Dear Sprocket Central Pty Ltd,

We have done a Data Quality Assessment on the Datasets and found the below issues which show bad data quality that needs to be addressed before further analysis of the datasets:

1) Data Completeness:

a) Transaction Dataset: There are missing data in seven columns of the dataset, kindly find below the column name and the sum of missing data in that column.

i) Online_order --------------------360
ii) brand ------------------------------197
iii) product_line---------------------197
iv) product_class-------------------197
v) product_size---------------------197
vi) standard_cost------------------197
vii) product_first_sold_date----197

b) CustomerDemographic: There are missing data in six (6) columns of the dataset, kindly find below the column names and the sum of missing data in that column.

i) gender--------------------------- 125 ii)
DOB------------------------------- 87 iii)
job_title--------------------------506 iv)
job_industry_category-------656
v) default---------------------------- 302 vi)
tenure--------------------------- 87

c) NewCustomerList: There are missing data in four (4) columns of the dataset, kindly find below the column names and the sum of missing data in that column.

i) last_name----------------------- 29
ii) ii) DOB------------------------------ 17
iii) iii) job_title------------------------ 106
iv) iv) job_industry_category----- 165 These missing data will affect the computation and analysis of the data, it can be mitigated by filling these missing data with actual data or they are dropped accordingly.

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

2) Data Accuracy:

a) Transaction Dataset: The product_first_sold_date column of the dataset doesn't have the right format which is Date format, it needs to be converted to the Date format, however, to convert this is challenging being that the data has missing data unless the missing data is resolved first.

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.

3)Data Validity:

a) NewCustomerList Dataset: This dataset has five (5) unnamed columns with non-missing data, these unnamed columns will affect the validity of the data and analysis, hence its either the right column names are provided, or they are dropped.
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.

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. 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.