Subject: Data quality assessment

Hi Client Name

Sprocket Central Pty Ltd,

Thank you for providing us with the Customer data for analysis.

The below table highlights the summary statistics from the datasets received. Please flag if the figures do not align with your understanding.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table Name** | **No. of records** | **Distinct Customer IDs** | **Date Data Received** |
| Customer  Demographic | 4000 | 4000 | 30/11/2022 |
| Customer Address | 3999 | 3999 | 30/11/2022 |
| Transaction Data | 20000 | 3494 | 30/11/2022 |
| NewCustomerList | 1000 | 1000 | 30/11/2022 |

I have done a quality assessment of the data provided.

I have mentioned below the identified quality issues.

Furthermore, recommendations have been provided to avoid the re-occurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

**Findings**

**CustomerDemographic**

1. Inconsistent gender column

Female- denoted by F, Female, Female

Male -denoted by M, Male

**Mitigation:** Introduce a dropdown box for selecting the options rather than free text to maintain uniformity.

1. For customer\_id 34, DOB is 1843 seems incorrect as deceased\_indicator is N

**Mitigation:** Include a dropdown calendar than free text.

1. Default column seems unnecessary

**Mitigation**: Delete the column

1. Missing data in many columns

**Mitigation**: Either remove the rows if fewer or impute using the data distribution.

**CustomerAddress**

1. Inconsistent state column

New South Wales also denoted by NSW

Victoria also denoted by VIC

**Mitigation:** Introduce a dropdown box for selecting the options rather than free text to maintain uniformity.

**Transactions**

1. Missing data in online\_order, brand, product\_line, product\_class, product\_size, standard\_cost, product\_first\_sold\_date

**Mitigation**: Either remove the rows if fewer or impute using the data distribution.

1. Standardise formatting for currency in List\_price, standard\_cost.

**Mitigation:** Introduce data type constraints

1. Incorrect date formatting for product\_first\_sold\_date.

**Mitigation:** Introduce data type constraints

**NewCustomerList**

1. Date format is inconsistent and has missing dates

**Mitigation:** Include a dropdown calendar than free text.

1. Some fields from the column last\_name and job\_title is missing.

**Mitigation**: Ask for the data or remove the columns and missing last name columns can be ignored as they won’t skew the analysis.

There are extra customer ids in CustomerAddress and Transactions which are not in the CustomerDemographic sheet.

**Mitigation:** Remove the extra rows as the data from CustomerDemographic will be considered as master data.

For effective data analysis, it is imperative that the data is complete, consistent, and accurate.

Moving forward, the team will continue with the data cleaning, standardisation, and transformation process for the purpose of model analysis. Questions will be raised along the way and assumptions documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central’s understanding.

Regards,

Purva R