To: [Client@gmail.com](mailto:Client@gmail.com)

Subject: Data quality assessment of Sprocket Central Pty Ltd.

Dear [Client point-of-contact],

I hope this mail finds you well. Thank you for providing us with the 4 datasets related to Sprocket Central Pty Ltd. Data quality assessment for the datasets have been performed based on a data quality framework that we follow here, which is also attached in the email. Data quality issues were found in all the datasets given to us which needs to be addressed. Furthermore, recommendations have been provided to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

1. **Outliers:**

Outliers make the data inconsistent to draw insights from them. In the dataset, eg: The customer ID “34” with the name of Jephthah Bachmann was born in 1843, meaning that he is 175 years old which is an error in the data in the Customer Demographic Table.

**Recommendation:** Eliminate the redundant data to get a more meaningful dataset.

1. **Missing values:**

The datasets had several missing values in various columns, which add no value to us. Columns like “Online Order”, “Brand Name”, “Product Line”, “Product Class”, “Product Size”, “Standard Cost”, and “product\_first\_sold\_date” in the Transactions table has missing values. Also, In the Customer Demographic “Job Title”, “Job Category” and “Tenure” columns some of the records were missing.

**Recommendation:** The percentage of missing values as compared to the whole size of the records is really low, hence they missing values can be removed from the dataset.

1. **Inconsistencies in the values of the dataset:**

The Transactions table has more customer\_ids than the Customer Demographic and Customer Address Table, which may skew the analysis results if there are any missing records. Therefore, it is important to ensure that the data is in sync with each other to avoid any inconsistencies.

**Recommendation:** Use data quality rules and standards to maintain consistency. Implement data quality controls and monitoring to detect and correct any issues.

1. **Duplicated values:**

In the "State" column of the Customer Address Table, there are multiple duplicate values such as "VIC" and "Victoria", and "NSW" and "New South Wales". Similarly, the "Gender" column of the Customer Demographic Dataset has some issues.

**Recommendation**: To use abbreviations of the states instead of full names for all the records to ensure consistency across addresses. For Gender Column, the records “U” can be imputed with the distribution of the dataset.

1. **Inaccurate values for the same attribute:**

In the transaction dataset, the values in the **product\_first\_sold\_date** indicate that all products have the same sold date but at different times. Moreover, for the same product, we find different first sold dates. Please make sure to check this. In customer demographics dataset, we notice that the **default** column contains some random characters so we decide to drop it.

**Recommendation:** Establish clear guidelines and procedures for data entry, on how to handle missing data and how to resolve discrepancies.

I hope that this addresses the data quality issues that were encountered. We are always striving to add value to our client’s business needs. With respect to the data quality issues, our team is relentlessly working towards it and always finds ways to make it better. Further to this, we would like to arrange a meeting with the data lead for this project so that we align on our expectations and achieve our goals together. Kindly document your queries associated with this and we shall discuss further before we proceed to phase two: Data Analysis.

Regards,

Sadhana Savvaser

Junior Data Analyst, KPMG