Dear Sprocket Central Pty Ltd,

I hope this email finds you well.

I have reviewed the following data sets, Transactions, Customer demographic and Customer address. I’ll highlight the recurring data quality issues, provide mitigation methods and recommendations on how the data accuracy can be improved.

* **Several columns throughout all datasets had missing values.**

Completeness issues can be found across two of the data sheets, transactions and customer demographic. The main concern here is when core values in the transaction sheet are missing such as brand and online\_orders. The blank values considered of less than 2% of the database, making the overall completeness of the database above 98%. The blank data is going to be removed in the analysis phase.

**Mitigation:** Ensure that core values have a requite to be filled to avoid, incomplete core values from entering the database. Non-core values can be filtered out e.g., online\_orders.

* **Inconsistent values for the same attribute**

Discrepancies between formats, spellings or units can cause several issues in the analysis phase. For example, the States Column in Customer address had several different values for the same state such as NSW or New South Wales. This will be standardised to the abbreviated version. Another example is the gender column in the Customer Demographic data set with entries such as Female, Femal and F.

**Mitigation**: Set up option limitations upon data collection where customers can only select a specific set of values as opposed to a free text field. A drop-down list would work very well here.

* **Inconsistent and incorrect data types**

Taking particular focus on the Transactions data set there are a couple of data type issues where numeric currency values for example standard\_cost has some values listed as a general data type leading to results such as 667.4000244.

**Mitigation:** Standardise existing data to a numeric data type. To ensure data type issues don’t occur again, create data type constraints within the database.

* **Data Accuracy**

One of the overarching issues across all data sets is accuracy.

***Mitigation****: Create option limitations for value selection in the data collection phase, particularly if its core data. Standardising data types and further limited what types of data can be inputted in the data collection phase. This will reduce the amount of data accuracy issues from occurring.*

We will continue to make clean, standardise and transform the data in preparation for analysis. If you have any questions or concerns, feel free to get in touch. We’ll be in contact again to discuss the insights gathered from data and how best practices can be implemented moving forward.

Kind regards,

Luis