Dear [Client point-of-contact],

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd.
This is Gautam Bhatia from KPMG Data Analytics (Virtual Internship) team. We have reviewed the data sets which were provided by your company and during the data quality analysis, we have found some errors in the data sets. The data quality analysis is the core phase and due to errors in the data set we suggest the following mitigates in order to improve the data quality which will eventually help us to driven the better analytics results for your company

● Additional customer_ids in the 'Transactions table' and 'Customer Address table' but not in '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" and Female being represented as "F", "Female")
Mitigation: Use regular expression to replace 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 makes it difficult to interpret results at the later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

CUSTOMER DEMOGRAPHIC (4000 records)

Field Name	**Errors**
DOB	01 record 1843 87 records Blanks
last_name	125 Blank Records
Gender	88 Records gender is “U” Values are not consistent Male is represented as “Male”, “M” Female is represented as “Female”, “F”
job_industry	656 records “NAN”
job_title	506 records blanks
Default	3317 records value ‘Special characters’ includes null and blanks
Tenure	87 Records blanks

Transactions(20000 records)

Field Name	**Errors**
Online Order	94 Records Blanks
brand	48 Records Blanks
product_line	48 Records Blanks
product_class	48 Records Blanks
product_size	48 Records Blanks
standard_cost	48 Records Blanks
product _first_solid_date	48 Records Blanks

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,
Gautam Bhatia