Dear client ,

This is Rime Fazilet Mehenni from KPMG Data Analytics (Virtual Internship) 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.

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

Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows. Furthermore, recommendations have been provided to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

1.**Worksheet name transaction where we identified Bank values for columns "online_order" and " brand" .the column for product_first_sold_date" was converted a date/ time format.**

a. We identified various blank values in the columns mentioned above, it is important to remove bank values from the dataset as the raise the data quality issue for completeness and many lead to inaccurate results while modelling.

b. The column for "product_first_sold_date" was converted into a date /time format which is easy to interpret , this problem may arise when exporting data from thrd party which may convert date value to integer , however they are not easy to interpret therefore changing in to date/time format.

2.Worksheet Name New Customer List where we identified bank values , there was also inconsistent values for gender .

a. As mentioned above , blanks values were discovered in the sheet for the column "second_name" however there were still blank values. There were followed by more blank and null values in columns " job_title" and "industry".
b. The column for " gender" which is a categorical varible has inconsistency, there are spelling errors for female, and some rows had abbreviations. This was changed to the columns being M for male and F for female. The column also consisted of an irrelevant variable "U" which was discarded from the column. However ,if more clarity could be provided on this it would be great or else for now it is irrelevant for the column.

3. Worksheet name customer demographic which has inconsistency for gender, there were missing values and irrelevant field called "default"

a. This worksheet was dealt with in a similar way to the worksheet for new customer list, the field for gender was changed to M / F. "U" which was an irrelevant value in the field was removed from the field.
b. Null Values were removed from "job_title" & "job_industry".
c. Irrelevant field called default was removed as it had no relationship to the data.

Moving forward, the team will continue with the data cleaning, standardisation 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

Rime Fazilet Mehenni