To: Sprocket Central Pty Ltd.
From: KPMG, Data Analytics Department
Subject: Data Quality Assessment
Date: 10/07/2023

Dear sirs,

Thank you once again for choosing our company to assist you with recommending a suitable marketing strategy and for your keen interest in learning more about our expertise in Analytics, Information and Modelling. We have now taken the time to review the datasets you have provided to us (**Customer Demographic**, **Customer Addresses**, and **Transactions data**), and have conducted a data quality assessment before analysis. As part of our assessment, we identified a number of issues with the dataset including issues with accuracy, completeness and consistency, which we have detailed below. Recommendations have also been included to avoid the re-occurrence of data quality issues and to improve the accuracy of the underlying data used.

Summary Table	Accuracy	Completeness	Consistency	Currency	Relevancy	Validity
Customer Demographic	Date of birth found for 21.12.1843 which would not be accurate. Spelling errors under Job title and Job industry.	Blanks removed for a few columns including job industry.	Lack of consistency for gender (F, M, Female and Male)	Deceased customers have been filtered out.	Default column removed as not relevant.	
Customer Address			There are inconsistencies with the state.			
Transaction Data	Three of the currencies in the standard cost column needed to be reformatted as currency.	There are blanks in the online order and brand column.	There are more customer ids which are inconsistent with the other sheets.	The product first sold date is from 1991-2016 which is a huge time period.		List price and product sold date needed reformatting.

Accuracy

The first issue we identified with accuracy was within the Customer Demographic data. We noted that one of the customers within the dataset (Jephthat Bachmann) has a date of birth listed as 21 December 1843. This is a clear error in the data as the customer would then be 179 years old. This is something that could be updated, otherwise it is best to remove this customer's details from the dataset using the filter option. Another issue identified was in relation to a number of spelling errors. Under the job title column, "Nurse Practitioner" and "Data Coordinator" are both incorrectly spelt. This was resolved by using the spell check option throughout the entire column. A similar issue was found in the job industry column where "Agriculture" was incorrectly spelt as "Argiculture". We have again resolved this issue using the spell check tool. We also noted a number of formatting issues. Within the transactions data, we noted that the 'product first sold date' column is formatted as numbers rather than dates. We have changed the format of that column to a short date format for a more accurate representation of the data. To avoid this re-occurrence, you could ensure that the data in the database have constraints on data types.

Currency

We also identified some issues with the currency of the data. We noted that deceased customers were included in the demographics data which affects the currency of the data. We have removed the details for the two deceased customers so that more accurate data is being analysed. Another issue with currency is under the 'product first sold date' where the data provided is from 1991 to 2016. This may cause errors during analysis as a lot would have changed over the last 30 years. Focusing on data within the last 10 years is likely to be more useful for our analysis. It would also be important to ensure that all tables are from the same period as we noticed additional customer ids in the transactions and customer address dataset. Analysis will be conducted using the list of customers within the customer demographic table.

Completeness

In relation to the completeness of the dataset, there were a number of issues which needed to be addressed. Across the three sheets there were a number of blank cells, for example within the customer demographic data a number of customers were missing data relating to their last name. As there is a first name and other details provided for the customers it is possible to still proceed without filtering out this data. There was also an issue with completeness when it came to the job title of the customers. An option would be to fill these blank cells with "not specified" or "n/a". An issue has also been identified in the default column in the 'Customer Demographic' data. This column has been deleted as it is not relevant and has many unidentifiable values.

Consistency

In relation to consistency there are also some issues to address. Within the Customer Demographic data, the gender column is inconsistent, with different inputs such as "F", "M", "Female" and "Male". We have changed all of the references from F to Female and M to Male using the find and replace tool to follow one consistent form. There is also a spelling error, "femal", which we have manually changed using the filter option. Within the 'Customer Address' data there is also inconsistency with the input values in the state column. This was again resolved using find and replace to replace values with abbreviations. References to Victoria have been replaced with Vic and references to New South Wales replaced with NSW to ensure consistency. We recommend a drop-down list for the user entering the data rather than a free text field. The final point on consistency relates to the transactions data. Under the standard cost, some of the data needed to be reformatted as currency, e.g., 312.7350159. Alongside this, the data under price list has now been formatted as a currency.

We also noted that there were no duplicates within the dataset, so no duplicate values needed to be removed.

We hope that we have been able to clearly highlight the issues identified with data quality and also the strategies to mitigate these issues.

Should you have any questions related to the above mentioned points, please do not hesitate to contact us. In the meantime, the team will continue with reviewing the data in preparation for analysis.

Kind regards

Maria Tayo

Data Analyst, KPMG Data Analytics Department