assignment-7-pyspark-dataframe-1

September 18, 2024

[1]:
```
sc
```

[1]:
```
<SparkContext master=local[*] appName=PySparkShell>
```

[2]:
```
spark
```

[2]:
```
<pyspark.sql.session.SparkSession at 0x7fd6540dc3c8>
```

[3]:
```
sc.stop()
```

a) Create a new Spark Session with new SparkConfig

[4]:
```
from pyspark import SparkConf, SparkContext
# setMaster() - set spark context manager which is local[cpu_cores]
config = SparkConf().setMaster("local[4]").setAppName("PySparkSession")
sc = SparkContext(conf=config)
```

[5]:
```
sc
```

[5]:
```
<SparkContext master=local[4] appName=PySparkSession>
```

b) Create new instance of Spark SQL session and define new DataFrame using sales_data_sample.csv dataset.

[6]:
```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("SparkSQLSession").getOrCreate()
```

[7]:
```
spark
```

[7]:
```
<pyspark.sql.session.SparkSession at 0x7fd638c0d7b8>
```

[3]:
```
sales_df = spark.read.csv("file:///home/hadoop/Downloads/sales_data_sample.
↪csv",header=True,inferSchema=True)
```

[4]:
```
sales_df.show()
```

```
+----------+--------------+--------+--------------+-------+--------------+-
------+------+--------+-------+-----------+----+-----------+--------------------
+---------------+------------------+-----------+-------------+--------+-----
```

```
-----+---------+---------+--------------+---------------+--------+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES|      ORDERDATE|
STATUS|QTR_ID|MONTH_ID|YEAR_ID|PRODUCTLINE|MSRP|PRODUCTCODE|
CUSTOMERNAME|           PHONE|        ADDRESSLINE1|ADDRESSLINE2|         CITY|
STATE|POSTALCODE|  COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTFIRSTNAME|DEALSIZE|
+-----------+--------------+---------+--------------+-------+--------------+-
------+------+--------+-------+-----------+----+-----------+--------------------
+---------------+-------------------+-----------+------------+--------+-----
-----+---------+---------+---------------+----------------+--------+
|      10107|             30|     95.7|              2| 2871.0| 2/24/2003
0:00|Shipped|     1|       2|   2003|Motorcycles|  95|   S10_1678|   Land of
Toys Inc.|      2125557818|897 Long Airport …|        null|          NYC|
NY|     10022|      USA|       NA|             Yu|            Kwai|   Small|
|      10121|             34|    81.35|              5| 2765.9|  5/7/2003
0:00|Shipped|     2|       5|   2003|Motorcycles|  95|   S10_1678|  Reims
Collectables|      26.47.1555|  59 rue de l'Abbaye|        null|        Reims|
null|     51100|   France|     EMEA|        Henriot|            Paul|   Small|
|      10134|             41|    94.74|              2|3884.34|  7/1/2003
0:00|Shipped|     3|       7|   2003|Motorcycles|  95|   S10_1678|     Lyon
Souveniers|+33 1 46 62 7555|27 rue du Colonel…|        null|        Paris|
null|     75508|   France|     EMEA|       Da Cunha|          Daniel|  Medium|
|      10145|             45|    83.26|              6| 3746.7| 8/25/2003
0:00|Shipped|     3|       8|   2003|Motorcycles|  95|   S10_1678|
Toys4GrownUps.com|      6265557265|  78934 Hillside Dr.|        null|
Pasadena|      CA|     90003|      USA|       NA|          Young|
Julie|  Medium|
|      10159|             49|    100.0|             14|5205.27|10/10/2003
0:00|Shipped|     4|      10|   2003|Motorcycles|  95|   S10_1678|Corporate Gift
Id…|      6505551386|     7734 Strong St.|        null|San Francisco|      CA|
null|      USA|       NA|          Brown|           Julie|  Medium|
|      10168|             36|    96.66|              1|3479.76|10/28/2003
0:00|Shipped|     4|      10|   2003|Motorcycles|  95|   S10_1678|Technics
Stores Inc.|      6505556809|   9408 Furth Circle|        null|   Burlingame|
CA|     94217|      USA|       NA|         Hirano|            Juri|  Medium|
|      10180|             29|    86.13|              9|2497.77|11/11/2003
0:00|Shipped|     4|      11|   2003|Motorcycles|  95|   S10_1678|Daedalus
Designs …|      20.16.1555|184, chausse de T…|        null|        Lille|
null|     59000|   France|     EMEA|          Rance|         Martine|   Small|
|      10188|             48|    100.0|              1|5512.32|11/18/2003
