Dear Manager,

After thorough research on data quality in conjunction with the data quality framework table provided to me, I was able to discern various data quality issues. Despite the large dataset from Sprocket Central Pty Ltd, there are a plethora of issues that need to be addressed prior to further data analysis. These are as follows:

## **Data Quality Issues**

### **Accuracy**

There is evidence of issues in accuracy of the dataset. For example:

- The product_first_sold_date covariate in the transaction data comprises of values which are not dates but rather integers.
- In the NewCustomerList dataset, we observe that for the customer Pauline Dallosso (the 987th data entry), their biological gender seems to be a value which is not Male or Female, thus thus value is not correct.
- Another obvious data quality issue can be seen from the customer demographic data where the “default” variable seems to have data entries which seem to be erroneous.

### **Completeness**

There is evidence of issues in completeness of the dataset. For example:

- Various missing values hinder the completeness of the data quality, for example in transactions there are missing values in the product_line, product_class and product_size.
- Other missing values are evidenced in NewCustomerList, where some individual customers do not have a job title.

### **Consistency**

There is evidence of issues in consistency of the dataset. For example:

- In the NewCustomerList, the property valuation for all the states in Australia are integers except a few values which are float numbers, showcasing that there is issues in consistency.
- In addition, the customer address data highlights that the values for the state of the customer address is mostly an acronym, except for new south wales which is abbreviated, thus making it inconsistent.

### **Currency**

There is evidence of issues in currency of the dataset. For example:

- Some of the values in the customer demographic dataset showcases that there are entries which are N/A values. This means that the data does not contain up to date values, such as the job industry category which each customer works in.

## Relevancy

There is evidence of data relevancy of the dataset. For example:

- The customer demographic dataset should showcase the information of the customer's personal information such as their name, date of birth, job title and so on. However, unknown variables such as "default" and "deceased_indicator" might not be relevant for KPMG given that all the customers are alive.
- Customer Address information does not necessitate property valuation for each customer, hence it is not so relevant for KPMG.

## Validity

The dataset overall has high validity.

## Uniqueness

There are various areas in the entire dataset where values are being repeated, for example the deceased indicator in the new customer list dataset, which indicates that all customers are alive. This is an indication that the dataset lacks uniqueness because these values are not that necessary and for KPMG they require information that is more imperative and unique for analysis.

I hope the data quality issues and strategies I have proposed to mitigate these issues are useful for you and KPMG. Feel free to contact me if you have any queries or comments.

Kind regards,
Faiyam Islam