|  |
| --- |
| **Module Task- 1**  Your task is to take a look at the following datasets provided by Sprocket Central Pty Ltd and identify all data quality issues. Once you've had a look at these datasets, draft an email to the client identifying all data quality issues. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Customer Demographic** | DOB: Inaccurate  Age: Missing | Job Title: Blanks  Customer Id: Incomplete | Gender: Inconsistency | Deceased Customer: Filter out | Default Column: Delete |  |
| **Customer**  **Address** |  | Customer Id: Incomplete | State: Inconsistency |  |  |  |
| **Transaction** | Profit: Missing | Customer Id: Incomplete  Online Order: Blanks  Brand: Blanks |  |  | Cancelled order status:  Filter out | List price: Format  Product Sold Date: Format |

Dataset 1: Transactions

1. We have seen that some data are missing all together. Like we note that if value is missing in column 'brand' than it is also missing in columns 'product line', 'product class', 'product size', 'standard cost' and 'product first sold date'. Hence we remove these entries.
2. Column ‘product first sold date’ is not in dates but in numbers. So to convert these integers to date we have to add these number to some base date, which we took 01/01/1900.
3. In column ‘online order’ 358 are missing values. We delete those entries.
4. Also we remove those data where transaction ‘order status’ is equal to cancelled. We only consider data where status is approved.
5. This data has maximum customer id of 3500 and one id=5034.

Dataset 2: Customer Demographic

1. We remove all those entries whose ‘DOB’ is missing which is around 2.1% of data.
2. We create the new column ‘age’ from DOB. One entry has DOB wrong as its age is showing 176 years. We removed it.
3. ‘default’ column has been dropped as it is of no use.
4. The ‘last name’ have some entries missing. We just replace them with empty string.
5. ‘Gender’ have male, female, U, F, M and femal. We combine them into three- Female, Male and U. Be consistent with either male/female or F/M but don’t use both.
6. ‘job title’ and ‘job industry category’ of individuals missing are replaced by the mode of that column.
7. There is only two rows where ‘deceased indicator’ is equal to Y. We remove them.

Dataset 3: Customer Address

1. This data has no missing values.
2. In column ‘state’ Victoria and VIC, New South Wales and NSW are same but both used which creates difficulty in analysis. So be consistent with VIC and NSW.

Data Present in one table but not in other

|  |  |  |  |
| --- | --- | --- | --- |
| Customer Id | Customer Demographic | Customer Address | Transaction |
| 3 | Yes | No | Yes |
| 10 | Yes | No | Yes |
| 22 | Yes | No | Yes |
| 23 | Yes | No | Yes |
| 4001 | No | Yes | No |
| 4002 | No | Yes | No |
| 4003 | No | Yes | No |
| 3501-4000 | Yes | Yes | No |
| 5034 | No | No | Yes |

>>Note>> we only consider only those customers ids which are present in customer demographic data.

We first left join the customer demographic data with customer address data and have 3911 unique customer ids where 4 entries are those whose address data is not given. So we remove those 4 entries. Hence at last we have 3907 unique customer ids.

Finally we left join this new customer data with transaction data and have 19308 entries with 3907 unique customers. We found that around 2.58% (499 entries) have not their transaction details. So we also remove them and left with 18801 entries and 3407 unique customer ids.