Dear Client,

We would like to thank you for trusting us and sharing the datasets with us for evaluation. We have spent some time understanding the datasets and came across a few issues which can affect the decisions made based on this data. Please find below the overview of the datasets along with their issues.

1. Incorrect Datatypes- There are several columns with wrong datatypes like past\_3\_years\_bike\_related\_purchases should be a numerical value, but the dataset has this value as text.
2. Incomplete data- There are several columns that have null values, in order to remove the null values, there can be huge data loss.
3. Age Validation- The dataset has a customer with an age value of 123 years which seems wrong.
4. Different names for the same category- Like Gender has F, Female, and Female which means the same, it can affect the analysis results.
5. Irrelevant data- The dataset has data that cannot be used because it is not in an understandable format like the column default in CustumerDemographic dataset.

There are various ways, how can this data be cleaned and how can we avoid these issues.

Mitigation strategies

1. Having a few data validation rules like a drop-down menu for gender and for DOB, the age to be greater than a certain date to avoid human errors.
2. Giving the option of ‘Others’, if customers do not belong to any of the mentioned categories like for job title or job industry.

Data Cleaning Process

1. Correct the datatypes for all the columns.
2. Remove the columns which are not relevant to our analysis, which is also called feature engineering.
3. Correcting the category names to one like changing F and Female to Female.
4. Filtering the blanks if few in a column and less data loss.
5. If there are a lot of blanks in a column, replace it with ‘Not Available’, so we do not lose the other column’s relevant data.

I hope these recommendations will help you in improving the data quality and integrity.

Please feel free to reach out to us in case of any queries.

Regards

Taslim Firdouse

Gender- U

Past\_3\_Years\_bike\_related\_purchases- formatted as text

Job title- Blanks

Online Order- Blanks

Brand- Instead of leaving it blank, can be substituted by others

Gender in demographics- either F or Female