Dear Customer,

I have analysed every single table and this is what I have found.

The data in the transaction table is quite accurate, although in some fields like product size, product class, list price around 200 values are missing but this is a bit from 20000 records so we can ignore them. I have also checked the class balance of many fields like online order which is as equal as other classes. The data is just for 2017 year so if we can get other year data it would be helpful. I found some class imbalance in product class as medium products are 13826 and other products are around 3000 so might this going to create some keypoint to our accuracy and most importantly there is many extra fields or fields which not going to have any effect in model prediction like transaction id, product id etc

In the Customer Address table , what I found is that this data is more specific to some places like our customers is just based on Australia, only from 5 states and having 873 different postcodes from 40000 records .

In Customer Demographic table there is many things which need to be clean before used to for prediction like in gender fields there are 6 different ways to write make or female but we need it in just binary order, there is also many missing value into different fields like first name,DOB, job_title which is around 3300 from 4000. And there are also a few fields which are extras like they are not going to help us to find any inside.

The NewCustomerList table is having similar issues like the customer demographic table. It also having some duplicates and missing value and there is also some extra fields

Best Regards