Dear Kathleen,

Thank you for providing us with the needed datasets from Sprocket Central Pty Ltd. The table and dotted points below highlight the issues in the three datasets that need to be modified to ensure that the data quality is acceptable for analysis and visualization. Please let us know if you have any problems with these issues presented.

Summary problem

	Accuracy	**Completeness**	**Consistency**	**Currency**	**Relevancy**	**Validity**
Customer Demographic	-Gender: inaccurate -DOB: inaccurate	Five columns have some missing data.	Gender: inconsistency		Default: delete all column	Default: Invalid
Customer Address			State: inconsistency			
Transactions		Seven columns have some missing data.			Order state: filter out cancelled values	Product first sold date: convert it to date format

As mentioned above you can see a summary of the data quality issues that were found during the analysis of the datasets. Below are more in-depth descriptions of the problems discovered and the method used to solve it. Following the instruction below will improve the quality and accuracy of the datasets used to make the business of Sprocket Central Pty Ltd. More productive. Please note that all the analysis and fixes were made using Python.

CustomerDempgraphic Dataset:

- There are some columns that have some missing data: (filled using ffill, bfill, mean, and mode)
  1. “last_name” column has 125 missing data.
  2. “DOB” column has 87 missing data.
  3. “job_title” column has 506 missing data.
  4. “job_industry_category” column has 656 missing data.
  5. “tenure” column has 87 missing data.

- In “gender” column there are some inconsistency values:
  1. “Femal” is consider as a typo, it should be “Female”. (Repeated 1 time)
  2. “M” and “F” should be Male and female for the Accuracy of the data. (Repeated 1 time for each)

3. "U" is repeated 88 times and needs to be more specific to be consistent.

- In "DOB" column there are some invalid data that need to be adjusted or removed:
  1. "1843" is invalid value, as the person age would be 177 nowadays and that is impossible.
- Most of "default" column values are not understandable and invalid, so we need to delete that column. The values of this column are irrelevant to the other columns and have a lack of consistency.

CustomerAddress Dataset:

- In the "state" column there are two values that needs to be reassigned to their shortcut:
  1. "New South Wales" values need to be assigned as "NSW" to serve the consistency of the dataset.
  2. "Victoria" values need to be assigned as "VIC" to serve the consistency of the dataset.

Transactions Dataset:

- There are some columns that have some missing data: (filled using ffill, bfill, mean, and mode)
  1. "online_order" column has 360 missing data.
  2. "brand" column has 197 missing data.
  3. "product_line" column has 197 missing data.
  4. "product_class" column has 197 missing data.
  5. "product_size" column has 197 missing data.
  6. "standard_cost" column has 197 missing data.
  7. "product_first_sold_date" column has 197 missing data.
- We may filter out the Cancelled values from "order_status" column as it is irrelevant to the transaction's dataset.
- Convert "product_first_sold_date" from number to date format to be relevant and valid to other data.
- We may add a new 'Profit' column to make the data more valuable and accurate.

If you faced any issue in any section above, please let us know to make it clear for you.

Best Regards,

Ahmed Soliman