Dear Kathleen *[client point-of-contact]***,**

Thank you for providing us with the datasets from Sprocket Central Pty Ltd.

As per our preliminary data exploration, we have identified specific data quality dimensions with the given datasets. The table below highlights the issue that needs to be modified to ensure that the data quality is acceptable for further analysis and visualization.

Please let us know if you have any questions regarding the same.

Data Quality Framework Table

Table Name	Accuracy	Completeness	Consistency	Relevancy	Validity
Customer Demographic	- DOB: Inaccurate - Gender: Inaccurate	Five columns have a few missing data	Gender: Inconsistent	Default: Not relevant.Delete all columns	Default: Invalid
Customer Address			State: Inconsistent		
Transactions		Seven columns have a few missing data		State: Filter out canceled values	Product first sold date: convert integer format to date time format

Below are in-depth descriptions of data quality issues that were encountered during preliminary analysis and the methods used to mitigate the identified data inconsistencies. Additionally, instructions have been added to improve the quality and accuracy of the data used to drive Sprocket Central Pty Ltd's business decisions.

Note: *Preliminary analysis and data cleaning were done using Python.*

i) CustomerDemographic Dataset

- Dropped some unnamed columns [from Unnamed: 13 to Unnamed: 25]
- Missing values in 5 columns
- ‘Gender’ column has inconsistent values
    - ‘Femal’ looks like a typo error and re-named it to ‘Female’.
    - For accuracy of the data ‘F’ and ‘M’ columns were renamed to ‘Female’ & ‘Male’ respectively.

- ‘Default’ column values are very inconsistent and invalid. So the ‘default’ column was removed.
- In ‘Gender’, the row ‘U’ has been repeated 88 times. This has to be more specific to be consistent.

ii) CustomerAddress Dataset

- Unnamed columns were dropped
- There are no missing values in the dataset
- In the ‘state’ column, the same state names were repeated. Therefore to serve the consistency ‘New South wales’ was renamed ‘NSW’ and ‘Victoria’ was renamed ‘VIC’.

iii) Transactions Dataset

- There were some unrequired/unnamed columns. Therefore it was eliminated/dropped.
- There are some missing values in 7 columns. They can be dropped or treated according to the nature of the analysis.
- There are zero duplicate values in the given dataset
- We can filter out the ‘Cancelled’ values from the ‘order_status’ column as it is not relevant to the transactions dataset
- Converted ‘product_first_sold_date’ from integer to date time format to be relevant and valid to match other data.
- ‘Online_order’ column has 360 missing data. Filled using bfill, ffill, mean & mode method.
- ‘Product_line’, ‘product_class’ & ‘product_size’ column has 197 missing data. It was filled using bfill, ffill, mean & mode method.

  If you have difficulty understanding the above pointers, please let us know to make it clear for you.

Thanks & Regards,
[Bhumika H Yogesh]