Dear Mr. Archie,

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Columns || Issues** | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| **IN\_Mobiles** | Currrent Age data format changed | All columns were complete | Converted Gender f(M) -> “1”  F(F) -> “0” Converted Purchase f(Yes) -> “1” & f(No) -> “0” | Segment Columns Added  Mobile category Added | Dropped the ID table | Converted date format  Converted annual income to currency |
| **CN\_Mobiles** | Currrent Age data format changed | All columns were complete | Converted Gender f(M) -> “1”  F(F) -> “0” | Segment columns Added Mobile category Added | Dropped the ID table | Converted date format Converted currency to INR |

Below are some detailed description of these key issues present in given three data sets and how we can resolve these key issues from occurrence

* **Accuracy Issues**

**Filter out outlier in [DOB] and create an [Age] column for “CustomerDemographic”. Create [profit] column for “Transactions”**

Creating age column will help out in better understanding for customer and profit column will help in better understanding and decision making in monetary policies.

*Suggestion: Use a dropdown list in DOB year entrance.*

* **Completeness**

**Excessive entries in [customer\_ids] is dropped because of incompleteness. [job\_title],[online\_order] and [brand] dropped nulls.**

Data of customer\_ids from 1 to 3500 is used for our model because upto 3500 customer\_ids are present in “Transactions” and “CustomerAddress”

Also some null values were present in columns [online\_order] & [brand] from “Transactions” and [job\_title] from “CustomerDemographic” which are removed to avoid incompleteness

*Suggestion: fetch data of same date from all data sets and put dropdown list in simplified forms*

* **Inconsistency**

**There is inconsistency in [gender] from “CustomerDemographic” & [state] from**

**“CustomerAddress”**

This inconsistency may corrupt our data model so it is been removed with

New South Wales -> NSW

Victoria -> Vic

Female,Femal -> F

Male -> M

*Suggestion: reduce manual entry by using dropdown list. Try putting sensitive information tag on data entry*

* **Currency**

**People with Y in [deceased]**

People which are deceased are not customers anymore so to avoid further ambiguities of our data model All [deceased] with value “Y” are dropped from our data

*Suggestion: Not holding record of deceased in datasets that are going to be used in our models and data visualization dashboards will increase currency of data*

* **Relevancy**

**Presence of metadata in given datasets, [order\_status] from “Transactions” and [default] from “CustomerDemographic”**

All the irrelevant data is dropped from datasets as cancelled orders and incompressible data from default column

*Suggestion: Check for not readability or meaningfulness of data present in datasets*

* **Validity**

**Format the [list\_price], [product\_sale\_date] for Transactions**

Changing format of [list\_price] to currency and [product\_sale\_date] to short date

*Suggestion: set up columns of database fields so when data comes in the decimals, commas, slashes and other signs are in their place in proper format and don’t lose the information*

Provided solutions will help company in data fetching retrieving and CRUD operations and also help the data analysis team to do further analysis. Highlighted Issues are needed to be resolved by every team to increase accuracy and precision of models they are going to generate. Resolving data conflicts will give better and more meaningful insights

Please let us know if there are any questions regarding provided solutions and suggestions

Kind regards,

Talal Manshoor