Subject: Data Quality Issues and Recommendations for Enhanced Insights

Dear [Name of the Client] at Sprocket Central Pty Ltd,

I hope this message finds you well. I am writing to discuss the quality issues we have identified in the three raw datasets recently received from SP Rocket Central Pty Limited. Our team has conducted a thorough analysis and discovered several areas that require attention to ensure the consistency and effectiveness of the data.

Outlined below are the key quality issues identified, along with our corresponding recommendations:

1. **Redundant Outliers:**

**Issue:** We have identified outliers in the data that may significantly impact the dataset's integrity. For instance, in the Customer Demographic Table, customer ID "34" belongs to Jephthah Bachmann, who supposedly has a birth year of 1843, making him 175 years old. This clearly indicates an error in the data.

**Recommendation**: We recommend removing redundant data points that may skew the overall dataset distribution, thus ensuring accuracy and reliability.

1. **Missing Values:**

**Issue:** We observed missing values in various attributes of the Transactions table, such as "Online Order," "Brand Name," "Product Line," "Product Class," "Product Size," "Standard Cost," and "product\_first\_sold\_date." Similarly, the Customer Demographic table shows missing records in the "Job Title," "Job Category," and "Tenure" fields.

**Recommendation:** As the percentage of missing values is relatively low compared to the dataset as a whole, we suggest proceeding by removing these records to maintain data completeness.

1. **Inconsistent Entries across the Datasets:**

**Issue:** The Transactions table contains a higher number of customer IDs than the combined count in the Customer Demographic and Customer Address tables, resulting in data skewness when missing records are encountered.

**Recommendation:** To ensure accurate analysis, we strongly advise using synchronized data across all three customer tables based on the customer ID.

1. **Multiple Datatypes for a Single Column:**

**Issue:** We discovered that the "Standard Cost" attribute in the Transactions table exhibits inconsistencies due to the presence of special string characters in certain records.

**Recommendation:** To ensure consistent data types, we recommend removing special characters from the affected records and converting all values to numeric data.

1. **Duplicate Values for the Same Column:**

**Issue:** The "State" column in the Customer Address table contains duplicate values such as "VIC" and "Victoria," as well as "NSW" and "New South Wales." A similar issue is observed in the "Gender" column of the Customer Demographic dataset.

**Recommendation:** To ensure uniformity across addresses, we suggest using state abbreviations instead of full names for all records. Regarding the Gender column, we recommend imputing the "U" records using the dataset's distribution.

We kindly request your attention to these data quality issues and urge you to implement the recommended changes. By doing so, we can ensure a consistent and reliable dataset that will enable us to extract meaningful insights for your esteemed company.

Thank you for your time and cooperation. Should you have any questions or require further assistance, please feel free to reach out to us. We look forward to your prompt response.

Best regards,

Shantam

Data Analyst