

*Good day,*

*I am reaching out today to submit my preliminary analysis findings regarding the review of the data quality to ensure that it is ready for the second part of the analysis.*

*The issue that we spotted with the datasets are as follows:*

**Transactions dataset**

- missing values  
Online order: 360  
Brand: 197  
Product line: 197  
Product class: 197  
Product size: 197  
Standard cost: 197  
Product first sold date: 197
- The product first sold date column is either not in an understandable date format or not a date at all.

**New customer data set**

- Last name column has 29 missing values  
Date of birth: 17 missing values  
Job title: 106 missing values  
Job industry category: 165.
- Columns 16 to 20 don't have any titles.
- The gender column has an unidentified entry type "U".

**Customer's demography data set**

- Missing values  
Last name: 125  
DOB: 87  
Job title: 506  
Job industry category: 656  
Default: 302  
Tenure: 87
- The gender column has conflicting entries: m and males, female and F, and error entries "femal" and "w".
- All the entries in the default column are entry errors.

- One of the customer age is greater than 110 (customer id = 34).

#### **Customer's addresses data set**

- The state column has conflicting entries: "New South Wales" and "NSW"; "VIC" and "VICTORIA".

#### **CONCLUSION AND RECOMMENDATIONS**

At this stage, the data cannot be used as they are for the next analysis part and we recommend:

- Trying fixing the missing values if possible.
- Fixing or removing the product first sold date column, and the default column.
- And fixing entries error or giving extra info about them.

*Thanks for trusting me with this task!*

*Kind Regards*

*Jeremie*