Hi ;

I’m Beytullah Söylev from KPMG Data Analytics (Virtual) team. 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.

The following are the details of analysis done on the dataset:

Table Name	**Table Records**		**Table Analysis**
	Before Data Cleaning	**After Data Cleaning**	
Transaction Data	20000 rows & 13 columns (1542 blank cells)	19445 rows & 14 columns (0 blank cell)	• Total profit: $10,930,284 (app.) • ‘Solex’ is the most purchased brand name • The most and least sold product line is ‘Standard’ and ‘Mountain’ respectively
New Customer List	1000 rows & 18 columns (152 cells)	878 rows & 18 columns (0 blank cell)	• Most new customers are from the New South Wales, Australia • Most customers own cars
Customer Demographic	4000 rows & 13 columns (806 blank cells)	3413 rows & 13 columns (0 blank cell)	• Most customers are ‘mass customers’ in wealth segment • Most customers are working in manufacturing and financial services industry
Customer Address	3999 rows & 6 columns (0 blank cell)	3999 rows & 6 columns (0 blank cell)	• Most customers are from New Sales Wales (NSW) • Most customers have post code between 2000 to 2190

Issue: Additional customer_ids in the ‘Transactions table’ and ‘Customer Address table’ but not in ‘Customer Master (Customer Demographic)’.

Recommendation: Ensure that all tables are from the same period. Only customers in the Customer Master list will be used as a training set for the model.

Issue: Various columns, such as the brand of a purchase, online order or job title, have empty values in certain records.

Recommendation: 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.

Issue: Inaccurate data in DOB (e.g. DOB is 1984 in NewCustomerList which is an incorrect value for DOB).

Recommendation: Ensure that the data provided is accurate as such inaccurate data can highly affect the training set for the model.

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

Recommendation: Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses. Enforce a drop-down list for the user entering the data rather than a free text field.

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

Recommendation: Convert selected records in characters to numeric. Remove non-numeric characters from string. Ensure that fact tables in the given database have constraints on data types.

Overall, the recommendations provided are sound and should help to improve the quality of the data. It is important to note that there is no one-size-fits-all approach to data cleaning and standardization, and the specific steps taken will vary depending on the specific dataset and the intended use of the data.

Best Regards;

Beytullah Söylev