Dear Client,

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Table name IDs** | **No. of records** | **No. Distinct Customer IDs** | **Date Data Received** |
| Customer  Demographic | 4000 | 4000 | 2017-12-30 |
| Customer Address | 3999 | 3999 | 2017-12-30 |
| Transaction Data | 20000 | 3494 | 2017-12-30 |
| New Customer List | 1000 | / | 2017-12-30 |

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.

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

*Mitigation: If only a small number of rows are empty, filter out the record entirely from the training set for prediction. Else, if it is a core field, impute based on distribution in the training dataset.*

* The transactions dataset, which is a key dataset for our analysis, has less than 3% of transactions have missing fields. These records have been removed from the training dataset.
* The customer demographics dataset contains some null values in fields like job\_title, job\_industry\_category, DOB. Since the most important columns don’t have missing values, we are not dropping any records.
* **Inconsistent values for the same attribute**

*Mitigation: Replace extended values into abbreviations to ensure consistency across fields.*

*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.

* In the customer address dataset, Victoria has been represented as “Vic” and “Victoria”, New South Wales also has been represented as “New South Wales” and “NSW”. To tackle this issue, we have represented Victoria across all records as “VIC”, New South Wales as “NWS”.
* In customer demographics dataset, the gender has six representations “Female”, “F”, “Femal”, “M”, “Male” and “U”. To tackle this issue, we have represented all females as “Female”, and all males as “Male”, and we replaced “U” by “Unspecified”. Note that gender records “Unspecified” can be replaced based on the distribution from the training dataset.
* **Inaccurate values for the same attribute**

*Mitigation: Determine the cause of the incorrect data and decide how to handle it. We may choose to remove the incorrect data, correct the data, or impute missing data.*

*Recommendation: Establish clear guidelines and procedures for data entry, including how to handle missing data and how to resolve discrepancies.*

* In the transaction dataset, The values in the **product\_first\_sold\_date** indicate that all products has the same sold date but at different times. Moroever, for the same product, we find different first sold dates. Please make sure to check this.
* In customer demographics dataset, We notice that the **default** column contains some random characters so we decide to drop it.
* **Additional customer\_ids in the ‘Transactions’ and ‘Customer Address’ datasets but not in ‘Customer Demographic’**

*Mitigation: Please ensure that all tables are from the same period. Only customers in the Customer Master list will be used as a training set for our model to ensure completeness.*

* 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.

We understand that ensuring the quality of the data is crucial to achieving accurate and reliable results. Therefore, we would like to work closely with you to address these data quality issues and implement the recommended strategies. We believe that by doing so, we can ensure that our analysis is based on high-quality data and provide you with valuable insights and recommendations.

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.

Please let us know your thoughts on this matter, and if you have any questions or concerns, do not hesitate to reach out to us.

Best regards,

Ashwini, Junior Data Analyst at KPMG.