**Subject: Data Quality Assessment Summary**

Hello Sprocket Central Pty Ltd,

I hope this message finds you well. As part of our data preparation process for analysis, we conducted a thorough assessment of the given data sets. Here is a summary of our findings and the actions taken to ensure data quality and completeness:

**For Customer Demographic Data Set:**

* **Data Overview:** Initial dataset has 4000 rows and 13 columns. Null values found in 6 columns, which were carefully addressed based on the nature of the data and analysis requirements.
* **Data Standardization:** We can transform all data values in to one case for consistency across records.
* **Data Consistency:**
  + Gender Categories: We standardized gender categories to 'Male' (M), 'Female' (F), and 'Unspecified' (U) for consistency.
  + Deceased Indicator: Records denoted as 'Y' were removed.
* **Age Calculation:** Age calculations were performed based on the Date of Birth (DOB) column. Customers with ages less than 65 were retained for analysis.
* **Column Removal**: The 'Default' column was removed as it did not provide meaningful insights.

**For Customer Address Data Set:**

* **Data Overview:** Initial dataset has 3999 rows and 6 columns. No null values were detected in the dataset, ensuring data completeness.
* **Data Standardization:** We can transform all data values in to one case for consistency across records.
* **Data Consistency:** In the 'State' column, we recognized that "New South Wales" and 'NSW' and "Victoria" or 'VIC' are referring to the same locations. Therefore, we standardized them as 'NSW' and 'VIC' for consistency.

**For Transaction Data Set:**

* **Data Overview** Initial Dataset: 20,000 rows and 13 columns.We identified null values in 7 columns, necessitating further action.
* **Handling Null Values:** Null values were addressed based on the nature of the data. For example, in the “standard\_cost” column, we replaced null values with the mean standard cost of that particular “product\_id”.
* **Data Standardization:** We can transform all data values in to one case for consistency across records.
* **Outliers:** In the "standard\_cost" column, we identified 195 outliers. These outliers were replaced with the mean standard cost of the corresponding product, ensuring a more representative dataset.
* **Data Format Correction:** We observed incorrect data in the "product\_first\_sold\_date" column. The column values were transformed into the proper date format to facilitate accurate analysis.

**For New Customers Data Set:**

* **Data Overview:** Initial dataset had 1000 records and 23 columns.4 columns contained null values, which required consideration for further analysis.
* **Handling Null Values:** Null values were addressed based on the specific requirements of the analysis, either by replacement or removal.
* **Data Standardization:** We can transform all data values in to one case for consistency across records.
* **Age Range Selection:** Considering the nature of the analysis and product-specific criteria, we filtered the dataset to include records for individuals aged 20 to 60 years.
* **Unnamed/Calculated Column:** We identified and removed 5 unnamed or calculated columns from the dataset, as they were deemed unnecessary for the analysis.

In conclusion, the data has been carefully assessed, cleaned, and prepared for analysis. We are confident that the dataset is now of high quality and ready for further exploration.

Should you require any additional information or wish to proceed with the analysis, please feel free to reach out. We look forward to your instructions on the next steps.

Best regards,

Shouvik Barik

Mail Id: [barikshouvik@gmail.com](mailto:barikshouvik@gmail.com)

For Detailed Analysis you can find the following github repository – Task 1 -> [click here](https://github.com/Half-boiled-egg/Sprocket-Central-Pty-Ltd-Data-Set-Analysis_KPMG-Virtual-Internship.git)