**DATASET**

**UCI Machine Learning Repository**

* [**Online Retail Dataset**](https://archive.ics.uci.edu/ml/datasets/Online+Retail)
* **Contains transactional data from a UK-based online retailer.**

The Online Retail Dataset from the UCI Machine Learning Repository is the dataset I am using for this project. It is well-suited for customer segmentation because it contains transactional data, customer behavior insights, and geographical information, all of which are important for my analysis.

**Why This Dataset?**

This dataset provides:

* Invoice and Product Information (InvoiceNo, StockCode, Quantity, UnitPrice) – Useful for analyzing purchasing behavior.
* Customer Identification (CustomerID) – Allows me to track individual spending habits and segment customers.
* Temporal Data (InvoiceDate) – Enables the calculation of Recency (days since last purchase).
* Spending Patterns (Monetary Value) – Helps classify high-value vs. low-value customers.
* Geographical Information (Country) – Can be used to study regional differences in customer behavior.

**How It Supports My Project**

Since my project involves comparing rule-based vs. machine learning-based segmentation, this dataset is ideal because it allows me to:

1. Manually define segmentation rules (e.g., frequent buyers, high spenders) for my baseline model.
2. Extract features (RFM: Recency, Frequency, Monetary) for clustering algorithms in the ML-based model.
3. Evaluate the effectiveness of both segmentation techniques using real-world customer data.

**Preprocessing Plan**

Before applying segmentation, I will:

* Remove duplicates and irrelevant columns.
* Filter out negative values (indicating returns).
* Aggregate data at the customer level to calculate RFM metrics.
* Normalize/scale features to ensure clustering works effectively.

This dataset provides a strong foundation for testing my hypothesis while also being widely recognized in academic and industry research, making it a credible choice for my project.