0:00|Shipped|     4|      11|   2003|Motorcycles|  95|   S10_1678|        Herkku
Gifts|   +47 2267 3215|Drammen 121, PR 7…|        null|       Bergen|    null|
N 5804|   Norway|     EMEA|         Oeztan|          Veysel|  Medium|
|      10201|             22|    98.57|              2|2168.54| 12/1/2003
0:00|Shipped|     4|      12|   2003|Motorcycles|  95|   S10_1678|     Mini
Wheels Co.|      6505555787|5557 North Pendal…|        null|San Francisco|
CA|      null|      USA|       NA|         Murphy|           Julie|   Small|
|      10211|             41|    100.0|             14|4708.44| 1/15/2004
0:00|Shipped|     1|       1|   2004|Motorcycles|  95|   S10_1678|    Auto Canal
```

```
Petit|  (1) 47.55.6555|   25, rue Lauriston|        null|        Paris|    null|
75016|   France|     EMEA|        Perrier|       Dominique|  Medium|
|      10223|             37|    100.0|             1|3965.66| 2/20/2004
0:00|Shipped|     1|       2|   2004|Motorcycles|  95|   S10_1678|Australian
Collec…|    03 9520 4555|   636 St Kilda Road|     Level 3|
Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10237|             23|    100.0|             7|2333.12|  4/5/2004
0:00|Shipped|     2|       4|   2004|Motorcycles|  95|   S10_1678|
Vitachrome Inc.|      2125551500|   2678 Kingston Rd.|   Suite 101|
NYC|      NY|     10022|      USA|       NA|          Frick|         Michael|
Small|
|      10251|             28|    100.0|             2|3188.64| 5/18/2004
0:00|Shipped|     2|       5|   2004|Motorcycles|  95|   S10_1678|Tekni
Collectable…|      2015559350|       7476 Moss Rd.|        null|       Newark|
NJ|     94019|      USA|       NA|          Brown|         William|  Medium|
|      10263|             34|    100.0|             2|3676.76| 6/28/2004
0:00|Shipped|     2|       6|   2004|Motorcycles|  95|   S10_1678|     Gift
Depot Inc.|      2035552570| 25593 South Bay Ln.|        null|  Bridgewater|
CT|     97562|      USA|       NA|          King|           Julie|  Medium|
|      10275|             45|    92.83|             1|4177.35| 7/23/2004
0:00|Shipped|     3|       7|   2004|Motorcycles|  95|   S10_1678|   La Rochelle
Gifts|      40.67.8555|67, rue des Cinqu…|        null|       Nantes|    null|
44000|   France|     EMEA|        Labrune|          Janine|  Medium|
|      10285|             36|    100.0|             6|4099.68| 8/27/2004
0:00|Shipped|     3|       8|   2004|Motorcycles|  95|   S10_1678|Marta's
Replicas Co.|      6175558555| 39323 Spinnaker Dr.|        null|    Cambridge|
MA|     51247|      USA|       NA|      Hernandez|           Marta|  Medium|
|      10299|             23|    100.0|             9|2597.39| 9/30/2004
0:00|Shipped|     3|       9|   2004|Motorcycles|  95|   S10_1678|Toys of
Finland, Co.|     90-224 8555|       Keskuskatu 45|        null|     Helsinki|
null|     21240|  Finland|     EMEA|      Karttunen|           Matti|   Small|
|      10309|             41|    100.0|             5|4394.38|10/15/2004
0:00|Shipped|     4|      10|   2004|Motorcycles|  95|   S10_1678|  Baane Mini
Imports|      07-98 9555|Erling Skakkes ga…|        null|      Stavern|
null|      4110|   Norway|     EMEA|     Bergulfsen|           Jonas|  Medium|
|      10318|             46|    94.74|             1|4358.04| 11/2/2004
0:00|Shipped|     4|      11|   2004|Motorcycles|  95|   S10_1678|Diecast
Classics …|      2155551555|    7586 Pompton St.|        null|    Allentown|
PA|     70267|      USA|       NA|             Yu|           Kyung|  Medium|
|      10329|             42|    100.0|             1|4396.14|11/15/2004
0:00|Shipped|     4|      11|   2004|Motorcycles|  95|   S10_1678|   Land of
Toys Inc.|      2125557818|897 Long Airport …|        null|          NYC|
NY|     10022|      USA|       NA|             Yu|            Kwai|  Medium|
+----------+--------------+---------+---------------+-------+---------------+-
------+------+--------+-------+-----------+----+-----------+-------------------
+---------------+-------------------+------------+------------+--------+-----
-----+---------+---------+---------------+---------------+--------+
```

```
only showing top 20 rows
```

c) Find the shape of DataFrame.

