

## 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:

Field Name	Type	Options
CustomerID	NHS Number	blank: 0% [Σ] [X]
Age	Number	min: 16 max: 70 decimals: 0 blank: 0% [Σ] [X]
Gender	Gender (Binary)	blank: 0% [Σ] [X]
Location	State	restrict states... Only US blank: 3% [Σ] [X]
MembershipLevel	Custom List	Bronze, Silver, Gold, Platinum random blank: 0% [Σ] [X]
TotalPurchases	Number	min: 1 max: 50 decimals: 0 blank: 0% [Σ] [X]
TotalSpent	Number	min: 10 max: 3000 decimals: 2 blank: 7% [Σ] [X]
FavoriteCategory	Department (Retail)	blank: 0% [Σ] [X]
LastPurchaseDate	Datetime	01/01/2023 to 12/31/2023 format: m/d/yyyy blank: 0% [Σ] [X]
Satisfaction	Number	min: 1 max: 5 decimals: 1 blank: 5% [Σ] [X]
MinutePerVisit	Number	min: 5 max: 200 decimals: 0 blank: 0% [Σ] [X]
PromotionPercent	Number	min: 1 max: 100 decimals: 0 blank: 0% [Σ] [X]
Comments	Boolean	blank: 0% [Σ] [X]
Churn	Number	min: 0 max: 1 decimals: 0 blank: 0% [Σ] [X]

Buttons: GENERATE DATA, PREVIEW, SAVE AS..., DERIVE FROM EXAMPLE..., MORE

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/1Q9DjruSQ2Ynp6NWz7ldTpDmym6oVMYJ/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