I would like to talk about the quality issues in the dataset. I am listing some points which need to be given emphasis on for the data to be ready for analysis-

1) There are many missing entries in the data which makes it difficult to analyze it. For example, in the transactions tab we have about 197 missing entries referred to as blanks in the list price, so it would be difficult to determine the revenue for as we don’t know the price of all the products. Similarly, we have missing values in some rows for the columns- brand, product\_line, product\_size, list\_price, standard\_cost and product\_first\_sold\_date. I would recommend to fill data for these missing rows.

2) The column product\_first\_sold\_date in the transactions tab has wrong formatting. The date format should be same as transaction\_date which is the standard date format of the country (MM-DD-YYYY).I would recommend to change the whole column with values in correct format.

3) The list\_price column should have a standard currency as it is price related to money. It should be listed as $ or pound or any currency in which we have listed the price. The currency should be specified in the column.

4) There are many irrelevant or wrong values in many columns which will reduce the accuracy of our model. They need to be corrected or removed. For example, The gender column in the CustomerDemographic tab has wrong values like F, U, M and Femal. I would recommend changing these values with correct ones.

5) There are some values which are not practically possible. For example, There is a value in DOB in the NewCustomerList tab which says the year of birth is 1843 but its not practically possible. I would recommend to check the database and give correct values for such columns.

6) The default column in CustomerDemographic tab has data in a non-understandable format. I think this is because of the source of data. Please check and change it into understandable format.

7) The data has some duplicate values in it which will lead to the analysis with duplicate values and wrong results. For example, transaction\_id in Transactions has some duplicate values but the transactions are unique and cannot have duplicate values. I would recommend deleting the duplicate values to maintain unique values in the dataset.