[5]:
```
# Number of rows
sales_df.count()
```

[5]:
```
2823
```

[6]:
```
# Number of columns
len(sales_df.columns)
```

[6]:
```
25
```

d) Find the Summary of DataFrame for all numerical data columns.

[7]:
```
from pyspark.sql.types import IntegerType, StringType
from pyspark.sql.functions import *
```

[8]:
```
numerical_cols = [field.name for field in sales_df.schema.fields if not␣
 ↪isinstance(field.dataType, StringType)]
```

[9]:
```
sales_df[numerical_cols].summary().show()
```

```
+-------+-----------------+-----------------+-----------------+--------------
--+-----------------+-----------------+-----------------+-----------------+-
----------------+
|summary|       ORDERNUMBER|  QUANTITYORDERED|         PRICEEACH|
ORDERLINENUMBER|             SALES|            QTR_ID|          MONTH_ID|
YEAR_ID|              MSRP|
+-------+-----------------+-----------------+-----------------+--------------
--+-----------------+-----------------+-----------------+-----------------+-
----------------+
|  count|             2823|             2823|             2823|
2823|             2823|             2823|             2823|
2823|             2823|
|   mean|10258.725115125753|35.09280906836698|
83.65854410201929|6.466170740347148|  3553.88907190932|2.7176762309599716|7.0924
548352816155|2003.8150903294368|100.71555083244775|
| stddev|  92.0854775957196| 9.74144273706958|20.174276527840536|
4.22584096469094|1841.8651057401842| 1.203878088001756|
3.656633307661765|0.6996701541300869| 40.18791167720266|
|    min|             10100|                6|             26.88|
1|            482.13|                1|                1|             2003|
33|
|    25%|             10180|               27|             68.8|
3|           2203.11|                2|                4|             2003|
68|
```

```
|    50%|             10262|               35|              95.7|
6|            3184.8|                3|                8|              2004|
99|
|    75%|             10334|               43|             100.0|
9|            4508.0|                4|               11|              2004|
124|
|    max|             10425|               97|             100.0|
18|           14082.8|                4|               12|              2005|
214|
+-------+------------------+-----------------+------------------+--------------
--+------------------+-----------------+------------------+-----------------+-
-----------------+
```

e) Identify and handle missing or null values in the columns.

[10]:
```
from pyspark.sql.functions import *

sales_df.select([sum(col(column).isNull().cast('int')).alias(column) for column␣
 ↪in sales_df.columns]).show()
```

```
+-----------+--------------+---------+--------------+-----+---------+------+--
----+--------+-------+-----------+----+-----------+------------+-----+----------
--+------------+----+-----+----------+-------+---------+--------------+--------
--------+--------+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|SALES|ORDERDATE|STATUS|QT
R_ID|MONTH_ID|YEAR_ID|PRODUCTLINE|MSRP|PRODUCTCODE|CUSTOMERNAME|PHONE|ADDRESSLIN
E1|ADDRESSLINE2|CITY|STATE|POSTALCODE|COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTF
IRSTNAME|DEALSIZE|
+-----------+--------------+---------+--------------+-----+---------+------+--
----+--------+-------+-----------+----+-----------+------------+-----+----------
--+------------+----+-----+----------+-------+---------+--------------+--------
--------+--------+
|          0|             0|        0|             0|    0|        0|     0|
0|       0|      0|          0|   0|          0|           0|    0|           0|
2521|   0| 1486|        76|      0|        0|             0|             0|
0|
+-----------+--------------+---------+--------------+-----+---------+------+--
----+--------+-------+-----------+----+-----------+------------+-----+----------
--+------------+----+-----+----------+-------+---------+--------------+--------
--------+--------+
```

[11]:
```
sales_df1 = sales_df.fillna("Nill")
```

[12]:
```
from pyspark.sql.functions import *
```

```
sales_df1.select([sum(col(column).isNull().cast('int')).alias(column) for␣
 ↪column in sales_df1.columns]).show()
```

```
+----------+--------------+---------+--------------+-----+---------+------+--
----+--------+-------+----------+----+----------+------------+-----+----------
--+------------+----+-----+----------+-------+---------+--------------+--------
--------+--------+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|SALES|ORDERDATE|STATUS|QT
R_ID|MONTH_ID|YEAR_ID|PRODUCTLINE|MSRP|PRODUCTCODE|CUSTOMERNAME|PHONE|ADDRESSLIN
E1|ADDRESSLINE2|CITY|STATE|POSTALCODE|COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTF
IRSTNAME|DEALSIZE|
+----------+--------------+---------+--------------+-----+---------+------+--
----+--------+-------+----------+----+----------+------------+-----+----------
--+------------+----+-----+----------+-------+---------+--------------+--------
--------+--------+
|          0|             0|        0|             0|    0|        0|     0|
0|       0|      0|          0|   0|          0|           0|    0|           0|
0|   0|    0|         0|      0|        0|              0|               0|
0|
+----------+--------------+---------+--------------+-----+---------+------+--
----+--------+-------+----------+----+----------+------------+-----+----------
--+------------+----+-----+----------+-------+---------+--------------+--------
--------+--------+
```

