KPMG,
Australia,
Lighthouse and Innovation Team.

24th October, 2023.

Administrator,
Sprocket Central Pty Ltd,
Australia.

SUBJECT: FEEDBACK ON THE DATASETS RECEIVED BY YOUR COMPANY FOR ANALYSIS

Dear Sir,

I am writing to provide feedback on the datasets we received from your company, Sprocket Central Pty Ltd, earlier this month on the 15th of October 2023. After a thorough scrutiny of the datasets, we encountered some issues with the datasets and I have decided to share these issues with you in this letter including the remediating measures we planned to take.

The issues with each table are highlighted below including the mitigating measures we plan to take as follows:

Customer Demographic Table

The table was formatted by setting the DOB column in an ascending order and from there, some notable errors, omissions, invalid values and some other data quality issues were sighted and they are mentioned below based on standard data quality dimensions. I will therefore be categorizing the issues with the dataset below. In addition, I will be giving mitigation measures to be taken in addressing these issues. The same would apply to other tables as shown in this letter:

1. Missing values – the data contains some missing values including, for example, in the Surnames column, row number 14 with the first name Corabelle, has its surname not registered in the dataset. This is notable in rows 20, 54, 121 and so on. The Job title column is seen to have missing values as seen in rows 5, 34, 46, 49, and other rows. The same applies to the DOB and Tenure columns where the missing values were noticed from rows 3915 to 4001.

Mitigating measures to be taken: The missing surnames, job titles, DOB and tenure columns would be regarded as missing values and filtered out of the dataset since they cannot be automatically replaced.

2. Invalid values – the dataset is also seen to contain values that are invalid and do not fit into the context of the dataset. For example, the column tagged Default contains invalid and unreadable figures.

Mitigating measures to be taken: the entire Default column would be expunged from the dataset as it looks irrelevant.

3. Inconsistent values – The dataset contains inconsistent representations as seen in the gender column, the representation of male and female is not consistent. Male, Female, F, and U were used to ascribe the gender types.

Mitigating measures to be taken: the inconsistency would be addressed by defining a representation for each gender.

4. Irrelevant values – the Deceased column is irrelevant as it is noticeable that all the customers are alive and would therefore not have any impact on our data analysis.

Mitigating measures to be taken: the entire Deceased column would be expunged from the dataset as it looks irrelevant or redundant.

5. Inaccurate/inconsistent value – A notable inaccurate value was noticed in the second row of the DOB column, showing a DOB of 1843-12-21, meaning that the age of the customer is 180 years which is very unlikely.

Mitigating measures to be taken: the inaccuracy would be addressed as this could be an error during data entry and we suppose that the 1843 should instead be 1943. This is much more consistent with the other Date of Birth.

Transactions Table

1. Wrong object format – the transaction date column is not in the DateTime object format required for dates for analysis.

Mitigating measures to be taken: the transaction date column would be set to the required DateTime format needed for analysis of the dataset.

2. Missing values – there are a number of missing values in the Online order, Brand, Product line, Product class, Product size, Standard cost and Product first sold.

Mitigating measures to be taken: The missing values of Standard cost could be filled by the arithmetic method looking at other values available in the same column. The Online order, Brand, Product line, Product class, Product size and Product first sold columns could also be filled depending on the context of the rows and table in general.

3. Inconsistent values – the data representation contains some inconsistencies in the list price and standard cost columns. The unit of the list price is not represented in the column's header while the unit ($) of the standard cost is written with the figures in the cells of the column instead of being represented in the header of the column alone. This special string character may affect numerical analysis.

Mitigating measures to be taken: the special string would be deleted from the entire column while it would only be represented in the column header.

New Customer List Table

1. Missing values – the table shows some missing values as depicted in the Last name, Job title, and DOB columns.

Mitigating measures to be taken: The missing Last name, job titles, and DOB columns would be regarded as missing values and filtered out of the dataset since they cannot be automatically replaced.

2. Inconsistent values – the Property validation, columns show inconsistency as some figures are written in two decimal places and many others in whole numbers.

Mitigating measures to be taken: the figures would be rounded up to a constant number of decimal places

3. Wrong value format – the Date format for DOB is not consistent – 1973-03-15 to 1979-02-26 was written in another format compared to the rest of the column. This was made visible when the table was filtered by the DOB column.

Mitigating measures to be taken: the DOB column should be set to the right format before any analysis is carried out.

4. Irrelevant values – the deceased column is irrelevant as it is noticeable that all the customers are alive and would therefore not have any impact in our data analysis.

Mitigating measures to be taken: the entire Deceased column would be expunged from the dataset as it looks irrelevant.

Customer Address Table

1. Missing values – the entire row of customer ID number 3 is missing including the values for address, postcode, state, country, and property valuation.

Mitigating measures to be taken: the missing values would be filtered out of the dataset.

2. Inconsistent value representations – In the State column, it was noticed that the representation of the states was not consistent. In most cases, the New South Wales state was written in abbreviations but the same state written in full was also identified.

Mitigating measures to be taken: The states would be set to a single format of representation, that is, use of acronyms and the fully named states would be set as such.

Therefore, pointing out these issues and mitigating measures to be taken, means that these would be the first steps we would take in optimizing the quality of the datasets we received from you.

The Data, Analytics and Modelling team is always your top choice in; data analytics strategy and roadmap, dashboards and visualization, execution support, scheduling/optimization, compliance (controls, fraud), data monetization, and bespoke models.

Thank you for trusting us with your data and we will remain professional in our dealings as always.

Sincerely,

Young Irivboje,

Lighthouse and Innovation Team,

KPMG