From: abdulrahmanyasser619@gmail.com

To: client@gmail.com

Subject: Data quality assessing

Hello, I've been asked to assess the quality of the data set sent to me and send you back the feedback. This is detailed feedback for every column in the datasets. Hope you bear it with me 😊.

I'll start with **Transactions dataset:**

Online order field: it's 360 missing data which represents 2% of the entire transaction dataset.

As for [brand – product line – product class – product price]: they all 've 197 missing data which represents less than 1% of the entire records.

Standard cost: currency format is not applied to all the data values, and all the data values need to be rounded to the same decimal points.

Product first sold date: this field has 197 missing values and the data values 're formatted in numbers and must be date.

The recommendation -> remove the missing values they only represent less than 1% of the entire data unless they 're very important to the company.

Secondly **NewCustomerList**:

First name: some names need to be fixed (ex. Angeta -> angela, adriena -> Adriana).

Last name: has 29 missing values.

Gender: has 17 data values that have U instead male/female.

Past 3 years bike related purchases: the field format is text and must be numbers.

DOB: the format is text and must be date and has 17 missing values.

Job title: has 106 missing data and the repetitive jobs ends with ranked roman numbers

(ex. Account Representative I - Account Representative II - Account Representative IV. etc.)

Job industry category: has 2 unknown values [N/A].

Post code: needs to be formatted into numbers.

Property valuation: convert to numbers and unify the number format to integers.

There are 5 columns named after column [1-5] it's not clear their purpose I recommend removing them.

The rank column and column 5: has the same data values [repeated column with different name].

Secondly **Customer Demographic**:

Last name: has 125 missing values which represents 3%.

Gender: unify all forms of male/female to [male, female] and there are 88 cells has the value [U].

DOB: has 87 missing data.

Job title: has 506 missing data which represents 13% of the dataset.

Job industry category: has 656 values of [N/A] which is the same as missing data and represents 16% of the dataset.

Default column: remove it.

The recommendation -> as for Inconsistent values for the same attribute I recommend making a drop-down list so the user can choose from to avoid misspelling/errors.

Missing values represent a significant proportion 13% & 16% so my recommendation is to fill 'em up

Lastly **Customer Address: the state field data values must be united.**

Important note: there is a difference between number of unique customer_id field in the three datasets **[CustomerAddress has 3999 / CustomerDemographic has 4000 / Transactions has 3494]**

Which tells that the records aren't from the same period since the **transaction_date** starts from **1/1/2017** and ends **12/30/2017** which may lead to difference in the other data values that exist in the common fields in the three datasets.

Kind regards, [Abdulrahman yasser Mahmoud].