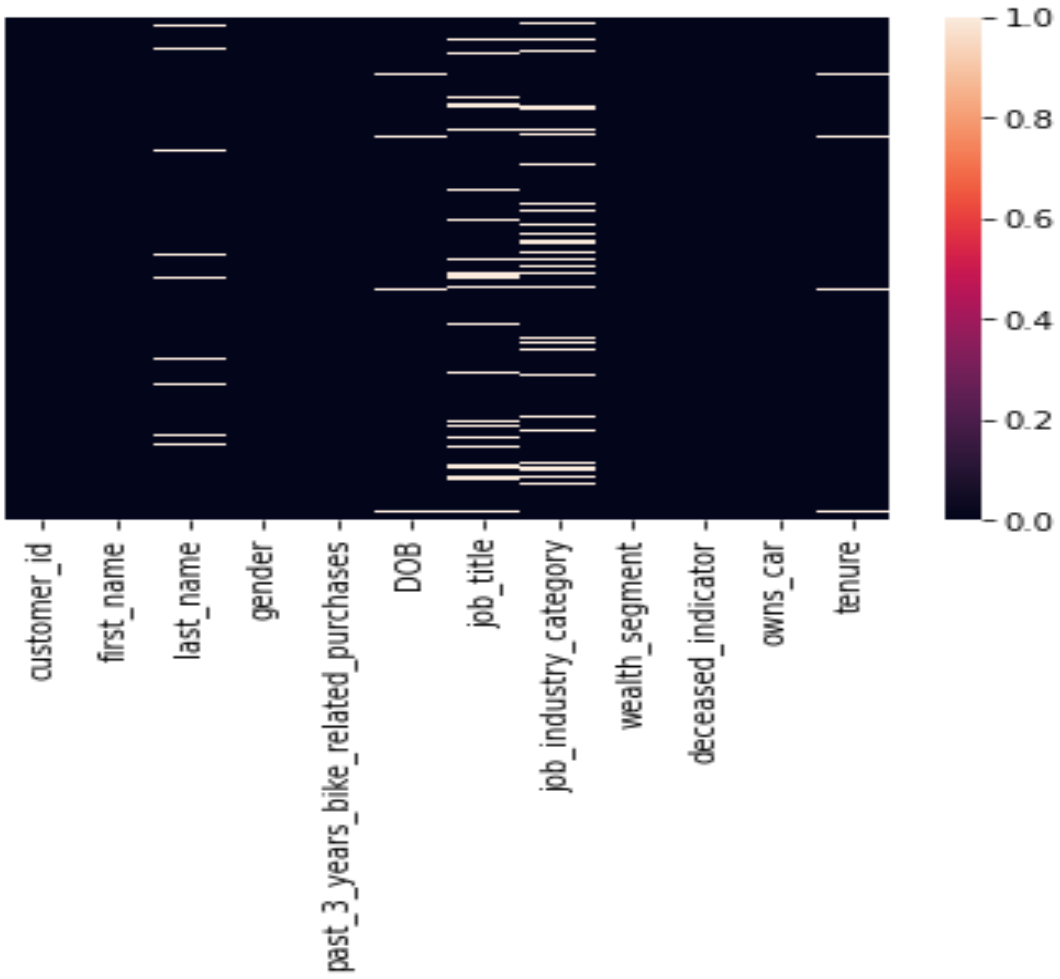
Sprocket Central Pty Ltd

DATA QUALITY AND RECOMMENDATIONS REPORT

CUSTOMER DEMOGRAPHICS TABLE

Completeness of the data

The data in this table is not complete as some of the data fields do not contain values as visualized below last_name, DOB, job_title, job_industry_category and tenure are the columns that contain fields with null values.



Consistency of the data

Secondly, the values in the gender column are not consistent since FEMAL, FEMALE AND F point to the same gender and the same case for MALE and M which point to the same thing.

Relevance of the data

Finally, the default column contains meta-data which may not be understood to be of importance in our current task.

