Dear James,

Thank you for providing us the datasets from Sprocket Central Pty Ltd.

The below table highlights the main info from the three datasets received. Please let us know if the figures are not aligned with your understanding.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tables** | No of records | Distinct id | Date received |
| **Customer Demographic** | 4000 | 4000 | May 25, 2023 |
| **Customer address** | 4000 | 4000 | May 25, 2023 |
| **Transactions** | 2000 | 2000 | May 25, 2023 |
| **New customer List** | 1000 | NA | May 25 |

Many quality issues were found in the Data provided:

**Multiple Missing data were found in many columns of all the tables except ‘Customer Address’ table.**

Mitigation: Please, make sure that the data is being collected correctly and all the table contains data from the same period.

**Many columns didn’t have a title they appeared as NAN.**

Everything we will need to do as we advance in the project will be based on what each piece of data represents. For that it is important to always have the column title. We cannot assess the quality of the data if we don’t know what it represents.

Mitigation: starting from the data collection process, please make sure to be recollecting the data we need, and protect the database from accidental deletion, also make sure to have in place disaster recovery alternatives.

**Missing Customer ID in the new customers list**

One very important thing to note was that the new customer list did not have customer ID assigned to the new customers, while this information won’t be very useful for modelling it helps a lot to have control over the data as it is the only unique feature the data process, formed by all the other features.

**Inconsistence In date formatting**

It is of crucial importance to have the date in a standardized format. For example, having day/month, day/month/year, day/month/year/time or even date cast as string in the same column is something make the effort to avoid. This is important because if in the future there is a need to perform calculations on them it would not be possible if they’re not in the same format.

**Inconsistent Value for the same attribute**

In columns where gender is represented, we found that the same gender was represented differently for example:

'F', 'Male', 'Female', 'U', 'Femal', 'M'.

Moving forward, the team will continue with the data cleaning, standardization and transformation process for the purpose of model analysis. Questions will be raised along the way and assumptions documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central’s understanding.

Regards,

Junior Consultant Yvens