Dear managers at Sprocket Central Pty Ltd

I am writing this email to inform you what we have learnt about the dataset you’ve sent us.

The dataset is thoroughly inspected. It is very comprehensive and would be great value for us to dive deeper into. But firstly, there are a few problems with the dataset. If these problems can be solved, the dataset can be used to its fullest extent.

For customer demographics, the customer ID is correct, there are no missing or duplicated IDs found. There some issues with DOB, customer ID 34 has the DOB of 1843-12-21, which makes him 177 years old, which means there is an error with this data point. There are also 87 customers whose DOB data is missing.

For the field of gender, there is one person with F instead of Female, and there are 88 people with no gender data associated with them. All the customers with the gender undetermined also do not have their DOB data, except for customer ID 34, who has his DOB wrong. In order to mitigate the problems with DOB and gender, survey emails can be sent out to the customers for the correction of information or those portions of the dataset can be simply ignored in the future analysis.

The default value field seems to be corrupted. Some of them are numerical values, some of them are commands. For example: for user id = 3958, default value is **1;DROP TABLE users**, which is an SQL injection attack. So, the whole dataset should be sanitized before it can be inputted into the database.

There are currently 505 people with job title missing. Other than the missing value, the data is reliable. There are 655 people with n/a as job industry category.  We suggest sending additional surveys to the people to fill the missing data.

There are no problems with the past 3-year bike-related purchases when viewed in combination with the transaction records. But the transaction record only goes back for a year.

For the customer addresses, the values of the state are not consistent. Some states are full names, while others are abbreviations. It is suggested that all abbreviations be substituted by the full names. The postcodes are correct as they do not associate with more than one states. In addition, the postcode can be used to verify that street addresses within these postcodes exist.

The property valuation does not have units associated with them; we need more information regarding what units are been used. It will also be helpful to include when the valuations are conducted to help keep the values up to date. There are 3 customers who do not have their address stored in the database.

For the transaction record. There is no transaction ID missing. The product IDs do not correspond well with the item brand and product line. One ID may be corresponding to multiple different items.

Transaction dates are correct.

There are 359 orders missing whether online purchased information. There are 196 purchases with **Brand, product line, product class, product size, standard cost and product first sold date** information missing. If you have that information in your database, please provide it to us. Otherwise, those missing values can be ignored in the future analysis.

The product first sold date is not accurate. Some products have two different first sold dates. We also cannot make sense of what those 5 digits numbers are supposed to represent, please provide us metadata regarding the product first sold date representation format.

Kind Regards

KPMG data analyst team

Peter Chen