Dear sir,

This is Satyam Rout from Data-Analytic team of KGMP(Virtual Internship). We have reviewed the following data-sets which are provided by your company and found some of the erroneous data which may effect the final model.

The data quality analysis is the vital phase and due to the errors in the data-sets, we suggest u the following mitigates in order to improve the data quality in order to better data analysis.

The below tables shows the characteristic of each data-sets with the improvements to be made in it.(please let us know for any alignment issues).

1. **Transaction:**

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Initial Values | Errors | Corrected values |
| Online\_order | 20,000 | FALSE(9811) and Blanks(360) | 9732 |
| Online status | 20,000 | Canceled(179) | 9732 |
| Brand, Product\_line, Product\_size, Product\_class | 20,000 | Blanks(197) | 9732 |
| Profit | Null | Null | 9732 - Added the value(list\_price - standard\_cost) |

1. **CostumerDemographic:**

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Initial Values | Errors | Corrected values |
| Last\_name | 4000 | Blanks(125) | 3912 |
| gender | 4000 | Typo errors | 3912 |
| DOB | 4000 | Blanks(87) | 3912 |
| Age | Null | Age > 100 | 3912 - Added the value(DOB - Current\_date) |
| Job\_title, Default | 4000 | unwanted | removed |

1. **CostumerAddress**

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | Initial Values | Errors | Corrected values |
| State | 3999 | Victoria(VIC),  New South Wales(NSW) | 3999 |

As shown in above observations the errors such as Blanks, Typo errors, age>100, deceased people, offline orders needed to be removed from the data-set in order to analyse it properly for model creation.

Regards

KGMP (Data-analytic team)