Dear Manager,

This is Arkadeep Mukherjee from the KPMG Data Analytics (Virtual Internship) Team. Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the three datasets received. Please let us know if the figures are not aligned with your understanding.

Table name	No. of records	Distinct customer IDs	Receive Date
Customer demographic	4000	4000	12-07-2020
Customer address	3999	3999	12-07-2020
Transaction data	20000	3496	12-07-2020

The following are the details of analysis done on the dataset:

Table Name	Table Records		Table Analysis
	Before Data Cleaning	After Data Cleaning	
Transaction Data	20000 rows & 13 columns (1542 blank cells)	19445 rows & 14 columns (0 blank cell)	• Total profit: $10,930,284 (app.) • 'Solex' is the most purchased brand name • Most Sold Product Line : Standard • Least Sold Product Line : Mountain
New Customer List	1000 rows & 18 columns (152 cells)	878 rows & 18 columns (0 blank cell)	• Most new customers are from the New South Wales, Australia • Most customers own cars • Most Customers belong to the mass customer category • Most Customers belong from the "Financial Services" job category
Customer Demographic	4000 rows & 13 columns (806 blank cells)	3413 rows & 13 columns (0 blank cell)	• Most customers are 'mass customers' in wealth segment • Most customers are working in manufacturing and financial services industry
Customer Address	3999 rows & 6 columns (0 blank cell)	3999 rows & 6 columns (0 blank cell)	• Most customers are from New Sales Wales (NSW) • Most customers have post code between 2000 to 2190

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 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 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, online order 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 0.7% of revenue) have missing fields. These records have been removed from the training dataset.

● Inaccurate data in DOB (e.g. DOB is 1984 in NewCustomerList which is an incorrect value for DOB)

Mitigation: Please ensure that the data provided is accurate as such inaccurate data can highly affect the training set for our model.

I have filtered out such inaccurate data from the dataset so that it makes the next processes easy to manage and shows correct results without generating any errors or outliers. Also, an additional table named 'default' have been removed as it consisted of trash values.

● 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.

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,
Arkadeep Mukherjee
KPMG Data Analytics (Virtual Internship) Team