Hi

Thank you for providing the datasets for Sprocket Central Pty Ltd. Below is the description of data quality issue discovered and method used for mitigation in the 4 datasets.

**Customer Demographic:**

1)Date of Birth (DOB)- Inaccurate date of birth and missing date of birth.

* Filter the date of birth from descending to ascending to remove the outlier and to remove the row with no date of birth
* Recommendation: Adding age column can help easily removable of outlier data (date of birth)

2)Deceased\_Indicator- Record with ‘Y’ indicator has been removed

* Filter out deceased record which is ‘Y’
* Deceased person is not the current customer filtering the data which is deceased ‘Y’ will increase the accuracy in further analysis

3) Last Name- Blank

* No need to filter the data with blank last name as it will not affect the further analysis

4) Gender – Inconsistence gender notation

* Inconsistence gender notation converted to proper notation. ‘ F’ is converted to ‘Female’ , ‘Famal’ is converted to ‘Female

**Customer Address:**

1. State: Record with full name of state converted to 3 letter Code names

* Replaced all ‘New South Wales’ to ‘NWS’ and ‘Victoria’ to ‘VIC’.
* Dropdown option help identify the inconsistency in abbreviation

**New Customer List:**

1. Gender – Inconsistence gender notation

* Inconsistence gender notation converted to proper notation. ‘ F’ is converted to ‘Female’ , ‘Famal’ is converted to ‘Female

1. Property Valuation, Value – Converted to Number format

* Converting ‘Property Valuation’ and ‘Value’ from text to number will help in easy analysis

**Transaction:**

1. Last\_Price, Standard\_Cost : Converted to currency

* ‘Last\_Price’ and ‘standard\_cost’ is converted to currency from text.
* Converting this column to currency will allow value to interpret easily and help in accurate analysis

1. Product\_first\_sold\_date: Converted to short date format

* Column is converted to short date format.
* It allows easy understandable format for ‘product\_first\_sold\_date’ column and can be used for meaningful analysis.

Above are all the data quality assessment issue and recommendation to mitigate the issue. Once datasets are cleaned further analysis will be easy.

Please let me know if you have any question on data quality assessment.

Regards

Abhisha Burande