Dear Manager,

Thank you for providing the datasets from Sprocket Central Ltd. Below is a table of summary for the assessment of data quality. Please let us know if the figures does not align with your understanding.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Datasets | Number of records | Unique Customer ID | Date Timestamp |  |
| 1. Customer Demographics |  |  |  |  |
| 2. Customer Address |  |  |  |  |
| 3. Transaction data |  |  |  |  |
| 4. New Customer |  |  |  |  |

There are a number of issues with the quality of the raw datasets. I will describe the methods used to mitigate any data inconsistencies and future recommendations the improve the accuracy of the data source.

‘Customer ID’ datapoints in dataset 2 and dataset 3 but missing in dataset 1.

Mitigation: Ensure that all datasets are from the same time period, otherwise any duplicate or missing data records may skew the data analysis. The datasets have been merged to a single master dataset in a consistent format.

Many missing datapoints across various columns.

Mitigation: If the number of null-value is small, these have been filled using appropriate statistical methods. Otherwise, if the number of null-value is significant, these records have been dropped from the master datasets. The only exception is if the sample size is small and the datapoints are critical.

Inconsistent data value and or format for the same attribute. ‘VIC’, ‘Victoria’, ‘V’.

Mitigation: Find all alternative values and replaced all alternative spellings into abbreviations to ensure consistency across the datasets. To avoid different representations of the same value, the data type should be categorical rather than a variable text field. Gender records have been filled according to the names.

Inconsistent data type for the same attribute e.g. integer for some fields, float for others which can introduce unintended bugs due to discrepancy in precision.

Mitigation: standardise all fields to have constraints on the permitted data types.

Moving forward the team will continue with the Extract, Transformation and Load process for the purpose of model analysis. Queries may arise on an ad-hoc basis and assumptions documented.

I would be happy to summarise this in a slide deck to ensure that all assumptions aligns with Sprocket Central Ltd’s understanding.

Kind regards,

Man Kit Yip