To: Sprocket Central Pty Ltd

Subject: Data Quality Issues and steps to mitigate these issues.

Dear Sprocket Central Pty Ltd.

Thank you for providing us with the dataset, we have reviewed the dataset and summarised the following data quality issues within the dataset. We have further provided with the steps to mitigate to move forward with the data cleaning process.

|  |  |  |  |
| --- | --- | --- | --- |
| **Sheet Name** | **Unique Column** | **No.of. rows**  **(Before cleaning)** | **No.of. distinct Customer ID’s** |
| Transactions | Transaction\_id | 20000 | 3494 |
| CustomerList | No unique column | 1000 | - |
| Customer Demographic | Customer\_id | 4000 | 4000 |
| CustomerAddress | Customer\_id | 4000 | 4000 |

**Date Data Received - 2017-12-30.**

**1. Accuracy:**

* NewCustomerList sheet*:* In the columns past\_3\_years\_bike\_related\_purchases, postcode, property\_valuation the data type is changed from text to Number.
* CustomerDemographic:
  + Filter out outliers in DOB.
  + Grammatical error in the job\_industry\_category. ‘Argiculture’ is changed to ‘Agriculture’.

**Steps to mitigate:**

The Data Handling team should Check the data type of columns to mitigate this type of issue.

**2. Completeness:**

* Transactions sheet: online\_order, brand, product\_line, product\_class, product\_size, standard\_cost, product\_first\_sold\_date
* NewCustomerList sheet: DOB, job\_title, job\_industry\_category
* CustomerDemographic sheet: last\_name, DOB, job\_title, job\_industry\_category, tenure.

**Steps to mitigate:**

The data can be improved by adding together previous data or appropriate data to fill in certain null values.

**3. Consistency:**

* CustomerDemographics sheet: In the gender column the values ‘F’ and ‘Femal’ are changed to ‘Female’ and ‘M’ to ‘Male’.
* CustomerAddress sheet: In the state column values are changed from "New South Wales" to "NSW" and “Victoria" to "VIC”.

Steps to mitigate:

The team can check the latest values used for fields to keep the database updated.

**4. Currency:**

* Transactions sheet: In standard\_cost and list\_price columns there are currency values which are not rounded to two decimals and unmentioned currency.

Steps to mitigate:

The team can check for currency values to avoid any errors and stay UpToDate.

**5. Relevancy:**

# Transactions sheet:

# The records with the ‘cancelled’ order\_status is irrelevant. Retain only those with the ‘Approved’ order\_status.

# The column product\_first\_sold\_date is converted into date format for easy interpretation.

* NewCustomerList sheet:
  + In the gender column the Value ‘U’ is deleted since it is unknown whether it is male or female.
  + The dataset contains an entry with a male whole Dob: 1843 which is impossible.
* CustomerDemographics sheet*:*
  + In the gender column the Value ‘U’ is deleted since it is unknown whether it is male or female.
  + Records with ‘Y’ in deceased\_indicator are irrelavent. Retain only those customers who have a ‘N’ in deceased\_indicator.
  + The column “default” is deleted since it cannot be decoded and irrelevant to the dataset.

**Steps to mitigate:**

The team can try to tally the records of "U" in gender to find out the correct values.

6. Uniqueness:

* CustomerDemographic sheet: Contains 4 duplicates, which are deleted.

We recognize that ensuring the quality of data is critical to delivering accurate and dependable outcomes. As a result, we would like to collaborate closely with you to solve these data quality concerns and put the recommended measures in place. By doing so, we will be able to ensure that our analysis is based on high-quality data and that we will be able to give you relevant insights and suggestions.

The team will continue to clean, standardize, and convert data in preparation for model analysis. Along the process, questions will be posed and assumptions will be documented. After we finish this, we'd like to meet with your data expert to check that all assumptions align with Sprocket Central's knowledge.

Please let us know what you think of this, and if you have any questions or issues, please do not hesitate to contact us.

Best regards,

Kavya Pabba, Junior Data Analyst at KPMG.