Dear client,

Greetings of the day, thank you for sharing these datasets, during the preprocessing of the datasets we have got the following insights, please let us know if you have any queries surrounding the issues presented.

**SUMMAY TABLE**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Datasets** | **No. of records** | **Unique id’s** | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **validity** |
| Transaction | 20002 | 3495 | Profit:  missing | Costumer id:  Incomplete  Online order:  Blanks  Brand:  blanks |  |  | Cancelled status order:  Filter out | List price:  Format  Product sold date:  format |
| Customer Demographic | 4002 | 4002 | DOB: inaccurate  Age:  missing | Job titles:  Blanks  Customer id:  incomplete | Gender:  inconsistency | Deceased customer:  Filter out | Default column:  delete |  |
| Customer Address | 4001 | 4001 |  | Costumer id:  incomplete | States:  Inconsistency |  |  |  |

Below are the more in-depth description of data quality issues and methods of mitigation used.

1. **Accuracy issues:**

* DOB was inaccurate for “Customer Demographic” and missing an age\_column;

Missing a profit column for “Transactions”.

*Mitigation:*

*Filter out outlier in DOB.*

*Create a age\_column and create a profit\_column in transaction*.

1. **Completeness:**

* In “Customer Demographic”,” Customer Address” and “Transactions”, additional customer\_ids were inconsistent.

*Mitigation:*

*Filter all customer\_ids from 1 to 3500*

* Blanks in job\_title for “Customer Demographic”, in online\_order and brand\_column for “Transaction”.

*Mitigation:*

*Filter out “blanks” for job\_title, online\_order, and brand\_column*.

Blanks are treated as incomplete data and can skew further analysis results.

The addition of dropdown option will allow to have more complete data and will result in more accurate analysis.

1. **Consistency**

* Inconsistency in gender and states for “Customer Demographic” and “Customer Address” respectively.

*Mitigation:*

*Filter all ‘M’ under the category ‘Male’, and filter all ‘Femal’ and ‘F’ under ‘Female’ for gender.*

*Filter all ‘New South Wales’ to ‘NSW’ and ‘Victoria’ to ‘VIC’ for states*.

It is recommended to create dropdown options as the y minimize human entry and human error. Allows for increase in consistency of terminology.

1. **Currency**

* People that are ‘Y’ in deceased\_indicator are not current customers for “Customer Demographic”.

*Mitigation:*

*Filter out customer that are indicated as ‘Y’ in deceased\_indicator*.

1. **Relevancy**

* Lack of relevancy and comprehensibility in default\_column for “Customer Demographic” and order\_status for “Transactions”

*Mitigation:*

*Delete metadata in default\_column.*

*Filter out ‘Cancelled’ order\_status*.

1. **Validity**

* Format of list\_price, product\_sale\_date for “Transactions”

*Mitigation:*

*Format product\_sale\_date to the short date format.*

*Format list\_price to currency.*

That summarizes all the data quality issues discovered through the first stage of data quality analysis. The mitigation strategies suggested are simple and effective way of improving data quality for future analysis. Please let us know if you have questions regarding mitigation or any data quality issues identified.

Kind regards,

Arya Sharma