Hello there,

This is Youssef Hussein from the KPMG Data Analytics team. We have reviewed the datasets that were provided by your company. During the data quality analysis, we've found some quality issues in these datasets.

For CustomerDemographic dataset:

Column	Issues
customer_id	Customers with IDs = [3,10,22,23] have no corresponding address data in 'CustomerAddress' dataset.
last_name	There's 125 missing value.
DOB	There's 87 missing value.
job_title	There's 506 missing value.
job_industry_category	There's 656 missing value.
default	There's 302 missing value.
tenure	There's 87 missing value. The column is irrelevant to the dataset.
gender	Inconsistent formats. The same value has more than one format. i.e. ['F', 'Female', 'Femal'] all refer to female gender!
default	The column has invalid values. The column is irrelevant to the dataset.
deceased_indicator	This column shows the data isn't up to date since it contains information about dead people!

Recommendations:

1. 'deceased_indicator' and 'owns_car' should be boolean values (True & False).
2. Delete 'default' and 'tenure' columns since they’re irrelevant to the dataset.
3. Filter out customers with deceased_indicator = 'Y' (dead people).

For CustomerAddress dataset:

Column	Issues
customer_id	Customers with IDs = [4,12,26,27] have no corresponding address data in 'CustomerDemographic' dataset.
online_order	There's 360 missing value.
brand	There's 197 missing value.
product_line	There's 197 missing value.
product_class	There's 197 missing value.
product_size	There's 197 missing value.
standard_cost	There's 197 missing value.
product_first_sold_date	There's 197 missing value.
state	Inconsistent formats. The same value has more than one format. i.e. ['New South Wales', 'NSW'] both refer to New South Wales state!

For Transactions dataset:

Column	Issues
customer_id	Customers with IDs = [3,10,22,23] have no corresponding address data in 'CustomerAddress' dataset.
last_name	There's 125 missing value.
DOB	There's 87 missing value.
job_title	There's 506 missing value.
job_industry_category	There's 656 missing value.
default	There's 302 missing value.
tenure	There's 87 missing value.
online_order	Inconsistent data types (i.e. object should be boolean)
product_first_sold_date	The column has an invalid format. The column is irrelevant to the dataset.
list_price	The values have unknown unit currency
standard_cost	Invalid data type (object should be float)

Recommendations:

1. Delete the 'product_first_sold_date' column since it's irrelevant to the dataset.
2. Insert a unit currency into column name of 'list_price' to be 'list_price_USD' or whatever the currency used.
3. Remove ($) sign for each value and replace it with USD in the column name as following: standard_cost_USD

Thanks for your reading, please contact us for any queries.

Best regards,

Youssef Hussein