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.

**Transactions 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 have attached a csv file of the list of datapoints that contained the above mentioned missing values

* Under the “online order” feature, I noticed that 360 data points had missing values, I have also included the csv file that contains only these data points. Note that only two of the data points overlap with the 197 rows from earlier.
* The “product sold date” feature does not seem to have the correct information as the data under the category is not in the correct date format.

The “online order” feature tells us whether the order was complete online or in person which is an important information to have if we want to explore investing into ecommerce.

There are 197 datapoints that missed 6 features of information. However, these data could still be useful to us should we decide to look at: the transaction dates to figure out seasonality of your business, average amount sold and costs to figure out profit margins in relation to dates.

We have 20 thousand data points to work with and 555 data points with missing values only represents for 2.8% of the data so we could also simply drop the data without making a noticeable impact of the result.

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**

* Between the property valuation and Rank columns, column names are missing for the data. Therefore, we cannot figure out what the data represents.
* The oldest timestamp for the DOB is 1938-06-08 seems reasonable as a 82 year old could purchase a bike as well.
* This worksheet has 285 data points with null values. The features that contain null values include:
  + Last name
  + DOB
  + Job title
  + Job industry category

They do seem to be important features to have since Sprocket Central wants to see the demographics of their customers to narrow the focus of their marketing strategies.

* I included a csv file of all the data points with missing values

There are five columns that are not labelled and requires further clarification

The dataset only contains 1000 datapoints so we cannot just drop the rows. The features we should focus on are the DOB, job title and job industry category.

**Customer Demographic worksheet**

* This worksheet contains a total of 1370 data points with missing values
* 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. 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” truly represents.
* Features within this worksheet that contain null values are:
  + Last name
  + DOB
  + Job title
  + Job industry category
  + Default
  + Tenure

There was one incorrect data entry for the person who has a DOB of 1843 which can be removed.

The “default” feature has inconsistent data types which made the data under the feature unreadable. Sprocket should ensure that the table for a database have constraints set on datatypes (ie. string, float, integer)

**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.
* There are no missing values in this worksheet.

Sprocket should place a greater emphasis on uniform naming conventions to employees. A way to force employees to use certain names is to give multiple choice survey style recording system so users can be restricted to using certain names for different states in Australia.

Customer ID’s are not in sync across the 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.