To,

Sprocket Central Pty Ltd

Subject - Data Quality issues and strategies to mitigate these issues.

Respected Sir/Madam,

Below are the identified data quality issues and strategies to mitigate these issues:

Exploring the Transactions Dataset

* Completeness

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Empty** | **Duplicate** | **Contain null** |
| transaction\_id | 0 | 0 | 0 |
| product\_id | 0 | 0 | 0 |
| customer\_id | 0 | 0 | 0 |
| transaction\_date | 0 | 0 | 0 |
| online\_order | 17 | 0 | 360 |
| order\_status | 0 | 0 | 0 |
| brand | 6 | 0 | 197 |
| product\_line | 6 | 0 | 197 |
| product\_class | 6 | 0 | 197 |
| product\_size | 6 | 0 | 197 |
| list\_price | 0 | 0 | 0 |
| standard\_cost | 6 | 0 | 197 |
| product\_first\_sold\_date | 6 | 0 | 197 |

* Try to tally the old data or relevant records to fill some null and blank values or remove them.
* Validity

|  |  |
| --- | --- |
| **Column name** | **Type** |
| transaction\_id | Int64 |
| product\_id | Int64 |
| customer\_id | Int64 |
| transaction\_date | date |
| online\_order | logical |
| order\_status | Text |
| brand | Text |
| product\_line | Text |
| product\_class | Text |
| product\_size | Text |
| list\_price | Number |
| standard\_cost | Number |
| product\_first\_sold\_date | Int64 |

* The ‘product\_first\_sold\_date’ column is stored as integers. This column should be changed to date.

Exploring the Customer Addresses

* Consistency

There is no uniformity in the way of data entry in the state column, so only one input method is fully recorded, such as New South Wales or NSW for short.

* Relevancy

The property valuation column is not appropriate in the customer address data table, so it should be removed

Exploring the Customer Demographic Dataset

* Accuracy

The DOB data of a customer with customer\_id of 34 is wrong because having a birth year 1843 means this person has an age of 175 years. So this is a bug and needs to be rechecked.

* Consistency

In the gender column, changed value "F" to "Female" and "M" to "Male" for consistency.

* The Data Handling team should check the latest values used for fields so as to keep

up to date database

* Completeness

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Empty** | **Duplicate** | **Contain null** |
| customer\_id | 0 | 0 | 0 |
| first\_name | 0 | 0 | 0 |
| last\_name | 30 | 0 | 125 |
| gender | 0 | 0 | 0 |
| past\_3\_years\_bike\_related\_purchases | 0 | 0 | 0 |
| DOB | 23 | 0 | 87 |
| job\_title | 119 | 0 | 506 |
| job\_industry\_category | 0 | 00 | 0 |
| wealth\_segment | 0 | 0 | 0 |
| deceased\_indicator | 0 | 0 | 0 |
| default | 62 | 0 | 240 |
| owns\_car | 0 | 0 | 0 |
| tenure | 23 | 0 | 87 |

* Try to tally the old data or relevant records to fill some null and blanks values or remove them.
* Relevancy

It is recommended to delete or replace the "U" value in the gender column as it is an undetermined value and does not support the analysis. In addition, the default column should also be removed because it does not contain clear data and does not contribute content to the analysis process.

Above are the errors and correction suggestions for each data table, if you have any questions or problems please contact us.

Best regard,

Le Tran Ai Sa