Subject: Data Quality Issues and Recommendations for Analysis

Hello,

This is Kamalesh from Data Analytics Team. I hope this email finds you well. We have reviewed the three raw datasets received from Sprocket central Pty Limited, and we have identified multiple data quality issues that need to be addressed. In order to improve the effectiveness of the data and ensure accurate analysis, we recommend the following strategies:

1. Redundant Outliers: Issue: Some data values are outliers and can disrupt the dataset. For example, customer ID "34" in the Customer Demographic Table has a birthdate of 1843, which would make the person 175 years old. Recommendation: Remove redundant data to avoid skewing the dataset distribution.

2. Missing Values: Issue: Several attributes in the Transactions table and Customer Demographic table have missing values. Recommendation: Since the percentage of missing values is low compared to the dataset, we suggest removing these records.

3. Inconsistent Entries across Datasets: Issue: There are more customer IDs in the Transactions table than in the Customer Demographic and Customer Address tables, causing skewed data. Recommendation: Perform the analysis only on the synchronized data across all three customer tables using customer IDs.

4. Multiple Datatypes for a Single Column: Issue: The "Standard Cost" attribute in the Transactions table contains special string characters, leading to inconsistency in the dataset. Recommendation: Remove special characters and convert all values to numeric data to ensure consistent data types.

5. Duplicate Values for the Same Column: Issue: The "State" column in the Customer Address table has multiple duplicate values, such as "VIC" and "Victoria," and the "Gender" column in the Customer Demographic dataset has inconsistencies. Recommendation: Use state abbreviations instead of full names to ensure consistency in addresses. For the Gender column, replace "U" with a distribution-based imputation approach.

In addition to these recommendations, we suggest the following actions to enhance the overall quality of the dataset:

- Merge all three datasets since they are relevant to each other.
- Replace missing DOB values using the mode of other customers' birthdates.
- Replace missing tenure values with the mean of available tenures for consistency.
- Verify the accuracy of the "product_first_sold_date" column in the Transactions dataset.
- Remove unnecessary columns like "default" in the Customer Demographic table.
- Use "Male," "Female," and "Unidentified" as values for the gender column to improve relevance and readability.
- Consider eliminating missing blank values as fake or incomplete entries.
- Make an effort to fill in missing values for more comprehensive analysis in phase two.

By implementing these strategies, we can mitigate the data quality issues and ensure a consistent dataset for further analysis. Should you have any questions or require additional information, please feel free to reach out. We look forward to proceeding with the analysis and uncovering valuable insights for the company.

Thank you for your attention to this matter.

Best regards,

Kamalesh K B

Data Analytics Team