Dear [Sprocket Central Pty Ltd Associate Director],

This is Abdelrahman Akmal from KPMG Data analytics team, Hope this mail finds you well. Regarding the datasets you have sent to us we have conducted a data quality assessment on each of the datasets to find issues that may obstruct us from proceeding with the analysis process to mitigate it. I have provided you with a list of the issues that we find and our recommendations for the procedures that we see that you should take.

1. **Issue: Wrong Data types of columns.**

   All Dates columns as “DOB” column in customer demographics sheet and new customers sheet, and “product_first_sold_date” column in transactions sheet doesn’t have datetime type which make their values inconvenient and make it hard to interact with those columns.

   Recommendation: The type of those columns should be changed to datetime type.

2. **Issue: Missing Values.**

   Some columns like “Online Order”, “Brand Name”, “Product Line”, “Product Class”, “ Product Size”, “Standard Cost”, and “product_first_sold_date” in the Transactions table had missing values. Also, In the Customer Demographic “Job Title”, “Job Category” and “Tenure” some of the records are missing, also for “gender” column some have “U” value for Undisclosed which we consider as a missing values.

   Recommendation: Either remove the records that have missing values as their percentage is small or find complete those records either by the real values or by finding a reasonable imputation technique based on the other attributes.

3. **Issue: Duplicated Values in same column.**

   Some columns like “state” in customer address table have duplicated values that have same meaning as “VIC” and “Victoria”, “NSW” and “New South Wales” , also “gender” column in customer demographics table have same issue as “F”, ”Female”, ”Femal” and “M”, ”Male” .

   Recommendation: Use abbreviations as the default values so “state” column values will be either “VIC”, “NSW”, and “QLD”, while “gender” values will be either “M”, or “F” .

4. **Issue: Inconsistent records over the datasets.**

   The “default” column in the customer demographics table doesn’t have any reasonable meaning that could affect our analysis.

   Recommendation: Drop “default” column in customer demographics table.

5. **Issue: Meaningless columns.**

   The is inconsistency between the number of records in “customer_id” through the datasets as transaction table has 20,000 unique customers record, while customer address table has only 3,999 unique customer record and customer demographics table has only 4,000 unique customer record.

   Recommendation: We will proceed only with the matched customer records between all datasets, which will be 3,999 customers.

Please look closely at the above-mentioned issues and send us the strategy that you want us to apply regarding those issues to proceed with the analysis.

Kind Regards,
Abdelrahman Akmal