Sprocket Central Pty Ltd

Data Quality Assessment

Review/Plan

Prepared by

Grace Ikechi,
11/3/2023

I. Overview

Project Name: Customer Transactions Data

Review Period:1st - 3rd Nov, 2023

Purpose: To identify data quality issues in dataset provided

II. Data Source and Description

Data Source: Txt files

Data Description: The table presented below provides an overview of the summary statistics derived from the three datasets we've received. Please inform us if the figures do not match your interpretation or expectations.

Table name	No. of columns	No. of records	Distinct Customer IDs
Customer Demographic, Master	14	4000	4000
Customer Address	6	3999	3999
Transactions	14	20,000	3494
New Customers List	23	1000	

III. Data Quality Assessment Report

We have completed a data quality assessment on your datasets and have identified issues and methods employed to address the observed data inconsistencies. Additionally, suggestions have been offered to prevent the recurrence of data quality issues and enhance the precision of the data that informs business decisions.

- Additional customer_ids in the 'Customer Demographic, Master' but not in 'Transactions table' and 'Customer Address table'

  Mitigation: Only customers in the Customer Demographic, Master list will be used as a training set for the model. This indicates that the data received may not be in sync with each other which may skew the analysis results if there are missing data records. Please refer to the Appendix section (IV) for the list of outliers between tables.

- Various columns, such as the brand of purchase, or job title, have empty values in certain records

  Mitigation: for rows with missing data that are not many, we will filter them out (remove) for the prediction.For core fields, we will impute based on the distribution in the dataset. For key datasets, such as transactions, less than 1% of transactions (totalling less than 0.1% of revenue) have missing fields. These records have been removed from the training dataset.

- Inconsistent values for the same attribute (e.g. Victoria being represented as “V”, “Vic” and “Victoria”)

  Mitigation: we will replace abbreviations with their regular form to ensure consistency across addresses. Also, gender records having ‘U’ have been replaced based on the distribution from the training dataset.
  Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field. In order to construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value.

- Inconsistent data type for the same attribute (e.g. numeric values for some fields and strings for others)

  Mitigation: we will convert selected records in characters to numeric and remove non-numeric characters from string.
  Recommendation: ensure that fact tables in the database have constraints on data types. Having different data types for a given field makes it difficult to interpret results at the later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

Moving forward, the team will continue with the data cleaning, standardization and transformation process for the purpose of model analysis. Questions will arise during this process, and we will maintain a record of assumptions. Once this phase is concluded, it would be valuable to schedule a session with your data Subject Matter Expert (SME) to confirm that all assumptions are in accordance with Sprocket Central's comprehension.

IV. APPENDIX

- Accuracy:
  - Transaction - Online Order: No information, lots of blank rows.
  - Product first sold date: Data not in date format, blank rows.
  - Gender (in New customer list): Some rows have 'U' as their gender.
  - Gender (in Cust Demographic): Data values not consistent, 'M', 'U', 'F', 'Femal'.
  - DOB (in Cust Demographic): '1' seems invalid, blank rows.
  - Property value: Mix of decimal and whole numbers.

- Completeness:
  - Brand: Blank brand names, blank rows.
  - Product line: Blank rows.
  - Product class: Blank rows.
  - Product size: Blank rows.
  - Standard cost: Blank prices.
  - Last name (in New customer list and Cust Demographic): Missing last name.
  - DOB (in New customer list and Cust Demographic): Blank rows.
  - Job title (in New customer list and Cust Demographic): Blank rows.

  - Job industry (in New customer list and Cust Demographic): Some rows have 'n/a' as data.
  - Address (in New customer list and Cust Add): Some have generic address info.
  - Address (in Cust Add): Incomplete addresses.
  - State (in Cust Add): Inconsistent naming.

- Consistency:
  - Gender (in New customer list and Cust Demographic): Data values not consistent, 'M', 'U', 'F', 'Femal'.

- Reliability:
  - Default (in New customer list): Data doesn't seem relevant.