Lab 4: Spark MLlib

On Google Colab and using the MLlib library:

1) Store the contents of all files in the **tp4_data** folder into a DataFrame.
2) Display the schema of the resulting DataFrame.
3) Fill the missing values (NaN) with the value 0.
4) Add a new column named "day_of_week". The value of this column is the name of the day of the week corresponding to the date in each row of the "InvoiceDate" column.
5) Split the data into a training set and a test set. Perform the split based on the **"InvoiceDate"** attribute: The training set contains purchases made before 2010-12-13, and the test set contains purchases made on or after 2010-12-13.
6) Create a **StringIndexer** to convert the days of the week in the **"day_of_week"** column into their corresponding numerical values.
7) Using the **StringIndexer**, Spark, for example, may represent Saturday as 6 and Monday as 1. However, with this numbering scheme, we implicitly indicate that Saturday is greater than Monday (based purely on numerical values). How can this issue be resolved?
8) Create a **VectorAssembler** containing three attributes: UnitPrice, Quantity, and day_of_week_encoded.

   Note: day_of_week_encoded is the result from question 7.

9) Create a pipeline configured with the results of steps 6, 7, and 8.
10) Our StringIndexer needs to know how many unique values it has to index. How can this issue be resolved?
11) Transform the training set data based on the stages of the pipeline.
12) Create an instance of KMeans, assuming the number of clusters is 20.
13) Train the KMeans model using the transformed data from step 11.
14) Make predictions on the test set.
15) Calculate the Silhouette coefficient.