Dear Client,

Thank you very much for providing us with your valuable data. I just did my preliminary analysis at the dataset and found that the data is relatively complete. Here are a summary and a few concerns about the information that I have seen. I have also mentioned some mitigation strategies for the same.

Transactions Dataset

	Non-Null Rows	Null Rows
transaction_id	20000	0
product_id	20000	0
customer_id	20000	0
transaction_date	20000	0
online_order	19640	360
order_status	20000	0
brand	19803	197
product_line	19803	197
product_class	19803	197
product_size	19803	197
list_price	20000	0
standard_cost	19803	197
product_first_sold_date	19803	197

- There are 20000 rows in the dataset. The total number of missing data is less than 2%. To address the issue, with this data, any row with missing data will be removed from the overall analysis.
- The transaction_date and produc_first_sold_date are not correctly formatted, so they are to be appropriately formatted into the proper data type.

Customer Address Dataset

	Non-Null Rows	Null Rows
customer_id	3999	0
address	3999	0
postcode	3999	0
state	3999	0
country	3999	0
property_valuation	3999	0

- Inconsistency in classification column **state.** This problem can be fixed by changing **New South Wale** as data point consistency to **NSW**

Customer Demographics Dataset

	Non-Null Rows	Null Rows
customer_id	4000	0
first_name	4000	0
last_name	3875	125
gender	4000	0
past_3_years_bike_related_purchases	4000	0
DOB	3913	87
job_title	3494	506
job_industry_category	3344	656
wealth_segment	4000	0
deceased_indicator	4000	0
default	3698	302
owns_car	4000	0
tenure	3913	87

- There is inconsistency in data; in the **gender** column, there may character strings for representing Male and Female such as M, Male, U, F, Femal, Female.
- In the **default** column, there are many data type such as Boolean, String, Integer etc. There are 4000 columns in the dataset and default does not appear to have any impact on our analytics. So the default column can be dropped.
- There is missing data in the last_name column, but as it will not have a significant impact on our process, we can remove it.
- The column DOB is not formatted correctly. And there are also specific dates which don't fit such as a customer's birth year is 1843 which is not at all correct. So I think we can drop DOB as well.
- The column job category has more than 10% missing data, and as it is an essential factor for analysis, we cannot drop it, So we can get job category from the job title.
- The job title has more 10 % of the data missing.

New Customer Dataset

	Non-Null Rows	Null Rows
first_name	1000	0
last_name	971	29
gender	1000	0
past_3_years_bike_related_purchases	1000	0
DOB	983	17
job_title	894	106
job_industry_category	835	165
wealth_segment	1000	0
deceased_indicator	1000	0
owns_car	1000	0
tenure	1000	0
address	1000	0
postcode	1000	0
state	1000	0
country	1000	0
property_valuation	1000	0
Unnamed: 16	1000	0
Unnamed: 17	1000	0
Unnamed: 18	1000	0
Unnamed: 19	1000	0
Unnamed: 20	1000	0
Rank	1000	0
Value	1000	0

- The scale is ambiguous for several columns including property_valuation, rank, wealth and value columns.
- There are five unnamed columns in data, so are they essential or they can be dropped ??

This my analysis till now feel free to reach me if any problem or questions.

Thanks

Vishvam Bhatt