Dear [Name of the Client],

I hope this message finds you well. We have received the three raw datasets from SP Rocket Central Pty Limited. As part of our preliminary data assessment, we've thoroughly reviewed the data for quality and have identified several key issues that warrant your attention. Below is a summary of the issues we've uncovered, along with our recommendations on how to address them to improve the overall data quality:

1. ## Incorrect Data Types

   Issues: In the Transaction table, the 'online_order' attribute has a data type of float64 but should be int64. The 'product_first_sold_date' column in the same table is currently stored as float64 but should be converted to a datetime type. In the Customer table, 'DOB' is currently an object type and should be converted to datetime. In the Demographic table, the 'gender' column is of object type and should be boolean. Additionally, 'DOB' is an object type and should be converted to datetime. The 'deceased_indicator' column has an object data type but should be boolean. The 'default' column has mixed types, from datetime to boolean. Lastly, 'owns_car' is of object type but should be be boolean. The 'tenure' column has a data type of float64 but should be int64.

   Recommendation: Address the above data type issues by converting the data types to their appropriate forms. It is also recommended to delete the 'deceased_indicator' variable as all rows have the same value ('N'). Columns 16 to 20 should be deleted from the Customer Data Set.

2. Missing Values.

Issue: Multiple attributes, such as 'Online Order,' 'Brand,' 'Product Line,' 'Product Class,' 'Product Size,' 'Standard Cost,' and 'product_first_sold_date' in the Transactions table contain blank values. Notably, the percentage of missing values in the 'brand' attribute corresponds to the same rows that have missing values in the columns 'product_line,' 'product_class,' 'product_size,' 'standard_cost,' and 'product_first_sold_date.' In the Customer table, 'last_name,' 'DOB,' 'job_title,' and 'job_industry_category' have some records with missing values. In the Demographic table, the columns 'last_name,' 'DOB,' 'job_title,' 'job_industry_category,' 'default,' and 'tenure' contain missing values. Of particular note is that the missing values in 'tenure' align with rows where 'DOB' is missing in this table.

Recommendation: Given the relatively low percentage of missing values in the datasets, we recommend proceeding by removing the records with missing data.

3. Cleaning Issues

Issue: In the 'state' column of the Address table, we observed variations such as 'New South Wales' appearing instead of 'NSW' and 'Victoria' appearing as 'VIC.' Similarly, in the 'Gender' column of the Demographic Dataset, variations like 'F' or 'Femal' occur instead of 'Female,' and 'M' instead of 'Male.'

Recommendation: To ensure consistency, we recommend using abbreviations for states and standardizing gender values.

4. Redundant Outliers.

Issue: The 'past_3_years_bike_related_purchases' column appears to have potential outliers based on standard deviation.

Recommendation:: Further investigation is needed to determine whether these values are genuine outliers or if they represent valid data points that don't significantly impact the data distribution.

5. Inconsistencies

Issue: We observed a discrepancy in the number of customer IDs between the 'customer list' and 'address' tables. Additionally, there is a customer with an age reported as 177, which is highly unlikely.

Recommendation: We recommend conducting the analysis solely on synchronized data from all customer tables, using the 'customer_ID' as the key. To resolve the discrepancy in customer IDs, the additional row in the demographic table should be deleted. Furthermore, customers with ages greater than 90 should be excluded from the analysis.

Please review the above-mentioned data quality issues and consider implementing the recommended changes to ensure the uniform quality of the dataset across all the tables. Once these issues are addressed, we can proceed with a more in-depth analysis to extract valuable insights for your company.

Should you have any questions or require further assistance, please do not hesitate to reach out to us. We look forward to your response and to assisting you with optimizing the quality and utility of your data.

Best regards,

Rúben Serpa