Dear Sprocket Central LTD,

This document tries to identify abnormalities present in the Sprocket Central Pty LTD datasets. Using a data quality framework as a guideline, Customer Demographic,Customer Addresses, Transaction data in the past three months which is a subset of the superset will be analysed.

Customer Demography

- The column 'default', lacks a clear understanding of what it stands for. Also, it contains characters that lack a clear description of what the column means.
- NAN values, the entire dataset has nan values of 1800. Columns; last_name (125), DOB(87), job_title(506), job_industry_category (656), default (339), tenure (87).
- The gender column has inconsistency ('F', 'Male', 'Female', 'U', 'Femal', 'M') were used
- The DOB date has both nan values, also the field contains ages of people who are supposed to be dead by now.
- Some of the details in the deceased indicator column is out of date, as most marked alive are supposed to be dead.
- Some columns has mixed of different characters in a single column

Customer Address

- The state column in the dataset has multiple values of storage for the value: Victoria ('VIC')

Transactions

- The dataset shows a total of null values amounting to 1,542. This values coming from

brand	197
product_line	197
product_class	197
product_size	197
standard_cost	197
product_first_sold_date	197
online_order	360

- The email specified that the transactions were within the period of 3 months but data shows a period exceeding 3 months period.
- Three figures in the standard_cost column has no currency symbol attached and also not rounded off by 2 decimals.
- Column product_first_sold_date storage format does not describe the purpose of this column. The format of storage showed not datetime format and cant be determined which is month, day, or year.

General Recommendation

- Authorized users who have accessed the database have to be cut down in order to ensure data integrity and data consistency, this will help with terms used when entering data.

- Some columns that showed mixed or different characters, should be converted to either numeric or particular data type for the purpose of analytical manipulation.
- As mentioned above, many of the columns show nan values (empty), this can be cleaned up or dropped if the number of rows that have nan values, otherwise it can be filled with median or mean values, well depends on their core values.
- Some columns can be given a function to transform data entered into the appropriate data.
- Transaction date which is supposed to be dated within 3 months, these other rows should be dropped to meet the requirement of this document and to enable accuracy in analysis.

Thanks
Kelvin Obed