**To Whom It May Concern,**

Thank you for providing us with the three datasets from Sprocket Central Ltd. I am writing this email to point out the data quality issues found as summarized on the table below.

***Data Quality Table Summary***

|  |  |  |  |
| --- | --- | --- | --- |
| **Metrics** | **Customer Demographic** | **Customer Address** | **Transaction** |
| Accuracy | *Inaccurate format:*  D.O.B |  | *Inaccurate format:*  Product First Sold Date |
| Completeness | *Missing (2 – 16% of total data):*  Last Name, DOB, Job Title,  Job Industry, Tenure |  | *Missing (0.9 – 1.8% of total data):*  Online Order, Brand, Product Line, Class, Size, Standard Cost, Product first sold Date |
| Consistency | *Inconsistency:*  Gender | *Inconsistency:*  State |  |
| Currency | Delete deceased customers & customers >100 yo |  |  |
| Relevancy | Delete default column |  |  |
| Improve | *Add:*  Age |  | *Add:*  Profit, Recency |

**Transaction dataset**

Transaction dataset is the backbone for analyzing customer segmentation accuracy. Hence, this dataset must be complete, accurate and up-to-date. The record provided only contains data from 2017 which is not sufficient for our models to analyze each customer. The ideal dataset should contain data from 2017 to recent.

Secondly, this dataset is not complete as the following columns are found to contain a certain number of missing values:

1. 360 missing value on the ***'online\_order'.***
2. 197 missing value on the ***'brand', 'product\_line', 'product\_class', 'product\_size', 'standard\_cost'*** and ***'product\_first\_sold\_date'.***

As customer tends to have different purpose on cycling and shopping preference, it is essential to fill up these missing values. However for this analysis, as it is insignificant (0.9 – 1.8% of total data), the following rows containing missing values were removed.

Thirdly, the values in ***'product\_first\_sold\_date'*** column is not accurate as it does not contain the appropriate date value. It would be favorable if the format is the same as the ***‘transaction\_date’***.

Finally, it will be helpful to provide profit and recency data for each customers thus allows the data analysis team to generate insights and see trends efficiently.

**Customer Demographic & Address**

Firstly, formatting error and inconsistency are found in customer demographic and address dataset. ***‘default’*** column contains a mixture of random string which is not applicable for this project hence deleted. Besides, the same date formatting error has been found on ‘***DOB’*** hence reformatted.

It is also practical to ensure overall dataset ‘s consistency. The ***‘gender’***column in Customer Demographic could be redefined as **F** = female, **M** = male and **U**= unknow. Additionally, the ***‘state’***column in Customer Address can recorded in state code only.

Missing values are detected in several data with most significant one found in Job Title and Job Industry (around 16%). Since random filling using pandas is misleading, we removed current data rows containing null values as customer segmentation might requires detail information. This is to be noted as future improvement for data collection and recording alignment.

**Equivalence of Transaction, Customer Demographic, and Customer Address dataset**

While aligning the three datasets, it’s discovered that many customers (through ‘***customer\_id’***) have missing demographic and address data. It is important to collect this data in the future hence company won’t lose any potential customers due to analysis inaccuracy rooting from insufficient data.

**Summary: Mitigation and Recommendation**

To mitigate data input inconsistency and accuracy, reformat date and alias in the beginning. Adding age and profit data to enrich future sales and segmentation analysis should also be considered. It is also recommended that company align data input based on single record ‘***customer\_id’*** and collect demographic data for every new customers.

These recommendations and mitigation strategies are effective and easy to implement. They will not only improve the analysis output that one can perform within the company but will increase the level of analysis that can be performed by KPMG and other hired analysis teams.

Please let us know if you have any queries surrounding the issues presented.

**Best regards,**

**Safirah Almira Nadyanti**