Subject: Discussion about the strategies to mitigate the data quality issues in the dataset

Dear Sir/Mam,

I am writing this email on behalf of the entire data analysis team. In this mail, I would like to discuss the data quality issues and the strategies to mitigate those.

At KPMG, we tend to do the data analysis on the basis of seven data quality dimensions: -

1. Correct Values or **Accuracy**
2. Data fields with values or **Completeness**
3. Values free from contradiction or **Consistency**
4. Values up to date or **Currency**
5. Data Items with Value meta-data or **Relevancy**
6. Data containing allowable values or **Validity**
7. Records that are duplicate or **Uniqueness**

In the following section, I would talk about the three datasets given by your company along with the data quality issues and strategies: -

1. Customer Address: -

Data Quality issues: - When our team analysed this dataset, we found that the data set does not satisfy the **Validity** dimension because some states are mentioned in Full name while some states are mentioned in abstract form. Along with this, the country for every entry is repeating.

Strategies to mitigate the issues: - The dataset needs to have valid input and different rows should not have different input type values. Also, the country column should not be present. It is suggested that since every entry is of Australia, it will be better to write the country in the beginning of the data set and to not include a column for country.

1. Customer Demographic: -

Data Quality issues: - When our team analysed this dataset, we found that dataset does not satisfy the **Accuracy** dimension because one of the DOB’s given is of the year 1843, which means the person is175 years old. This is an error. The data also does not satisfy the **Completeness** dimension because several fields of the columns are not filled and are left empty. The Data is not **Relevant** as well because the values in the default column is not relevant to this dataset. The dataset also does not satisfy the **Validity** dimension because some data values indicated in the gender column are not written properly and are arbitrary and cannot be understood properly. Lastly, the dataset does not satisfy the **Currency** dimension because the new customers are not added in the demographic sheet.

Strategies to mitigate the issues:- The dataset needs to have accurate and errorless information which will not contradict the meta-data. Also, there should be proper metadata and relevant information about the data in the meta-data. The dataset should be updated with every required information and should not be left empty. The values in a column should be written in one format so that it is easier for the reader to understand the data. Also, new customers should be added in the demographic list and should not be left.

1. New Customer List :-

Data Quality Issues :- When our team analyzed this dataset, we found that dataset does not satisfy the **Currency** dimension because several fields for records are not filled and are left empty. Also, the dataset does not satisfy the **Validity** dimension because for the gender column, the data provided is arbitrary for some records and is not consistent. It is difficult to understand the meaning of “U” in the field.

Strategies to mitigate the issues :- The dataset need to be complete. This means that every record needs to have complete field values. Also, the filed values should not be arbitrary for any field and should be written in a consistent manner.

1. Transactions :-

Data Quality Issues :- When our team analyzed this dataset, we found that dataset does not satisfy the **Completeness** dimension because some fields of records are not filled and are left empty. Also, the dataset does not satisfy the **Currency** dimension because the transactions are given for year 2017 only.

Strategies to mitigate the issues:- The dataset needs to be complete. This means that every record needs to have complete field values. Also, the transactions should be present for years 2018 and after.

I hope that this analysis and suggestions will help your company in improving the quality of your company’s data. Once the data quality issues are fixed, my team then be ready for analysis in second phase.

Regards,

Nimit Sharma