KPMG

**Subject: Regarding the data quality issues**

Dear Client,

I have been through all the datasets. And here I have mentioned all the issues in the data. Also, I have mentioned how we can overcome the issues. Kindly check and let me know if you have any doubt.

**1) Sheet 1: Transactions TABLE**

|  |  |
| --- | --- |
| Sheet 1 : Transactions |  |
| 1) COMPLETENESS – | 1) Many product ids are zero which does not describe any product’s id which makes data incomplete.   |  | | --- | | Solution – We should delete all the rows having 0 as a product id to reduce redundant data. | |
| 2) ACCURACY – | 1) column online\_order should be either true or false. But some cells also contain null data.  2) brand, product\_line, product\_class, product\_size also has null data.   |  | | --- | | Solution – we should avoid taking null data as an input | |
| 3) CONSISTENCY – | 1) many transactions show approved status. But they do not have any product\_id, standard cost. Such transactions should not be approved. |
| 4)RELEVANCY – | 1) list\_data does not specify exactly what it is trying to convey. Also, it does not have any currency sign.   |  | | --- | | Solution – We should delete this column. | |
| 5) VALIDITY – | 1) product\_first\_sold\_date should have data type as a date instead of a whole number. |

**2) 2nd sheet : New Customer list**

|  |  |
| --- | --- |
| 2nd sheet |  |
| 1) COMPLETENESS – | 1) column last\_name has many null values.  2) DOB has null values.  3) Job\_title, job\_industry\_category has null values. |
| 2) ACCURACY – | 1) Gender column has 3rd value called as U. It clearly does not specify the actual meaning of U.   |  | | --- | | Solution – Should only accept categorical variable. We can change U into a new name which can be easily understandable. | |
| 3) CONSISTENCY – | New\_customer\_list sheet does not have customer\_id which is a primary key in other tables. New\_customer\_list does not have any primary key which can be shared as a foreign key.   |  | | --- | | Should include a primary or foreign key in this sheet. | |
| 4)RELEVANCY – | 1) **value** column does not show much relevancy to the data.   |  | | --- | | Solution – We should delete this column. | |
| 6) UNIQUENESS –  DUPLICATED | 1) column **Rank** has given a particular rank to many customers. Also, many in between ranks are missing. |

**3) 3rd Sheet : Customer Demographic TABLE**

|  |  |
| --- | --- |
| 3rd Sheet - |  |
| 1) COMPLETENESS – | 1) last\_name have null values.  2) job\_title has null values.  3)job\_industry\_category has n/a values.  4) tenure has null values. |
| 4)RELEVANCY – | 1) **default** column has redundant data which is not at all useful or should not give such injection of any unnecessary values.   |  | | --- | | Solution – We should delete this column. | |
| 5) VALIDITY – | 1) Gender column has multiple values. It should be categorical and should accept only categorical data. It has F, M,femal like categories.  2) DOB has type alphanumeric. But it should have date as a data type.  3) DOB has one value which shows that a customer has age 180 but practically it is not possible. So we should remove an outlier.   |  | | --- | | Solution – We should remove all the outliers in data preprocessing step. | |

4) 4th sheet which is Customer Address has high quality data and is free from data issues.

Regards,

KPMG Employee.