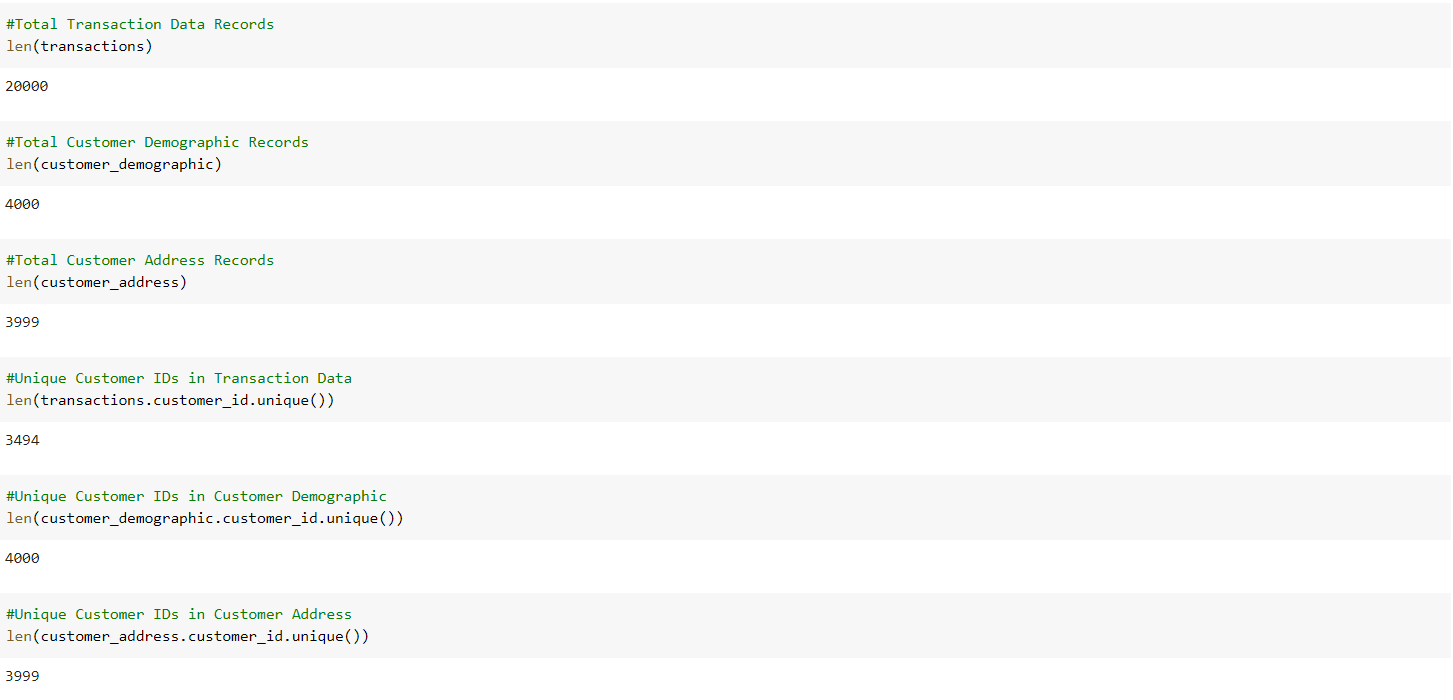
Warm Greetings,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below screenshot from Python Notebook provides descriptive details to understand the nature of data we have on our hands.



Besides our descriptive analysis, there are few data discrepancies which I would like to put across. In addition, we have tried our best to provide relevant suggestions to address the data quality issues and enhance the overall accuracy of the model which we implement to provide apt business solutions.

● There is a visible gap in the amount of Customer IDs in the ‘Transactions Table’, ‘Customer Address Table’, and ‘Customer Demographic’.   
**Solution:** We need to ensure the entries across each table is consistent. Citing the credentials, Customer Demographic will be treated as a training set for our model, indicating that the data received may not be in sync with each other which may skew the analysis results if there are missing data records. Please refer to excel file ‘data\_outliers.xlsx’ for the list of outliers between tables.

● Various columns, such as the brand of a purchase, job industry, or job title have missing entries in certain records.  
**Solution:** If only a small number of rows are empty, we can drop those data points directly before constructing the prediction model. Else, if it is a significant feature, we need to impute those entries based on distribution in the training dataset. For key datasets, such as transactions, less than 1% of transactions (totalling less than 0.1% of revenue) have missing fields. These records have been removed from the training dataset.

● Inconsistency in entry of values for the same attribute – for instance, Victoria being represented as “V”, “Vic”, and “Victoria”.   
**Solution:** Using regular expression for text cleaning. 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. 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.

● Lack of uniformity in data type for the same attribute, for instance – the numeric values for some fields and strings for others.   
**Mitigation:** Convert the records of character data type 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 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.

Continuing the progress, the team will initiate the data cleaning, standardisation, and transformation process to make the dataset ready for model building.

Sincere Regards,

Aditya Singh Kashyap