Dear [Sprocket Central Pty Ltd],

Thank you for providing us with the three data sets from Sprocket Central Pty Ltd.

Notable data quality issues were encountered and the methods that can be used to mitigate the identified data inconsistencies are presented as follows.

* **Additional Customer\_id in ‘Transactions’ and ‘CustomerAddress’ but not in ‘CustomerDemographic’.**

*Mitigation: All tables are supposed be confirmed to be in same period in order to avoid inconsistency. Only those records in ‘CustomerDemographic’ would be used as a training set for our model.*

(This indicates that the received data set may not be sync with each other which may skew the analysis results if there are some missing values.)

* **Some columns, such as ‘job\_title’, ‘last\_name’, have some missing values.**

*Mitigation: If the amount of missing values is very small, we can directly drop them from the training set for prediction. Else, it is a core filed, we should impute the empty values based on distribution in the training dataset.*

* **Some same attributes have different expressions(e.g.,NSW, New South Wales)**

*Mitigation: Avoid mutiple representations for the same value .Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses. In addition, gender records where ‘U’ should be replaced.*

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

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

*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.

Moving forward, the team will continue with the data cleaning, standardization 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,

Qingyi