Dear [client name]

Thank you for providing us with the three spreadsheets from Sprocket Central Pty Ltd. I'm from KPMG Data Analytics team and I have analysed the data provided. Based on the analysis done, I've grouped the data quality issues in three main groups listed below:

A. Missing information

B. Not standardised structured data

C. (Not) relevant information

**A. Missing information**

*Transactions spreadsheet:*

- 555 transactions with at least one missing information.

- 358 transactions with no information if the order was made online

- 197 transactions with missing information about the product purchase: brand, line, class, size, standard cost and date that it was first sold.

- I believe that the company should be worried about missing information related to product in the Transactions sheet. Those information are quite important to understand the customer preferences, the revenue, and the product's stock control. Those information should be tracked (maybe checking the invoices) and transferred into the Transactions sheet.

- Besides that, I recommend to think in ways to avoid this kind of situation (missing important information). For example, is the data entered manually into this sheet? If so, we could automate this process.

*CustomerDemographic spreadsheet:*

- 125 customers with missing last name

- 87 customers with no DOB information

- 506 customers with missing job title

- 87 customers with no tenure value

*NewCustomerList spreadsheet:*

- 152 (15%) of the spreadsheet with at least one missing data: 29 new customers have missing information about their last name, 17 new customers with no DOB, and 106 new customers with missing job title. I consider DOB and last name important data. Age can be correlated with customer behaviour and could be a feature to 'build' a machine learning algorithm to predict if the customer is going to accept or not a offer, for example. And last name can be one of the informations used to avoid duplicated record (e.g.it could be used along DOB).

**B. Not standardised structured data**

*Transactions spreadsheet:*

- I noticed that the 'list\_price' columns has no currency sign. but the standard\_cost contains a dollar sign before the amount. The amounts in the sheet should be consistent, so all the monetary data should be free of currency sign (and a metadata file should be created indicating detailed information, such as currency).

*NewCustomerList spreadsheet:*

- 1000 records of new customers with no customer\_id. I suggest to assign to each of them a customer\_id as many of them purchase frequently. Besides that, it is important to have information about transactions made to understand their behaviour. With no customer\_id is not possible to check the transactions made. For example, there a customer who purchased 94 times, I cannot check what items s/he purchased.

- There are 3 genders: Male, Female and U. Not clear if U is for Unkown or Undefined. As gender in the sheet was not entered by code, it should be standarlised: or codes (F,M,U) or written(Female, Male, Unknown/Undefined).

- Keep an standarlised name for column. Usually they are all in lowercase. So, I suggest that the columns 'Rank' and 'Value' are written in lowercase (not capitalised).

- Agriculture in job industry column is mispelled (Argiculture).

*CustomerAddress spreadsheet:*

- 29 transactions made by customers with no address information. If the purchase was made online, it could be tracked.

- Nearly 35000 dollar from products sold to customer with missing address information.

- 3 customers (with id number 4001, 4002, and 4003) in the CustomerAddress sheet have no info in the CustomerDemographic sheet

- 4 customers (with id number 03, 10, 22, and 23) in the CustomerDemographic sheet have no info in the CustomerAddress sheet.

This indicates that the data received may not be in sync with each other which may skew the analysis results if there are missing data records.

*All spreadsheets - Dates:*

- The date format of tramsaction\_date in Transactions sheet is DD/MM/YYYY

- DOB in NewCustomerList sheet and in CustomerDemographic is YYYY-MM-DD

I suggest to follow the local Date format, such as the date and time Style manual from Australian Government (https://www.stylemanual.gov.au/style-rules-and-conventions/names-and-terms/dates-and-time). Therefore, dates should be DD/MM/YYY (numbers separated in a numeric date with an unspaced slash).

**C. (Not) relevant information**

*NewCustomerList spreadsheet:*

- I suggest to create a column containing the tenure of customers. That way, you can easily know if the customer is a 'new' one or not. By doing that you don't need to keep a separated spreadsheet only for new customers.

Feel free to reach me anytime if you have any query. Besides that, I would be available to spend some time with your data SME to ensure that all assumptions are aligned with your company's understanding.

Kind Regards,

Noa Prada Schnor | Data Analyst |KPMG