Hello,

This is Prakhar Srivastava from KPMG Data Analytics Team. Data quality assessment is very important in ensuring a good analysis outcome. Hence, as part of our analysis, we performed a data quality assessment of the data provided by your team. During the assessment we have found some data quality issues.

The following are the details of error encountered in the data set :

Transaction data in the past three months (Total records 20000)

Field Names	**Errors**
online_order	360 blank records
product_line	197 blank records
product_class	197 blank records
Product_size	197 blank records
standard_cost	197 blank records
product_first_sold_date	197 blank records
brand	197 blank records

Customer Demographic (Total records 4000)

Field Names	**Errors**
last_name	125 blank records
DOB	87 blank records
gender	Values are not consistence M, Male, F, Female, Femal, U
job_title	506 blank records
job_industry_category	656 blank records
tenure	87 blank records
default	302 blank records, inconsistent values

We suggest the following mitigates in order to improve the data quality which will eventually help us in better analytics results for your company.

- We can take mode value to fill online_order in order to maintain the consistency of data.
- 197 common records contains no information about product_line, product_class, product_size, standard_cost, product_first_sold_date, brand can be treated as invalid transactions and can be dropped (i.e. 0.009% of the original data) will help to maintain the verity of the data.
- We can take a mode year value for the missing records of customers DOB.
- In gender column following replacement can be done
    - 'U' with Unspecified
    - 'Femal' and 'F' with Female
    - 'M' with Male.
- 105 records miss both job_title and job_industry can be taken as unemployed, others can be filled with the help of the corresponding columns .
- For values missing in tenure, we can replace it with the mean of rest of the values in order to maintain the consistency of data.
- default column contains metadata and nothing special can be derived from it , can be dropped.

Regards,

KPMG(Data Analytics Team)