f) Calculate the total revenue generated per country by combining the columns QUANTITY-ORDERED and PRICEEACH using Spark DataFrame operations?

[13]:
```
sales_df2 = sales_df1.withColumn('Total␣
 ↪Revenue',col('QUANTITYORDERED')*col('PRICEEACH'))
sales_df2.groupBy('COUNTRY').agg(sum('Total Revenue').alias('Total Revenue')).
 ↪show()
```

```
+-----------+------------------+
|    COUNTRY|     Total Revenue|
+-----------+------------------+
|     Sweden|174264.10000000006|
|Philippines| 80291.16999999998|
|  Singapore| 227985.5000000001|
|    Germany|         178689.08|
|     France| 919257.8499999997|
|    Belgium|          94528.88|
|    Finland|268714.70000000007|
|      Italy| 309402.8699999999|
|     Norway| 246115.8000000001|
|      Spain|1021705.9700000002|
|    Denmark|         192747.63|
|    Ireland|          43237.24|
```

```
|        USA|2986425.2099999995|
|         UK|413203.33999999997|
|Switzerland| 93344.90999999999|
|     Canada|193504.34000000003|
|      Japan|153076.68999999994|
|  Australia|521598.45999999985|
|    Austria|172793.05000000002|
+-----------+------------------+
```

g) Determine the top 5 products with the highest total sales revenue using Spark DataFrame?

[14]:
```
product_sales = sales_df1.groupBy("PRODUCTLINE").agg(sum("SALES").
↪alias("total_sales"))


sorted_sales = product_sales.orderBy(col("total_sales").desc())
top_5_products = sorted_sales.limit(5)
top_5_products.show()
```

```
+---------------+------------------+
|     PRODUCTLINE|       total_sales|
+---------------+------------------+
|    Classic Cars| 3919615.659999997|
|    Vintage Cars|1903150.8399999992|
|     Motorcycles|1166388.3400000003|
|Trucks and Buses|1127789.8399999996|
|          Planes| 975003.5700000001|
+---------------+------------------+
```

h) Find the average order quantity for each product using groupBy and agg operations?

[15]:
```
sales_df1.groupBy('PRODUCTLINE').agg(mean('QUANTITYORDERED').alias('Avg order␣
↪quantity')).show()
```

```
+---------------+------------------+
|     PRODUCTLINE|Avg order quantity|
+---------------+------------------+
|     Motorcycles|35.235649546827794|
|    Vintage Cars| 34.71004942339374|
|           Ships| 34.73076923076923|
|Trucks and Buses| 35.80398671096346|
|    Classic Cars|35.152016546018615|
|          Trains| 35.22077922077922|
|          Planes| 35.05555555555556|
+---------------+------------------+
```

i) Using Spark DataFrame, filter orders where the SALES value exceeds $10,000 and sort the results by the ORDERDATE column?

