Dear Manager,

Thank You for providing us with the three datasets from Sprocket Central Pty Ltd. The summary table below highlights key quality issues that we discovered within the three datasets. Please, let us know if you have any queries:

Summary Table:

Dataset	**Accuracy**	**Completeness**	**Consistency**	**Currency**	**Relevancy**	**validity**
1. Customer demographic	▪ DOB: inaccurate ▪ Age: Missing	▪ Job title: blanks ▪ Customer ID: incomplete	▪ Gender: Inconsistence	▪ Deceased customers: Filter out	▪ Default Column: delete	
2.Customer address		▪ Customer ID: incomplete	▪ States: inconsistence			
3. Transactions	Profit: missing	▪ Customer ID: Incomplete ▪ Online order: blanks ▪ Brand: blanks			Cancelled Status: filter out	List price: format Product Sold date: format

Table name	No. of records	Distinct customer ids	Date data recieved
Customer demographic	4000	4000	07-06-2020
Customer address	3999	3999	07-06-2020
Transaction data	20000	3496	07-06-2020

We have tried to figure out the limitations and the corrections of the dataset. Following recommendations will improve accuracy of data to influence business decisions of Sprocket Central Pty Ltd in future.

Accuracy Issue:

- **DOB was inaccurate for "customer demographic" and missing an age column; missing a profit column for "Transactions"**
- **Mitigation:** filter out outlier in DOB
  Recommendation: Create and age column, allowing for more comprehensible data and easier to check for error. Create a profit column in "transaction" data.

Completeness:

- **Additional customer_ids were inconsistent in the "customer demographic", "customer address" and "Transactions"**

  Mitigation: Filter out all customer_ids until 3500
  Recommendation: Ensure tables are up to date

● 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 with each other which may skew the analysis results if there are missing data records. Please refer to excel file 'data_outliers.xlsx' for the list of outliers between tables.

● 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 (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 expression to replaced extended values into 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. 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 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.

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, 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,

Zian Md Afique Amin