Dear Manager,

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

| **Table name** | **No. of records** | **Distinct Values** | **Date data recieved** |
| --- | --- | --- | --- |
| Customer demographic | 4000 | 4000 | 07-29-2020 |
| Customer address | 3999 | 3999 | 07-29-2020 |
| NewCustomerList | 1000 | 1000 | 07-29-2020 |
| Transaction data | 20000 | 20000 | 07-29-2020 |

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 that used to drive business decisions. Finally, a table is showcased below to check whether the datasets meet the data quality parameters.

Datasets can be joined with the key as customer Id except for NewCustomerList(Id is missing)

1. NewCustomerList table is added without the customer Id details. Moreover, there are several unlabeled columns in the same table, which has some values.

**Mitigation**: Please ensure that customerID is added, then only merging is possible . Only customers in the Customer Demographic list will be used as a training set for our model. This indicates that the data received may not be in sync with each other and consider without leaving any customer,which may skew the analysis results if there are missing data records.

**Recommendation:**Makesure that column names are meaningful and relevant that comply with the data given.

2.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. For key datasets, such as transactions, less than 1% of transactions (totalling less than 0.1% of revenue) have missing fields.

3.Inconsistent and invalid values for the same attribute (e.g. Female is represented as ‘F’and ’Femal’. Also in state Victoria is given as ‘VIC’)

**Mitigation**: Use regular expressions to replace extended values with abbreviations to ensure consistency across addresses.

**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.

4.Inconsistent data type for the same attribute (e.g. numeric values for some fields ,text for others and any type for dates)

**Mitigation**: Convert selected records in characters to numeric. Remove non-numeric characters from string.

**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.

**Data Quality Framework Table**

| **Tables** | **Accuracy** | **Completeness** | **Consistency** | **Orderlines** | **Validity** | **Uniqueness** |
| --- | --- | --- | --- | --- | --- | --- |
| Customer demographic | ✔ | ⨯ | ✔ | ⨯ | ⨯ | ✔ |
| Customer address | ✔ | ✔ | ✔ | ⨯ | ✔ | ✔ |
| NewCustomerList | ✔ | ⨯ | ✔ | ⨯ | ⨯ | ✔ |
| Transaction data | ✔ | ⨯ | ✔ | ⨯ | ⨯ | ✔ |

Note: These parameters are for overall tables not for single columns.

Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only. 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.

Kind regards,

Siji S