[16]:
```
sales_df1 = sales_df1.withColumn("ORDERDATE", to_timestamp(col("ORDERDATE"), "M/
↪d/yyyy H:mm"))
sales_df1.filter(sales_df1.SALES > 10000).orderBy(sales_df1.ORDERDATE).show()
```

```
+-----------+---------------+---------+--------------+-------+-----------------
--+----------+------+--------+-------+------------+----+-----------+-----------
--------+---------------+--------------------+------------+-------------+-----+-
---------+---------+---------+--------------+----------------+--------+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES|
ORDERDATE|    STATUS|QTR_ID|MONTH_ID|YEAR_ID| PRODUCTLINE|MSRP|PRODUCTCODE|
CUSTOMERNAME|           PHONE|         ADDRESSLINE1|ADDRESSLINE2|
CITY|STATE|POSTALCODE|
COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTFIRSTNAME|DEALSIZE|
+-----------+---------------+---------+--------------+-------+-----------------
--+----------+------+--------+-------+------------+----+-----------+-----------
--------+---------------+--------------------+------------+-------------+-----+-
---------+---------+---------+--------------+----------------+--------+
|      10127|             46|    100.0|             2|11279.2|2003-06-03
00:00:00|   Shipped|     2|       6|   2003|Classic Cars| 207|   S12_1108|
Muscle Machine Inc|     2125557413|   4092 Furth Circle|   Suite 400|
NYC|   NY|     10022|      USA|       NA|          Young|            Jeff|
Large|
|      10150|             45|    100.0|             8|10993.5|2003-09-19
00:00:00|   Shipped|     3|       9|   2003|Classic Cars| 214|   S10_1949|Dragon
Souveniers…|   +65 221 7555|Bronz Sok., Bronz…|        Nill|    Singapore|
Nill|     79903|Singapore|    Japan|      Natividad|           Eric|   Large|
|      10247|             44|    100.0|             2|10606.2|2004-05-05
00:00:00|   Shipped|     2|       5|   2004|Classic Cars| 207|   S12_1108|
Suominen Souveniers|+358 9 8045 555|Software Engineer…|        Nill|
Espoo| Nill| FIN-02271|  Finland|     EMEA|       Suominen|           Kalle|
Large|
|      10304|             47|    100.0|             6|10172.7|2004-10-11
00:00:00|   Shipped|     4|      10|   2004|Classic Cars| 214|   S10_1949|  Auto
Assoc. & Cie.|     30.59.8555|67, avenue de l'E…|        Nill|   Versailles|
Nill|     78000|   France|     EMEA|         Tonini|          Daniel|   Large|
|      10312|             48|    100.0|             3|11623.7|2004-10-21
00:00:00|   Shipped|     4|      10|   2004|Classic Cars| 214|   S10_1949|Mini
Gifts Distri…|     4155551450|     5677 Strong St.|        Nill|   San Rafael|
CA|     97562|      USA|       NA|         Nelson|         Valarie|   Large|
|      10322|             50|    100.0|             6|12536.5|2004-11-04
00:00:00|   Shipped|     4|      11|   2004|Vintage Cars| 127|   S18_2325|Online
Diecast Cr…|     6035558647|2304 Long Airport…|        Nill|       Nashua|
NH|     62005|      USA|       NA|          Young|         Valarie|   Large|
|      10333|             46|    100.0|             2|11336.7|2004-11-18
00:00:00|   Shipped|     4|      11|   2004|Vintage Cars|  99|   S18_3320|
```

```
Mini Wheels Co.|     6505555787|5557 North Pendal…|        Nill|San Francisco|
CA|      Nill|      USA|       NA|         Murphy|           Julie|   Large|
|      10339|             55|    100.0|             13|10758.0|2004-11-23
00:00:00|   Shipped|     4|      11|   2004|Vintage Cars|  88|   S24_3151|Tokyo
Collectable…|+81 3 3584 0555|      2-2-8 Roppongi|        Nill|    Minato-
ku|Tokyo|  106-0032|    Japan|    Japan|      Shimamura|           Akiko|
Large|
|      10375|             43|    100.0|              2|10039.6|2005-02-03
00:00:00|   Shipped|     1|       2|   2005|      Planes|  72|   S24_4278|   La
Rochelle Gifts|     40.67.8555|67, rue des Cinqu…|        Nill|       Nantes|
Nill|     44000|   France|     EMEA|        Labrune|          Janine|   Large|
|      10388|             46|    100.0|              2|10066.6|2005-03-03
00:00:00|   Shipped|     1|       3|   2005|      Planes|  91|  S700_1691|
FunGiftIdeas.com|     5085552555|   1785 First Street|        Nill|  New
Bedford|   MA|     50553|      USA|       NA|        Benitez|         Violeta|
Large|
|      10403|             66|    100.0|              9|11886.6|2005-04-08
00:00:00|   Shipped|     2|       4|   2005| Motorcycles| 193|   S10_4698|UK
Collectables, …| (171) 555-2282|Berkeley Gardens …|        Nill|
Liverpool| Nill|   WX1 6LT|       UK|     EMEA|          Devon|       Elizabeth|
Large|
|      10405|             76|    100.0|              3|11739.7|2005-04-14
00:00:00|   Shipped|     2|       4|   2005|Classic Cars| 140|   S24_3856|
Mini Caravy|     88.60.1555|    24, place Kluber|        Nill|   Strasbourg|
Nill|     67000|   France|     EMEA|        Citeaux|      Frederique|   Large|
|      10406|             65|    100.0|              1|10468.9|2005-04-15
00:00:00|  Disputed|     2|       4|   2005|Classic Cars| 141|   S18_3685|Danish
Wholesale …|     31 12 3555|        Vinb'ltet 34|        Nill|    Kobenhavn|
Nill|      1734|  Denmark|     EMEA|       Petersen|           Jytte|   Large|
|      10407|             76|    100.0|              2|14082.8|2005-04-22
00:00:00|   On Hold|     2|       4|   2005|Vintage Cars| 170|   S18_1749|The
Sharp Gifts W…|     4085553659|      3086 Ingle Ln.|        Nill|     San
Jose|   CA|     94217|      USA|       NA|          Frick|             Sue|
Large|
|      10412|             60|    100.0|              9|11887.8|2005-05-03
00:00:00|   Shipped|     2|       5|   2005|Classic Cars| 169|   S18_3232|Euro
Shopping Cha…| (91) 555 94 44|  C/ Moralzarzal, 86|        Nill|       Madrid|
Nill|     28034|    Spain|     EMEA|         Freyre|           Diego|   Large|
|      10424|             50|    100.0|              6|12001.0|2005-05-31
00:00:00|In Process|     2|       5|   2005|Classic Cars| 214|   S10_1949|Euro
Shopping Cha…| (91) 555 94 44|  C/ Moralzarzal, 86|        Nill|       Madrid|
Nill|     28034|    Spain|     EMEA|         Freyre|           Diego|   Large|
+-----------+--------------+---------+---------------+-------+----------------
--+----------+------+--------+-------+------------+----+-----------+-----------
--------+--------------+--------------------+------------+-------------+-----+-
---------+---------+---------+--------------+---------------+--------+
```

