KPMG Virtual internship- documentation

Problem statement: Sprocket Central Pty Ltd is a medium size bikes and cycling accessories organisation who needs help with their customer and transactions data to optimise their marketing strategy.

3 data sets have been provided -

- Customer Demographic
- Customer Addresses
- Transactions data in the past 3 months

Task 1: Draft an email to the client identifying the data quality issues and strategies to mitigate these issues. Refer to 'Data Quality Framework Table' and resources below for criteria and dimensions which you should consider.

Transactions dataset :

Blanks were found in the following columns -

- Online_order
- Brand
- Product_line
- Product class
- Product_size

1. The datatype of the list_price column was changed to currency

2. A profit column was also added as (List_price - Standard_cost)

3. Incorrect format and values for the column 'product_first_sold_date'

NewCustomerList Dataset

1. **Past_3_years_bike_related_purchases** column value datatype was changed from text to number. The same was carried out from postcode and property_valuation

2. There were hidden columns between P and V with numbers but without context.

3. Columns 'Rank' and 'Value' need more context.

CustomerDemographic dataset:

- **Non-Uniform format for Gender**

Solution for this : Select column -> editing -> find and select -> replace option -> Input the wrong formats identified and what you what it replaced with

Adding a column for 'Age' would be useful for further analysis.
For this use **=(NOW() - *Cell with the DOB in that row*)/365** and change the decimals

In the **customer demographic dataset**, by sorting, it was found out that Customer Jephthah Bachmann with customer ID 34 Had a date of birth dating 1843-12-21. This false value is to be changed by updating the entry and for the purpose of analysing the data, this entry is deleted.

This data set also had a column titled 'Default' with corrupt data, which is also deleted.

The Job_title column has blanks but they are not deleted just yet as we are not yet aware if they are needed.
For a cleaner dataset, the deceased personals and one's with job_titles as blank are simply filtered out (not deleted)

CustomerAddress dataset:

No duplicates but there are missing customer data. Final Customer ID and total number of customers do not match.

The states column does not have a uniform format.
Example- New South wales has been represented as 'New South Wales' as well as 'NSW'
Victoria has been changed to VIC
And New South Wales to NSW using the replace option

To whomsoever it may concern,

Thank you for providing us with the datasets from Sprocket Central Pty Ltd. After assessing these datasets for data quality, we have noted data quality issues that we have encountered. The nature of quality issues, and mitigative strategies have been recommended to improve the quality of the data used for decision making.

- **In the transactions dataset,**

- Blanks were noted in the following columns- Online_order, Brand, Product_line, Product class and Product_size. Since only a small number of these have been found, it is advisable to delete these records entirely.
- The data type of the column List_price was changed from Text to currency
- Incorrect format and values for the column 'product_first_sold_date' was observed.

- **In the NewCustomerList dataset**

- Hidden columns between P and V with numbers but without context was observed. Addition of context to these in the master sheet is recommended.
- The datatype in columns Past_3_years_bike_related_purchases, postcode and property_valuation was incorrect.

- **In the CustomerDemographic dataset,**

- Non-Uniform format for Gender was observed. ‘F’ and ‘Femal’ were changed to ‘Female’ and ‘M’ was changed to ‘Male’ for uniformity.
- Issues of data accuracy was observed in the column ‘DOB’ where it was found out that Customer Jephthah Bachmann with customer ID 34 had a date of birth dating 1843-12-21.
- Possible corrupt data was observed in the column titled ‘Default’ which was deleted.
- Blanks were observed in the column ‘Job_title’.

- **In the CustomerAddress dataset,**

- Non uniform format for the column ‘State’ was observed. In the given records, New South wales has been represented as ‘New South Wales’ as well as ‘NSW’ and Victoria has been changed to VIC. For further data collection process, it is recommended that uniformity be followed.
- Missing values were also observed in the given dataset.

We wish to spend some time with your data management team to understand various context and aspects of the given datasets to move forward with analysing this data for gathering insights.

Regards,
KPMG