Hello,

This is Zeel from 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.

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

Table Name	Table Records	
	Before Data Cleaning	After Data Cleaning
Transaction Data	20000 rows & 13 columns (1542 blank cells)	19445 rows & 14 columns (0 blank cell)
New Customer List	1000 rows & 18 columns (152 cells)	878 rows & 18 columns (0 blank cell)
Customer Demographic	4000 rows & 13 columns (806 blank cells)	3413 rows & 13 columns (0 blank cell)

- 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 (totalling 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 sameattribute (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.

Kind regards,
Zeel