Dear Sprocket Central Pty Ltd team:

Thank you for providing us with three datasets from Sprocket Central Pty Ltd. Here is the brief summary of the datasets that we received. Please let us know if you have any questions about the figure.

Table name	No. of records	Distinct customer IDs
Customer Demographic	4,000	4,000
Customer Address	3,999	3,999
Transaction Data	20,000	20,000

The following list some noteworthy data quality issues. The reference mitigation approach was also noted and some recommendations were made to prevent recurrence of similar data quality problems in the future.

- **The data accuracy need to be verified**
    - Locate problem:
        - The ‘DOB’ of customer with customer_id 34 is abnormal. It shows ‘1843-12-21’ which means he is 177 years old. It is obviously inaccurate information.
        - The ‘default’ column in Customer Demographics is meaningless since it fills with messy code.
    - Mitigation: Given the very small number of these types of problems, we could remove this inaccurate record.
    - Recommendation: We cannot completely rule out inaccurate data at this stage, but ensuring data accuracy is one of the most fundamental parts of data quality. We strongly recommend your company enhance data verification in the future.
- **Many columns of data have varying degrees of data missing**
    - Locate problem:
        - There exist over 500 rows of missing data in column ‘job_title’ and ‘job_industry_category’ in Customer Demographics.
        - 7 columns in Transaction contain missing data. Especially, 197 rows of missing data are highly correlated in ‘brand', 'produce_line', 'produce_size', 'product_class', 'standard_cost', and 'product_first_sold_date'.

    - Mitigation: It is normal to have a small number of missing values in a dataset. In this case, we do not have to include these values in the training set since the vacant values are less than 1%.
    - Recommendation: If the missing value is less than 3% of the dataset, we could filter out these records from the training set. Otherwise, if it is a core feature, missing values need to be estimated and filled by existing data.
- **Inconsistent value for the same attribute**
    - Locate problem:
        - Victor being represented as “Victor” or “Vi” in Customer Demographics.
    - Mitigation: Use regular expression to replace extended values in abbreviations to ensure consistency.
    - Recommendation: Enforce a drop-down list for users entering the data.
- **Inconsistent data type**
    - Locate problem:
        - The ‘product_first_sold_date’ in Transaction is not a date, but numbers instead.
    - Mitigation: Convert selected records into correct data type.
    - Recommendation: Ensure the fact that tables in the given database have constraints on data types.

Moreover, that will be great if you are willing to provide documents that explain each column of data to improve data understanding. For example, Column ‘tenure’ in Customer Demographics has no unit. We assume that this column shows how many years of these customers in their current position.

Moving forward, we are still working on the data cleaning, transformation and feature selection. More questions will be raised during the process. If possible, we would like to spend some time with you to make sure our working processes are aligned with Sprocket Central’s understanding.

Best regards,
Yujing Song