Dear Manager,

We appreciate you sharing the four datasets from Sprocket Central Pty Ltd with us. The summary statistics from the three received datasets are highlighted in the table below. If the numbers don't match your understanding, do let us know.

|  |  |  |
| --- | --- | --- |
| Table name | No. of records | Distinct customer ids |
| Transaction | 20000 | 3494 |
| New Customer List | 1000 | NA |
| Customer demographic | 4000 | 4000 |
| Customer address | 3999 | 3999 |

The following is a list of notable data quality concerns discovered, as well as the techniques used to address the discovered data inconsistencies. Furthermore, recommendations have been made to prevent the recurrence of data quality issues and to improve the accuracy of the underlying data that serves as the foundation for business decisions.

* The "Transactions table" and the "Customer Address table" have fewer customer ids than the "Customer Demographic table." Please ensure that all tables correspond to the same time period as mitigation. This suggests that the data received may be out of sync with one another, which could distort the analysis's findings if there are missing records from the data set. Please see the table for a list of outliers between tables.
* Many columns in some records, such as the brand of purchase or the job title, have empty values. Mitigation: If only a few rows are empty, remove the record entirely from the dataset for prediction. Otherwise, if it is a core field, impute based on the distribution in the dataset. Missing fields are present in less than 1% of transactions (or less than 0.1 percent of revenue) in important datasets such as transactions. These records have been removed from the training dataset.
* Inconsistent values for the same property, such as Victoria being represented by "V," "Vic," and "Victoria" To ensure address consistency, use a regular expression to convert extended values to abbreviations. It is recommended that users enter data into a drop-down list rather than a free text box. To create useful variables for the model, the data was cleaned to remove duplicate representations of the same value. Furthermore, gender records where "U" has been substituted based on the distribution of the training dataset.
* Property that is identical but has an inconsistent data type (e.g. numeric values for some fields and strings for others) Mitigation: Some records have their characters converted to numbers. Remove all non-numeric characters from the string. Recommendation: Ensure that the provided database's fact tables have data type constraints. When different data types are present for a given area, it is difficult to understand the results later on. The appropriate data transformations are used to provide consistent data types for a given field.

Moving forward, the team will continue to clean, standardize, and transform data in preparation for model analysis. Along the way, questions will be raised and assumptions will be documented. After we finish this, we'd like to meet with your data expert to ensure that all assumptions are in line with Sprocket Central's understanding.

Best Regards,

Rolly Mougoue