Dear Client,

Regarding the datasets which were provided to us, we looked through it and discovered some key quality issues that will affect our analyzing the data. Below are the framework descriptions of the data quality issues and methods of mitigation used. Please follow the recommendations provided to improve the accuracy of data so that we will be able to make better business decisions for you.

***List of Issues***

**Accuracy**

Some data from Data of Birth in “Customer Demographic” is inaccurate; missing profit columns in “Transactions”.

*Mitigation:* Filter out the outlier in DOB.

*Recommendation:* Create a “Profit” column in the Transactions dataset to check accuracy. It will also help in future analysis. Creating a column for age will also allow for easier identification of errors.

**Completeness**

Additional customer IDs were inconsistent among “Customer Demographic”, “Customer Address” and “Transactions”.

*Mitigation:* Filter all customer ids.

*Recommendation:* Take out all the unique ids from the datasets given by the filter option at the top which also completes the dataset with unique values.

**Consistency**

Few blank values were present in the three datasets which are inconsistent values in the dataset.

*Mitigation:* Filter out the blank values from the following columns.

*Recommendation:* Remove all the blank values for better performance of the dataset and easy purpose in the future to calculate the transactions.

**Currency**

The prices in the “Transactions” dataset with integer values changed into currency.

*Mitigation:* Cells with integer values representing prices were changed into currency values with their proper denotation of signs.

*Recommendation:* Using the currency values we are finding out the “profit” columns in transactions datasets to improve their quality.

**Relevancy**

Few data in the “New Customer List” the relevant values are only shown rest are minimized into a merged data form.

*Mitigation:* Data from the New Customer List like property validation and the rest 3 columns are not relevant which had similar values columns that were deleted and minimized when in use.

*Recommendation:* The irrelevant values can be hidden into one to show a better version of understanding of the dataset which is needed after changing the values into numbers and errors formed in it.

**Validity**

In the address column of a dataset had irrelevant values which has to be deleted.

*Mitigation:* The text values which are not useful can be removed to mitigate the irrelevancy from the data to protect other data present in it.

*Recommendation:* Error values in a few columns of the data can be filtered out into numbers or test values to use them as allowable ones.

**Uniqueness**

Values from Customer IDs which are unique keys could be found in the all dataset provided

*Mitigation:* Few duplicate values which are missing or similar are removed through unique keys in the Customer IDs column.

*Recommendation:* Improve the accuracy through the uniqueness of the values of the datasets by removing the outliers. The states represented in “Customer Address” and the gender column in “customer demographic” had few duplicate values which were filtered and replaced into useful data only.

Moving forward, the team will continue with the data cleaning, standardisation and transformation process for the purpose of model analysis. Questions will be raised along the way and assumptions documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central’s understanding.

Kind Regards,

Gloria Sara Joji