Dear Manager,

After reviewing the datasets provided by Sprocket Central Pty Ltd, some data quality issues were encountered. The methods used to mitigate the inconsistent data are as follows;

	Accuracy	Completeness	Consistency	Relevancy	Validity
Customer Demographic	DOB: *Inaccurate*	Customer Id: *Incomplete* Last Name: *Blanks* Job title: *Blanks*	Gender: *Inconsistent*	Default: *Deleted*	
Customer Address		State: *Inconsistent*			
Transactions		Online order: *Blanks* Brand: *Blanks*		Cancelled order status: *Filtered out*	List Price: *Format* Standard Cost: *Format* Product Sold Date: *Format*
New Customers		DOB: *Blanks* Job title: *Blanks*			

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 re-occurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

- **Additional customer_ids in the 'Transactions table' and 'Customer Address table' but not in 'Customer Master (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.
  This indicates that the data received may not be in sync, 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: If only a small number of rows are empty, filter out the record entirely from the training set for prediction. If it is a core field, impute based on the 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: 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.

  To construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value. Additionally, gender records where 'U' have been replaced based on the distribution from the training dataset.

- **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 a string. Recommendation: Ensure that fact tables in the given database have constraints on data types.

  Having different data types for a given field makes it difficult to interpret results at a later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

The team will continue with the data cleaning, standardization, and transformation process for 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,
Mevhare Afe.