Dear Sir,

I am Ruchita Shah from KPMG Data Analytics (Virtual Internship) team. We have reviewed the 3 datasets provided to us for analyzing in order to optimize the marketing strategy and help Sprocket Central Pty Ltd grow its business.

Several data quality issues were encountered at the preliminary data exploration phase. Various data quality issues and ways to clean the underlying data issues so to mitigate these issues are recommended below for each dataset.

Transactions:

Issues	Recommendations to mitigate the issues
• There are 5 irrelevant fields named as "unnamed" with null values.	• Drop the irrelevant fields to make the dataset more manageable and efficient for data analysis.
• Invalid data types of following attributes, ○ product_first_sold_date ○ online_order ○ List_price	• Maintain consistency and validity of data values by casting, ○ product_first_sold_date to date format. ○ online_order to str format. ○ List_price in correct currency format
• Missing completeness of the data as the following field contains several null values. ○ brand ○ product_line ○ product_class ○ product_size ○ standard_cost ○ product_first_sold_date	• Make the fields entry compulsory so as to ensure completeness of data.

CustomerDemographic:

Issues	Recommendations to mitigate the issues
• Irrelevant field "Default" contains random data.	• Maintaining the relevancy in data by dropping the field "Default" for efficient data analysis
• Missing completeness of the data as the following field contains null values. ○ last_name ○ DOB ○ job_title ○ job_industry_category ○ tenure	• Make the fields entry compulsory so as to ensure completeness of data.
• Inconsistent data type and invalid values for, ○ DOB ○ Gender	• Maintain consistency and validity of data values by casting, DOB to date format • Ensure validity of DOB by computing the age of a person. • Place a list-box for Gender to select either the values Male, Female or others.

CustomerAddress

Issues	Recommendations to mitigate the issues
• Inconsistent values for "State" field.	• Replace short forms for state to full form in the dataset. • Place a list-box to select the state to maintain consistency of the state values.

Apart from the above listed issues, customer_id which is a common field between all the 3 datasets has several irrelevant values. Certain values of customer_id present in “transactions” table are not present in the other two datasets. So there must be common values for customer_id in all the 3 datasets to conduct efficient data analysis.

The quality of data plays an important role as the better the data quality more easily and efficiently data analysis can be conducted which will eventually help us to derive the insights from the data beneficial to the company. I hope the above-mentioned issues and recommendations will be taken into consideration and necessary steps will be taken.

Yours sincerely,

Ruchita Shah