Subject: Data Quality Assurance Report

Dear Sir/Mam,

I hope this email finds you well. I am writing to provide you with an update on the data quality assurance process for the given datasets and to propose strategies for addressing the identified issues.

Firstly, I am pleased to inform you that our data quality assessment process has been successfully completed, and the datasets have undergone rigorous scrutiny to ensure accuracy, completeness, and consistency. The majority of the data has met our stringent quality standards, which is a testament to the efforts put forth by the team involved.

However, during the assessment, we have identified several issues that, if rectified, will further enhance the quality and reliability of the data. After careful consideration, we have listed the issues and devised the following strategies to address these concerns:

**Summary of given datasets:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Number of rows** | **Number of columns** | **Total number of entries** |
| Transactions | 20000 | 14 | 280000 |
| New Customer List | 1000 | 18 | 18000 |
| Customer Demographic | 4000 | 13 | 52000 |
| Customer Address | 3999 | 6 | 23994 |

**Major issues in the given datasets:**

1. There are few entries within the dataset which contain missing values. These missing values indicate the absence of data for certain attributes or variables. It is crucial to address this concern as missing values can significantly impact the accuracy and reliability of the dataset.

**Summary of missing values in the given datasets:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Total number of entries** | **Number of missing entries** | **Total number of rows** | **No. of rows with missing values** |
| Transactions | 280000 | 1542 | 20000 | 555 |
| New Customer List | 18000 | 317 | 1000 | 285 |
| Customer Demographic | 52000 | 1763 | 4000 | 1370 |
| Customer Address | 23994 | 0 | 3999 | 0 |

1. In the customer demographic table, it is observed that the gender column exhibits inconsistencies. Specifically, we have identified three variations in which female gender is represented: "femal", "F" and "female" being used in different entries and two variations in which male gender is represented: "M", "Male" Such inconsistency can lead to data interpretation challenges and hinder accurate analysis.
2. In the customer demographic table, there is an entry in the dataset where the birth year is recorded as 1843. This raises concerns about the accuracy and reliability of the data, as it seems unlikely for an individual to have a birth year that far in the past.
3. In the customer address table, the state "Victoria" is represented in two different formats: "VIC" and "Victoria."
4. In the transactions table, there are several instances where referential integrity is compromised, indicating discrepancies and inconsistencies in the relationships between tables. Specifically, we have discovered the following issues:

* In the transaction table, there is a record with a customer ID of 5034, which is not present in the customer demographic table. This inconsistency indicates a breakdown in the relationship between the transaction and customer demographic tables, potentially leading to incorrect or incomplete analysis.
* The customer address table is missing address details for customers with the IDs 3, 10, 22, and 23. This absence of address information for specific customers raises concerns about data completeness and accuracy. It is essential to address this issue to ensure a comprehensive understanding of customer addresses.
* The customer address table includes details for customer IDs 4001, 4002, and 4003, which do not exist in the customer demographic table. This discrepancy indicates a violation of referential integrity and highlights the need to synchronize data across related tables for accurate and reliable analysis.

**Strategies to mitigate the issues in the given datasets:**

1. To address the issue of missing values in the dataset, our strategy involves employing suitable imputation methods instead of dropping the affected records. Since a significant number of records have missing values, it is essential to preserve the data. We propose utilizing imputation techniques such as mean, median, mode imputation to fill in the missing values. By applying these methods, we aim to retain the integrity of the dataset, enable comprehensive analysis, and minimize the loss of valuable information. Documentation of the imputation methods used will ensure transparency and reproducibility in future analyses.
2. To overcome issues with referential integrity in the dataset, we can establish mechanisms for data integration and synchronization to ensure consistent and accurate data across all related tables. We can maintain proper documentation of the relationships and constraints to facilitate ongoing data management and future analysis. By implementing these steps, we aim to restore and maintain referential integrity, ensuring the reliability and usability of the dataset.

These strategies are designed to tackle the specific issues we have encountered and will contribute significantly to improving the overall quality of the datasets.

Should you require any additional information or assistance regarding the suggested strategies, please do not hesitate to reach out to me. I am here to provide guidance and address any questions you may have.

Thank you for your attention to this matter. Your commitment to data quality is instrumental in our collective efforts to make informed decisions based on reliable information.

Moving forward, our analysis team will help you in analyzing your datasets and providing valuable insights and business strategies to become successful among the competitors.

Best regards,

Roshankumar S