Hello,

Hope you are doing well! Thank you for reaching out to KPMG’s Analytics, Information & Modelling team for data exploration support! Let me share the result of my data quality analysis regarding the given datasets as well as recommendations to mitigate current data quality concerns with you in this email.

* **Customer Demographic**

I reviewed the dataset of Customer Demographic carefully, and here are my comments based on my thorough data analysis of the dataset using Python.

1. **Need to record the missing values.** There are 125 missing values in column of last\_name, 87 missing values in column of DOB, 506 missing values in column of job\_title, 656 missing values in column of job\_industry\_category, 302 missing values in column of default, and 87 missing values in column of tenure.
2. **Need to rename values in gender column.** Some of the values in the column of gender are not properly recorded. Need to rename "F" and "Femal" with "Female", and "M" with "Male".
3. **Need to fix the unknown gender.** There are 88 rows of information about customers with unknown gender.
4. **Drop the column of default.** the column of default has inconsistent values which are not properly recorded.

* **Customer Addresses**

This sheet looks good. There are no missing values in each column, and no duplications in rows. And after investigating the columns, the values in each are looking good.

* **Transaction data in the past three months**

1. **Need to record the missing values.** There are 360 values missing in column of online\_order, and there are 197 values missing in column of brand, product\_line, product\_class, product\_size, standard\_cost and product\_first\_sold\_date respectively.
2. **Need to convert the values in column of product\_first\_sold\_date into proper date format.** The values in product\_first\_sold\_date are all integers. Need to be converted to "datetime" object.
3. **Need to verify the values in column of product\_first\_sold\_date.** As the date values show that the product's first sold dates are all on the same day, just on different times of the day.

* **New Customer List**

1. **Need to drop the columns with unexpected errors.** There are columns with names of "Unnamed 16" "Unnamed 17" "Unnamed 18" "Unnamed 19" "Unnamed 20".
2. **Need to record the missing values.** There are 29 missing values in column of last\_name, 17 missing values in column of DOB, 106 missing values in column of job\_title, and 165 missing values in column of job\_industry\_category.
3. **Need to fix the unknown gender.** There are 17 rows of information about customers with unknow gender.

The above are my comments and recommendations regarding the current data quality issues I’ve discovered. I’ve also attached the modified data excel preserving the original tabs in the email. Please let me know if you have any questions.

Best Regards,

Roxanna