Hello

Thank you for providing us with the datasets. The summary table below highlights key quality issues that we discovered.



Below are more in-depth descriptions of data quality issues discovered, with some comments:

1. Accuracy:

-CustomerDemographic:

* Correct DOB, or can be excluded from any age analysis.

2. Completeness:

- CustomerDemographic:

* Missing last names for 4 customers.
* Add missing DOB and tenures for 87 customers -about 2% of dataset- or they can get filtered out.
* 506 missing job\_title about 12.65% of the dataset can result in skewness in any job\_title analysis so we recommend adding them.
* 656 job\_industry\_category set to n/a about 16.4% of the dataset can result in skewness in any job\_industry\_category analysis so we recommend adding them.

- NewCustomerList:

* Add 29 missing last names.
* Add gender and DOB got 17 missing records -about 1.7% of dataset- or they can get filtered out.
* 'job\_title' got 106 missing records -about 10.6% of the dataset- or they can get filtered out.
* 'job\_industry\_category' got 22 n/a -about 2.2% of the dataset- or they can get filtered out.

-Transactions:

* Add 360 missing records in 'online\_order' -about 1.8% of the dataset- or they can be excluded from any order type analysis.
* 197 missing records 'brand','product\_line', 'product\_class', 'product\_size', 'product\_first\_sold\_date', 'standard\_cost' (they are the same) – about 1% of the dataset -.

3. Consistency:

- CustomerAddress:

* Inconsistency in the 'state' names, sometimes full name others a 3 Letter word, we recommend to use the 3-letter word.

-CustomerDemographic:

* Inconsistency in the 'gender' sometime full word sometimes one Letter "F" or "M", we recommend to use the full word, also one missing last letter 'e' in female (ID 54).

- NewCustomerList:

* 5 customers with property\_valuation got a decimal space; we recommend this to get rounded to 0 decimal.
* 'Value' inconsistent decimals, we recommend to round to 1 decimal place.

- Transactions:

* Inconsistent decimals in 'list\_price' and 'standard\_cost', we recommend no changes.

4. Currency:

- CustomerDemographic:

* 2 customers with deceased\_indicator set to Y (ID 753& 3790); we recommend that they get filtered out.

5. Relevancy:

-CustomerAddress:

* -'property\_valuation' is unclear, we would like more explanation on it.

-CustomerDemographic:

* We are unable to find any meaning for the 'default' column, we recommend to delete it.

-NewCustomerList:

* 'Rank' &'property\_valuation' are unclear, we would like more explanation on it.

-Transactions:

* 'product\_first\_sold\_date' it not in proper date format, also data does not seem to follow any other usual date formats.

6. Validity:

-CustomerAddress:

* Some 'address' has an extra zero (362 address) or a double zero at the beginning of the address (37 address). We recommend to remove the extra zeros.

-CustomerDemographic:

* Only one customer first name has a dash (-) (ID 5), another has a first name and 2nd name (ID 1174), we recommend to fix those.

-NewCustomerList:

* 5 address start with double zero or more and 995 start with a zero, we recommend to remove the extra zeros.

This summarizes our findings, our analysis was done on an independent sheet level, we have not looked at the relationship between them.

Please let us know if you have any queries surrounding the issues presented.

Best Regards

Shehab