Final Email regarding the Data Quality Assessment

Hi there,

As per our previous conversation, I had received the datasets. It was mentioned there would be three datasets but I found there were 4 datasets in the Excel sheet that was shared with me.

* Transactions
* New Customer List
* Customer Address
* Customer Demographics

Anyhow, I’ve assessed the data quality for each of the four datasets provided and here are the issues I’ve found.

In the Transactions dataset:

* There are around 361 missing data entries under ‘online\_order’ column. The values being Boolean values – either True or False
* There are 198 missing values under ‘brand’, ‘product\_line’, ‘product\_class’, ‘product\_size’, ‘standard\_cost’, ‘product\_first\_sold\_date’ for **‘product\_id’ = 0**
* The values in the column – ‘list\_price’ should converted to currency values
* The values in the column – ‘product\_first\_sold\_date’ are supposed to be in Date format but instead they are in general format. Kindly convert them into ‘Date’ format

In the New Customer List dataset:

* In the ‘gender’ column, there are three types of values that the dataset contains – Male, Female, U. The value – ‘U’ could be replaced with appropriate value for better understanding and further analysis of data.
* The column – ‘past\_3\_years\_bike\_related\_purchases’ has values stored as Text. Convert it into numerical form for further analysis of data.
* The column ‘DOB’ has 18 missing values and they are unregistered when ‘gender’=U
* There are 106 missing values under the column – ‘job\_title’
* The values in ‘postcode’ and ‘property\_valuation’ are in Text format. Convert them into numerical form. Also, some of the values in ‘property\_valuation’ are in decimal form while others are in integer form. Convert every value in either decimal or integer form.
* Split the ‘address’ column into different columns like ‘street\_number’, ‘street\_name’, ‘area’ for better insights.
* There are 5 columns in between ‘property\_valuation’ and ‘Rank’ with no headers and some of the values in them are in incorrect format. Convert all of the values into decimal values and round them off with a specific value for better appearance and provide headers to those columns to generate insights during analysis of data.

In Customer Demographics dataset:

* In the ‘gender’ column, there are multiple values like F, Femal, Female, Male, M, U. Categorize the values – [F, Femal, Female] as either Female or F. Likewise categorize the values – [M, Male] as either Male or M and provide an appropriate value for U since it’s unclear what the value represents.
* There seem to be an outlier in the ‘DOB’ column with the value being – ‘1843-12-21’ for ‘customer\_ID’ = 34
* There are 88 missing values in ‘DOB’ when ‘gender’ = U.
* There are 498 missing values in ‘job\_title’ column
* The data in the ‘default’ column doesn’t make any sense with respect to analysis of data. Mention what the data means exactly. If there’s no appropriate form of data values for the column, then remove the data.
* There are 88 missing values in ‘tenure’ when ‘gender’ = U.

In the Customer Address dataset:

* Split the ‘address’ column into different columns like ‘street\_number’, ‘street\_name’, ‘area’ for better insights.
* There are 5 different values in ‘state’ column – NSW, VIC, QLD, New South Wales, Victoria. Replace New South Wales with NSW and Victoria with VIC to make it unique and accurate.

Kindly go through the above points and make necessary changes for the following datasets to assure the Quality of the Data for picking it forward for analysis.

Let me know if you have any questions.

Kind Regards,

XXXXXXXXX