**Dear sir/ma,**

Thank you for trusting us at KPMG to handle your datasets with utmost professionalism. This is to inform you that the datasets you sent to us from your company Sprocket Central Pty Ltd have been received and properly assessed for data quality issues. Based on our quality framework table, we have assessed these datasets (Transactions, New Customer List, Customer Demographics, and Customer Address), and the results are summarized below for each dataset received.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table name** | **No. of records** | **Distinct Customer IDs** | **Date Data Received** |
| Transactions | 20,000 | Yes | 5/15/2023 |
| New Customer List | 1,000 | Not Available | 5/15/2023 |
| Customer Demographics | 4,000 | Yes | 5/15/2023 |
| Customer Address | 3,999 | Yes | 5/15/2023 |

**Transactions**

* Presence of null or missing values within many fields of columns such as the online order, brand, product line, product class, product size, standard cost, and product first sold date.
* The data type format for some columns are inaccurate. E.g., transaction date, list price, standard cost should always be in the correct format etc.

**New Customer List**

* 5 missing column names
* Presence of missing values within the dataset
* Inconsistency with the gender column, different values are being inputted such as (female, male, U, F)
* Inaccurate data type format for some of the columns

**Customer Demographics**

* Presence of missing values within the dataset
* Inconsistency with the gender values
* Inaccurate DOB for customer id 34
* The default column is irrelevant to the dataset.
* Inaccurate data type format for some of the columns

**Customer Address**

* Inconsistency with the state names; while some are spelled in full, others are abbreviated (New South Wales, NSW, VIC, QLD).
* Inaccurate data type format for some of the columns

**Below are some recommendations that could help your company to mitigate the current data quality concerns and improve decision making.**

**Accuracy** – In terms of accuracy, a quality data set should always contain the correct values and must correspond to reality to prevent issues that may arise during the analysis.

**Completeness** – Data fields must also be filled with values, in order words for a more accurate analysis its best to avoid having missing information within the dataset. If only a small number of rows are empty, filter out the record entirely from the training set for prediction. Else, if it is a core field, impute based on distribution in the training dataset.

**Consistency** – This is a question of whether the data is consistent with the rest of the data provided or if it’s an outlier. Basically, the values in the dataset must be free from contradictions. For example, in the gender column rather than have some values abbreviated and others written in full, it is better use regular expressions to replace extended values into abbreviations to ensure consistency across addresses.

**Currency** – This refers to the age of the data, is the data old or is it up to date?

**Relevancy** – When gathering data, the relevance of that data becomes important to consider whether the information is really needed or not. Irrelevant information doesn’t need to be included to prevent time wasting.

**Validity** – All data types must be stored in a valid and useful format to prevent data quality issues. To mitigate this issue, convert selected records in characters to numeric, and remove non-numeric characters from string.

**Uniqueness** – The uniqueness of a dataset is important, so it is imperative that the records are checked for any duplicates that may exist. All data points under consideration should be completely unique within the database.

Moving forward, the team will continue with the data cleaning, standardization, and transformation process for the purpose of model analysis. Again, thank you for trusting us to handle your datasets. If you have any issues or concerns do not hesitate to reach out to us.

Kind Regards,

Junior Consultant,

Ugo Ejidoh