Subject: Assessment of Data Quality and Recommendations for Sprocket Central Pty Ltd

Dear Client,

I hope this email finds you in good health. First off, all thank you for sharing the information related to the customer demographics, customer address and the transactions for the past 3 months. I was able to conduct a thorough review of the available dataset and was able to assess the quality of the data and identify potential issues that could affect the analysis.

1. Transaction Data:

In the transactions table I observed a total of 20,000 entries and upon assessment I discovered that 197 rows had at least 3 columns with null value which could affect the results. I am happy to conclude that all columns have properly defined null values. I noticed the column named ‘product\_first\_sold\_date’ was in a numeric format and needed conversion to a datetime format. Furthermore, I found no outliers or issues with the columns 'transaction date', 'list price', and 'standard cost' columns.

1. New Customer data:

Although you didn’t mention this dataset in the email, I still performed the analysis on it. And after reviewing the table, I noticed that the table had no index defined and I was able to set customer\_id as the index, this could prove useful when trying to combine with tables such as transactions. In the gender column I observed values ‘Female’ and ‘Male’ but then there was another value ‘U’ which if I’m not mistaken is undefined. The values in the DOB column were not all the same type, some were datetime but others object. I was able to successfully convert the non datetime objects to datetime for easier analysis. Diseased column had only the value N and it seems redundant as the whole column was filled with N and nothing else. The customers are all from Australia, this could skew the results. I found no outliers in the value column at the end.

1. Demographics data:

In the gender column I observed values ‘Female’ and ‘Male’ but then in-addition to these there were values such as F, M and Femal which I have converted to follow the same standards. The 'Diseased' column contained only the value 'N' and appeared redundant as the entire column was filled with this value. I calculated the age of customers based on the date of birth and found 1 customer that was born in 1843 which is not possible, and I think this could skew the results. The column named default was confusing and I couldn’t decipher it.

1. Upon closer inspection of the Customer Addresses data, I found that one customer's information was missing from the dataset. I also conducted an analysis of the 'property valuation' column and did not identify any outliers.

Based on the assessment, I have identified several data quality issues that require attention. I recommend the following methods to resolve these issues:

1. Perform proper conversion of the 'product\_first\_sold\_date' column in the Transaction data table to a datetime format.
2. Ensure that all data columns in the Customer Demographic and Customer Addresses datasets are correctly defined and contain the expected values.
3. Consider providing clarification on any undefined values, such as 'U' in the gender column of the Customer Demographic dataset.
4. Validate the accuracy and consistency of the 'DOB' column in the Customer Demographic dataset to eliminate any outliers that could impact the analysis.
5. Provide clarification regarding the purpose and meaning of the 'default' column in the Customer Demographic dataset to ensure accurate interpretation.

Please note that addressing these issues could go above and beyond when performing the analysis at a later stage. More optimizations can be performed during the analysis based on needs there and then.

Thank you for your cooperation and support.

Kind Regards,

My Name

My Pos

KPMG Analytics, Information & Modelling Team