Dear Kathleen,

Sprocket Data Quality Assessment

I hope this email finds you well. I'm writing to convey the data quality issues discovered with the Sprocket Central Pty Ltd datasets. The company provided us with an excel file with 4 sheets contained with the following summary statistics.

Name of sheet	Number of rows	Number of columns	Count of Distinct Customer_Id
Transactions	20000	13	3492
NewCustomerList	1000	23	Not a column here.
CustomerDemographic	4000	13	4000
CustomerAddress	3999	6	3999

Upon analyzing the datasets, I found a couple of issues which I would be listing below with possible solutions as pertaining to each sheet.

Transactions sheet.

1. Incompleteness: 5 columns had null values, 4 of which i would advise are just dropped cause they seem to be canceled or incomplete transactions. With the 5th column, online_order, the null values should be filled with either true or false depending on which these null values is most likely from.
2. The product_first_date column is unclear as the name clues it should be a date column but the values do not fit any date format. Also this column should be well defined and well named.

NewCustomerList Sheet

There are 5 unnamed columns which contain numerical values. These columns should be named appropriately if details are available or better still removed before performing further analysis.

1. Incompleteness: The sheet contains null values in the DOB, job_title, jo_industry and last_name columns. The rows with null values should be dropped from all these columns except the last_name column. Also these columns should be marked as compulsory to fill when collating the data so they are not left null. For the job_title and job_industry column an option of unemployed or unknown should be made available for it.

2. The dataset contains redundant columns like country, address, postcode and state. All these columns are related and provide similar info. The country column especially is not needed since all the values are within Australia.

Customer Demogaphic sheet

1. There is an entry of DOB with year 1834, and the customer is not deceased. This entry is most likely a data entry error. To avoid such new entries should have a year restriction.
2. The gender column has inconsistent value labeling. E.g F, Femal, Female all for the feminine gender. Also it contains U which is not a known gender value
3. There is a column named default which contains gibberish data values. This should be checked whether it's a format error.

In the **Customer_Address** sheet, I discovered that the State column has states written in full words and abbreviation e.g VIC, Victoria. This increased the categories for this column and affects the distribution count. The abbreviations format should be stuck with as this same format is used in the NewCustomerList sheet.

These are the quality issues discovered within the datasets and I advise the company should review the form or database rules and restrictions for these columns during the data entry process. They should make necessary details like DOB be compulsory to fill and also create default categories for columns like gender and online_order to avoid discrepancies.

I'd be grateful for the opportunity to discuss these findings with you in more detail. Your insights and expertise in this matter would be invaluable as we work to ensure the highest standards of data quality within our processes.
Thank you for considering my inquiry. I look forward to the possibility of collaborating and contributing to the enhancement of data quality processes.

Warm regards,
Data analyst,
Akinyemi Anjola.