Hi Tony,

Thank you for providing us with the three datasets from Sprocket Central Pty ltd. The below table highlights the summary statistics from the three datasets. Please let us know if there is any mismatch in figures with your understanding.

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | Total no of records | Distinct Customer\_id | Date data received |
| Transaction | 20000 | 3494 | 20000 |
| Customer Demographic | 4000 | 4000 | 3912 |
| Customer Address | 3999 | 3999 |  |

Preliminary data exploration has been notified and the methods used to mitigate the identified data inconsistencies are listed below. In addition to that, recommendations have been provided to avoid the reoccurance of data quality issues and improve the accuracy of underlying data quality to drive business decisions.

* Accuracy in DOB column for the customer\_id =34 must be verified as Jephthah Bachmann’s DOB was mentioned as 1843.
* **Various columns, such as the brand of a purchase, or job title, or lastname have empty values in certain records.**

**Mitigation: If only small values are missing, then we filter out those records in the training data set for prediction. Else if it’s a core field then data imputation will be done.**

* **Inconsistent values for the same attribute (e.g. in the gender column F, Female)**  
  **Mitigation**: Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses*.*  
  **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 replaced based on the distribution from the training dataset.
* **Inconsistent data type for the same attribute (e.g., numeric values for some fields and strings for others)**   
  **Mitigation**: Convert selected records in characters to numeric. Remove non-numeric characters from string.
* **Recommendation**: Ensure that fact tables in the given database have constraints on data types.   
  Having different data types for a given field makes it difficult to interpret results at the later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.   
  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.

Regards,

Myvizhi (Data Analyst)