Greetings Sir,

Below is a full description on the **Data Quality Assessment** made on the three datasets provided by the **Sprocket Central Pty Ltd**. These datasets include **Customer Demographics**, **Customer Addresses**, and **Transactions data in the past 3 months**. This description covers the observed data quality issues, its impact on our analysis and recommendations to mitigate such issues.

1. **ACCURACY**

* **Customer Demographics Dataset**

*Quality issues*:

* The Gender Column contains incorrect values such as F, U, Femal, M.
* DOB column contains some questionable values indicating ages as high as 120 to 177 years.

*Impact on Analysis*: This can lead to inaccurate data output, which prohibit or contaminate data-driven decisions.

*Recommendation*: Its usually advisable to get the data directly from the client’s or customer’s website.

1. **COMPLETENESS**

* **Transactions dataset**

*Quality issues*:

* **1542** are number of missing values from columns such as: online\_order, brand, product\_line, product\_class, product\_size, standard\_cost, and product\_first\_sold\_date.
* **Customer Demographics Dataset**

*Quality issues*:

* **1763** are number of missing values from columns such as: last\_name, DOB, job\_title, job\_industry\_category, default, and tenure.

*Impact on Analysis*: This data quality issue makes it difficult to generate accurate insights from the data. It equally makes the rest of the information provided to be less useful.

*Recommendation*: It is advisable to make sure same data are collected from each customer.

1. **CONSISTENCY**

*Quality issues*:

* The dimensions of the three datasets are not the same.
* There is a significant difference in the number of columns and their labeling.

*Impact on Analysis*: The analysis is can be uncoordinated because of inconsistence of the data. This implies that, actions implemented on data might indirectly affect the other datasets.

*Recommendation*: Attention should be paid in storage of data in multiple places, the dataset should be consistent across all of them.

1. **CURRENCY**

*Quality issues*:

* From the Transaction date, the data is 3years backward
* There might be possible changes in the customer’s address in this past years.

*Impact on Analysis*: Using outdated data can lead to inaccurate results and taking actions that don’t reflect the current reality.

*Recommendation*: Data should be imported as soon as possible and used from within a pre-determined timeframe.

1. **RELEVANCY**

* **Customer Demographics Dataset**

*Quality issues*:

* The default column is irrelevant to our analysis
* The NewCustomerList is not need since it was not part of the three datasets mentioned earlier.

*Impact on Analysis*: This can lead to an increase in operational costs. It is equally a share waste of time and resources in gathering such datasets.

*Recommendation*: Only the datasets that are related to the customer should be collected.

1. **UNIQUENESS**

* **Transactions dataset**

*Quality issues*: There are duplicated values in columns such as product\_id and customer\_id.

* **Customer Demographics Dataset**

*Quality issues:* There are duplicated values in columns such as first\_name and last\_name.

*Impact on Analysis*: This data quality issue makes it difficult to generate accurate insights from the data

*Recommendation*: A piece of data should be recorded once. Then if there is no single, then it has to be separated.