

## **RE: Sprocket Central – Data Quality Issues and Recommendations for Data Treatment**

Dear Mr./Mrs. \_\_\_\_\_:

As per your request, our team had carefully assessed the dataset that you provided. Here is a summary of the data quality issues that might have negative impacts on the analysis going forward. Mitigation strategies are also recommended.

A Github repository was created to house the documents and results related to this work. Please refer to the Jupyter Notebook for the full data quality assessment.

- Module 1 Notebook: [Link](#) and Dashboard File: [Link](#)
- Project Github: [Navyhoang/KPMG Virtual Internship Challenge: Customers and transactions analysis \(github.com\)](#)

### **Customer Demographic Data:**

- **Data type: customer\_id is in integer data type, it can be converted to object/text/str data type as this column is not used for computations.**
- Columns with missing values:
  - Last\_name, job\_title\_, job\_industry\_category: 'nan' values can be filled with 'Unknown'.
  - Tenure: a conservative approach is to fill 'nan' with the minimal value of this column.
  - Age (calculated from DOB): filled with the mode value of the column.
- Mistyped data:
  - **First\_name, last\_name : have values with less than 3 characters. There are chances that these values are mistyped and thus results in multiple representations of the same value. However, this issue will not affect the model in the next stage as we will use customer\_id to identify unique customers, the first\_name and last\_name columns will be dropped.**
  - Gender: contains mistyped data such as 'F', 'M', 'U', 'Femal'. They can be simplified to "Female", "Male", or "Unknown".
- Irrelevant data:
  - Default: this column does not add any value to the analysis, it can be dropped.
  - Deceased\_indicator: almost all values are 'N'. This column doesn't add much differentiating factor to the ML model in the next stage. It can be dropped.
- Invalidated/ contradicting data:
  - Age (calculated from DOB): max value is 177. This is out of the acceptable range since the world's oldest person record is 122. Values larger than 177 can be replaced by 122.

### **Customer Address Data:**

- **Data type: customer\_id is in integer data type, it can be converted to object/text/str data type as this column is not used for computations.**
- **Missing records: There are 4 customers (id 3, 10, 22, 23) that are presented in the Customer Demographic data but not in the Customer Address data. Vice versa, there are 3 customers ( id 4001, 4002, 4003) that are presented in the Addresses data but not in the Demographics data. This indicates that the data received may not be from the same period. Only customers from the Demographic dataset will be used for machine learning model training.**

- Mistyped data:
  - Address: 399 rows either have missing street numbers or numbers start with '0'. This is not a normal practice in real life. 0's can be stripped out. Those addresses with missing street numbers can still be kept for the analysis since the information from other columns can still be useful to answer other questions.
- Irrelevant data:
  - Country: all customers are from Australia. This column does not add any value to the analysis and can be dropped.

### **Transaction Data:**

- Data type:
  - transaction\_id, product\_id, customer\_id are in integer data type, they can be converted to object/text/str data type as these columns is not used for computations.
  - Product\_first\_sold\_date: this column shows the number of days counting from Jan 1, 1900. The data can be converted to date\_time data type.
- Missing records: There are 507 unique customers that are presented in the Customer Demographic data but not in the Transactions data. Vice versa, there is 1 unique customer ( id 5034) that is presented in the Transactions data but not in the Demographics data. This indicates that the data received may not be from the same period. In the next stage, these records inconsistency will be cleaned up prior to train the model. Only customers from the Demographic dataset will be used for machine learning model training.
- Columns with missing values:
  - Online\_order: 360 rows with missing values. They can be dropped since they only make up ~1.8% of the dataset.
  - Brand, product\_line, product\_class, product\_size, standard\_cost, product\_first\_sold\_date: 197 rows with missing values. They can also be dropped since they only make up <1% of the dataset.

Thank you for choosing KPMG for this exciting project work. We are happy to receive your feedbacks and iterate the analysis until it meets your expectations. Please let me know if you have any questions or concerns.

Best regards,  
Thao Hoang

### **Data Analyst**

Cell: (306)715-3789

Email: [navy.hoang@mail.utoronto.ca](mailto:navy.hoang@mail.utoronto.ca)