**Data Quality Assessment - draft Email**

Good morning,

We've done some initial exploration of your three datasets. The number of transactions and customers is impressive, and the data structure aligns with your goals. There are some issues and doubts that emerged from this preliminary analysis we want to share with you.

**Transaction dataset**

1. In the column online\_order, there are 360 rows (1.8% total columns) with null values. It does not impact too much but keep monitor if this issue is persistent on new data.
2. The columns brand, product\_line, product\_class, product\_size, standard\_cost and product\_first\_sold\_date have 197 null values. Every row with these null values has the product\_id = 0; this product\_id is present in other rows with these columns filled with different values and prices. Is it normal that the same product\_id can have different prices?
3. The column product\_first\_sold\_date is a number. This number is the Excel number that represents a date. Better to convert into DateTime format.
4. Some equal product\_id have different list\_prices and standard\_costs.
5. Some customer\_id are present only in these datasets.

**Customer demographic dataset:**

1. The columns last\_name, DOB contain missing values. Explore this lack of data, and mitigate this event with mandatory fields.
2. Job\_title, job\_industry\_category, default, tenure contain missing values.
3. The gender column has six values (Female, Male, U, F, M, Femal). You need to understand what means "U" and correct the other three misspelled errors. It can be possible to force the input to four different fixed categories: "Male", "Female", "Unknown", "Other."
4. The default column contains values encoded incorrectly. Specify the correct encoding when ingesting or when export the data.
5. The max values of "past\_3\_years\_bike\_related\_purchases" is 99. Is it a cap for the column?

**Customer address dataset:**

1. The column state has two values written in an extended and abbreviated way. Fix these values from the default list with the same format in the data collection phase.
2. The "property" column has values from 1 to 12. Is it a range? What does it mean 1? What does it mean 12?

Thanks in advance for your meaningful answers. It is a starting point to make your datasets cleaner and useful for data/business analysts.

Should you need any further clarification, please do not hesitate to contact us.

Regards,

Simone Coslovich