Hi Sir,

We have received the three raw datasets from SP rocket central Pty Limited.

* Customer Demographic
* Customer Addresses
* Transaction data in the past three months

As per the preliminary task, we have analyzed the quality of the raw data and found multiple quality issues in the data that’s needed to be addressed and we’ve suggested recommendations to mitigate the quality issues and improve the effectiveness of the data.

1. Accuracy: The term “accuracy” refers to the degree to which information accurately reflects an event or object described.

* Observation: In Transaction data the column **product\_last\_sold\_dates** values doesn’t match up with the date value rather than seems to be some other feature labeled wrongly.

Recommendation: To re-check what column **product\_last\_sold\_dates** data actually represents in the dataset.

* Observation: In Customer Demographic data the **DOB** columns has a date look at the customer (Jephthah Bachmann, we'll notice that they were born in 1843, meaning that they are 175 years old! This is obviously an error in the data.

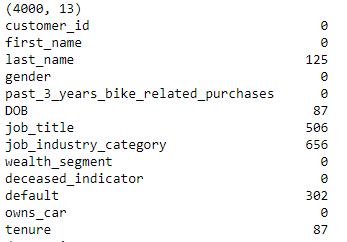


Recommendation: To remove the outliers or incorrect data from the dataset to lower the standard deviation.

1. Completeness: This dimension can cover a variety of attributes depending on the entity. For customer data, it shows the minimum information essential for a productive engagement.

* Observation: *Customer Demographic* Dataset have some few missing values in features like**job\_title , job\_industry\_category ,DOB** and last name.

*Missing Value*



Recommendation: We cannot remove all the empty values from the data as it will reduce the amount of data present with us but we can fill the empty values with mean, mode or median as per the columns.

eg. **tenure** can be filled with mean and **job\_title** and **job\_indusry\_category** can be filled with mode.

1. Consistency: At many companies, the same information may be stored in more than one place. If that information matches, it’s considered “consistent.”

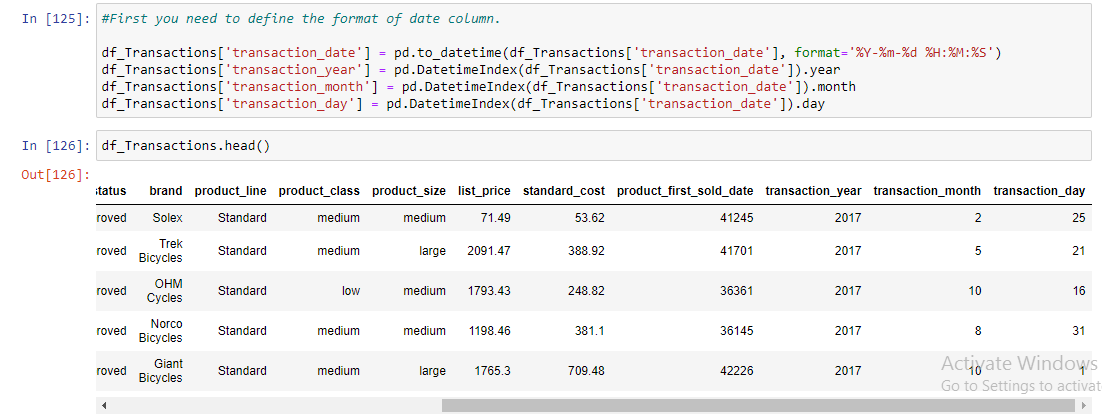
* Observation: There are an additional number of entries in **customer\_ids** in the Transactions table than *Customer Demographic and Customer Address* Table.

Recommendation: We can perform the complete analyses only on the data for which **customer\_ids** matches on all three tables.

1. Validity: It is a data quality dimension that refers to information that doesn’t conform to a specific format or doesn’t follow business rules.

* Observation: In Transaction data the column **transaction\_date** contains complete date.

Recommendation: The date data is correct but it needs to be broken down to day month and year to perform better analysis on the data.



* Observation: default column in Customer Demographic doesn’t contain the valid information.

Recommendation: To remove the column from the dataset.

1. Uniqueness “Unique” information means that there’s only one instance of it appearing in a database.

* Observation: In Customer Demographic the column gender contains some additional category which are similar to other category but just annotation is different.

Recommendation: The single annotation to be used for one category



Please look into the above-mentioned quality issues along with the recommended changes to ensure the consistent quality of the dataset across all the tables. If all the suggestions are matched we can proceed with further analysis of the data to find some suitable insights for the company.

Regards,

Abhijeet Shete