KPMG Virtual Internship – Data Analytics

Data Quality Assessment

1. Data Missing Values Check:

Customer Demographic:

There are over 500 missing values in the job title and category columns, which is indicates that a lot of customers do not wish to provide their occupancy information. Potential reason could be privacy concerns. There are 125 missing values in the Last Name column, which show some ambiguity in the customer identification and comparison to the transaction data. The Default column has 302 missing values and since most of the values in this column is not recognisable, the usefulness of it is suspected and the client is suggested to take a closer look at this feature to assess its accessibility. There are also 87 missing values in the Tenure column which may of concern.

Transaction:

In the transaction data, there are 360 missing values in the Online Order column, and the same number of missing values (197) in the Brand, Product Line, Product Class, Product Size, Standard Cost and Product First Sold Date columns, which indicates they are from the same transactions' records.

New Customers:

In the New Customers data table, there are only 29 missing values in the Last Name column, while still over 100 missing values in the Job-related columns. There are also 17 missing values in the Date of Birth (DOB) column.

Address:

In the Address table, there is no missing value which indicates the data table is complete.

2. Data Accuracy Check

Customer Demographic:

In summary, the date of birth column in the customer demographic table has one incorrect value, and there are some expressions in the Gender columns need to be replaced with correct format.

Transactions:

In summary, the numeric and categorical variables in the Transaction table seems good without incorrect expressions, extreme values or duplications. The distribution of categorical variables are looking good. The whole table seems well on accuracy. Next, it comes to the accuracy check of the New Customer data

New Customers:

In summary, the data table do not have duplication rows and most of the values in the numerical variables seems correct, while there are four unnamed columns with randomized values should be more considered as their future's productivity. For categorical variables, in the Gender column, there are 17 rows having 'U' ('Unknown') which may need to be adjusted to correct value. The other categorical variables in the table seems make sense and accurate.

Address:

The numerical variables in the Address data seems accurate without duplication and extreme values. For categorical variables, in the State column, some expressions such as "New South Wales" and "Victoria" should be adjusted to the correct expression "NSW" and "VIC".