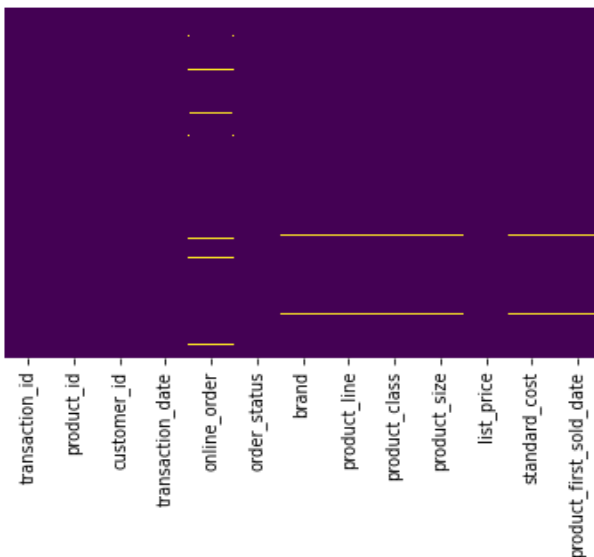Hello,

I am Soumit from the KPMG Data Analytics (Virtual Internship) team. We have reviewed the data sets which were provided by your company and during the data quality analysis we have found some error in the data sets .

Transactions data in past 3 months :

- In this data set **Online order** has 360(0.018%) missing values, **brand**, **product_line**, **product_class**, **product_size**, **standard_cost**, **product_first_sold_date** have the same missing values i.e. 197(0.009%) .
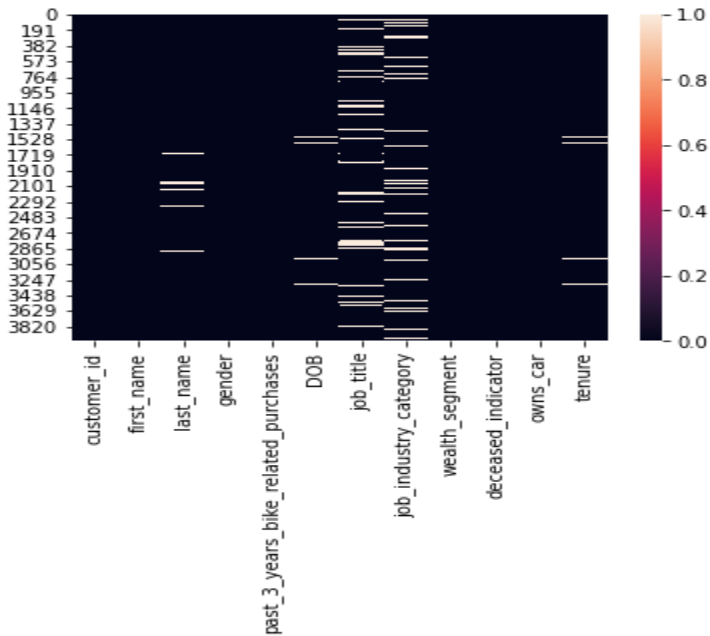

- ❏   As there is less than 1% data missing, the records have been removed .

New Customer List:

- In this sheet **last_name** has 29, **DOB** has 17, **job_title** has 106 , **job_industry_category** has 165 missing values.
- There are 5 unnamed columns.

CustomerDemographic:

- In this sheet **last_name** has 125(0.03%), **DOB** has 87(0.02%) , **job_title has** 506 (0.12%), **job_industry_category**has 656(0.16%), **tenure** has 87 (0.02%) missing values.
- There is inaccuracy in the **default** column .

[Missing values heat map]

❏ We can assign a uniform last_name of customers for missing values.
❏ We can take a mode year value for missing data in DOB.
❏ Replace gender 'U' with reference to customer names and make a consistency.
❏ Drop the 5 unnamed columns.
❏ Eliminate the blank Online Orders considered as fake.
❏ We can take the mean of rest of values and fill the missing values for tenure.

Customer Address:

❏ Inconsistent values as 'VIC' for Victoria state and 'NSW' for New South Wales .
❏ As per Inconsistent data type there is some data type transformation required (i.e string to integer/ integer to string).

Regards,
 Soumit Kar
[KPMG (Data Analytics Team)]