# Dataset Description: eda\_sample\_customers.csv

This is a synthetic customer dataset created for practicing Exploratory Data Analysis (EDA), feature engineering, and machine learning (ANN, regression/classification).

It simulates a subscription-based company (like telecom, SaaS, or online retail) and contains messy, real-world–like data with missing values, outliers, rare categories, and text notes.

**What the dataset is about**

Each row = **a customer profile**  
Each column = **a customer attribute or outcome**

**Key columns:**

* **customer\_id** → unique ID (shouldn’t be used as a feature).
* **signup\_date** → when the customer joined. Useful for tenure features.
* **age** → customer’s age (has missing values + unrealistic outliers like 5, 120).
* **income** → annual income, skewed (log-normal), with some extreme outliers.
* **gender** → categorical, includes missing values.
* **city** → customer’s location (categorical, with some rare categories).
* **has\_car** → Yes/No/Unknown (inconsistent category).
* **purchases\_last\_6m** → number of purchases in last 6 months (integer, with many zeros).
* **avg\_session\_mins** → average app/website session time (has outliers and missing values).
* **credit\_score** → derived from income + some noise, with missing values.
* **satisfaction** → ordinal rating (1–5).
* **referral\_source** → how they found the company (Ads, Friend, Search, etc., with missing values).
* **zip\_code** → customer’s postal code (some missing).
* **notes** → messy free-text notes (e.g., “VIP customer!!!”, “prefers SMS”).
* **churned** → **target variable** (0 = still customer, 1 = churned).

**What does it predict?**

The target is **churned** → whether the customer has **left (1)** or is still **active (0)**.

* This makes it a **binary classification problem**.
* Input features (age, income, credit score, city, satisfaction, etc.) are used to **predict churn**.
* In real life, this is crucial for companies: **predict churn early → take action (discounts, retention campaigns).**

**Why this dataset is useful**

* **Messy like real-world data** → has NaNs, outliers, duplicates, inconsistent categories.
* **Mixed data types** → numeric, categorical, datetime, free-text, ordinal.
* **Perfect for EDA** → histograms, boxplots, correlations, missing value analysis.
* **Great for feature engineering** → imputing, encoding, scaling, log transforms, rare category grouping, text cleaning.
* **Binary target** → fits neatly into ANN classification workflows.