Sprocket Central Pty Ltd , a medium size bikes & cycling accessories organisation.

In particular, they speak about how the team can effectively analyze the datasets to help Sprocket Central Pty Ltd grow its business.

Primarily, Sprocket Central Pty Ltd needs help with its customer and transactions data. The organisation has a large dataset relating to its customers, but their team is unsure how to effectively analyse it to help optimise its marketing strategy.

However, to support the analysis, “*the importance of optimising the quality of customer datasets cannot be underestimated. The better the quality of the dataset, the better chance you will be able to use it drive company growth*.”

The client provided 3 datasets:

* Customer Demographic
* Customer Addresses
* Transactions data in the past 3 months

Then decide to start the preliminary data exploration and identify ways to improve the quality of Sprocket Central Pty Ltd’s data.

We have received the three raw datasets from SP rocket central Pty Limited. As per the preliminary task, in the below-mentioned list, we have analysed the quality of the raw data and we found multiple quality issues that need to be addressed. Also, we have suggested recommendations to mitigate the quality issues and improve the effectiveness of the data.

1. **Redundant Outliers.**

**Issue**: Some of the data values are outliers and can disrupt the whole dataset. For example, The customer ID “34” with the name of Jephthah Bachmann was born in 1843, meaning that he is 175 years old which is an error in the data in the Customer Demographic Table.

**Recommendation**: Remove the redundant data as it may skew the distribution of the dataset.

1. **Missing Values.**

**Issue**: Multiple attributes like “Online Order”, “Brand Name”, “Product Line”, “Product Class”, “ Product Size”, “Standard Cost”, and “product\_first\_sold\_date” in the Transactions table had blank values. Also, In the Customer Demographic “Job Title”, “Job Category” and “Tenure” some of the records are missing.

**Recommendation:** As the percentage of missing values in the datasets is low as compared to the whole dataset we can go proceed by removing them.

1. **Inconsistent Entries across the datasets.**

**Issue**: There are an additional number of entries in customer\_ids in the Transactions table than Customer Demographic and Customer Address Table. Hence, the skewed data cannot be used if there are any missing records.

**Recommendation:** We will only perform the analysis on the synced data of all the three customer tables across the customer\_ID.

1. **Multiple DataTypes for a Single Column.**

**Issue:** For the attribute “Standard Cost” in the Transaction table there are some records with special string characters which causes inconsistency in the dataset.

**Recommendation:** Remove the special characters from the records and convert all the characters into numeric data to ensure consistent data types.

1. **Duplicate values for the same column.**

**Issue**: In the “State” Column of the Customer Address Table multiple duplicate values were found such as “VIC” & Victoria, “NSW” & “New South Wales”. Also, the issue is in the “Gender” column of the Customer Demographic Dataset.

**Recommendation**: To use abbreviations of the states instead of full names for all the records to ensure consistency across addresses. For Gender Column, the records “U” can be imputed with the distribution of the dataset.

The above-mentioned quality issues along with the recommended changes ensure the consistent quality of the dataset across all the tables.

**Data visualization finding insights.**

Sprocket Central Pty Ltd has given us a new list of 1000 potential customers with their demographics and attributes. However, these customers do not have prior transaction history with the organisation.   
  
The marketing team at Sprocket Central Pty Ltd is sure that, if correctly analysed, the data would reveal useful customer insights which could help optimise resource allocation for targeted marketing. Hence, improve performance by focusing on high value customers.

Generating new columns like transaction month, profit.