Hi! Hope so you are fine.

I am working as a data analyst to check the Data quality of three datasets provided by Sprocket Central Pty Ltd

* Customer Demographic
* Customer Addresses
* Transaction

Although datasets provided are in good form they have data about customers and transaction details and metadata about these details but there are several data quality issues about data.

There is Data Accuracy issues in Customer Demographic and Customer Addresses. In Customer Demographic column **gender** in some places Male is represented as **M** and in some places as **Male**, in some places Female is represented as **F** and in some places as **Female** and in some places as **Femal.** In Customer Addresses column **state**, some places Victoria is represented as **Victoria** but some places it is represented as **VIC**, in some places New South Wales is represented as **New South Wales** but some places as **NSW**, in some places.

There is Data Completeness issues in data, we have DOB column in Customer Demographic table but don’t have **age** column and we list price column in Transaction table but don’t have **profit** column.

There are several **Data validity** issues because different columns have invalid datatype like **product first sold date, list price and standard cost** column of Transaction table has invalid datatype.

There is Data consistency issues because several columns like **job title, job industry category** and many more in different tables that have missing values.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Customer Demographic** | **Customer Addresses** | **Transactions** |
| Data Accuracy | In **gender** column we make three categories male ,female and U while replace M as **male** and F, femal as **female** | In **state** column we kept 3 categories **NSW, VIC,QLD** while replace New South Wales with **NSW** and Victoria as **VIC** | Not required |
| Data Completeness | We extract **age** of customer at the time of transaction by subtracting **transaction date** column into **DOB** column in the form of **year** | Not required | We extract **profit** by subtracting **list pric**e column into **standard cost** column |
| Data consistency | We replace missing value of columns job title, job industry category as **Not Listed** it is helpful in our analysis**.** In last name column we replace blanks with space because we use first name of customer as customer representation. | Not required | In online order and brand, we remove blanks by using filter. Because online order as two categories of true and false and we can’t replace it with any other word. |
| Data validity | Not required | Not required | We change datatype of product first sold **date** into Date, list price and standard cost into **currency**. |
| Data Currency | Merge first name and last name column into **Customer column** | Not required | Not required |
| Data Relevancy | We delete **Default,** **Deceased indicator** and **tenure** column because they are irrelevant of analysis. | We delete **country** and **postcode** column because they are irrelevant of analysis. | We delete **product id** column because it is irrelevant of analysis. |

These are the all data quality issues and their solution which are required for analysis.

Kind Regard

Wajiha Ahmed(Data Analyt)