Dear John,

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table name** | **Accuracy** | **Completeness** | **Consistency** | **Relevancy** | **Validity** |
| **Customer**  **Demographic** | Inaccuracy: DOB | Missing Values in some features:  DOB, Job Title, Job Industry Category, Tenure | Inconsistency: Gender | Deceased customers: removed |  |
| **Customer Address** |  |  | Inconsistency: State |  |  |
| **Transaction Data** | Profit : Missing | Missing values in some features:  Brand, Product Line, Product Class, Product Size, Product first sold date, Online Order |  | Cancelled orders: removed | Format: List price, Standard cost , Product First sold date |

**DATA INFORMATION**

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.

* **ACCURACY ISSUES**

The data in DOB was inaccurate in the Customer Demographics region.

Mitigation: Resolved the DOB by filtering out the inaccuracies.

Recommendation: Create a profit column in the Transactions Data in order for easy calculation of sales and keep track of monetary issues.

* **COMPLETENESS ISSUES**

There were some issues in Customer Demographics as some values were missing such as-DOB, Job Title, Job Industry Category, Tenure.

Some more issues found in Transaction data with some features missing as Brand, Product Line, Product Class, Product Size, Product First Sold Date, Online Order.

Mitigation: Resolved the issues by removing the missing values so that every dataset corresponding to every transaction, customer is complete.

Recommendation: In order to keep the data uniform with all information complete, blanks field should be removed.

* **CONSISTENCY ISSUES**

There were some issues in Gender feature in the Customer Demographics such as Female, Male, F, Femal, U.

There were also some issues in the State Feature in the Customer Address sheet such as NSW, VIC, QLD, New South Wales, Victoria.

Mitigation: Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses and genders.

Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field. In order to construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value. Additionally, gender records where ‘U’ have been removed from the training dataset and Femal and F have been replaced with Female.  
The states representing the Victoria, New South Wales have been replaced with VIC and NSW respectively in order to maintain uniformity.

* **RELEVANCY ISSUES**

The inclusion of deceased customer in Customer Demographics is not suited for the future analysis. The data of alive customer is more important for predictions.

The Cancelled orders should be removed from the Transaction sheet in order to keep track of the orders Approved and in order to calculate the Profit.

Mitigation: The deceased customer data should be removed.

The cancelled order should be removed from the database.

Recommendation: Keep the data up-to-date with removing by removing the details of deceased customer regularly.

Also cancelled data can be taken care of as suggested above.

* **VALIDITY ISSUES**

The format of List Price and Standard Cost is currency but it is not containing any currency symbol, rather just contains the numeric value.

The format of Product first solved is also incorrect as it doesn’t contain the date in any of the DD/MM/YYYY format or MM/DD/YYYY format.

Mitigation: Keep the transaction details such as list price, standard cost with the currency symbol for good visibility.

Recommendation: In order to keep track of the Product first sold date, it should be stored in the above format as discussed rather than just numeric values.

Moving forward, the team will continue with the data cleaning, standardization 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,

Shubham Gupta