Dear Sprocket Central Pty Ltd,

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

Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows. Furthermore, recommendations have been provided to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

|  |  |  |  |
| --- | --- | --- | --- |
| Sheet Name | Column Name | Error Type | Reason |
| Transactions | Online Orders | Completeness | Null Values |
|  | Brand | Completeness | Null Values |
|  | Product first sold date | Relevancy | To convert from Integers to Date |
| New Customer List | Gender | Consistency | We found that many row had irrelevant Values |
|  | DOB | Completeness | We found many date doesn’t had correct values |
|  | Job Title | Completeness | Null Values |
|  | Job Industry Category | Completeness | Null Values |
|  | Postcode | Validity | Some values were not assigned with correct data type |
|  | Property valuation | Validity | Some values were not assigned with correct data type |
| Customer Demographic | Gender | Consistency | U - irrelevant value, (Male, Female, Femle) converted to M and F |
|  | Job title | Completeness | Null Values |
|  | Job industry | Completeness | Null Values |
|  | Deceased indicator | Completeness | Y - its irrelevant |
|  | default | Relevancy | Unknown Values |
| Customer Address | State | Consistency | (New South Wales, Victoria) converted to NSW and VIC |

* **‘Transactions Sheet’ in the Data Set has 3 Errors :-**
* **Online Orders & Brand :-**

This two columns had null values where the null values have been removed.

* **Product first sold date :-**

**(e.g. numeric values for some fields and strings for others)**

Having different data types for a given field make it difficult to interpret results at the later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

* **New Customer List sheet had 6 columns which contains errors :-**
* **Gender :-**

It had values ‘U’ which was a unknown value so the rows were removed.

* **Date of Birth (DOB) :-**

It had values were the total date of birth was not give so it has been removed from there.

* **Job Title & Job Industry Category :-**

This two columns had null values where the null values have been removed.

* **Postcode & Property valuation:-**

**(e.g. numeric values for some fields and strings for others)**

Having different data types for a given field make it difficult to interpret results at the later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

* **Customer Demographic List Sheet had 5 Columns which Contained Errors :-**
* **Gender :-**

It had values ‘U’ which was a unknown value so the rows were removed and also this columns had values **(e.g. F for female had been represented as “Femal”, “Female”)** which are converted to same as the other values in the column.

* **Job title & Job industry :-**

This two columns had null values where the null values have been removed.

* **Decreased indicator:-**

There were values which were named as “Y” which showed that the customers are not there. So, this columns were not needed and hence, this columns had been removed.

* **Default :-**

The whole column had unknown values which are of no use for further analysis and hence the whole column has been deleted to clean the data.

* **Customer Address List Sheet had 1 Column which Contained Errors :-**
* **State:-**

**(e.g. Vic for Victoria being represented as “V” and “Victoria”)**

In order to construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value.

Moving forward, the team will continue with the data cleaning, standardisation and transformation process for the purpose of model analysis. Questions will be raised along the way and assumptions documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central’s understanding.

Kind regards,

Shwetabrata Basu Ray

Junior Data Consultant