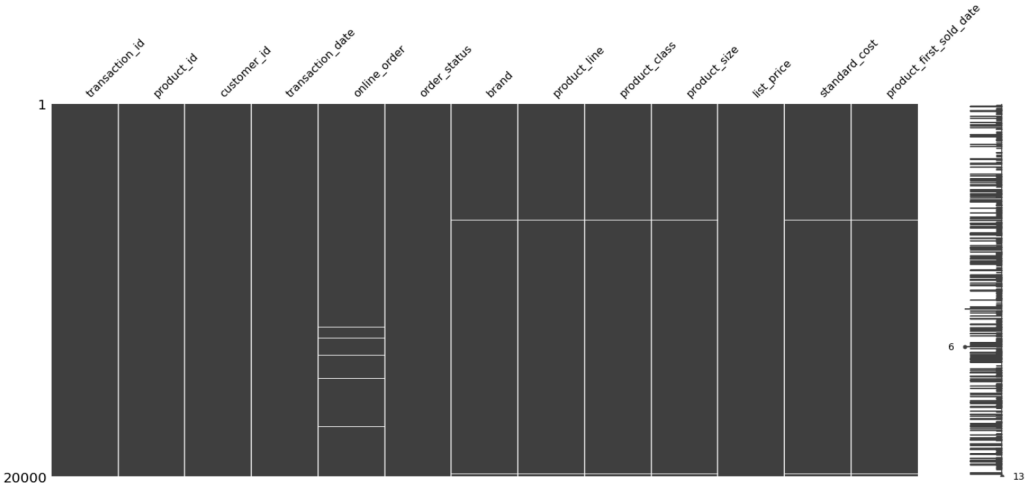
Muhoza U Bizumuremyi
Address: 400 Running Water Trl, Fort Worth, TX
Phone: (469) 639-5382
Email: writetoursus@gmail.com
Linkedin: linkedin.com/in/muhozaursus/

Dear Customer,

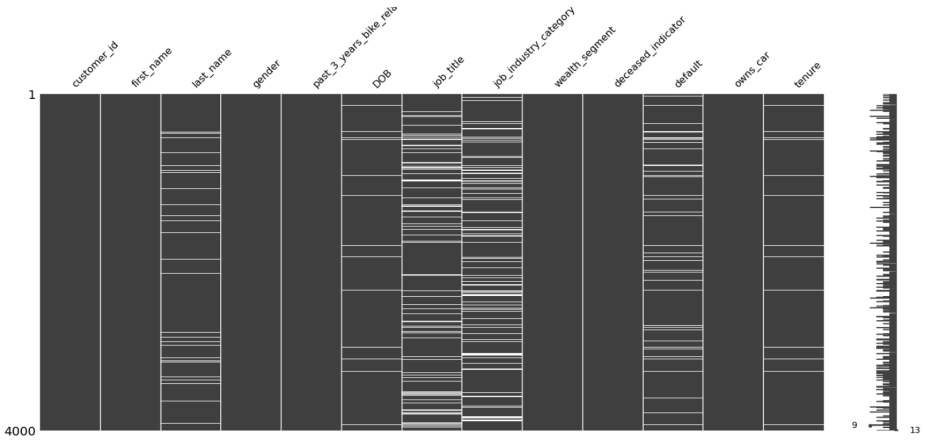
Based on the industry metrics (Accuracy, Completeness, Consistency, Currency, Relevancy, Validity, and Uniqueness), we have assessed the datasets you've sent us and found some problems. We also gave some recommendations on how to solve those problems.
Please review this document and give us feedback if we should continue to the next step.

DataSets	Total Missing Values	Columns with Missing Values	Total rows with missing values
Transactions	1542	online _order, brand, product_line, product_class, product_size, standard_cost, product_first_sold_date	555
Customer Demographic	1763	Last_name, DOB, job_title, job_industry_category, default, tenure	1370
Customer Addresses	0	None	0

Observation 1:
Transactions dataset's columns with missing values (except online_order) have missing values that are actually on the same rows (check from row 97 up to row 19871). The horizontal lines on the picture below show where the missing values are.

