Part 1: Dataset Generation

The dataset applied for this assessment is generated from an online data generator named Mockaroo (mockaroo.com). There are 1000 rows generated for this dataset, which is the maximum number of rows able to be generated on this website without a paid subscription.

While designing the data, the dataset “E-commerce Customer Behavior Dataset” by Laksika Tharmalingam on the Kaggle website is referred to. The designation of data is as follows:

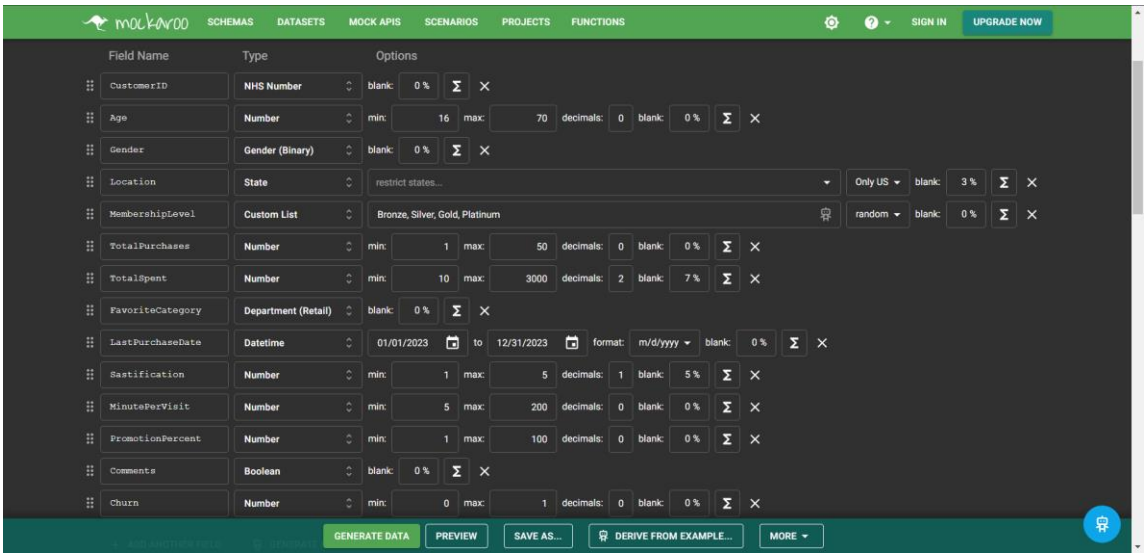

Figure 1: Dataset designation.

Besides the listed columns in the question, there are four columns created in this dataset, making the total column numbers to 14. The additional columns and their description are listed as follows:

Satisfaction: The average satisfaction of the customer for every purchase. The range is from 1 (Very unsatisfied) to 5 (Satisfied)
MinutePerVisit: The time stay on the website in minutes per visit.
PromotionPercent: The proportion of items purchased during promotion.
Comments: Whether the customer has given comments after purchase before.

Noted that the columns Location, TotalSpent and Satisfaction (after renaming) are designated to have missing values for 3 to 7% of the rows.

Dataset reference:
https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset

Dataset generated:
https://drive.google.com/file/d/1Q9DjruSQt2Ynp6NWz7ldTpDmym6oVMYJ/view?usp=sharing

Before entering the data, I separate a few columns from the dataset manually into a reduced dataset. The list of columns for the dataset that needed to load into the Talend Data Integration is as follows:

D7005AA1_DataA	D7005AA1_DataB
CustomerID Age Gender Location	CustomerID MembershipLevel TotalPurchases TotalSpent FavouriteCategory LastPurchaseData Satisfaction MinutePerVisit PromotionPercent Comments Churn