j) Filter out rows where the STATUS is 'Cancelled' and calculate the total sales from the remaining orders?

[17]:
```
sales_df1.filter(sales_df1.STATUS != 'Cancelled').agg(sum('SALES').alias('Total␣
↪Sales')).show()
```

```
+-----------------+
|      Total Sales|
+-----------------+
|9838141.370000018|
+-----------------+
```

k) Use Spark Data Frame transformations to derive the yearly sales for each customer (CUSTOMERNAME) based on the ORDERDATE column?

[18]:
```
from pyspark.sql.functions import *
sales_df1 = sales_df1.withColumn("ORDERDATE", to_timestamp(col("ORDERDATE"), "M/
↪d/yyyy H:mm"))

sales_df1 = sales_df1.withColumn("YEAR", year(col("ORDERDATE")))


yearly_sales_df = sales_df1.groupBy("CUSTOMERNAME", "YEAR").agg(sum("SALES").
↪alias("TOTAL_SALES"))

yearly_sales_df.show()
```

```
+--------------------+----+------------------+
|        CUSTOMERNAME|YEAR|       TOTAL_SALES|
+--------------------+----+------------------+
|  Baane Mini Imports|2003|56176.659999999996|
|Stylish Desk Deco…|2004|13739.900000000001|
|Marseille Mini Autos|2003|52481.840000000004|
|Danish Wholesale …|2004|          60157.62|
|Toms Spezialitten…|2003|          31363.18|
|Australian Collec…|2004|140859.56999999998|
|Dragon Souveniers…|2004|           3127.88|
|    Super Scale Inc.|2003|          42498.76|
|Collectables For …|2004|15110.800000000001|
|Royal Canadian Co…|2004| 74634.84999999999|
|Online Diecast Cr…|2003|           76114.7|
|   Gifts4AllAges.com|2005|          48316.89|
|        Herkku Gifts|2004|           16363.1|
|Diecast Classics …|2004|115971.34000000001|
|Motor Mint Distri…|2003|          27398.82|
|Daedalus Designs …|2003|48874.280000000006|
|Stylish Desk Deco…|2003|           75064.6|
```

```
|         Mini Caravy|2005|          35680.35|
|     Mini Wheels Co.|2004|          30348.72|
|Scandinavian Gift…|2005|          31606.72|
+--------------------+----+-------------------+
only showing top 20 rows
```

l) Add a new column to the DataFrame that categorizes orders as “High”, “Medium”, or “Low” sales based on the SALES value?

