To,
Sprocket Central Pty Ltd.

Dear Team,

I hope this email finds you in great spirits!

Thanks for sharing the comprehensive datasets with us. We are thrilled about the prospect of helping you achieve the best results in identifying new potential customers through analysis of the collected data.

I performed a detailed analysis of the datasets that were provided, from a data quality perspective and wanted to highlight a few issues, along with my suggestions regarding the type and quality of certain fields in the data so that these issues are addressed before the data is taken forward for analysis. These measures will help in gleaning and transforming the dataset so that it is ready for smooth analysis to drive company growth.

I'll report my findings in three different sections, one for each of the worksheets in the provided workbook. First, I will provide a summary statistic of the data received. Please let us know about any mismatches that you find in the summary.

Customer Demographic

Summary Statistics –

No. of Records: 4000
No. of Columns/Features: 13
Distinct Customer ID's – 4000

This has been considered as the Master Table for the customer data.

Issues –

1. In the **gender** column, multiple inconsistent notations are observed to indicate the same value. To indicate Female, the columns has 3 different values – F, Femal and Female. This can cause confusion at the time of analysis. Values/Notations for the same data need to be single and atomic.
   Mitigation Scheme – By creating regex patterns, or string formulae in excel, extended values can be turned into abbreviations like F for Female and M for Male. We have performed that conversion for further analysis
   Recommendation to avoid this – To avoid such erroneous entries, a drop down should be provided on the user onboarding webform for users to fill their data in.

2. Blank Records in columns (Dates, Product_Class, Brand, etc.)
   Recommendation and Mitigation Scheme – If the number of blank records is less than 1% of the entire dataset, the blanks can be removed before proceeding with the analysis. If the blanks consist of PI details, those can't be imputed with average values, so such rows with multiple blank columns can be removed from the dataset.

3. Permissible date ranges for DOB.
   Recommendation and Mitigation Scheme – A validation for minimum permissible DOB must be set to avoid getting erroneous values in the date fields. Over here, rows with DOB < (Current Date – 100 years) have been removed.

4. Deceased Customers – A Value of Y in the deceased indicator column, indicates that the corresponding individual is no longer a customer of Sprockets. Including data for these customers can reduce the currency of our data and make it skewed.
   Mitigation Scheme – We have removed such customers who have deceased from our analysis to make it more relevant.
   Recommendation – A pre-check should be imposed on the database to disallow including rows that have a 'Y' in the deceased column to preclude this data from entering these records.

Customer Addresses

Summary Statistics –

No. of Records: 3999
No. of Columns/Features: 6
Distinct Customer ID's – 3999

Issues –

1. The **customer_id** column in Customer Addresses has gaps in between, meaning for a few customers (four customers with customer_ids – 3, 10, 22, 23), which are present in the Customer Demographics table, their corresponding address details are not available.
   Also, a few customer_ids present in this table (4001, 4002, and 4003) have no records in the CustomerDemographics table.
   This could mean that data has been captured incorrectly, or the timeline of spooling the 2 record logs is inconsistent. Such missouts in the data can skew our resulting analysis.

   Mitigation Scheme – To avoid such inconsistencies, we will be considering the customer id's in demographics as the master record and would only be incorporating those into analysis, joining them on all other tables for more details.

   Recommendation to avoid issue – Ensure that reports are collated for the same timeline.

2. In the **state** column, for the same state, multiple notations are used.

   To indicate New South Wales, some rows have New South Wales in their state column, while some have NSW. Similarly, to indicate Victoria, some rows have Victoria written while some store data as VIC. This difference of notation can be harmful for the location analysis of customers, which will be an important part of our analysis. Single notations should be used to indicate the same value.

   Mitigation Scheme – By creating regex patterns, or string formulae in excel, extended values can be turned into abbreviations. We have performed that conversion for further analysis.
   Recommendation to avoid this – To avoid such erroneous entries, a drop down should be provided on the user onboarding webform for users to fill their data in.

Transaction data in the past three months

Summary Statistics –

No. of Records: 20000
No. of Columns/Features: 13
Distinct Customer ID's – 3494

Issues –

1. Blank Values.
   In the **online_order** column, apart from TRUE (presumably indicating online orders) and FALSE (presumably indicating offline orders), there are 360 blank values.
   In all, a total of 107 transactions have no Product details. For all these transactions, the columns – brand, product_line, product_class, product_size, standard_cost, and product_sold_date are all blank.

   Recommendation and Mitigation Scheme – If the number of blank records is less than 1% of the entire dataset, the blanks can be removed before proceeding with the analysis. If the blanks consist of PI details, those can't be imputed with average values, so such rows with multiple blank columns can be removed from the dataset.

2. Inconsistent data types.
   product_first_sold_date is not in an appropriate date format.
   Costs are not in a unanimous format.
   Inconsistent formats can cause hindrances in the analysis process. To avoid the same, data needs to be transformed for consistency across values.

Mitigation Scheme -
Entire columns have been formatted for consistency across values in the sheet. This would help in smooth analysis of provided data in further steps
Recommendation to avoid issue – Impose schema constraints on the tables while storing values. This would ensure data is cleaned while storing itself.

3. Irrelevant values.
   Orders that are cancelled shouldn't be included in our analysis since they will cause erroneous overestimation of sales if included in the sales record analysis.
   Mitigation Scheme – These rows with a Cancelled value in the order_status column have been removed/excluded from our analysis in further stages.
   Recommendation – While spooling this report, an option should be given to filter only approved records.

These are our findings with regard to a qualitative dataset acquisition. Please write back with your views regarding the same, and the measures being taken to mitigate the issues. Our team will keep you informed new findings or changes that might get discovered during further standardization and analysis phases. Any new assumptions will be documented and discussed.

Looking forward to discussing on this further with your data management team. Have a great day!

Thanks & Regards,
Karan Gupta.