Hello sir/madam,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. In this email I highlight some of the issue encountered with the data and some possible solutions to them. But first, I would like to confirm whether the basic summary of the data is to your understanding.

|  |  |  |
| --- | --- | --- |
| Dataset | No. of IDs | No. of Records |
| Transactions | 3494 | 20000 |
| Customer Address | 3999 | 3999 |
| Customer Demographic | 4000 | 4000 |

To note is that there are different total number of customer IDs per dataset, therefore, increasing the chance of having missing values after merging the datasets.

Some other issues encountered:

**Transactions Dataset**

Some columns have been presented in wrong data types such as online\_order which should be a non-numeric column.

Missing values in columns such as brand, product\_line, etc. A good remedy would be to make sure that all columns are filled during data entry in future. However, for the current data, methods like filling with mean, or median for numerical columns and mode for categorical columns or dropping the records will be adopted.

Product first sold date is an irrelevant column as the data presented is not understandable.

**Customer Address Dataset**

There were some recording issues in state column where some records were recorded as New South Wales and Victoria instead of NSW and VIC respectively. I recommend having set options or a dropdown list of states in future in order to avoid such inconsistencies.

**Customer Demographic Dataset**

There were some recording issues in the gender column: instances of Female recorded as F and Femal, and Male recorded as U. I also recommend having a dropdown list for this column in future too avoid errors.

There were some missing values. Which will be dealt with by filling with mean or median for numerical columns and mode for categorical columns or dropping the records will be adopted.

DOB was recorded as a string instead of a date datatype. Also to note is, a customer has a recorded DOB of 1843, making him 180 years old which is improbable. This record is replace with a more realistic value of 1943.

The ‘default’ column has irrelevant information and should not be included in the analysis and in future.

Moving forward the team will proceed with data cleaning, standardization and transformation for purposes of analysis. All issues encountered, questions addressed and assumptions made will be documented. After that it would be great to have a meeting with Sprocket Central’s SME (Subject Matter Expert) to confirm whether all is aligned to the company’s understanding and goals.

Kind regards,

Nick Kimani