RECOMMENDATIONS

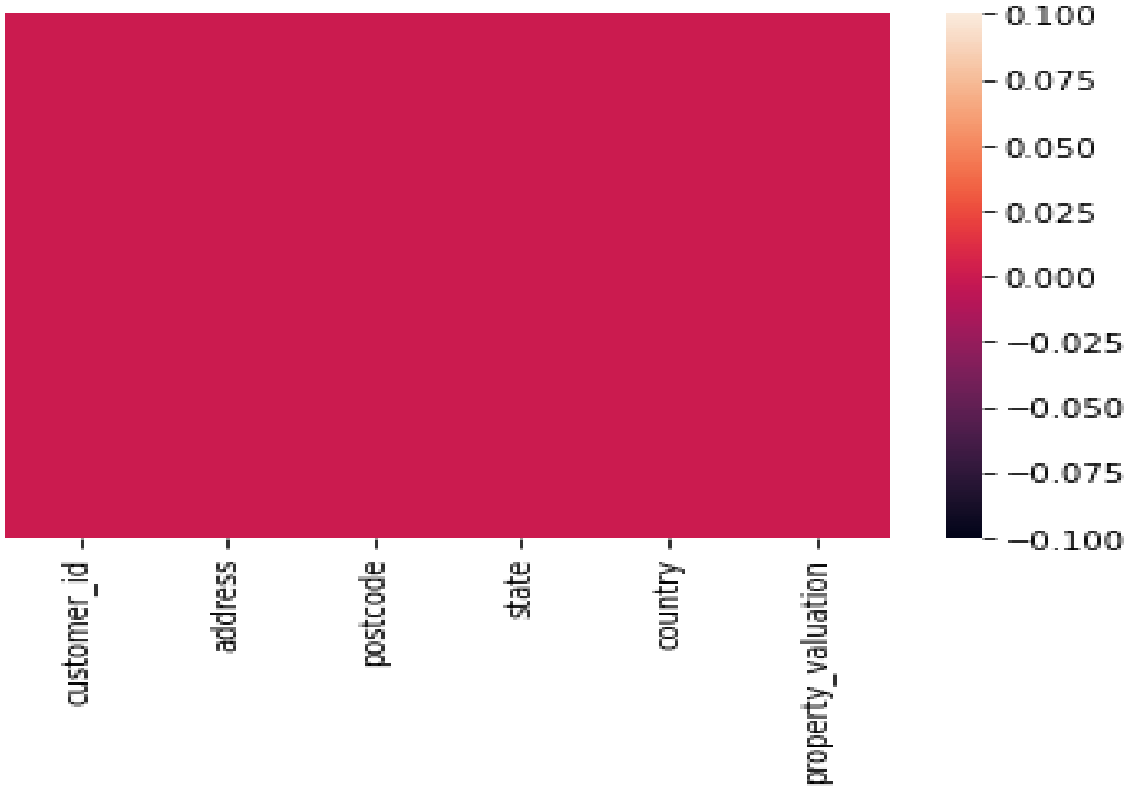
1. Drop all nulls in the table since the missing data is less than 1 percentile of the total data.
2. Use of abbreviations in the gender column. M to represent MALE and F to represent FEMALE and also retain the 'U' gender to represent those who don't identify as male or female.
3. Drop the whole default column since the data may not have the information we need.
4. The last_name column will be maintained as is since the first_name column is complete with no nulls.

CUSTOMER ADDRESS TABLE

Completeness of the data

The data in the table passes the test of completeness since all data fields have values.

Below is a plot of table columns against nulls



Consistency of the data

The data in the state column is not consistent since VIC and Victoria point at the same place also New South Wales and NSW

