Greetings,

We are excited to be working for Sprocket Central Pty Ltdand can’t wait for further meaningful engagement.

I had a chance to go through the data provided to identify some of the data quality issues. Kindly note that the quality of data provided is very key as it informs the kind of results and insights to be generated from our analysis. We take steps to ensure that we have correct, consistent and relevant data for optimized results.

That being said I noticed some data quality issues from the data provided. They include:

1. Completeness

There were several instances of missing data in the table including but not limited to data on whether orders were online or not in the Transactions table for transaction ID’s 167, 170, 251, 301 and a couple others

Additionally, there was a big chunk of missing data on details about the purchased products on the transaction table for several transaction IDs. After the order status column only the list-price is provided and there lacks information about the brand, product line etc. for a couple of entries.

The lack of this data could significantly impact the kind of results.

1. Accuracy

I noted one entry on the customer demographic table with the gender F. The other entries are classified as either male, female or U. Confirm that this might have been a mistake on your side and intended for female?

1. Relevancy / Validity

The default column in the customer demographic table lacks data values and the relevancy of the data is not quite clear.

1. Consistency

The customer IDs in the Transactions table jump from 3500 to 5034 revealing an inconsistency in the records. Furthermore, in the customer address table there are records of customer IDs up to 4003. Having this kind of data may make it extremely hard to combine the tables and gather insights for instance matching customer IDs to names may prove to be a dauting task.

In the new CustomerList table there are records in the same column with different data types e.g postcodes of numbers and others text, value with different decimal places. This could be a slight error from your side but also questions the accuracy of data and a confirmation of the same would be helpful.

These are a few of the data quality issues I noticed from the data provided. We hope to get clarity on the same and overall optimize our data analysis process. Looking forward to hear from you.

Salome Mbithe,

KPMJ Intern