Dear Sprocket Central Pty Ltd,

Below are some of the discoveries we made about the datasets, as well as strategies we could use to mitigate the issues before moving forward to the next stages of this project. In order for us to design a meaningful marketing strategy, we need to ensure the quality of our data before forming any decisions.

**Transaction worksheet**

* This worksheet contains 555 rows of data that contain missing values. There were 197 rows of data that were missing the same data which were: Brand, Product line, Product class, Product size, Standard cost, Product first sold date
* I can provide you with a file of these 197 rows that include the above-mentioned missing values
* The “online order” feature had 360 missing values. Note that only two of the data points overlap with the 197 rows from earlier. This shows that we do not information about this order was taken. The “online order” feature tells us whether the order was completed online or in person which is an important information to have if we want to explore investing into ecommerce.
* The “product sold date” feature does not seem to have the correct information as the data under this column is not in the correct date format.
* The product sold date feature is stored in an incorrect format. I expected a date, but the data shown were in number format.

**New Customers List worksheet**

This worksheet has 285 data points with null values. The features that contain null values include: Last name, DOB, Job title, Job industry category. These rows can not be dropped as there are only 1000 data points in this sheet.

These features are important as Sprocket Central will need to see the demographics of their customers to narrow the focus of their marketing strategies.

**Customer Demographic worksheet**

* There is an error with the dataset where one person has a date of birth of 1843-12-21 which would make the person way over 100 years old and it also shows that this person has made a purchase within the last 3 years. When sorted the dataset by the DOB, the person that have the lowest DOB is 1931-10-23 which still seems reasonable. It can be said that the datapoint with the 1843 DOB is an incorrect data entry.
* The “default” feature is filled with unreadable text symbols. Under this feature, there are datetime objects and numbers so it is unclear what “default” represents.
* Features within this worksheet that contain null values are: Last name, DOB, Job title, Job industry category, Default, Tenure.

**Customer Address worksheet**

The state column contains the state where the customer is located. However, this column used different abbreviations to describe the same state. For instance, New South Wales was described as “New South Wales” and “NSW”. These two groups of customers are from the same state but they may be segmented into different groups when we perform data analysis on this dataset. Another state that had more than one name is Victoria. “Victoria” and “VIC” were used to describe customers from the Victoria state. Note that there are no missing values in this worksheet.

Customer ID’s are not in sync across the three datasets. For instance, the transactions worksheet has up to 5034 customer ids, the customer demographics worksheet has up to 4000 customer ids and the customer address worksheet has 4003 customer ids. Data may become skewed if the data received are not in sync. Sprocket should ensure that only customers in the master list are used for the dataset prepared to build the business models.

Please let me know if you need any further clarification.

Kind Regards,

Huda