

Sir/Ma'am,

The following were the issues with the Sprocket Central Pty Ltd. data received from your side:

1. **Missing column values:**

- Missing values in Customer Demographic sheet were found in following columns:  
Last Name, DOB, job title, job category, deceased indicator and tenure.
- Missing values in Customer List sheet were found in following columns:  
Last Name, DOB, job title, job category.
- Missing values in Transactions sheet were found in following columns:  
Online order, brand, product line, product class, product size, standard cost, product first sold date.
- No missing values found in Customer Address sheet.

**Solution:**

- Dropping columns mentioned is not considered in this scenario as a lot of data can be lost.
- First Name and Last Name columns can be combined into a single column "Name".
- DOB can be filled by creating a new column "Age" which calculates the age of the person. Mean of the age can be taken and DOB can be filled accordingly.
- The columns which have missing values can be removed if they have no importance in the current scenario.

2. **Duplicated values:**

- No duplicated values found in the sheets.

3. **Data Inconsistency:**

- Joining tables was not successful as there were additional Customer Ids in Transactions sheets which are not found in other sheets. These are known as outliers.
- Some values in columns have different data types for eg some contain integer type or string type.
- In the gender column, gender was also defined as 'F','M','U' in some rows.
- In DOB, year was also defined like 1890 which is not true.

**Solution:**

- If the data should not contain outlier, make sure that the data was gathered from the same source at same time.
- For inconsistent data type, it is recommended to convert the data type of all the values in the column to string type.
- In gender column, 'F' to be replaced by 'Female', 'M' to be replaced by 'Male' and 'U' to be replaced by 'Unknown'.
- For DOB, a range should be defined in advanced so that there should be no inconsistency. If the year is below the lower limit or above the upper limit, a mathematical formula can be defined to bring it back to the range defined.

This was the Quality Assessment done by the team. The team has already started performing data cleaning. Further improvements can be discussed and we can draw the best solution for this.

Regards