

Good day,

I am writing to you to get some clarifications on the data sets you provided to the KPMG Analytics Team. While we were doing some data review and exploration, we observed some errors as well as some information we assumed might be an error but will need more clarification on. Here's a quick complications and questions for each meta-data;

#### CUSTOMER DEMOGRAPHIC DATA

1. Gender has some incorrect values. Distinctively, there should only exist female and male in the gender column but some were represented by 'U'. What can assume for those with 'U' gender? I assumed we could delete them to prevent affecting our analysis. A total of 87 cells affected by this out of 4000 cells.
2. Date of birth of some are missing. I searched for any inconsistency and validity of age searching for outliers. One of it would be a customer age recorded for 178 years. Is this an error? But we suggest a deletion.
3. In the job industry category, for unknown data 'n/a' was included which contradicts the contents so I suggest a categorical impute of "Others" to classify other job industries instead of removing the data.
4. In the deceased indicator column, I assumed the 'N' and 'Y' is a representation of No and Yes if the customer is dead or not. So, for better understanding we'd be replacing them correctly by using N as 'Not dead' and Y for 'Dead' and also filter out dead customers.
5. The default seems to be irrelevant to the data so we deleted it.
6. Some DOB are missing

#### CUSTOMER ADDRESS DATA

1. My first observation was trying to compare the total number of customers with the customer demographic table and I observed 3 extra customer was present here. The customer with the ID 4001, 4002 and 4003. A customer with just address information but no other details information could proof useless for analysis so we will be mitigating it.
2. In the state column I observed NSW and VIC is the same as New South Wales and Victoria in Australia so I corrected them to their full name. Also, I replaced QLD with Queensland.

#### TRANSACTION DATA

1. A customer with ID 5034 doesn't appear in the customer demographic table or address table which gets me wondering if it's an error or not.
2. The online order column contains either TRUE or FALSE but there still exists blank cells. Should I assume a false answer for the affected?
3. I observed products without brands exist in the transaction data alongside an empty product line, product class, product size and standard cost with first date the product the was sold. Should I assume the sets of data incorrect?

#### Recommendations;

- Enforce a drop-down list for the user entering the data rather than a free text field.
- Ensure that fact tables in the given database have constraints on data types.

Kindly let us know what you do think to the above corrections and clarifications. Thank you and have a nice day.

Kind regards,

Adejayan Ebunoluwa