Dear Mr. Steve Smith,

The three datasets- Customer Demographic, Customer Address and Transactions provided by Sprocket Central Pty Ltd have been analysed. Given below are the data quality issues with respect to the data quality dimensions that were discovered in your datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| **Quality**  **Dimensions** | **Customer Demographic** | **Customer**  **Address** | **Transaction** |
| Accuracy | DOB  Added age column |  | Added profit column |
| Completeness | job\_title | customer\_id | online\_order, brand |
| Consistency | gender | state |  |
| Currency | deceased\_indicator |  |  |
| Relevancy | default column |  | order\_status (Cancelled) |
| Validity |  |  | list\_price, product\_first\_sold\_date |

**Accuracy**

In **CustomerDemographic** DOB is inaccurate, an age column is missing, profit column missing in **Transactions.**

Strategy to mitigate: Filter abnormal values from DOB column. Create an age column - it will help out clearing such values. Create a profit column – it will give a better idea about profits made during sales.

**Completeness**

1. Blanks in job\_title in **CustomerDemographic**; in online\_order and brand in **Transactions**.

*Strategy to mitigate*: Filter out blanks from job\_title, online\_order and brand columns. Simplify job titles. Provide drop down list for online\_order and brand

Column.

1. By cross-referencing the datasets complete data is present for customer\_ids from 1 to 3,500. For analysis only this must be used else analysis results will be affected.

Strategy to mitigate: Filter customers from customer\_id 1-3,500. Keep customer information updated across all datasets.

**Consistency**

**CustomerDemographic** contains inconsistent data in gender column, and **CustomerAddress** contains inconsistent data in state column.

*Strategy to mitigate*: Filter all M in category of Male, filter all F and femal in category of Female. Filter all New South Wales and Victoria under their respective state abbreviations. Create a drop-down list for gender and state to avoid inconsistent data.

**Currency**

In **CustomerDemographic** customers with data deceased\_indicator = ‘Y’ not required for analysis.

*Strategy to mitigate*: Filter out customers with deceases\_indicator = ‘Y’. Deceased customers are not active customers and hence must be filtered out from the dataset. Always update such information and filter the data accordingly.

**Relevancy**

The default column in **CustomerDemographic** and order\_status in **Transaction** contains data that is not relevant.

*Strategy to mitigate*: Delete default column. Filter out all order\_status = ‘Cancelled’. Always filter out data that may be unintelligible.

**Validity**

In **Transactions** the data format for list\_price and product\_first\_sold\_date is not set.

*Strategy to mitigate*: Change format of list\_price to currency and product\_sold\_first\_date to Short date format. Use a single format for data related to both date and price.

This concludes all the issues that were discovered during analysis of the datasets. The mitigation strategies to get rid of such issues have also been provided. This will improve the data quality of datasets and also lead to better data analysis results.

Please let me know if you need anything further and I would be happy to help.

Kind regards,

Ankit Taklikar