**Data Quality Assessment**

For *Sprocket Central Pty Ltd*

To make things faster and clearer, I separated the explanations/suggestions with concrete issues. I assessed each sheet separately while considering overarching (global) issues too. I categorized the suggestions and labeled them with numbers corresponding to the explanations below.

**Reasons and advice:**

1. In terms of ***completeness***, CustomerDemographic, NewCustomerList and Transactions contain some empty cells. This can cause issues when a specific piece of data is queried and there is nothing to return. If such cells are referred in other tables, an error can occur. Therefore, I suggest leaving no empty cells or, if there is no data, to denote it explicitly by inserting *“n/a”* like in other cells where data is not provided.
2. ***Accuracy*** could be improved in CustomerDemographic and Transactions as there exist some typos and junk values. This is to ensure data can be sorted and classified correctly, as well as to assure the correct values are returned when data is queried. Junk values take up space and are unnecessary. They do not provide any useful information. I suggest rechecking the pointed-out cells for any typos and replace or remove any rows or columns with junk data.
3. ***Consistency*** should be kept within columns in CustomerDemographic, CustomerAddress, NewCustomerList and Transactions. It is best to keep a single name for values of specific category because it makes things easier to manage as data is in one place. Same format across the values within a column allows correctly comparing data when doing analysis. I suggest picking one or the other naming for a certain class and sticking with only one cell format within a single column and every other column of the same type.
4. ***Validity*** should be checked for some data in CustomerDemographic, CustomerAddress, NewCustomerList and Transactions. Values must be within appropriate bounds and use the correct data type and format. Having incorrect type of values can cause errors and inconsistencies when data is handled, as well as provide falsifying information. It is best to look at each case separately and fix and approve the data reflects the reality and is formulated correctly.
5. ***Relevancy*** must be ensured in CustomerAddress and Transactions. Irrelevant data consumes extra space and requires extra management when a dataset is expanded. Having old data might not reflect the information correctly if it is not used anymore. I suggest getting rid of columns or rows that are not required.

Below, the specific problems are identified. Although it was said only 3 datasets will be provided there is a 4th dataset, NewCustomerList, which I have also checked. Since it contains some duplicate problems with CustomerDemographic and CustomerAddress, I have not pointed them out and therefore the datasheet should be checked for them. The non-duplicate issues are identified.

**Customer Demographic**

* Columns *last\_name*, *DOB, job\_title, default* and *tenure* have cells with ***missing values*** **(1)**. For example, the row with a value of *“4”* in *customer\_id* has no data in *last\_name* and *job\_title.*
* Columns *gender* and *DOB* have ***inaccuracies*** **(2)**.
  + One cell has a value of *“Femal”,* lacking *“e”* at the end.
  + The value of “*1843-12-21”* in *DOB* column is most likely mistyped.
* Columns *gender,* *DOB*, *job\_title* and *deceased\_indicator* with *owns\_car* have ***inconsistencies*** **(3)**.
  + *“Male”/ “Female”* and “*M”/ “F”* are used to denote the same thing. I suggest using “Male”/ “Female” to be consistent with the data in NewCustomerList.
  + Row with *DOB* value of “*1843-12-21”* has a different cell format from the other values within that column.
  + The last word of the value *“Analog Circuit Design manager”* in *job\_title* is not capitalized which makes this value inconsistent with other ones.
  + *“Yes”/ “No”* and *“Y”/ “N”* are used to express the same thing between columns *deceased\_indicator* and *owns\_car.*
* Column named *past\_3\_years\_bike\_related\_purchases* is ***valid (4)***, but it could be improved (shortened and simplified) as it would make it easier to perform any queries to this name.

**Customer Address**

* Columns *address* and *state* have ***inconsistencies (3)***.
  + Some addresses use different notation for street numbers, for example, number 7 is sometimes written as *“7”*, sometimes as *“07”*. There should be no *“0”* in front of the number if the amount of digits in every street number is different. Otherwise, 0’s could be added for formatting and sorting reasons.
  + *“New South Wales”*/ *“Victoria”* and *“NSL”*/ *“VIC”* are used to denote the same thing. I suggest sticking to the short version as it will be consistent with NewCustomerList too.
* Some values in columns *customer\_id* and *address* are ***not valid (4)***.
  + When comparing IDs in CustomerAddress with IDs in CustomerDemographic, some of them are not included in one or the other dataset. For example, ID *“4003”* is only found in CustomerAddress.
  + Address numbers usually do not have more than 3 digits and they never have a value of 0.
* Column *country* might be ***redundant (5)*** for the dataset if all the customers use addresses in Australia.

**New Customer List**

* Columns *past\_3\_years\_bike\_related\_purchases*, *postcode* and *property\_valuation* have ***invalid (4)*** cell formats and *customer\_id* is missing.
  + Many cells with numbers are given text format. As well as that, *past\_3\_years\_bike\_related\_purchases and postcode* cell formats are inconsistent with the ones in CustomerDemographic and CustomerAddress.
  + The table should also contain IDs of each new customer for them to be properly linked in other tables they may be added to.

**Transactions**

* Columns *online\_order, brand, product\_line, product\_class, product\_size, standard\_cost* and *product\_first\_sold\_date* contain some ***empty cells*** ***(1)***.
* Column *customer\_id* contains ***inaccurate (2)*** value of *“5034”*. No customer has this ID in CustomerDemographic.
* Columns *transaction\_date, list\_price* and *standard\_cost* have ***inconsistencies*** ***(3)***.
  + *transaction\_date* has different date notation than in other dataset sheets. DD/MM/YYYY should be changed to YYYY-MM-DD. It would also be wise to use a custom cell format rather than date format to keep it consistent with the other sheets.
  + *list\_price* has different price notation than in *standard\_cost.*
  + Some values in *standard\_cost* have different format than the majority, for example, *“312.7350159”* has more than 2 decimal values and does not have a dollar symbol in front.
* Columns *product\_id,* *product\_first\_sold\_date* and *standard\_cost* contains some values that are ***not valid*** ***(4)***.
  + It would be good to check whether *product\_id* denotes a unique item. There are some duplicate values for *product\_id* with different *brands* or *product\_lines*. One ID should refer to only one item, however, if there are same IDs but for different brands, a composite key should be used to uniquely identify the product.
  + Values (or their notation) in *product\_first\_sold\_date* should be changed to represent the date like in other columns of the same type. If the values do not refer to date, the column should be renamed.
* Some records might not be ***relevant (5)*** anymore as it was mentioned that the dataset should contain only the transactions within the last 3 months.