**TO: SPROCKET CENTRAL PTY LTD**

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the three datasets received. Please let us know if the figures are not aligned with your understanding.

|  |  |  |  |
| --- | --- | --- | --- |
| **TABLE NAME** | **NO OF RECORDS** | **DISTINCT CUSTOMER IDs** | **DATE** |
| Customer Demographic | Rows: 4000  Columns: 13 | 1 - 4000 | 16/08/2023 |
| Customer Address | Rows: 3999  Columns: 6 | 1 - 4003 | 16/08/2023 |
| Transactions | Rows: 20000  Columns: 13 | 1 - 5034 | 16/08/2023 |

Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows. Furthermore, recommendations have been provided to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

* Additional customer\_ids were found in the ‘transactions table’ and ‘customer Address table’ but not in ‘customer demographic [Main File]’.

**Mitigation:** It was ensured that all tables are from the same period, and only customers in the customer demographic [Main File] will be used as a training set for our 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.

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

**Mitigation:** Small number of rows which were empty were filtered out from the training set for prediction. If it is a core field, values were imputed based on distribution in the training dataset. For key datasets, such as transactions, less than 1% of transactions (totaling 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:** Replaced abbreviations with extended values to ensure consistency across addresses. In order to construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value.

Additionally, gender records having ‘U’ have been replaced with ‘Undecided’.

**Recommendation:** Enforce a drop-down list for the user entering the data rather than a free text field.

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

**Mitigation:** Normalize the datatype of columns to the appropriate datatype.

**Recommendation:** Ensure that fact tables in the given database have constraints on data types. Having different data types for a given field make it difficult to interpret results at the later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

* Columns such as ‘DOB’ and ‘Address’ were not explicitly well named.

**Mitigation:** Rename ‘DOB’ and ‘Address’ to ‘date\_of\_birth’ and ‘street\_address’