[71]:
```
quantiles = sales_df1.approxQuantile('SALES',[0.33,0.67],0.0001)
val1 = quantiles[0]
val2 = quantiles[1]

sales_df1 = sales_df1.withColumn('CATEGORY',when(col('SALES') <= val1,'Low').
 ↪when(col('SALES') <= val2,'Medium')
                                  .otherwise('High'))
sales_df1.show()
```

```
+-----------+--------------+---------+--------------+-------+-----------------
--+-------+------+--------+-------+-----------+----+-----------+---------------
----+----------------+--------------------+-----------+------------+--------+-
---------+---------+---------+--------------+----------------+--------+----+---
-----+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES|
ORDERDATE| STATUS|QTR_ID|MONTH_ID|YEAR_ID|PRODUCTLINE|MSRP|PRODUCTCODE|
CUSTOMERNAME|           PHONE|        ADDRESSLINE1|ADDRESSLINE2|         CITY|
STATE|POSTALCODE|
COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTFIRSTNAME|DEALSIZE|YEAR|CATEGORY|
+-----------+--------------+---------+--------------+-------+-----------------
--+-------+------+--------+-------+-----------+----+-----------+---------------
----+----------------+--------------------+-----------+------------+--------+-
---------+---------+---------+--------------+----------------+--------+----+---
-----+
|      10107|             30|     95.7|             2| 2871.0|2003-02-24
00:00:00|Shipped|     1|       2|   2003|Motorcycles|  95|   S10_1678|   Land of
Toys Inc.|      2125557818|897 Long Airport …|        Nill|          NYC|
NY|     10022|      USA|       NA|             Yu|            Kwai|
Small|2003|  Medium|
|      10121|             34|    81.35|             5| 2765.9|2003-05-07
00:00:00|Shipped|     2|       5|   2003|Motorcycles|  95|   S10_1678|  Reims
Collectables|      26.47.1555|  59 rue de l'Abbaye|        Nill|        Reims|
Nill|     51100|   France|     EMEA|        Henriot|            Paul|
Small|2003|  Medium|
|      10134|             41|    94.74|             2|3884.34|2003-07-01
00:00:00|Shipped|     3|       7|   2003|Motorcycles|  95|   S10_1678|     Lyon
Souveniers|+33 1 46 62 7555|27 rue du Colonel…|        Nill|        Paris|
Nill|     75508|   France|     EMEA|       Da Cunha|          Daniel|
```

```
Medium|2003|  Medium|
|      10145|             45|    83.26|              6| 3746.7|2003-08-25
00:00:00|Shipped|     3|       8|   2003|Motorcycles|  95|   S10_1678|
Toys4GrownUps.com|      6265557265|  78934 Hillside Dr.|        Nill|
Pasadena|      CA|     90003|      USA|       NA|          Young|
Julie|  Medium|2003|  Medium|
|      10159|             49|    100.0|             14|5205.27|2003-10-10
00:00:00|Shipped|     4|      10|   2003|Motorcycles|  95|   S10_1678|Corporate
Gift Id…|      6505551386|     7734 Strong St.|        Nill|San Francisco|
CA|      Nill|      USA|       NA|          Brown|           Julie|
Medium|2003|    High|
|      10168|             36|    96.66|              1|3479.76|2003-10-28
00:00:00|Shipped|     4|      10|   2003|Motorcycles|  95|   S10_1678|Technics
Stores Inc.|      6505556809|   9408 Furth Circle|        Nill|   Burlingame|
CA|     94217|      USA|       NA|         Hirano|            Juri|
Medium|2003|  Medium|
|      10180|             29|    86.13|              9|2497.77|2003-11-11
00:00:00|Shipped|     4|      11|   2003|Motorcycles|  95|   S10_1678|Daedalus
Designs …|      20.16.1555|184, chausse de T…|        Nill|        Lille|
Nill|     59000|   France|     EMEA|          Rance|         Martine|
Small|2003|     Low|
|      10188|             48|    100.0|              1|5512.32|2003-11-18
00:00:00|Shipped|     4|      11|   2003|Motorcycles|  95|   S10_1678|
Herkku Gifts|   +47 2267 3215|Drammen 121, PR 7…|        Nill|       Bergen|
Nill|    N 5804|   Norway|     EMEA|         Oeztan|          Veysel|
Medium|2003|    High|
|      10201|             22|    98.57|              2|2168.54|2003-12-01
00:00:00|Shipped|     4|      12|   2003|Motorcycles|  95|   S10_1678|     Mini
Wheels Co.|      6505555787|5557 North Pendal…|        Nill|San Francisco|
CA|      Nill|      USA|       NA|         Murphy|           Julie|
Small|2003|     Low|
|      10211|             41|    100.0|             14|4708.44|2004-01-15
00:00:00|Shipped|     1|       1|   2004|Motorcycles|  95|   S10_1678|    Auto
Canal Petit|  (1) 47.55.6555|   25, rue Lauriston|        Nill|        Paris|
