Dear Sprocket Central Pty Ltd,

I am writing this email to point out the data quality issues found on the dataset. The main focus of the following discussion will be on the data which will affect the model performance on providing a legitimate marker strategy in the later stage.

**Transaction dataset**

The transaction dataset could help provide insight into each customer behavior to improve the accuracy of customer targeting. Hence, this dataset must be complete, accurate and up-to-date.

First of all, the overall dataset contains only the 2017 transaction record, 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.

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 the format is the same as the ***‘transaction\_date’***. Furthermore, the products with id = 0 does not have consistent information.

Finally, it will be helpful to provide the currency unit for the ***'list\_price'***. The currency unit allows the data analysis team to evaluate the validity of each price value and the relevancy between each feature.

**NewCustomerList dataset**

First of all, the ***‘first\_name’*** and ***‘last\_name’*** are not relevant to the analysis and should be taken away to prevent violation of the GDPR act. A unique ***'customer\_id'*** could be assigned to each customer for identification purpose instead of using their first and last name.

Other than that, 4 headers are missing between the ***‘property\_valuation’*** and ***‘Rank’***which makes it difficult to speculate the usage of this information. Apart from missing header, the following columns also contain missing values:

1. 17 missing values on the ***‘DOB’.***
2. 165 missing values on the ***‘job\_industry\_category’***.
3. 106 missing values on the ***‘job\_title’***.

**CustomerDemographic and Customer Address:**

First and foremost, the ***‘default’*** feature contains a mixture of random string which is not applicable for this project.

Besides that, the CustomerDemographic dataset does not match with the CustomerAddress dataset. When merging two of the dataset, 4 customers are missing from the dataset due to the **‘customer\_id’** mismatch. The mismatch customer ids are 3, 10, 22, 23, 4001, 4002 and 4003.

Apart from that, an error was found in the CusomerDemographics dataset as there is a 175 years old customer (customer\_id = 34) who purchased 59 bike-related products over the past three years. Furthermore, there are 87 missing value on the ***‘DOB’*** feature.

Additionally, a duplicate data was found where the same address appear at different postcode and state. The duplicate address are 3 Mariners Cove Terrace, 3 Talisman Place and 64 Macpherson Junction.

Finally, it is practical to ensure the format consists of the overall dataset. The ***‘gender’***column in CustomerDemographic could be redefined as **F** = female, **M** = male and **U**= unknow. Additionally, the ***‘state’***column in CustomerAddress could follow the NewCustomerList dataset where each state is recorded in state code only.

**Equivalence of NewCustomerList and CustomerDemographic + CustomerAddress dataset:**

To analysis how each feature from the current customer will affect the sales and market, the number of features (i.e., columns) in the CustomerDemographic and CustomerAddress must match with the NewCustomerList dataset. However, it is found that the ***‘Rank’*** and ***‘Value’*** of the customer property are missing from the CustomerAddress dataset.