Hi, John Scott

Thank you for providing us the three Datasets from Sprocket Central Pty Ltd. The below table summarises the three Datasets provided. Please check if the information contained in the table is in sync with your analysis.

|  |  |  |
| --- | --- | --- |
| **Table Name** | **Number of Records** | **Distinct Customer IDs** |
| Transactions | 20000 | 3494 |
| Customer Demographic | 4000 | 4000 |
| Customer Address | 3999 | 3999 |

We faced some issues related to Data Quality while analysing the data. Below are mentioned some of the problems, mitigation methods and recommendations to avoid reoccurence of issues and improve overall Data Quality.

* **Additional customer IDs in Customer Addresses table and Transactions table other than Customer Demographics table.**

*Mitigation – Taking IDs only that are present in the customer Demographics dataset and removing customer\_id greater than 4000 from Customer Addresses table and Transactions table so as to keep customers consistent across the three Datasets.*

*Recommendations – Update all the tables simultaneously so that if any new customer does transaction his or her information is recorded in the customers Demographics table.*

* **Various features such as online\_order, brand, job\_title, job\_industry\_category have missing values.**

*Mitigation – Please ensure* ***completeness of data****. Remove the missing values if they are around 1% of the whole dataset but if there are a large number of missing values, it is better to impute them appropriately based on data types and distribution of dataset as removing them will lead to a lot of loss of information.*

*Recommendation –* **Don’t allow data to be left blank**. It is better to have some information rather than no information.

* **Some of the features such as gender and state have inconsistent information(eg F, Femal, Female for Female or M, Male for Male or NSW, New South Wales for New South Wales.)**

*Mitigation – Make values consistent across a feature because inconsistent values could create new categories that would distort information. We Replaced Female for Female and Male for Male in gender and NSW for New South Wales, VIC for Victoria in states to* ***ensure consistent values in a an attribute****.*

*Recommendations –* Create drop down lists for users to select facts such as Gender and States so as to minimise mistakes.

* **Features such as default in Customer Demographics table and deceased\_indicator have irrelevant information such as deceased customers and incomprehensible values in default feature.**

*Mitigation – Remove default column and filter out deceased customers as we cannot target customers who are no more in this world. This irrelevant data is noise in the dataset and hampers analysis.*

*Recommendations – Data should be upgraded continously to keep it relevant to the customers that can be targetted.*

* **Irrelevant datatypes in product\_first\_sold\_date. Date feature should have date datatype but it had integer datatype.**

*Mitigation – Changed datatype from int to date.*

*Recommendation –* Please ensure that datatype of a feature is relevant to the information the feature wants to provide otherwise it would make interpretation of information provided by the feature difficult.

This concludes all the Data Quality issues that were encountered through the first phase of analysis of data. Mitigation and Recommendations are easy to implement strategies that can be used to enhance the quality of data to be used for analysis for better business decision making. We will keep on exploring and transforming the data as we move move ahead so as to get the best out of it.

Kind Regards

Ashish Siwach