender indicates “F” which should be changed to an existing option. Customer ID 57’s gender indicates “M” which should be changed to an existing option.
* Inconsistent data type in “property valuation”. 5 cells are set to number type while the rest of the cells are of text type.

CustomerAddress:

* Inconsistent naming in “state” column. Should combine “New South Wales” and “NSW” and combine “Victoria” and “VIC”.

Solution: Do data cleaning to ensure that all naming remain consistent, and no non-numeric characters are present in numeric entries. For inconsistent data type, convert all records to one consistent data type after obtaining all data.

**Number of data entries mismatched**:

CustomerAddress:

* Number of customer ID in CustomerAddress and Transaction do not match with customer ID in CustomerDemographic.

Solution: Enter entries of other tables according to the key attribute (Customer ID) in the main table (CustomerDemographic).

**Error data entry**:

* There might be an error with the customer ID 34’s DOB. Data indicates that he was born on 1843-12-21.

Solution: Add constrain when trying to add data to the table to make sure that no errors like this happen.

Moving forward, the team will continue with the data cleaning, standardisation, and transformation process. Please contact me if you have any questions or want to make any clarifications.

Kind regards,

[My name]