Dear Sprocket Central Pty Ltd,

I have looked through the three datasets provided and the table below summarizes the statistics from each dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **No of records** | **No of unique customer IDs** | **Date data received** | **No of rows with invalid entries** |
| Transactions | 20000 | 3496 | 26/06/2020 | 555 |
| Customer Demographic | 4000 | 4000 | 26/06/2020 | 689 |
| Customer Addresses | 3999 | 3999 | 26/06/2020 | 0 |

* **There are more customer\_ids in Customer Demographic than in transactions and customer addresses.** Customer addresses has 1 less customer\_id than Customer Demographic and transactions has 504 less customer\_ids than Customer Demographic. This means that the datasets are inconsistent and may be from different times.
* **Certain rows have empty field values.** In Transactions there are 555 empty cells and in Customer Demographic there are 689 empty cells. Ideally we can go through the dataset and if the empty cell value is essential to the data analysis we can ignore that entire row and if the empty cell value was nonessential we can choose to ignore that empty cell and continue with the rest of row data. Empty field values may skew the data and make it inaccurate.
* **Default in Customer Demographic has useless data.** The default column consists of special characters and doesn’t give us any information. I think it would be safe to discard this column entirely as it provides us with no new information.
* **Date in transactions.** Date column is empty for all entries so it would be safe to delete the column.
* **Invalid cell values.** There is a customer with ID 34 in Customer Demographic with age over 110. This is very unlikely so we can safely discard this entry.
* **Inconsistent values from the same attribute.** For example, Victoria being represented as “V”, “Vic” and “Victoria”. To counteract this we can use a regular expression to replace these values with just one so it will be more consistent.

In general the dataset quality is good if we can fix the issues addressed above.

From,

Anirudh Lakra