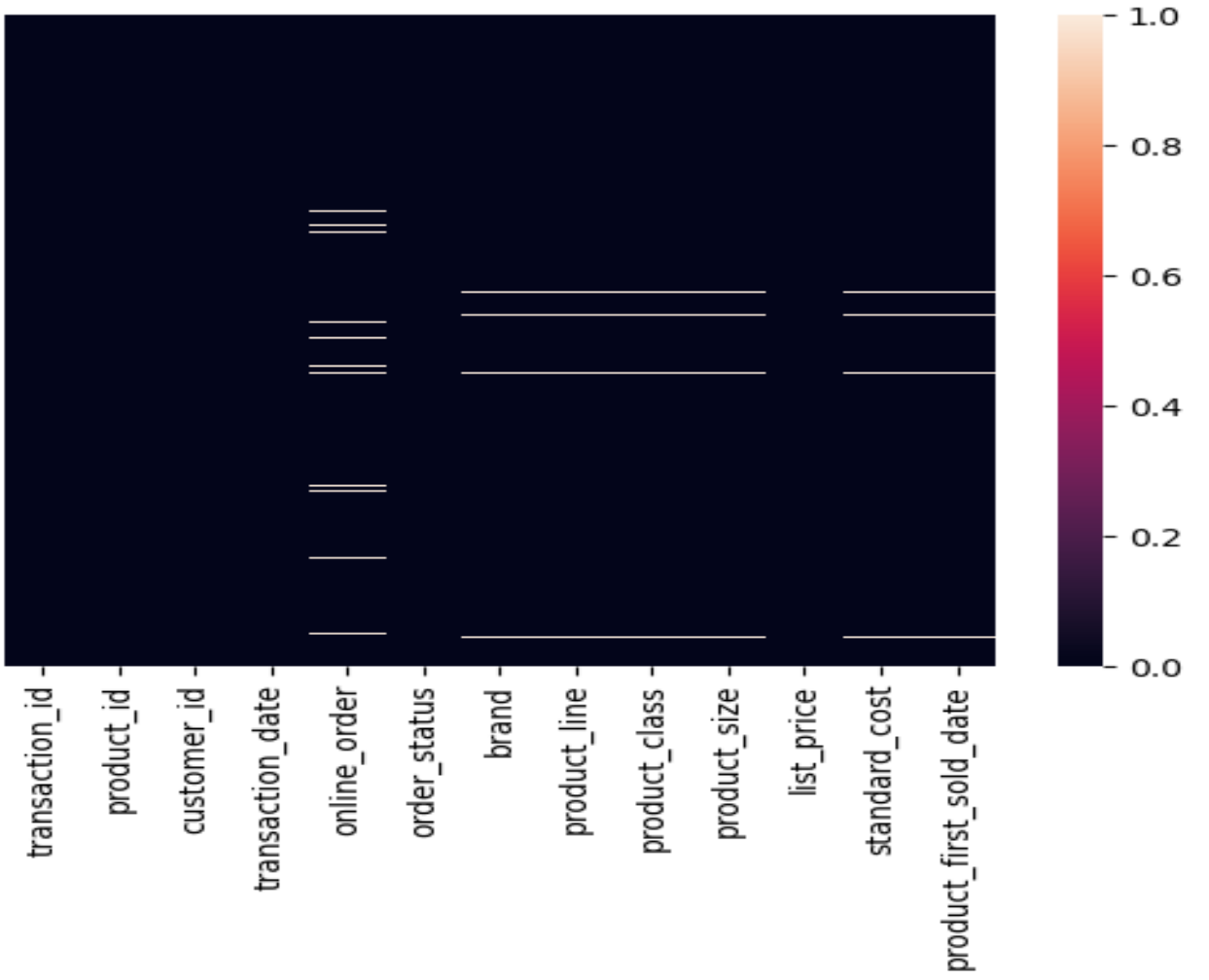
Recommendation

Use abbreviation for the state column to ensure uniformity, NSW for New South Wales and VIC for VICTORIA

TRANSACTIONS TABLE

Completeness of the data

The data in this table is not complete, it contains nulls in the following columns online_order, brand, product_line, product_class and product_size as visualized below



VALIDITY OF THE DATA

The data contained in the product_first_sold_date is not valid since its entered as a general number and if converted to date contains dates that are not within the dates of transactions.

There are additional customer_id in the table only customer_id present in customer Demographics will be used.

Recommendations

1. Drop all nulls in the columns with null values.
2. Drop the product_first_sold_date column.
3. For relational purposes all customer_id that are not shared with the Customer demographic table will not be used.