**Mail:**

Hi there,

Thank you for giving us the opportunity to analyses and identify the data quality issues and strategies to mitigate these issues. Below is a list of the Data Quality issues which we discovered from datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| **Standard**  **Data Quality Dimensions** | **Transactions** | **Customer**  **Demographic** | **Customer**  **Address** |
| **Accuracy** |  | **- D.O.B:** Inaccurate |  |
| **Completeness** | **- Online\_order:**  94 records blanks  **- Brand:**  48 records blanks  **- Product\_line:**  48 records blanks  **- Product\_class:**  48 records blanks  **- Product\_size:**  48 records blanks  **- Standard\_cost:**  48 records Blanks  **- Product\_first\_sold\_date:**  48 records blanks | **- D.O.B:**  87 records blanks  **- Job\_title:**  506 record blanks  **- Tenure:**  87 records blanks |  |
| **Consistency** |  | **- Gender:**  Inconsistency | **- States:**  Inconsistency |
| **Currency** | **- Online\_order**  Remove “False” value  **- Order\_status**:  Remove “Cancelled” value | **- Deceased\_customer:** Remove “Y” value |  |
| **Relevancy** |  | **- Default\_column:**  Remove |  |
| **Validity** | **- List\_price:**  Incorrect format  **- Product\_first\_sold\_date:** Incorrect format |  |  |
| **Uniqueness** |  |  |  |

Detail description of data quality issues and there are some recommendations to mitigate data quality issues.

**Accuracy:**

* DOB was inaccurate for the “Customer Demographic” dataset.
* **Mitigate:** Filter out DOB to date.
* **Recommendations:** Add age column in the “Customer Demographic” dataset for better insight of sales data by use formulas “(NOW()-F)/365)”

**Completeness**

* Record blanks in dataset.
* **Mitigate:** 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. For key datasets, such as transactions, less than 1% of transactions have missing fields. So all records blanks in dataset have been removed.
* **Recommendations:**  Blank values in columns affect data analysis in a negative way, so avoid null values.

**Consistency**

* Inconsistency in gender column for “Customer Demographic” dataset.
* Inconsistency in states column for “Customer Address” dataset
* **Mitigate:** Filter out ‘M’ as ‘Male’ and ‘F’ as ‘Female’ in gender column. Filter out ‘New South Wales’ as ‘NSW’ and ‘Victoria’ as ‘VIC’ in the states column.

**Currency**

* Deceased\_indicator column having value ‘Y’ in “Customer Demographic” dataset is not a current customer.
* Online\_order and status\_order having value ‘False’ and ‘Cancelled’ in “Transactions” dataset is not need statistics
* **Mitigate:** Filter out deceased column having value ‘Y’, ‘False’ and ‘Cancelled’.
* **Recommendations:** Up to date data of deceased customer will increase efficiency of data

**Relevancy**

* Lack of relevance in default\_column for “Customer Demographic” dataset and order\_status column in “Transactions” dataset.
* **Mitigate:** Delete default column in “Customer Demographic” and remove the ‘cancelled’ value from order status column in “Transactions”.

**Validity**

* Format of list price; and product first sold date column in “Transactions” dataset.
* **Mitigate:** Change format of list price column to currency and product first sold date to short date.
* **Recommendations:** Assign suitable format for any column to make data more readable and understandable.

The recommendations and mitigation strategies are effective and easy to implement. I hope it will be helpful in improving data quality

Please contact us for any queries and issues regarding the data quality framework table.

Best regards,

Tuan Pham Dinh