**MAIL**

Hello [ ClientX ],

Thank you for providing datasets.

Below are details of Data Quality Assessment:

**Common Data Quality Issues observed & verified in the provided datasets :**

1) Unnecessary/Unwanted columns

2) Inconsistent column names

3) Missing values of unimportant columns

4) Inconsistent values in columns (Accuracy)

5) Duplicate Records (Uniqueness)

|  |  |  |
| --- | --- | --- |
| **CustomerDemographic** | | |
| **Column name** | **Issues (Yes/No)** | **Issue (I), Assumptions(A) , Mitigation (M)** |
| customer\_id | No |  |
| first\_name | No |  |
| last\_name | Yes | Missing values  There may be cases where customers prefer to use first names only. |
| gender | Yes | Data is inconsistent with Male, Female, M, F, U. It must be either F,M,U or Male, Female, Unavailable. |
| past\_3\_years\_bike\_related\_purchases | No |  |
| DOB | Yes | DOB is an important client detail, with its missing values, we can’t focus on the type of population we are interested in. |
| job\_title | Yes | Missing values can be **ignored as this field is not as important as other fields** |
| job\_industry\_category | Yes |
| wealth\_segment | No |  |
| deceased\_indicator | No |  |
| default | Yes | The column is unimportant as the values can’t be understood. |
| owns\_car | No |  |
| tenure | Yes | Missing values (NaN) are replace by 0 as no tenure data will be considered as no tenure. |

|  |  |  |
| --- | --- | --- |
| **CustomerAddress** | | |
| **Column name** | **Issues (Yes/No)** | **Issue Description** |
| customer\_id | Yes | Customer id is not matching with customer ids in Customer Demographic (assuming CustDemo is the master dataset) |
| address | No |  |
| postcode | No |  |
| state | Yes | Data is inconsistent with abbreviations of Aus cities |
| country | No |  |
| property\_valuation | No |  |

|  |  |  |
| --- | --- | --- |
| **Transactions** | | |
| **Column name** | **Issues (Yes/No)** | **Issue Description** |
| transaction\_id | No |  |
| product\_id | No |  |
| customer\_id | No |  |
| transaction\_date | No |  |
| online\_order | Yes | I don’t understand the significance of the column. So, column can be removed/dropped |
| order\_status | No |  |
| brand | Yes | Missing values are removed |
| product\_line |
| product\_class |
| product\_size |
| list\_price | No |  |
| standard\_cost | Yes | Mis |
| product\_first\_sold\_date | Yes | Format changed to short date in excel.  Missing values are removed along with other fields |

|  |  |  |  |
| --- | --- | --- | --- |
| **NewCustomerList** | | | |
| **Column name** | **Issues (Yes/No)** | | **Issue Description** |
| first\_name | No |  | |
| last\_name | Yes | Missing values but they are ignored. | |
| gender | Yes | Data is inconsistent with Male, Female, M, F, U. It must be either F,M,U or Male, Female, Unavailable. | |
| past\_3\_years\_bike\_related\_purchases | No |  | |
| DOB | Yes | Blanks are removed as DOB is a required field. | |
| job\_title | No |  | |
| job\_industry\_category | No |  | |
| wealth\_segment | No |
| deceased\_indicator | No |  | |
| owns\_car | No |  | |
| tenure | Yes | Dat1type changed to float to match with other fields. | |
| address | No |  | |
| postcode | No |  | |
| state | No |  | |
| country | No |  | |
| property\_valuation | No |  | |
| Unnamed | Yes | The filed is unnamed & needs to be removed | |
| Rank | No |  | |
| Value | No |  | |

**Approaches adopted :**

* Data analytic tool **Python** has been used for Data Cleaning.
* Below are the steps followed :
* **Step 1** : Install Packages
* **Step 2** : Reading Dataset & Loading to Data frame
* **Step 3** : Remove Data Quality Issues

Further, the team will be following methods of data cleaning and transformation for the model analysis. Please confirm that our assumptions are in-line with yours. We will send you our queries if we find any.

PFA the python code and clean dataset for the analysis.

Thanks & Regards,

Shweta Parida