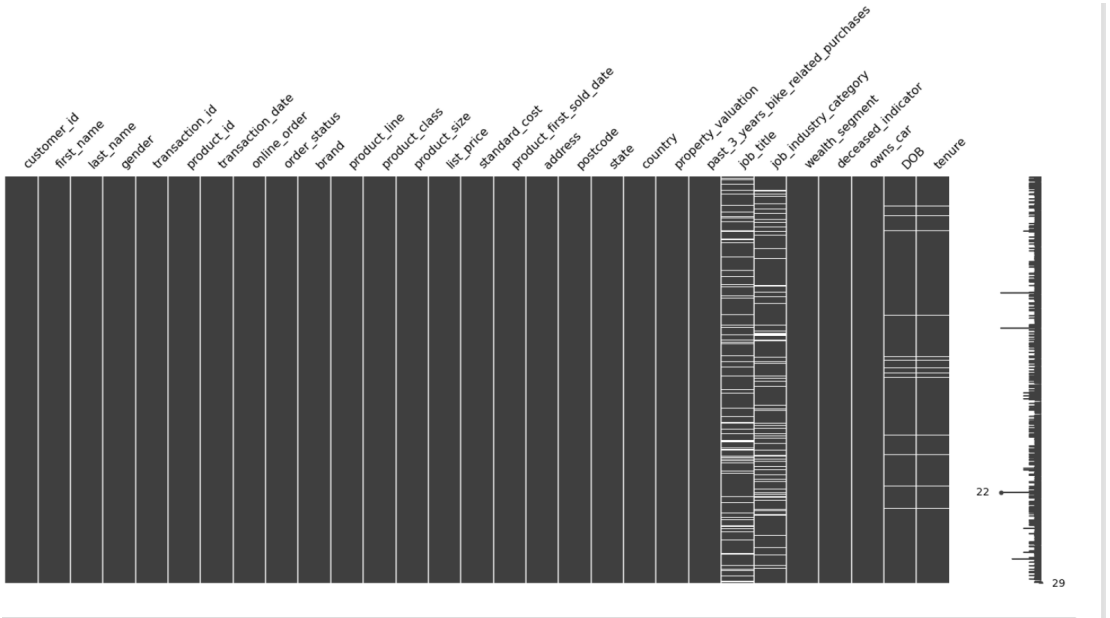


In the customer demographic dataset, we have even more missing values, however, according to the picture below we see no clear connection between missing values of different columns.



Recommendations for dealing with missing values in the transactions dataset:

- In the email sent to me, transactions data that was needed is for the last 3 months, since the last transaction date was on 2017-12-30, the data we need will start from 2017-09-30. This will leave us with 383 total missing values from 1542. Total rows with missing values also decreased up to 143 rows from 555.
- We can delete any rows with missing values since it won't affect highly the number of rows that are needed to continue(from 5128 rows to 4985 rows).
- We can merge the 3-month transaction dataset with the demographic and customer address dataset having up to 31 columns.
- We deleted the default column because it contained unknown characters
- Replace missing values in the last_name column with 'U' which means 'unknown'
- By looking at the DOB and Tenure columns in the picture below we can see how missing values are on the same rows, because they are not too many deleting those missing values won't cause a big problem



Looking at the picture above we see columns like job_title and job_industry_category have a lot of missing values, this might be because when collecting data, it's not mandatory to fill in about your job and that in the job_title column people filled in whatever they wanted.

To better collect data about these two columns, a person has to choose from his/her job title to a limited list for example: 'Accountant', 'Engineer', 'Manager',...

- For the moment, we suggest deleting those two columns for a better analysis.
- We also suggest not to use the country column since all these transactions are from Australia (Eastern Australia to be specific) remaining with 26 columns.

Observation 2:

For the gender column, there are 4 values collected: *Male, Female, F, Femal, and M,* there should be only 2 values Male and Female or M and F

For the state column, there are 5 values collected there should be only 3 values; NSW, VIC, and QLD

Observation 3:

For the column product_first_sold_date, this column has serial excel date format for a better analysis we will have to convert these dates into normal dates

Finally, we have highlighted problems found in the datasets such as missing values, unknown characters, untrue characters,... and we have given suggestions on how we can resolve these issues in order to continue to the next step, however, there is a possibility that also you can look back into the database and correct them or even change the methodology used to collect data.

Thank You
Muhoza Bizumuremyi