Dear [Client point-of-contact],

Thank you for providing us with the four datasets from Sprocket Central Pty Ltd. The below information highlights the summary statistics from the four datasets received.

Following are the initial observation of the files sent by you:

1. The **CustomerAddress** sheet contains following issues:

- Inconsistent values for the same attribute (e.g. Victoria being represented as "Vic" and "Victoria", New South Wales being represented as "NSW" and "New South Wales")

Mitigation Strategy:

- In order to construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value. Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses.

2.The **Transactions** sheet contains following issues:

- 'list_price' and 'standard_cost' columns have different formatting. The latter contains dollar ($) sign while former do not;
- 'product_file_sold_date' column is incorrectly formatted at number, however, it needs to be in a date format;
- Also, few columns in this sheet contains missing values, like in online_order, brand, product_line, product_class, product_size .

Mitigation Strategy:

- Filled in all null values with the backward and forward propagation, i.e. with next and last valid observations respectively, except for 'standard_cost' and 'product_first_sold_date' columns in which null values are filled using their respective columns mean or average.

3. The **NewCustomerList** sheet contains following issues:

- This sheet contains hidden columns at column-labels Q to U which have been dropped;
- Few columns in this sheet contains missing values, like in last_name, DOB, job_title, job_industry_category.

Mitigation Strategy:

- Filled in all null values with the backward and forward propagation, i.e. with next and last valid observations respectively and in last_name , DOB filled none values.

4. The **CustomerDemographic** sheet contains following issues:

- This sheet contains one column named 'default' is incorrectly formatted.
- However, this sheet contains also contains null values, like in last_name, DOB, job_title, job_industry_category, tenure.

- Inconsistent values for the same attribute (e.g. Male being represented as “Male” and “M”, Female being represented as “F” and “Female”)

<u>Mitigation Strategy:</u>

- Drop the ‘tenure’ column as it contains incorrectly formatted values
- Filled in all null values with the backward and forward propagation, i.e. with next and last valid observations respectively.
- In order to construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value. Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses.

Please let us know, if you need any clarification.

Regards
[Junior Consultant Name]