Dear Client,

This is Shantanu Jadhav from KPMG Data Analytics (Virtual Internship) team. We have reviewed the data sets which were provided by your company and during data quality analysis, we have found some errors in the data sets

First Of all, we would like to thanks you for providing us with datasets. From the datasets we could derive following insights.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Total No of Records** | **Unique Records** | **Receiving Date** |  |
| **Transaction** | **20002** | **3495 (Duplicates in Customer\_Id)** | **25/06/2020** |  |
| **Customer Demographic** | **4002** | **4002 (No duplicates)** | **25/06/2020** |  |
| **Customer Address** | **4001** | **4001** | **25/06/2020** |  |

The data quality analysis is the core phase and due to errors in the data set we suggest following mitigates in order to improve data quality.which will eventually help us to driven the better analytics,results for your company

The Follwing are the details of the existing errors inside dataset and their respective mitigation.

**Transaction -**

**.** Firstly we removed the unnessary header row present at the top.

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Errors** | **Mitigation** | **Data Quality Dimension** |
| **Online\_order** | **Blanks-360 , False Value- 9811** | **Removing Blanks and False values** | **Completeness** |
| **Order\_Status** | **Canceled Orders- 179** | **Removing canceled orders** | **Validity** |
| **Brands** | **Blanks- 197** | **Removing blanks** | **Completeness** |
| **Product\_First\_sold date** | **Blanks-197**  **Given in number fotmat** | **Removing Blanks**  **Changing it into Date Format** | **Completeness** |
| **Profit (New column)** |  | **Creating New column from Standard\_cost and List\_Price** | **Accuracy** |

**Customer Demographic -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Errors** | **Mitigation** | **Data Quality Dimension** |
| **Gender** | **Inconsistant values** | **Converting Female, Male into F,M for uniformity**  **Removing ‘U’ values** | **Consistency** |
| **Age (New column)** | **Creating new Age column from DOB column** |  | **Accuracy** |
| **Deceased** | **Deceased - Yes** | **Removing Yes columns** | **Currency** |
| **Default** | **Irrelevant Values** | **Deleting whole column** | **Relevancy** |

**Customer Address -**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Error** | **Mitigation** | **Data Quality Dimension** |
| **State** | **Inconsistent State Name** | **Converting New south wales - NSW**  **Victoria- VIC** | **Consistency** |
|  |  |  |  |

This summarises all data quality issues in the first stage of the Analysis. Mitigation are suggested to improve future data for better analysis.

Let us know if you have any question regarding mitigation and data quality issues identified.

Regards,

Shantanu Jadhav