Dear client,

I am writing to you in response to the Sprocket Central Pty Ltd Datasets that you shared with our team.

* Transaction Dataset
* Customer Demographic dataset
* Customer Addresses dataset

We have analysed the quality of the raw data found several quality issues in the datasets.

|  |  |  |
| --- | --- | --- |
| Datasets | Total Number of Records | Unique ID Records |
| Transaction | 20000 | 3427 |
| Customer Demographic | 4000 | 4000 |
| Customer Addresses | 4003 | 4003 |

We believe that these issues needed to be addressed before we continue this further processing of the data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy | Completeness | Relevancy | Validity | Consistency |
| Transaction | **Profit:** Missing | **Online order:**  Blanks  **Brands:** Blanks | **Cancelled status:** Filtered Out | **List price and date product sold:** Format |  |
| Customer Demographic | **Age:** Missing  **DOB:**  Inaccurate | **Job title:**  Blanks | **Default column:** Irrelevant |  |  |
| Customer Addresses |  |  |  |  | **States:**  Inconsistence |

Below we’ve suggested ways on how you can mitigate those data quality issues in the future.

1. **Accuracy Issues:**

* Customer Demographic dataset

Some cells in the DOB column had inaccurate values and some of the cells in the age columns had missing values.

**Recommendation:** Filter out outliers in DOB column and double check the age and the profit columns to ensure there are no missing values.

1. **Completeness Issues**

* Customer Demographic dataset:

Some cells in the job\_title was Blanks.

* Transaction datasets:

Both the online\_order and the brands were Blanks.

**Recommendation:** Double checking the columns to ensure that the columns don’t have blank spaces.

1. **Consistency Issues:**

* Customer Demographic datasets:

Both the Male and Female genders were inconsistent.

* Customer Addresses datasets:

In states column both the name New South Wales and Victoria are inconsistent.

**Recommendation**: Using selective options minimizes manual entry which in return reduces human error.

1. **Relevancy issue:**

* Customer Demographic datasets:

Irrelevancy of data values in the default column.

* Transaction data:

Irrelevancy of data values in the order\_status column.

**Recommendation:** Recheck and delete all the irrelevant data values.

1. **Validity:**

* Transaction data:

Wrong data format in both the list\_price and product\_first\_sold\_date.

**Recommendation:** Ensure that the column data values are in the correct format.

Please investigate the above-mentioned quality issues along with the recommended changes. If all the suggestions are matched, we can proceed with further analysis of the data.

Regards,

Junior Consultant