TASK1:

Dear Client,

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

We found the following datasets:

* Transaction dataset
* Customer Demographic dataset
* Customer Addresses dataset

We have analyzed the quality of the raw data and found several quality issues in the dataset

|  |  |  |
| --- | --- | --- |
| DATASET | TOTAL NUMBER OF RECORDS | UNIQUE ID RECORDS |
| TRANSACTION | 20000 | 3490 |
| CUSTOMER DEMOGRAPHIC | 4000 | 4000 |
| CUSTOMER ADDRESSES | 4003 | 4003 |

We believe that these issues needed to be addressed before we continue with the processing of the data, we focused mainly on following aspects:

* Accuracy
* Completeness
* Relevancy
* Validity
* Consistency
* Duplicate and spelling check

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| DATASET | ACCURACY | COMPLETENESS | RELAVANCY | VALIDITY | CONSISTENCY |
| CUSTOMER DEMOGRAPHY | Age : Missing  DOB :  Inaccurate | Job Title : Blanks values | Default :Irrelevant column |  | Gender :Inconsistent |
| TRANSACTION | Profit :  Missing values | Online Orders :  Blank values  Brands :  Blank Values | Order Status:  Cancelled status (filtered out) | List Price and Product cost : Format |  |
| CUSTOMER ADDRESSES |  |  |  |  | States:  Inconsistent |

Below are the methods we have suggested on mitigating the data quality issues in the future.

**i)Accuracy Issues:**

* **CUSTOMER DEMOGRAPHIC DATASET:**

Some cells in the DOB column have inaccurate values and some of the cells have missing values.

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

**ii)Completeness Issues :**

* **CUSTOMER DEMOGRAPHIC DATASET:**

Some cells in the Job Title column are blank

* **TRANSACTION DATASET**:

Online orders and Brand columns have blanks

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

**iii)Consistency Issues** :

* **CUSTOMER DEMOGRAPHIC DATASET:**

Both the Female and Male genders were inconsistent

* **CUSTOMER ADDRESSES DATASET:**

Both the New South Wales and Victoria states in the states column were inconsistent.

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

**iv) Relevancy Issues :**

* **CUSTOMER DEMOGRAPHIC DATASET:**

Irrelevant data values in default column

* **TRANSACTION DATASET**:

Irrelevant data values in the order status column

**Recommendation:** Recheck and delete all the irrelevant columns

**v) Validity Issues:**

* **TRANSACTION DATASET**:

Wrong data format in both the List prices , product cost and product first sold date columns

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

**vi) Spelling check & Duplications:** No spelling mistake were found and there are duplications in the transaction dataset which is not a issue please check if there are duplications which effects the data

Please take into consideration above mentioned data quality issues along with the recommended mitigation methods . If all the suggestions are matched we can proceed with further analysis of the data.

Thanks & regards,

BASHUTKAR SHILPA BAI

(KPMG DA TEAM)