Hello,

This is Filbert Miko Linardi from the KPMG Data Analytics (Virtual Internship) team and I would like to thank you for providing us with the datasets. We have reviewed the datasets and during the analysis we have found errors that exist in the datasets that may have become a problem later on.

Summary Table

	Accuracy	Completeness	Consistency	Currency	Relevancy	Validity
Customer Demographic	- DOB: inaccurate - Age: Missing	- Job title: Blanks - customer_id: incomplete	- Gender: inconsistency		- Default_ column: delete	
Customer Address			- States: inconsistency			
Transactions	- Profit: missing	- Online_order : blanks - Brand: blanks			- Status order: filter out "Canceled"	- Product sold date: format - List price: format

I have created a more in depth descriptions of the data quality analysis along with issues that I have discovered and the recommendations to avoid and prevent further data quality issues in the future. These recommendations will improve the accuracy data used to influence business decisions of Sprocket Central Pty Ltd in the future.

Customer Demographic

Issues

- DOB was inaccurate, Age column was missing, Job title had many blanks, Customer_id was incomplete, Gender column was inconsistent and Default_column should be deleted.
    - There were some outlier in the DOB column especially a record had 1843 as the birth year.
    - Age column should exist as this allows the dataset to be more comprehensible and easier to check for errors
    - Additional customer_id from the New Customer List is not present.
    - There are inconsistencies in the Gender column. Errors like "Femal", inconsistencies between "M" and "Male" as well as "U" for "Unspecified"
    - The default column in Customer Demographic was redundant as it consisted of random letters, numbers and symbols

Recommendations

- Create an Age, Dropdown list for gender
    - Age column should exist as this allows the dataset to be more comprehensible and easier to check for errors
    - The value type for gender should be consistent. Whether it be all "F", "M", and "U" or "Female", "Male" and "Unspecified", the values within this column should be consistent. A drop down list could be implemented to allow more complete data.

By creating a column for Age checking for errors would be easier and will assist for future analysis. The consistency in dataset is vital because it will lead to longer time when analysing the quality of the data. Each of this recommendation will help for future analysis to be more accurate and in-depth.

Customer Address

Issues

- States column was inconsistent
    - There are inconsistencies in the States column. Some of the format for the states were different, i.e. “NSW” and “New South Wales”, “Victoria” and “VIC”

Recommendations

- Dropdown list for states
    - The value type for states should be consistent. So it should all be the whole name of the state or abbreviations like “NSW” and “VIC”. A dropdown list could be implemented to avoid this issue which will allow more complete data

The consistency in dataset is vital because it will lead to longer time when analysing the quality of the data as well as higher accuracy for the dataset.

Transactions

Issues

- Profit column missing, blanks in Online_order and Brand column, “Cancelled” order status is redundant and wrong data format for both product_sold_date and list_price
    - Profit column should exist as this allows the dataset to be more comprehensible and easier to check for errors
    - Blanks in online_order and brand column meaning there are redundant data
    - “Cancelled” order status is redundant as you didn’t make the sales and there are no cashflow activity with these cancelled orders
    - The data in both product_sold_date and list_price are wrong.

Recommendations

- Create a Profit column, remove redundancies and reformat wrong data type to the correct ones.
    - Profit column will allows easier to check the accuracy of sales
    - Filter out any redundant records
    - Reformat data types to ensure better and correct analysis

By creating a column for Profile, filtering out redundant records and reformatting the data types, checking for errors would be easier and will assist for future analysis to be more accurate and correct.