To: Client, Sprocket Central Pty Ltd,
Subject: Quality issues pertaining to customer data - reg

Dear Manager,

Thank you for providing the dataset containing data related to the customers of Sprocket Central Pty Ltd. The dataset was received on December 3 2022.

I would like to raise the following issues pertaining to the quality of the dataset and propose some strategies to mitigate them:

For the Transactions Sheet:

- Data is outdated, it was collected in the year 2017.
  Solution: Kindly provide the latest customer data.
- The online order column contains 360 entries with missing values.
- For 197 entries, values are missing under the columns named ‘brand’, ‘product line’, ‘product class’, ‘product size’, ‘standard cost’ and ‘product first sold date’.

New Customer List:

- For 17 entries, the date of birth is not given.
- For some customers, the date of birth appears to be incorrect. For example, Laura Fawdrie was born in 2002, but she is working as a VP as of 2017.
  Solution: Data can be recorded again or the column can be dropped.
- For 106 entries, job title is not given, and for 165 entries, job industry category is not given.
- Under the date of birth column, not all the values are not formatted in the same way, some are formatted as dates and some are formatted as general text.
  Solution: The values formatted as general text are reformatted as dates.

Customer Demographics:

- There is a data entry where the year of birth was entered as 1843, which is unrealistic. Solution: drop the column.
- Under the gender column, multiple category variables are used for the same gender. For example: F, Femal and Female are used to represent female gender.
  Solution: Corrections can be made in the entries.
- A column named ‘default’ contains garbage values.
  Solution: Drop the column.
- For 506 entries, job title is not given, and for 656 entries, job industry category is not given.
  Solution: Check for correlation between the number of sales and the job industry category, if there is not much correlation, then the column can be dropped.
  The job title column contains several categories, so here either the missing values can be filled or the column can be dropped.

Customer Addresses:

- Some extra customer IDs are given (4001-3) and some IDs are missing (3, 10, 22, 23) with respect to the customer demographics sheet.

> Solution: Drop the entries in the demographic sheet for there is no corresponding entry in the address sheet and vice versa.

For the entries containing missing values, it has been decided to drop the entries as the number of such entries is small compared to the overall number of entries in each table.
Note: The data and information in this document is meant only for the intended recipient. The purpose of this document is to help the client avoid the data quality issues raised above while recording data in the future. The data analysis team will go ahead with the data cleaning, and transformation procedures for further analysis. The data analysis team will spend some time with the client to ensure that the assumptions made by the team align with the client’s understanding of the business.

Thank you,
Regards,
Akhil