Hi team,

Thank you for providing the datasets. Please find below summary of data quality assessment:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** | **Uniqueness** |
| Transactions | Missing Profit | **online\_order** – 1.8% missing values  **brand**, **product\_line**, **product\_class**, **product\_size**,  **standard\_cost**, **product\_first\_sold\_date**  - 0.98% missing values |  |  |  | **product\_first\_sold\_date** – Format invalid | Yes |
| CustomerDemographic | Missing Age | **last\_name** - 3.12%  **DOB** - 2.17%  **job\_title** - 12.65%  **job\_industry\_category** - 16.4%  **tenure** - 2.17% | gender | deceased\_indicator – 2024 deceased | default – irrelevant column |  | Yes |
| NewCustomerList |  | **last\_name** - 2.9%  **DOB** - 1.7%  **job\_title** - 10.6%  **job\_industry\_category** - 16.5% |  |  | Unnamed: 16  Unnamed: 17  Unnamed: 18  Unnamed: 19  Unnamed: 20  Irrelevant columns |  | yes |
| CustomerAddress |  |  | State |  |  | **State** |  |

Accuracy issues:

Issue #1 - Transactions contains missing profit column

Solution – Adding Profit column

Profit = list\_price - standard\_cost

Issue #2 – CustomerDemographic contains missing Age column

Solution – Adding Age column

Completenessissues:

Issue #1 – Transactions, CustomerDemographic, NewCustomerList

contain missing values

Solution – Dropping the missing values

Consistency issues:

Issue #1 – The dataset contains these categories - Female, Male, U, F, Femal, M

Solution – converting into consistent categories F, M, U where F is Female, M is Male and U is Unspecified

Issue #2 – CustomerAddress contains categories New South Wales, Victoria, QLD, VIC, NSW

Solution – Converting New South Wales – NSW, Victoria - VIC

Currency issues:

Issue #1 – CustomerDemographic contains 2024 deceased which depicts that the data is not current

Solution – Considering the data with Deceased = ‘N’

Relevancy issues:

Issue #1 – CustomerDemographic contain irrelevant column default and NewCustomerList contain Unnamed: 16, Unnamed: 17, Unnamed: 18, Unnamed: 19, Unnamed: 20 irrelevant columns

Solution – Dropping the irrelevant columns

Validity issues:

Issue #1 – Transactions contain product\_first\_sold\_date column which has invalid date format

Solution – Correcting the date format

We would proceed by applying the solutions and cleaning and transforming the data for better analysis.

Thanks and Regards,

Mandar Borhade

KPMG intern team