**Owino Otieno Victor**

**Data Quality Issues**

The data provided by Sprocket Central Pty Ltd suffers from various data quality issues which makes the extraction of insightful information close to impossible and biased in a sense. The various problems are;

1 Out of the four sheets provided namely, Customer Demographic, Transactions, New Customer list and Customer Address, three of them had missing values which turned out to be the biggest issue with the data. In addition five columns in the new customer list dataset were unnamed thus not being far from missing themselves.

2. A case in hand was that of a person born way back in 1843 and according to the data the person is still alive. The data is in accurate for it would render the person to have close to 180 years which is impossible.

3. The “default” column in the Customer Demographic data contains values whose datatypes, formats and ranges are unspecified. Thus no insightful meaning can be drawn from the said column.

4. There was inconsistency in the Customer Demographic data, gender column. There were many variations of female which was the cause of the said quality barrier.

5. However, it was positive to note that none of the dataset had any duplicate value.

**Recommendations.**

1. In the case of the Missing Values the below approach should be applied:

a. In the case of missing values in the last name column, the null values should be replaced with a “Doe” like name for anonymity. This is because the last name doesn’t bear any weight in analysis and is supplemented with the other column like gender making it not that important..

b. In transactions datasets, for the missing values in the standard cost and product first sold date columns which are of the float datatypes, the null values should be replaced with the means of the specific columns for the column follows a normal distribution thus the mean is not far from the median as many value cluster around the median so will those that are missing.

c. Still in the transactions dataset, the missing values in the brand, product line, product class and product size which are categorical in nature should be replaced with the modes in each specific column. Because, the null values are less significant in the data ad there is a huge disparity between the top and the second largest categorical value that any addition of the most occurring values to the missing data still maintains the structure of the data.

2. The cell with the inaccurate date of birth is appended to the Customer Demographic dataset which is to be returned to the client for clarification among the others with missing values.

3. The “default” column with the unspecified data types, range and format should be dropped from the Customer Demographic Dataset.

4. Choosing one form of female and making the other variants of the female similar to the one chosen will remedy the inconsistency in the column.