```
# Install PySpark
!pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.11/dist-packages (3.
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.11/dist-packa
```

```
# Import and create Spark session
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("Task02_Preprocessing") \
    .getOrCreate()
```

```
from google.colab import files
uploaded = files.upload()
```

Browse... Online.csv

Online.csv(application/vnd.ms-excel) - 45580670 bytes, last modified: n/a - 100% done

```
Saving Online.csv to Online.csv
```

```
# Load CSV into Spark DataFrame
df = spark.read.csv("Online.csv", header=True, inferSchema=True)

# Preview data
df.show(5)
df.printSchema()
```

```
+---------+---------+--------------------+--------+--------------+---------+---------
|InvoiceNo|StockCode|         Description|Quantity|   InvoiceDate|UnitPrice|CustomerI
+---------+---------+--------------------+--------+--------------+---------+---------
|   536365|   85123A|WHITE HANGING HEA...|       6|12/1/2010 8:26|     2.55|     1785
|   536365|    71053| WHITE METAL LANTERN|       6|12/1/2010 8:26|     3.39|     1785
|   536365|   84406B|CREAM CUPID HEART...|       8|12/1/2010 8:26|     2.75|     1785
|   536365|   84029G|KNITTED UNION FLA...|       6|12/1/2010 8:26|     3.39|     1785
|   536365|   84029E|RED WOOLLY HOTTIE...|       6|12/1/2010 8:26|     3.39|     1785
+---------+---------+--------------------+--------+--------------+---------+---------
only showing top 5 rows

root
 |-- InvoiceNo: string (nullable = true)
 |-- StockCode: string (nullable = true)
 |-- Description: string (nullable = true)
 |-- Quantity: integer (nullable = true)
 |-- InvoiceDate: string (nullable = true)
 |-- UnitPrice: double (nullable = true)
```

```
 |-- CustomerID: integer (nullable = true)
 |-- Country: string (nullable = true)
```

```
# Drop nulls in key columns
clean_df = df.dropna(subset=["Quantity", "UnitPrice", "Country"])
```

```
from pyspark.sql.functions import col

clean_df = clean_df.withColumn("TotalValue", col("Quantity") * col("UnitPrice"))
clean_df.select("Quantity", "UnitPrice", "TotalValue").show(5)
```

```
+--------+---------+------------------+
|Quantity|UnitPrice|        TotalValue|
+--------+---------+------------------+
|       6|     2.55|15.299999999999999|
|       6|     3.39|             20.34|
|       8|     2.75|              22.0|
|       6|     3.39|             20.34|
|       6|     3.39|             20.34|
+--------+---------+------------------+
only showing top 5 rows
```

```
from pyspark.ml.feature import VectorAssembler

assembler = VectorAssembler(
    inputCols=["Quantity", "UnitPrice"],
    outputCol="features"
)

final_df = assembler.transform(clean_df)
final_df.select("features", "TotalValue").show(5)
```

```
+----------+------------------+
|  features|        TotalValue|
+----------+------------------+
|[6.0,2.55]|15.299999999999999|
|[6.0,3.39]|             20.34|
|[8.0,2.75]|              22.0|
|[6.0,3.39]|             20.34|
|[6.0,3.39]|             20.34|
+----------+------------------+
only showing top 5 rows
```

```
pandas_df = final_df.select("Quantity", "UnitPrice", "TotalValue").toPandas()
```

```
pandas_df.to_csv("Cleaned_Online.csv", index=False)

from google.colab import files
files.download("Cleaned_Online.csv")
```