Dear Client,

This is Yuanzhe Li from the KPMG Virtual Internship team. I hope all is well with you.
We have reviewed and checked the datasets provided from your team, and have finished a data quality inspection on that. Below are the key findings and the related mitigations suggested by us for a better solution.

General information of the main tables and keys:

| Table | No. of records | Unique Customer_id | Maxi_Customer-id |
|---|---|---|---|
| Transactions | 20000 | 3494 | 5034 |
| CustomerDemographic | 4000 | 4000 | 4000 |
| CustomerAddress | 3999 | 3999 | 4003 |

**Relevancy:**
There are some Customer_id found in the Transactions or CustomerAddress table but failed to be found in the Customer Demographic table. Eg, ‘5034’ and ‘4003’.

Mitigations: To make sure the 3 tables are from the same time range, or to exclude the distinct Customer_id.

**Inconsistency:**
CustomerAddress-state: different representations for the same attribute across down the column. Eg, ‘New South Wales’ and ‘NSW’;
CustomerDemographic-gender: different representations for the same attribute across down the column. Eg, ‘Female’ and ‘F’, ‘Male’ and ‘M’

Mitigations: To fix those, using a consistent representation for each attribute.

**Wrong Data Type:**
Transactions - ‘Product_first_sold_date’ column is with data type of numeric;
New Customer List - ‘Past_3_year_bike_related_purchases’ and ‘property_valuation’ columns are with datatype of text.

Mitigations: To change the data type of ‘Product_first_sold_date’ from numeric to date;
To change the data type of ‘Past_3_year_bike_related_purchases’ and ‘property_valuation’ from text to numeric.

**Data Missing:**

| Transactions | Blank cells |
|---|---|
| online_order | 360 |
| brand | 197 |
| product_line | 197 |
| product_class | 197 |
| product_size | 197 |

| | |
|---|---|
| standard_cost | 197 |
| product_first_sold_date | 197 |

| NewCustomerList | Blank cells |
|---|---|
| first_name | 29 |
| DOB | 17 |
| job_title | 106 |

| CustomerDemographic | Blank cells |
|---|---|
| last_name | 125 |
| DOB | 87 |
| job_title | 506 |
| tenure | 87 |

Mitigations:

1. 197 missing value transactions can be excluded from the dataset as they account for less than 1% of the numbers of the transactions.
2. Other missing values in name, DOB, job_title can be assigned with values imputed based on distribution.

We will proceed with data cleaning and manipulating for your datasets, if you have any questions regarding the data quality assessment, please do not hesitate to contact us.

Regards,

Yuanzhe Li