Dear Client,

I’m thankful to you for providing us with three datasets from Sprocket Central Pty Ltd. The Summary able given below provides some key quality issues that we discovered within three datasets:

Summary Table

|  |  |  |
| --- | --- | --- |
| **Datasets** | | |
| **Standard Data Quality Dimensions** | **Customer Demographic** | **Customer Addresses** | **Transaction’s data in the past 3 months** |
| Accuracy | * DOB: Inaccurate * Age: Missing | Additional Customer ID’s | * Additional Customer ID’s * Product First Sold Date * Profit: Missing |
| Completeness | There are blanks in the columns-   * Job Title * Job Industry category (N/A data) * Last Name * DOB * Tenure |  | There are blanks in the columns-   * Online Orders * Brand * Standard Cost |
| Consistency | * Gender | * State: Inconsistency |  |
| Currency | * Deceased Indicator: Deceased People are also included. |  |  |
| Relevancy | * Default Column is irrelevant |  | * Order Status: Cancelled orders |
| Validity |  |  | Wrong format of -   * Product first sold date * List Price. |
| Uniqueness | All three datasets contain unique data. | | |

Detailed description of each of the above listed quality issues are provided below. Along with the issues, some suggestions are also provided in order to improve the quality if the data.

**1. Accuracy Issues:**

* DOB was inaccurate and age column was missing in “Customer Demographic”.
* Product First Sold Date was given in wrong format and Profit column was also missing from “Transactions”.
* Additional customer\_id’s were present in “Transactions” and “Customer Address” dataset

Corrections –

1. Remove the outlier in DOB column to get accurate data.
2. Correct the format of ‘Product first sold date’ column so that its correctly recognized.
3. Kindly ensure that all tables are from the same period so that the data will be in sync with each other and we can get better analysis results.

It’s advisable to add an ‘Age’ column in “Customer Demographic” which makes the data more relevant. Also, include a ‘Profit’ Column which gives a meaningful aspect to data in “Transactions” dataset.

**2. Completeness Issues:**

* There are blanks in the columns - Job Title, Job Industry category (N/A data), Last Name, DOB and Tenure in “Customer Demographic” dataset.
* There are blanks in the columns - Online Orders, Brand and Standard Cost in “Transactions” dataset.

Correction – Remove the blanks so that the data becomes clean.

**3. Consistency Issues:**

* ‘Gender’ and ‘State’ columns are inconsistent in “Customer Demographic” and “Customer Addresses” dataset.

Corrections –

1. For the ‘Gender’ column, replace all the misspelt words with the correct ones
2. For the ‘State’ column, we should replace all the State full Names to their corresponding abbreviations.

Above corrections will make the data look a lot cleaner.

**4. Currency Issues:**

* “Customer Demographic” should contain current data so that it can provide reliable information but the ‘Deceased Indicator’ column indicates that the data also contains the information for those who aren’t alive.

Corrections –

1. Remove the information regarding dead people from the dataset by filtering out the “Y” cells from ‘Deceased Indicator’ column.

**5. Relevancy Issues:**

* ‘Default’ column in “Customer Demographic” includes irrelevant symbols and characters.
* ‘Order Status’ column in “Transaction” also includes cancelled orders which does not contribute significantly to the data.

Corrections –

1. It’s better that we exclude or delete ‘Default’ column from “Customer Demographic” dataset.
2. Exclude cancelled order from ‘Order Status’ column in “Transaction” because we could get skewed and inappropriate results if we include them.

**6. Validity Issues:**

* In “Transaction” dataset, the columns Product first sold date and List Price are in wrong format.

Corrections –

1. Convert the ‘Product first sold date’ column data from text to date format so that we can access correct information.
2. Also, we can add currency symbol to the ‘List Price’ column so as to correctly identify its data.

Kindly provide any suggestions needed for this assessment. After completing this, it would be great to if we would be given a chance to spend some time with your data SME to ensure that all the assumptions and suggestions made are aligned with Sprocket Central’s understanding.

Kind regards,

Dhriti Vohra