Hello sir,

This is Nithish Kumar. Analyzing and working through the 4 datasets, it is highly conspicuous that there are many week points in the datasets which could make the model development and interpretation to an unintended direction. I have made a detailed analysis report below, regarding the quality of the data and also have made the possible mitigation to the dataset in practical.

All datasets provided are handled with acute precision individually to make it highly noiseless.

TRANSACTION

- ➔ Firstly, the column 'transaction_id' has no use in future because, there is no other data source provided to correlate with this dataset.
- ➔ Next, taking a look at null values, it is obvious that all columns are having the same number of null values except 'online_order'.
    - ❖ So, it is decided to remove all voids which are in same numbers and the column 'online_order' is going to have a value of 'unknown', which is going to be conservative, reducing the possibility of losing more data.
- ➔ The column 'birth_date' is having an inappropriate datatype. So, it is first changed into proper datetime datatype and later the dates are changed to age of a person, which is more meaningful.
- ➔ Lastly, in columns 'standard_cost' and 'brand', some string replacements and formatting are to be done in order to correct them with appropriate datatypes.

NEW CUSTOMER LIST & CUSTOMER DEMOGRAPHIC

- ➔ Initially, 'first_name' and 'last_name' are merged to a single column 'full_name'.
- ➔ Removing unknown attributes completely and getting rid of null values.
- ➔ Converting the column 'birth_date' into age.
- ➔ The column 'gender' had its text formatted differently, which is reformatted to correct form.
- ➔ Removing columns which are having the same values trough the table.
- ➔ Finally, renaming and correcting the datatypes certain columns.

CUSTOMER ADDRESS

- ➔ The column 'state' had some formatted text within them, which are reformatted to their shortened form.
- ➔ The column country is been removed completely because, they had the same values all over the table.
- ➔ Null values has been eliminated at the last.

The above are the brief conclusions drawn from analyzing the quality of the datasets.

SIMPLE MITIGATIONS

- ➔ It is better that the unique columns like transaction_id, customer_id, etc. are having some meaningful value which are alphanumeric.
- ➔ Avoiding irrelevant formatting of strings in various object columns are necessary.
- ➔ While initiating the database itself it is better to name the attributes with more meaning, instead of the names which are not familiar.

With Regards,

Nithish Kumar S