**SUBJECT:** **Quality Assessment of the given datasets.**

Dear Client,

I’m pleased to share the following details with you. I have assessed the datasets provided by your organization for quality and tidiness.

The three given datasets were:

* Transactions data (past 3 months)
* Customer demographic
* Customer address

Please find the attachment of the jupyter notebook in which the quality issues were assessed.

A brief summary of the issues identified is listed below.

**Assessment of the “Transaction” dataset**

* About 360 missing values were found in the “online\_order” attribute followed by 197 NaN under “brand”,”product\_class”,”product\_size”,” product\_line”, “standard\_cost”,”product\_first\_sold\_date”.
* Most of these missing values correspond to the value “0” under the attribute “product\_id”. We could mitigate this issue by ensuring proper value checks during the data entry process.
* The data type of the column "product\_first\_sold\_date" datatype has to be changed from float to “DateTime” format for analytical purposes.
* The attribute “product\_id” does not provide any valuable insight because several of them correspond to the similar product brand, line, class, size. The product\_id as it is does not provide any value to the dataset. It may correspond to a different trait such as style but that trait needs to be recorded to ensure detailed analysis.

**Assessment of the “Customer Demographic” dataset**

* Missing values were found under the “last\_name”, “job\_title”, “job\_industry\_category”,” tenure”.
* These empty slots can be filled by using mandatory questions during data collection.
* Some column names like “past\_3\_years\_bike\_related\_purchases” could be altered into a shorter form for better analysis.
* There are some value inconsistency errors found under the “gender” column. The values “Male”, “Female” are found to be misspelled or in their abbreviated format. We can alleviate this error by using a multiple-choice type question during the data collection.
* An outlier is found in the DOB column that corresponds to an unreasonable age. This can be avoided by ensuring allowable error and value checks.
* There is an unwanted column named 'default' filled with gibberish and could be removed for better data exploration.

**Assessment of the “Customer Address” dataset**

* The data type for the “postcode” column has to be changed. (It should be recorded as a string)
* Unnecessary zeroes are present in the address column. A value check will do the job of mitigating this issue.
* The “state” column is filled with the abbreviation of the respective states but in few rows, it is represented in the fully expanded form. We could avoid this error by using a multiple-choice question during data collection.

**Tidiness issues:**

* There are some extra customer\_ids present in the “transactions” table and “customer\_address” table which results in loss of synchronization with the master (Customer Demographic) table during the data exploratory and data modelling process. We could mitigate this issue by ensuring that all the tables come from the same timeframe. This is an integral step because only the customers present in the master(Customer Demographic) table will be used in the training set during modelling.

Please let me know if you need anything further and I would be happy to help.

Thank you,

Derin Ben Roberts