Good morning,

Hope this email finds you well. First and foremost, thank you for choosing us as a trusted partner. Regarding the dataset we received last week, our team spent time to perform a data quality assessment. Please find below for the feedback as well as advisor to improve your data collector pipeline

- Customer Demographic
    - **NULL** value as well as **INVALID** value (negative age) appears in Date of Birth column. This is an important column where we should try our best to collect valid information because it will help a lot with customer segment analysis later -> change to calendar option to improve the correctness
    - Values in Gender column are **NOT CONSISTENCY**, we must decide either using abbreviate "F" and "M" or using "Female" and "Male". Also, there is some typo of 'U' in the value. -> change to drop down list for users to use
    - Default column contains all the irrelevant data -> going to be dropped
    - Also, missing value in Last Name, Job Title, and Job Industry columns. In general, the more completeness of the data, the more insight and value we can drive. -> cannot make any assumption so going to be dropped
- Transaction data
    - **NULL** value in Brand, Product Line, Product Class columns, and Online Order. -> cannot make any assumption so going to be dropped
    - Standard Cost column's values are inconsistency (contains $ sign and 'comma' for millions or thousands), take more time to data preprocessing and cleaning. -> should have data type constraint, going to be processed and convert to float
    - There are 4 rows where customer IDs are invalid and not associated with data information from "Customer Demographic" table -> going to be dropped. It may happen because of different timeframe where data is collected
- Customer Address
    - Values in STATE column are **NOT CONSISTENCY**, we must decide either using abbreviate "NSW" and "VIC" or using "New South Wave" and "Victoria". -> change to drop down list for users to use. We will use abbreviate.
    - There are 4 rows where customer IDs are invalid and not associated with data information from "Customer Demographic" table -> going to be dropped. It may happen because of different timeframe where data is collected

Thank you very much again and looking forward hearing from you soon

Sincerely,

Analytics Team
KPMG AUS