Date: Sept 6th, 2020

To: Evaluator, KPMG

From: Rupali Shekhawat

Subject: Data quality issues and strategies regarding ‘Data Quality Framework Table’

Dear Evaluator,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the three datasets received. Please let us know if the figures are not aligned with your understanding.

|  |  |  |  |
| --- | --- | --- | --- |
| Table name | No. of records | Distinct Customer IDs | Date Data Received |
| Customer Demographic | *4002* | *4002* | *04/09/2020* |
| Customer Address | *4001* | *4001* | *04/09/2020* |
| Transaction Data | *20002* | *20002* | *04/09/2020* |

Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows. Furthermore, recommendations have been provided to avoid the re- occurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

**Accuracy**

* DOB was inaccurate in “Customer Demographics”

*Mitigation: Filter out outlier in DOB*

Only relevant DOB should be present in the data for better analysis

**Completeness**

* Blanks in various sections of “CustomerAddress”, “CustomerDemographic”, “NewCustomerList” and “Transactions”

*Mitigation: Filter all values and remove blanks*

With incomplete data, some issues might arise during analysis hence we remove blanks

**Consistency**

* Customer\_ids were inconsistent in “CustomerAddress”, “CustomerDemographic” and “Transactions”

*Mitigation: Convert selected records in characters to numeric. Remove non-numeric characters from the string.*

*Recommendation: Ensure that fact tables in the given database have constraints on data types.*

Having different data types for a given field make it difficult to interpret results at the later stage.

Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

* Inconsistency was found in ‘state’ section of “CustomerAddress” and ‘gender’ section of “CustomerDemographic”

*Mitigation: Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses.*

*Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field.*

To construct meaningful variables for the model, the data has been cleaned to avoid

multiple representations of the same value. Additionally, gender records where ‘U’ have been

replaced based on the distribution from the training dataset.

**Currency**

* People indicated as Y in the ‘deceased\_indicator’ section of the “CustomerDemographic” are not current customers

*Mitigation: Filter out the customers checked as ‘Y’ in the deceased\_indicator*

Removal of Deceased Customers is necessary as they are not current customers and their removal will lead to more accuracy in data during analysis.

**Relevancy**

* No significance of ‘default’ section of “CustomerDemographic”

*Mitigation: Deletion of meaningless data from ‘default’ column*

Irrelevant information can result in data inconsistency hence we remove redundant values

**Validity**

* Format of ‘list\_price’ and ‘product\_first\_sold\_date’ sections in “Transactions” is not valid

*Mitigation: Format the ‘list\_price’ section to currency format and ‘product\_first\_sold\_date’ section to date format*

Formatting the following will increase consistency as it will make interpreting the data easy

**Uniqueness**

* Duplicate values were found in ‘unnamed’ and ‘rank’ sections in “NewCustomerList”

*Mitigation: Removal of duplicate values is important to improve consistency*

Duplicate records have been removed from the training dataset.

Strategy to improve data quality for future analysis would be to solve the Quality Issues stated above.

These are the results that have been predicted from the ‘Data Quality Framework Table’.

Kindly let us know if you have any questions regarding the same.

Regards,

Rupali Shekhawat