Hello Mr./Mrs. XYZ,
Manager,
Sprocket Central Pty Ltd.

Thank you for coming to KPMG for analyzing the data to optimize marketing strategy.

Below table summarizes the Statistics of the data that has been received.

Dataset	Number of Records	Unique Customer ID
Transactions	20000	3500
Customer Demographic	4000	4000
Customer Address	4003	4003

The Data Quality Issues that were found during analysis are given in table below:

Dataset	Accuracy	Completeness	Consistency	Currency	Relevancy	Validity	Uniqueness
Transactions	There is one Customer ID: 5034 which is inaccurate as the values range from 1 to 3500.	Missing values in online_order, brand, product_line, product_class, product_size, standard_cost, product_first_sold_date				Product First Sold Date was not in a Proper Format and same with the list price.	
Customer Demographic	DOB: Inaccurate	Missing values in last_name, job_title, default, DOB, tenure	Gender values are inconsistent	Deceased customers are included in the data	Default column itself is not relevant and the values indicated have no meaning whatsoever.		
Customer Address			State values are inconsistent				

In-depth Analysis of Above Data Quality Issues:

Accuracy:

In Transactions data, I observed that the customer id ranges from 1 to 3500. However, there is one customer id: 5034 which I think is either wrong or it may be typed incorrectly. So, that customer id could be 3045. Since there is a zero in it, we can surely say that it is not 345, 543, 453, 354, 435 or 534. Also, get updated data to resolve above issues.

In Customer Demographics data, there is one customer id: 34 with First Name: Jephthah & Last Name: Bachmann whose DOB is 1843-12-21 i.e., 21ST December 1843 which gives his total age as per 2022 to be 179 years and yet deceased indicator shows him as NOT DECEASED. So, this a data quality issue. The DOB may be wrong since we have all the DOB starting from 1931 or the person may have been deceased.

Completeness:

In Transactions data, there are lots of missing values in online_order, brand, product_line, product_class, product_size, standard_cost, product_first_sold_date. Here, for categorical columns we can take the mode of the values and fill the missing data and for numerical columns we can use mean value to fill the data.

In Customer Demographic dataset, there are lots of Missing values in last_name, job_title, default, DOB, tenure. So, make sure the data is updated with these values otherwise there will be many difficulties in analyzing the data further.

Consistency:

In Customer Demographic data, the gender values are inconsistent. There are values like F, Femal that should be replaced by Female and values like M should be replaced with Male. Also, there are some values with U which I am considering as Undefined. If there is an updated data for these values, please let me know so that I can replace those values with correct one.

In Customer Address data, the state values are inconsistent. There are values like New South Wales, Victoria which needs to be replaced with their respective abbreviations NSW and VIC.

Currency:

In Customer Demographic data, the deceased persons are also included. These values must be filtered out since our goal is to optimize marketing strategy, we do not need records that contains deceased persons.

Relevancy:

In Customer Demographic data, there is no meaning of default column and also the values in this column are not relevant. Those are just some special characters with no meaning whatsoever. We need to remove that column as it is not necessary for our analysis.

Validity:

In Transactions data, Product First Sold Date was not in a DATE format and also list_price was not in a number format. So, we need to convert those into their proper format for our analysis.

Above are the all-probable data quality issues in datasets. Following the proper data cleaning methods will not only improve data quality for future analysis but also it will improve the business strategies after analysis.

Furthermore, our team will continue the data cleaning and preparation process for model analysis. Any issues encountered will be documented. After completion it will be great to have a meeting with your SME to ensure all the assumptions are aligned with Sprocket Central Pty Ltd.

Please let us know if you have any queries regarding the mitigation and recommendations mentioned in the report.

Regards,

Akshay Paunikar
Junior Data Analyst
KPMG