Dear Manager,

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 received. Please let us know if the figures are not aligned with your understanding.

| **Table name** | **No. of records** | **Distinct customer ids** |
| --- | --- | --- |
| **Customer demographic** | 4000 | 4000 |
| **Customer address** | 3999 | 3999 |
| **Transaction data** | 20000 | 3494 |

Data quality issues were encountered.The methods used to mitigate the identified data inconsistencies are as follows:

Recommendations have been provided to avoid the recurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

**Transactions Data**

The data has 20k rows and 13 columns.

1)In-Correct Data Type

* Transaction\_date - the data type of the attribute is int64
* Product\_first\_sold\_date - the data type of the attribute is int64

Solution:-Convert the type of attribute to date\_time

2)Missing Values

7 columns have missing values.

Solution:-We can treat missing values based on our analysis objective.

**Customer Demographic Data**

The data has 4000 rows and 13 columns.

1)Missing Values

6 columns have missing values.

Solution:-We can treat missing values based on our analysis objective

2)Data in default column

The data in default doesn't seem right,and would cause issue during analysis.

Solution:-Drop the column

3)Inconsistent Values in Gender Column

Solution:-Replace with M with Male,F with Female,U with Unspecified.

**Customer Address Data**

The dataset has 3999 rows and 6 columns.

We don't see any missing values in the dataset.

**General Observations**

● Various columns, such as the brand of a purchase, or job title, have empty values in certain records

Solution: If only a small number of rows are empty, filter out the record entirely from the training set for prediction. Else, if it is a core field, impute based on distribution in the training dataset.

● Inconsistent values for the same attribute (e.g. Victoria being represented as “V”, “Vic” and “Victoria”)

Solution: Use regular expressions to replace 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.

● Inconsistent data type for the same attribute (e.g. numeric values for some fields and strings for others)

Solution: Convert selected records in characters to numeric. Remove non-numeric characters from the 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.

Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only. Moving forward, the team will continue with the data cleaning, standardisation and transformation process for the purpose of model analysis. Qu

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.

Note:-Please check the attached analysis report for better understanding.

Thanks and Regards,

Jayadeva Javali

KPMG Virtual Internship Team