Dear Ma’am,

I am Gayatri Krishna from KPMG Data Analytics Team (Virtual Internship). I have reviewed the datasets that were shared to me. During the Data Quality Analysis, it was observed that the data set had a lot of errors.

Data Quality is very important. To draw valuable conclusions from the provided data, we need to remove these errors.

1. **Transaction:**

* The transaction data has 20000 observations and 13 colums. The data has 6 variables that are categorical in nature.
* Here we observe the values in the (product\_first\_sold\_date) column are not correct as it shows everything happening the same day at different times.
* The variable order\_status is a categorical variable (ie) Approved and Cancelled. Further this variable has no missing values. It is observed that most of the orders are "Approved".
* Further we have checked which all columns have missing values and also if there are any duplicate values.
* The organisation Sprocket Central Pty Ltd deals with 6 brands and the top selling brand was Solex. However, this variable has 197 missing values.
* Missing value columns:

|  |  |
| --- | --- |
| Field Name | Missing values |
| Online\_order | 360 |
| brand | 197 |
| product\_line | 197 |
| product\_class | 197 |
| product\_size | 197 |
| standard\_cost | 197 |
| product\_first\_sold\_date | 197 |

1. **New Customer List:**

* The NewCustomerList data has 1000 observations and 23 colums.The data has 6 variables that are categorical in nature. There are few columns that are unwanted, so we tend to drop them. Thus the dimension of the data becomes 1000, 18.
* Missing values exist in this data too but no duplicate values exist.
* Gender is a categorical variable. There are 3 categories for the same. Most of the new customers are females. There are no missing value for this column. However, the 3rd category "U" is quite misleading. So we will replace the same as "Unspecified".
* All these customers are from Australia and there are 3 states. Most of them are from NSW.
* Missing value columns:

|  |  |
| --- | --- |
| Field Name | Missing values |
| last\_name | 29 |
| DOB | 17 |
| job\_title | 106 |
| job\_industry\_category | 165 |

1. **Customer Demography:**

* The CustomerDemographic data has 4000 observations and 13 colums.
* we observe that the data has a column named 'default'. This particular column gives us no information, so we will drop that column. Thus, the dimension is reduced to 4000 observations and 12 colums.
* Next we see that gender is classified into 6 types. During the data entry time, certain categories are not correctly titled. The names in these categories are re-named. Thus, there are 2039 females, 1873 males and 88 unspecified customers. Further there are no missing values for this column.
* The entire data has few missing values but there are no duplicate values.
* Missing value columns:

|  |  |
| --- | --- |
| Field Name | Missing values |
| last\_name | 125 |
| DOB | 87 |
| job\_title | 506 |
| job\_industry\_category | 656 |
| tensure | 87 |

1. **Customer Address:**

* The CustomerAddress data has 3999 observations and 6 columns. The dataset has no missing values. All the columns appear to have consistent and correct information.
* There are 5 states in this data, however they are not correctly titled. VIC stands for Victoria and NSW stands for New South Wales. So we need to replace these 2 categories.

Regards

Gayatri Krishna