Dear Sir/Madam,

We have received the raw datasets from Sprocket Central Pty Ltd. As instructed, we have analysed the quality of the data and we have found several quality issues that need to be addressed. Below are the identified data quality issues and corresponding recommendations.

1. Missing values are present:

|  |  |  |
| --- | --- | --- |
| Sheet Name | Column Name | N° of Blank Cells |
| Transactions | Online\_order | 360 |
| Brand | 197 |
| Product\_line | 197 |
| Product\_class | 197 |
| Product\_size | 197 |
| Standard\_cost | 197 |
| Product\_first\_sold\_date | 197 |
| NewCustomerList | Last\_name | 29 |
| DOB | 17 |
| Job\_title | 106 |
| Job\_industry\_category | 165 |
| CustomerDemographic | Last\_name | 125 |
| DOB | 87 |
| Job\_title | 506 |
| Job\_industry\_category | 656 |
| Default | 302 |
| Tenure | 87 |

*Recommendations: Since the percentage of missing values is low compared to the whole datasets, we can delete them without having it affecting the datasets. However, it would be best to understand why some values are missing and if it’s possible to fetch them back.*

1. Irrelevant columns

* In the NewCustomerList sheet, there are 5 unnamed columns.

*Recommendation: If you know what they represent please specify it, if not we will delete them.*

* In the CustomerDemographic sheet, the ‘default’ column is not readable.

*Recommendation: Since it’s not readable it’s best to delete it.*

1. Inconsistent data entries across datasets

* In the customer\_ids column, there are more data entries in the Transactions sheet than the CustomerDemographic and CustomerAdress sheets.

*Recommendation: We will only conduct analysis on the synced data of all the datasets.*

1. Format issue

* In the Transaction sheet, values in the product\_first\_sold\_date column are not into date format.

*Recommendation: Convert the values in that column into a datetime format.*

* In the NewCustomerList and CustomerDemographic sheets, in the gender column, it is unclear what the ‘U’ value mean.

*Recommendation: We suppose it means ‘Undefined’ so we will replace ‘U’ into that name.*

1. Duplicated values for the same column:

* In the CustomerDemographic sheet, there are 3 columns representing ‘Female’ and 2 columns representing ‘Male’ in the ‘gender’ column.

*Recommendations: Replace ‘F’ and ‘Femal’ into ‘Female’, and ‘M’ into ‘Male’.*

* In the CustomerAdress sheet, we encounter the same problem in the ‘state’ column.

*Recommendations: Replace ‘New South Wales’ into ‘NSW’ and ‘Victoria’ into ‘VIC’.*

As we will soon be proceeding to these changes, we wanted to make you aware of it and its potential impacts.

If you have any questions or concerns, please let me know.

Kind regards,

Eden

Data Analytics Team