Dear [Client point-of-contact],

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below table are the summary statistics from the received datasets. Please let us know if the figures are not aligned with your understanding.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset name | Number of records | Unique customer IDs | Date data received |
| Customer Demographic | 4000 | 4000 | 03/09/2021 |
| Customer Addresses | 3999 | 3999 | 03/09/2021 |
| Transaction data in the past three months | 20000 | 3495 | 03/09/2021 |

The data quality issues that were encountered as well as the methods used to mitigate the identified data inconsistencies are as listed below. On top of that, recommendations have been provided to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

For the ‘Customer Demographic’ dataset:

● **Inconsistent values for the same attribute**

For example: For the column ‘gender’, Female is being represented as “F” and “Femal” and Male is being represented as “M”

Mitigation: Replaced abbreviations into either Male or Female to ensure consistency of data.

Recommendation: Have a checkbox where users are able to choose their respective gender instead of having them manually entering the data via free text field.

● **Various columns, such as the last\_name, DOB, job\_title, job\_industry\_category and tenure have empty values in certain records**

Mitigation: If there is only a small number of rows are empty, we are able to filter out the record entirely from the dataset for prediction. However, what I did was to replace all of the NaN values with the last observed non-null value forward until another non-null value is encountered. But for job\_title column, I replaced all of the NaN elements with the first observed non-null value backward until another non-null value is met.

Recommendation: Must ensure that all of the data is filled in before submitting their information. In other words, users are not allowed to submit if all of the fields are not filled in and data must be in the right formatting.

For the ‘Customer Addresses’ dataset:

● **Inconsistent values for the same attribute**

For example: Victoria being represented as “VIC” and “Victoria” and New South Wales as “NSW”.

Mitigation: Replaced abbreviations into either Victoria, New South Wales, Queensland to ensure consistency of data.

Recommendation: Have a checkbox or a drop out list where users are able to choose their respective state instead of having them manually entering the data via free text field.

For the ‘Transactions’ dataset:

● **Various columns, such as the online\_order, brand, product\_line, product\_class, product\_size, standard\_cost and product\_first\_sold\_date have empty values in certain records**

Mitigation: If there is only a small number of rows are empty, we are able to filter out the record entirely from the dataset for prediction. However, what I did was to replace all of the NaN values with the last observed non-null value forward until another non-null value is encountered. But for job\_title column, I replaced all of the NaN elements with the first observed non-null value backward until another non-null value is met.

Recommendation: Must ensure that all of the data is filled in before submitting their information. In other words, users are not allowed to submit if all of the fields are not filled in and data must be in the right formatting.

Moving forward, the team will continue with the data cleaning, standardisation 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.

Kind regards,

Ashley Ooi