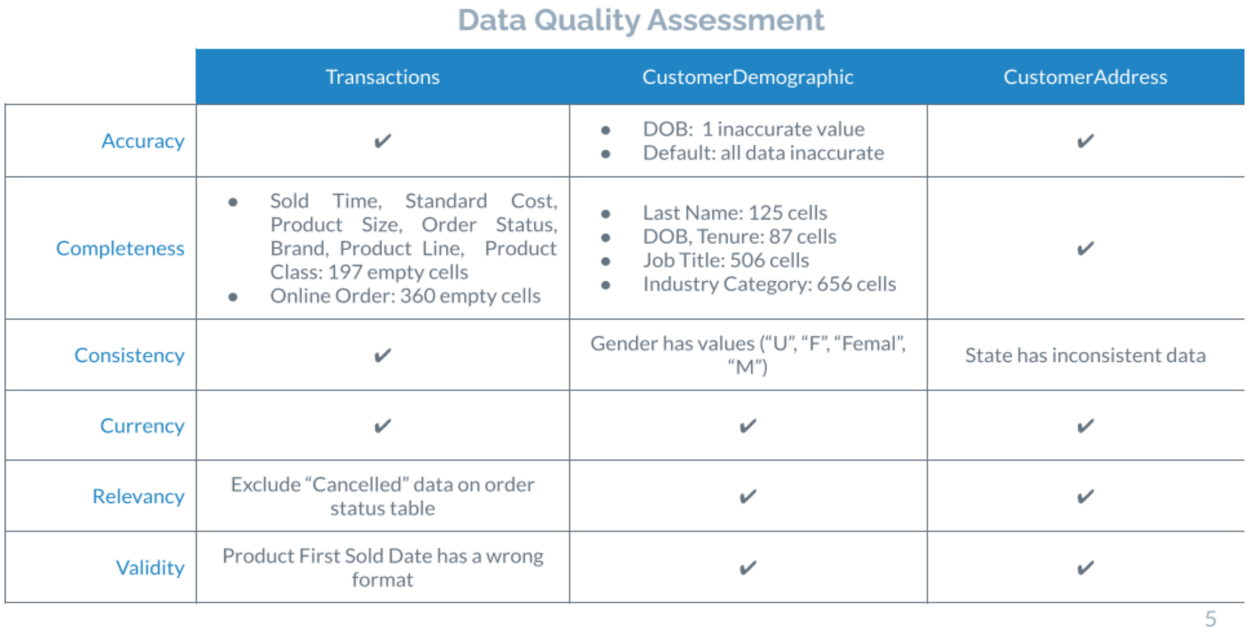
Dear John Doe,

Thank you for choosing us as your partner, the three datasets that you sent were amazing and insightful, especially for me. After reviewing your datasets, this table below is the highlight of your dataset quality that has been assessed. Please, let us know if there is another issue(s) or any following question.



Data Quality Assessment

	Transactions	CustomerDemographic	CustomerAddress
Accuracy	✔	• DOB: 1 inaccurate value • Default: all data inaccurate	✔
Completeness	• Sold Time, Standard Cost, Product Size, Order Status, Brand, Product Line, Product Class: 197 empty cells • Online Order: 360 empty cells	• Last Name: 125 cells • DOB, Tenure: 87 cells • Job Title: 506 cells • Industry Category: 656 cells	✔
Consistency	✔	Gender has values (“U”, “F”, “Femal”, “M”)	State has inconsistent data
Currency	✔	✔	✔
Relevancy	Exclude “Cancelled” data on order status table	✔	✔
Validity	Product First Sold Date has a wrong format	✔	✔

5

Explanations have been written below for more comprehensive understanding. Recommendations also provided to avoid reoccurence data assessment, increase the accuracy of data, and also ease of data visualization .

- **Inaccuracy of data (i.e. Default on CostumerDemograhic)**
  Accuracy means the data has a correct value based on their column, for example, if an age column should have 32, it must be written 32, if it was written 34 the data was wrong/inaccurate. On your dataset, there are 2 columns which had inaccurate values. First, Date of Birth (DOB), it has a value of a person aged 170+ years old. Then, the Default column was abstract, we can’t get a clear insight from that.
  Mitigation: Drop out the DOB’s outlier and clear all values of default column
  Recommendation: Create new columns named age and age_group which can be more helpful for visualizing.
- **Lot of missing values**
  2 out of 3 datasets contain null values (incomplete) almost in every columns, and some of them have a big number of null values.
  Mitigation: Let it be if it just a small number, use data centering (mean, modus, median) for numerical values.
- **Data contain inconsistency of values (i.e. Gender has values M, F, Femal)**
  As we all know, gender is only known for Female and Male, but in CustomerDemographic there are M, F, and Female. Also at CustomerAddress,

we know NSW stands for New South Wales and VIC stands for Victoria. If we divide those values, the data become inconsistent.
Mitigation: We can change New South Wales to NSW, and also with Victoria.
Recommendation: Using data validation on Ms. Excel or Spreadsheet to avoid human error when entering data manually.

- **Invalid data format**
  On Transactions dataset, there’s a column had a wrong data format, it is product_first_sold_date.
  Mitigation: change the format from numeric to datetime.
  Recommendation: same as State column, we can use data validation on spreadsheet or Ms. Excel

Moving forward, we will analysis your dataset, so we can get insight from the datasets provided, and it can be used for future decision making. Once again, thank you for having us as your partner. It is nice to work with you.

Best Regards,
Azriel Akbar Al Fajri