**Data Quality Assessment**

Dear Manager,

We have reviewed the three datasets previously provided by Sprocket Central Pty Ltd.

The summary table below highlight data quality issues we have discovered in the data quality assessment process.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** | **Uniqueness** |
| **Customer Demographics** | Wrong Value:  DOB, Gender. | Missing Value: Last Name, DOB, Job Title, job Industry, Tenure. | Format: Gender. | - | Irrelevant:  Default. | Data type: DOB. | - |
| **Customer Address** | - | - | Format: State. | - | - | - | - |
| **Last three months transactions (Transactions)** | New Column: Profit (Price-Cost). | Missing Value: Online Order, Brand, Product Line, Product Class, Product Size, Standard Cost, Product First Sold Date. | - | - | - | Data type: Transaction Date, Product First Sold Date, Price and Standard Cost. | - |

In the following, we have carried out a deeper analysis of the data quality issue and strategies that can be taken to mitigate the same error in the future. Before going into a deeper analysis, here is a brief understanding of the terms in the data quality framework that we use. If you need a more detailed framework, we will be happy to provide it to you.

* Accuracy : Correct Values
* Completeness : Data Fields with Values
* Consistency : Values Free from Contradiction
* Currency : Values up to Date
* Relevancy : Data Items with Value Meta-data
* Validity : Data Containing Allowable Values
* Uniqueness : Record that are Duplicated
* ***Accuracy Issues***

In Customer Demographics dataset, DOB column has one wrong value that indicated customer was born in 1843 (180 years old per 2022). Gender column also has more than one wrong value, e.g., Femal, F, M, U.

In Transactions dataset, there is one column that can be extracted (profit) from an existing column (price and cost) to facilitate further analysis.

*Mitigation:* Data filtering needs to be done before analyzing the data, to check the accuracy (correct values) of the data.

* ***Completeness Issues***

Customer Demographic and Transactions datasets have many columns that contain null values.

*Mitigation:* Provide drop-down option for categorical data because the null values mostly appeared in categorical data.

* ***Consistency Issues***

For Customer Demographic dataset, inconsistency appears in column Gender (Female, F, M, U). While in Customer Address dataset, inconsistency appears in column State (use of abbreviation).

*Mitigation:* These two columns are categorical data. To avoid inconsistency in categorical data, drop-down option could eliminate human error in manual data entries.

* ***Relevancy Issue***

Default column in Customer Demographics dataset contains strange value which does not provide information to the dataset. This column seems corrupted.

*Mitigation:* If this column is corrupt, then the company should have a data backup so that if there is an important column that is corrupted, it can be recovered.

* ***Validity Issues***

Data containing date in the given dataset is still of type object, not datetime, so there is the possibility of input errors due to human error such as input 1843 in the DOB column.

Product First Sold Date column has a mismatch between its name and value.

List Price and Standard Cost should be in the same format (currency mode or general mode).

*Mitigation:*

Data containing date should be changed to datetime type.

Columns containing currency values must be uniform, the same as using currency mode or general mode.

The above summarize our highlight on data quality assessment. If you have further comments or questions regarding the data quality assessment, we will be happy to discuss them to ensure that we are on the same page and that our team can continue the analysis process to the next stage.

Kind Regards

Tony Smith

KPMG’s Lighthouse & Innovation Team