Dear Sprocket Central Pty Ltd,

Thank you for giving us the opportunity to work with you, for providing the dataset. We have reviewed the dataset and summarized the following data quality issues with the dataset.

Work Sheet Name	Data Quality Issues
Transactions	Null Values & Relevancy
NewCustomerList	Null Values & Consistency
CustomerDemographic	Null Values, Consistency & Relevancy
CustomerAddress	Consistency

We further describe the data quality issues in detail with our comments as mentioned below :

Work Sheet Name	Column Name	Null Count	Data Quality Issues	Comments
Transactions	online_order	360	Null Values	
	brand	197	Null Values	
	product_line	197	Null Values	
	product_class	197	Null Values	
	product_size	197	Null Values	
	standard_cost	197	Null Values	
	product_first_sold_date	197	Null Values	
	product_first_sold_date	NA	Relevancy	The format was in Text, instead of Data Format.
NewCustomerList	last_name	29	Null Values	
	DOB	17	Null Values	
	job_title	106	Null Values	
	gender	NA	Consistency	U is deleted from the dataset.
CustomerDemographic	last_name	125	Null Values	
	DOB	87	Null Values	
	job_title	506	Null Values	
	tenure	87	Null Values	
	gender	NA	Consistency	spelling error, U is deleted from the dataset.
	default	NA	Relevancy	Deleted the whole column as the data was not readable.
CustomerAddress	state	NA	Consistency	Combination of aberration and full form.

The table above gives some Data Quality issues with respect to the dataset given. And below we have explained the issues in detail and recommended some solution as well.

1. **Transaction worksheet** have Null values and Relevancy data quality issues. There were null values present in the online_order,brand,product_line,product_class,product_size, standard_cost and product_first_sold_date. product_first_sold_date column was converted into integer.

    a. **Null Values** : There were null values present in the mentioned columns. Many machine learning algorithms fail if the dataset contains missing values. However, algorithms like **K-nearest and Naive Bayes** support data with missing values. You may end up building a biased machine learning model, leading to incorrect results if the missing values are not handled properly.

b. **Relevancy** : The column “product_first_sold_date” was in text format. Have converted the text format into date format. These issues may have occurred, If you import data into Excel from another source, or if you enter dates with two-digit years into cells that were previously formatted as text, you may see a small green triangle in the upper-left corner of the cell. This error indicator tells you that the date is stored as text.

2. **NewCustomerList worksheet** have Null values and Consistency data quality issues. There were null values present in the last_name, DOB and job_title. U is used for unspecified.

    a. **Null Values** : There were null values present in the mentioned columns. Many machine learning algorithms fail if the dataset contains missing values. However, algorithms like **K-nearest and Naive Bayes** support data with missing values. You may end up building a biased machine learning model, leading to incorrect results if the missing values are not handled properly. There were null values present in the last_name as well, but not deleted the cells as it does not make any huge impact on the data.

    b. **Consistency** : In the gender columns, U was mentioned for Unspecified. It always recommended to use same pattern. For example if you are using aberration then only go for “F” for female, “M” for male and “U” for Unspecified. And if you are going for full form then use Female, Male and Unspecified. Don’t mix aberration and full form together.

3. **Customer Demographic worksheet** have Null values, Consistency and Relevancy data quality issues. There were null values present in the last_name,DOB,job_title and tenure. Gender columns has consistency issues and default column has relevance issue.

    a. **Null Values** : There were null values present in the mentioned columns. Many machine learning algorithms fail if the dataset contains missing values. However, algorithms like **K-nearest and Naive Bayes** support data with missing values. You may end up building a biased machine learning model, leading to incorrect results if the missing values are not handled properly. There were null values present in the last_name as well, but not deleted the cells as it does not make any huge impact on the data.

    b. **Consistency** : In the gender columns, U was mentioned for Unspecified. Female was written in 3 ways, “F”, “Femal” and “Female”. Male was written in 2 ways, “M” and “Male”. It always recommended to use same pattern. For example if you are using aberration the only go for F for female, M for male and U for Unspecified.

And if you are going for full form then use female, male and unspecified. Don't mix aberration and full form together. And spell check should be carried out.

c. **Relevancy** : The column "Default" has some data which was not readable that's why deleted the whole column. The data should be at least readable and relevant for the research work. If the data is not relevant, there is no use of such data.

d. **deceased_indicator :** The Column **deceased_indicator** has two entries as Y for yes and N for No. we have deleted the Y, as we only want to work upon the customer which are alive.

4. **CustomerAddress worksheet,** There were no null values found in the dataset but the column "state" has Consistency issue.

   a. Consistency : The column "State" has some Consistency issue. There were combination of aberration and full forms. For example, New South Wales was written in 2 ways " New South Wales " and " NSW" and Victoria was written in 2 ways "Victoria" and "VIC". It always recommended to use same pattern, Either use aberration or full form, don't mix it.

Added columns

We have added two columns as transaction_month and Profit for better understanding of the dataset.

Wrok Sheet Name	Columns Added	Comment
transaction_id	transaction_month	columns were added for understanging the data.
	Profit	

As this is the Task 1, we have reviewed the data and observed the data quality issues and summarized it as well. The corrected excel sheet is also attached for your reference. We are now excited to work upon the data and visualizing it well.

Dhwani Gupta
Thanks and Regards