**Dataset: Online Retail Dataset (UCI Machine Learning Repository)**

* **Link**: [Online Retail Data](https://archive.ics.uci.edu/ml/datasets/Online+Retail)

This dataset contains transactional data, including invoice numbers, product descriptions, quantities, invoice dates, unit prices, customer IDs, and countries of sale. It’s suitable for tasks related to order analysis, sales trends, and customer behaviour.

### Questions:

**Q1.** Load the Online Retail dataset using SparkSession as a DataFrame. (PySpark - DataFrame)  
**[10 Marks]**

**Q2.** Remove any rows where the Quantity or UnitPrice is negative or zero. Create a new DataFrame called Cleaned\_Data.  
**[10 Marks]**

**Q3.** Extract year, month, and day from the InvoiceDate column and add them as separate columns (Year, Month, Day). Print the schema and display the top 5 records.  
**[10 Marks]**

**Q4.** Find the top 3 countries (excluding the UK) with the highest total sales (use Quantity \* UnitPrice as the sales amount).  
**[10 Marks]**

**Q5.** Create a new column called Order\_Value that contains the total value of each invoice (Quantity \* UnitPrice).  
**[10 Marks]**

**Q6.** Identify the **customer with the highest total spending** across all invoices.  
**[10 Marks]**

**Q7.** Use a **window function** to calculate the cumulative sales per country. Display the cumulative sales for the top 5 countries.  
**[10 Marks]**

**Q8.** Count the total number of **unique customers** who made purchases in December 2010 and check how many returned to shop in each subsequent month of 2011.  
**[10 Marks]**

**Q9.** Save the output of **Q8** as a returning\_customers.json file.  
**[10 Marks]**