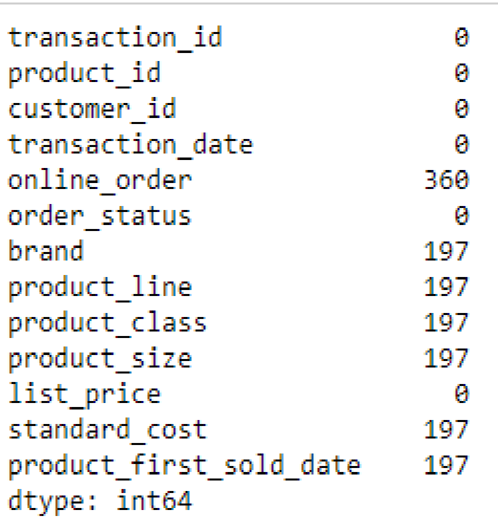
Dear Sprocket Central Pty Ltd,

Thank you for providing KPMG with the three datasets. We have identified the data quality issues associated with these datasets and have made recommendation for each issue to mitigate it going forward.

Summary Statistics:

Dataset	Number of Records	Distinct Customer ID
Transactions	20000	3494
Customer Demographics	4000	4000
Customer Address	3999	3999

The issues related to Transaction dataset and our recommendations:



```
transaction_id              0
product_id                  0
customer_id                 0
transaction_date            0
online_order              360
order_status                0
brand                     197
product_line              197
product_class             197
product_size              197
list_price                  0
standard_cost             197
product_first_sold_date   197
dtype: int64
```

Completeness: The pictures shows the number of missing values in each of the columns from the transaction dataset. Here we can delete most of the missing values as they are irrelevant for further analysis like online_order, brand, product_line, product_class and product_size. But standar_cost is needed in calculating product for that particular product and product_first_sold_date also is needed. For these two the recommendation would be to impute the data.

Accuracy: Some important correct values are missing in the dataset like profit. We can create a new column profit from list_price and standard_cost.

Relevancy: The order_status column is not relevant at all. It should be removed. Doing analysis for cancelled orders doesn't make sense.

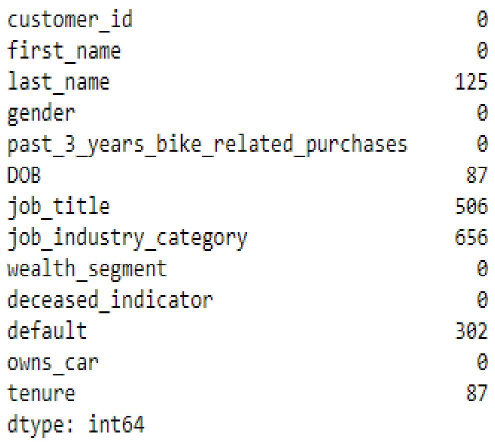
Validity: Some columns does not have proper format. Like the product_sold_date and the list_price are not in proper date and price format. We need to change them for further analysis.

The issues related to Customer Address dataset and our recommendations:

Completeness: The customer_id column from all three datasets don't match, meaning there are missing. We can remove those customer_id which are extra as some information about them is missing and they might not be useful for further analysis.

Consistency: The state column is inconsistent. For example, New South Wales and NSW, both represent the same. We can use VLOOKUP or find and replace function to overcome this and make it consistent.

The issues related to Customer Demographic dataset and our recommendations:



```
customer_id                            0
first_name                             0
last_name                            125
gender                                 0
past_3_years_bike_related_purchases    0
DOB                                   87
job_title                            506
job_industry_category                656
wealth_segment                         0
deceased_indicator                     0
default                              302
owns_car                               0
tenure                                87
dtype: int64
```

Accuracy: The DOB column has some incorrect values. They should be cross checked and rectified or should be removed. Age column is missing from the dataset. Age is an important feature for any demographic. So, we can create age column using current date and their Date of birth.

Completeness: As you can see from the picture, the dataset has missing values. We can remove missing values as they are not needed for further analysis from this dataset.

Consistency: The gender column from this dataset is inconsistent. For example, Female and F. We can use VLOOKUP to overcome this issue.

Relevancy: The deceased column is irrelevant as no need of doing analysis on deceased people. So deceased people can be removed.

The default column does not make any sense. It can directly be removed for our analysis.

Let me know if you have any questions. Thanks for coming to KPMG.

Kind Regards,
KPMG Analytics, Information & Modelling team