Dear Sprocket Central Pty Ltd,

Thank you for sharing the Sprocket dataset with us. After a thorough review, we have identified several data quality issues within the dataset. Below, we outline the problems we encountered and explain how we addressed them. Additionally, we present our plan for moving forward with the data cleaning process.

Sheet name	Column Name	Data Quality issue
Transactions	online_orders	Null Values
	brand	Null Values
	product_first_sold_date	Relevancy (int to Datetime format)
NewCustomerList	last_name	Null Values
	gender	Unknown value (U)
	job_title	Null Values
	job_industry_category	n/a
CustomerDemographic	Last_name	Null Values
	gender	F, Female & M, Male & U
	Job_title	Null Values
	job_industry_category	n/a
	deceased_indicator	Removed Yes (As died)
	default	No Relevancy to our Analysis

The table above highlights several data quality issues present in the Sprocket Central Pty dataset. We have carefully examined these issues and developed recommendations to prevent their recurrence in the future. Please find our suggestions below for addressing these data quality concerns effectively.

1. In the "transaction" worksheet, we observed the presence of blank values in the "online_order" and "brand" columns. Additionally, we performed a conversion of the "product_first_sold_date" column into a date/time format.

To address this data quality issue, we recommend taking the following steps:

a. **Handling blank values**: Blank values in the "**online_order**" and "**brand**" columns should be removed from the dataset. These empty fields can introduce data quality problems and may lead to inaccurate results during modeling. By eliminating these blank values, we can enhance the overall completeness of the dataset.
b. **Conversion of "product_first_sold_date" column**: The original format of the "**product_first_sold_date**" column was difficult to interpret or might have been incompatible. This issue can potentially occur when exporting data from a third-party source, where the date values might be converted into integers. Therefore, **we converted it into a date/time format**

that is more intuitive and easier to work with. This ensures better data comprehension and facilitates meaningful analysis.

2. In the "New Customer List" worksheet, we encountered two data quality issues: Blank values & inconsistent values for gender.

a. **Blank values** were identified in the "**second_name**" column. While this might not be a critical issue as we can use the first name instead, it is still important to note the presence of blank values. Additionally, further blank and null values were found in the "**job_title**" and "**job_industry_category**" columns. Addressing these blank values is essential to ensure the dataset's completeness and reliability.
b. The "**gender**" column, being a categorical variable, displayed inconsistencies, an **irrelevant variable "U"** was removed from the column. Further clarification on this issue would be helpful; otherwise, it is considered irrelevant for the column at this stage.

3. The "customer demographic" worksheet presented several data quality issues, including gender inconsistency, missing values, and an irrelevant field called "default." To address these concerns, the following actions were taken:

a. **Gender inconsistency**: The "**gender**" column, being a categorical variable, displayed inconsistencies such as spelling errors for "female" and some rows with abbreviations. To standardize the gender representation, we made the necessary corrections, using **"M" for male** and **"F" for female**. Any irrelevant values, such as **"U"** were removed from the gender column to ensure data consistency.
b. **Removal of missing values**: Null values in the "**job_title**" and "**job_industry**" columns were identified and subsequently removed.
c. **Removal of irrelevant field**: The "**default**" field, which had no relationship to the dataset's information, was identified as irrelevant and removed from the dataset.

Moving forward, the team will proceed with the data cleaning and transformation process in preparation for modeling. Throughout this process, any questions that arise will be addressed, and all assumptions made will be thoroughly documented. To ensure alignment with the understanding of Sprocket Central Pty Ltd, it would be beneficial to collaborate with your data Subject Matter Expert (SME) closely to ensure this alignment.

Best Regards,
Anand Kumar Singh