Dear Sprocket Central Pty Ltd,

I have taken the time to deeply examine the dataset, and there are issues that I have identified below. I have also provided some recommendations to rectify these shortcomings.

Before that, I have drawn up a summary statistics of the three datasets below:

|  |  |  |  |
| --- | --- | --- | --- |
| Sprocket Central Pty Ltd. Data Set | No. of Records | Distinct Customer IDs | Date Data Received |
| Transaction Data | 2001 | 2000 | 2017 |
| Customer Demographic | 4001 | 4000 | 2017 |
| Customer Address | 4000 | 4001 | 2017 |

The dataset is currently not clean, and therefore, not ready for analysis. Below, I have listed these discrepancies, and suggested ways to improve and mitigate the data quality to make it analysis ready.

The datasets have the following issues:

1. **Accuracy**: The ‘CustomerDemographic’ data, the ‘DOB’ column has a date 1843, which would mean that the customer is 180 years old. This is an error, and it makes the entry inaccurate.

**Recommendation**: The entry should be removed, as leaving it there can skew the result of the analysis.

1. **Missing values**: Across the three sheets of the data set, there are some blank cells and missing values. Transaction\_id 37, for instance, has several missing variables. This therefore makes the data incomplete.

**Recommendation**: The blank cells should be deleted from the dataset.

1. **Inconsistency:**

* The values in the Gender column of the ‘CustomerDemographic’ data are not consistent. There are male, female, F, M and U.
* For the ‘customerAddress’ data, in the ‘state’ column, some entries were abbreviated like ‘NSW’, while some were ‘New south wales’. There is also ‘QSD’, and ‘VIC’
* The distinct number of Customer IDs are different for the three data. Transaction data has 2000 distinct customer IDs, Customer Demographic has 4000, while Customer Address has 4001.

**Recommendations:**

* The cells with ‘F’ should be corrected to ‘Female’ and the ones with ‘M’ should be corrected to ‘Male’, so as to ensure uniformity with the rest of the values.
* The abbreviated entries should be written in full. Or the ones in full should be abbreviated. Either should be done to ensure consistency across the column.

1. **Currency:** As at the time of this task, which is July, 2023, this dataset is not up to date, as the ‘Transaction’ data, which was the only one that had a date, shows the transactions happened in 2017. That was six years ago.

**Recommendation:** A much more current and relevant dataset may be added as a lot would have changed within 6 years such as the cost and the overall buying pattern of customers.

Summarily, the data is a dirty one, and these recommendations will clean it up and make it analysis ready. Proceeding to analyse dirty data will skew the analysis result, which will in turn lead to poor business decisions.

The datasets have been looked into, and corrections would be made accordingly, after which we can now proceed with analysis, and subsequently, visualisation. Please, let us know what so you think, and if we can proceed to with cleaning the data.

Regards,

Salami Opeyemi.