Dear (Sprocket Central Pty Ltd),

Thank you for providing us with the 3 datasets from Sprocket Central Pty Ltd. Our team had gone through dataset and found some data quality issues. We would like to reach out to you for further information and clarification required about the attributes of the dataset.

Here are the summary statistics of the 3 datasets. Please let us know if the figure is not aligned with your understanding:

Table name	**No. of records**	**No. of columns**	**No. of columns having missing values**	**Date data received**
Transaction	20000	13	6	12/06/2020
CustomerDemographic	4000	13	6	12/06/2020
CustomerAddress	4000	6	0	12/06/2020

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.

- **Lack of completeness of certain columns across the table**

  Various columns have empty values in certain records. Summary statistics of records having empty values are as follows:

Table name	Column name	Count	Percentage
Transaction	online_order	360	1.8%
Transaction	brand	197	0.985
Transaction	product_class	197	0.985
Transaction	product_size	197	0.985
Transaction	standard_cost	197	0.985
Transaction	product_first_sold_date	197	0.985
CustomerDemographic	last_name	125	3.125
CustomerDemographic	job_title	506	12.650
CustomerDemographic	DOB	87	2.175
CustomerDemographic	job_industry_category	656	16.5
CustomerDemographic	default	302	7.55
CustomerDemographic	tenure	87	2.175

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.*
Recommendation: Recheck the process of constructing the database. Missing data having very less percentage (1% or 2%) can be removed from training dataset.

- **Additional Customer Ids in the 'Transactions table' and 'Customer Address table' but not in 'Customer Demographic table'**
  Mitigation: *Please ensure that all tables are from the same period. Only customers in the Customer Demographic list will be used as a training set for our model.*
  Please refer to excel file 'missing_customer_id.xlsx' for the list of Customer Id between tables

- **Inconsistent of values of certain columns across table**
  (e.g. "standard_cost" column in Transaction table, 99 percent of the data lies below 1610.90, except some record values 1759.85 which counts to 1 percent)
  Mitigation*: Some values seem to be an outlier and need to be rechecked whether it is a genuine input, indicator of empty space or a technical glitch.*
  Recommendation: See whether all the values lies in accepted range.
  Please refer to excel file 'data_outliers.xlsx' for the list of outliers between tables.

- **Inconsistent in labels of same categories of certain columns across table**
  (e.g. gender column in CustomerDemographic table have 'F' and 'Female' for female category)
  Mitigation: *Remap the labels of the categorical columns.*

- **Relevant columns showing strong correlation**
  (e.g. Pearson correlation coefficient is 0.55 of column "list_price" and "standard_cost" in Transaction table)
  Mitigation*: Check for derivatives or duplicated columns.*
  Recommendation: Plot graphs between columns to better visualise the linear relation.

- **Unexplained columns**
  Some columns were unnamed or the values were un explainable

(e.g. column ‘default’ in CustomerDemographic table)
Mitigation*: Label the unnamed columns, check for unexplainable data and provide with additional information.*
Recommendation: Could be either due to technical error or database software incompatibility issues (if using one). Need to contact IT department for software incompatibility issue.

- **Invalid records of certain column**
  (e.g. Jephthah Bachmann birth year 1843 in DOB column in CustomerDemographic table)
  Mitigation*: Reach out to the customer for valid information. Or try to contact domain expert for guidance.*

  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,
  Prashant Lal