Nill|     75016|   France|     EMEA|        Perrier|       Dominique|
Medium|2004|    High|
|      10223|             37|    100.0|              1|3965.66|2004-02-20
00:00:00|Shipped|     1|       2|   2004|Motorcycles|  95|   S10_1678|Australian
Collec…|    03 9520 4555|   636 St Kilda Road|     Level 3|
Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|2004|  Medium|
|      10237|             23|    100.0|              7|2333.12|2004-04-05
00:00:00|Shipped|     2|       4|   2004|Motorcycles|  95|   S10_1678|
Vitachrome Inc.|      2125551500|   2678 Kingston Rd.|   Suite 101|
NYC|      NY|     10022|      USA|       NA|          Frick|         Michael|
Small|2004|     Low|
|      10251|             28|    100.0|              2|3188.64|2004-05-18
00:00:00|Shipped|     2|       5|   2004|Motorcycles|  95|   S10_1678|Tekni
```

```
Collectable…|      2015559350|       7476 Moss Rd.|        Nill|       Newark|
NJ|     94019|      USA|       NA|          Brown|         William|
Medium|2004|  Medium|
|      10263|             34|    100.0|             2|3676.76|2004-06-28
00:00:00|Shipped|     2|       6|   2004|Motorcycles|  95|   S10_1678|     Gift
Depot Inc.|      2035552570| 25593 South Bay Ln.|        Nill|  Bridgewater|
CT|     97562|      USA|       NA|           King|           Julie|
Medium|2004|  Medium|
|      10275|             45|    92.83|             1|4177.35|2004-07-23
00:00:00|Shipped|     3|       7|   2004|Motorcycles|  95|   S10_1678|   La
Rochelle Gifts|      40.67.8555|67, rue des Cinqu…|        Nill|       Nantes|
Nill|     44000|   France|     EMEA|        Labrune|          Janine|
Medium|2004|    High|
|      10285|             36|    100.0|             6|4099.68|2004-08-27
00:00:00|Shipped|     3|       8|   2004|Motorcycles|  95|   S10_1678|Marta's
Replicas Co.|      6175558555| 39323 Spinnaker Dr.|        Nill|    Cambridge|
MA|     51247|      USA|       NA|      Hernandez|           Marta|
Medium|2004|    High|
|      10299|             23|    100.0|             9|2597.39|2004-09-30
00:00:00|Shipped|     3|       9|   2004|Motorcycles|  95|   S10_1678|Toys of
Finland, Co.|     90-224 8555|       Keskuskatu 45|        Nill|     Helsinki|
Nill|     21240|  Finland|     EMEA|      Karttunen|           Matti|
Small|2004|  Medium|
|      10309|             41|    100.0|             5|4394.38|2004-10-15
00:00:00|Shipped|     4|      10|   2004|Motorcycles|  95|   S10_1678|  Baane
Mini Imports|      07-98 9555|Erling Skakkes ga…|        Nill|      Stavern|
Nill|      4110|   Norway|     EMEA|     Bergulfsen|           Jonas|
Medium|2004|    High|
|      10318|             46|    94.74|             1|4358.04|2004-11-02
00:00:00|Shipped|     4|      11|   2004|Motorcycles|  95|   S10_1678|Diecast
Classics …|      2155551555|    7586 Pompton St.|        Nill|    Allentown|
PA|     70267|      USA|       NA|             Yu|           Kyung|
Medium|2004|    High|
|      10329|             42|    100.0|             1|4396.14|2004-11-15
00:00:00|Shipped|     4|      11|   2004|Motorcycles|  95|   S10_1678|   Land of
Toys Inc.|      2125557818|897 Long Airport …|        Nill|          NYC|
NY|     10022|      USA|       NA|             Yu|            Kwai|
Medium|2004|    High|
+----------+--------------+---------+--------------+-------+----------------
--+-------+------+--------+-------+-----------+----+-----------+---------------
----+---------------+--------------------+------------+-------------+--------+-
---------+---------+---------+--------------+---------------+--------+----+---
-----+
only showing top 20 rows
```

m) Assume , If you have another DataFrame with customer demographic data, how would you perform a join to compute the total sales per demographic group?

[101]:
```
customer_demographics_data = [
    ("Land of Toys Inc.", "25-34", "High", "VIP"),
    ("Reims Collectables", "35-44", "Medium", "Regular"),
    ("Lyon Souveniers", "45-54", "Medium", "Regular"),
    ("Toys4GrownUps.com", "25-34", "High", "VIP"),
    ("Corporate Gift Ideas", "35-44", "High", "VIP"),
    ("Technics Stores Inc.", "45-54", "Medium", "Regular"),
    ("Daedalus Designs Inc.", "55-64", "Low", "New"),
    ("Herkku Gifts", "65+", "Norway", "New"),
    ("Mini Wheels Co.", "25-34", "Medium", "Regular"),
    ("Auto Canal Petit", "35-44", "Medium", "Regular"),
]
customer_demographics_columns = ["CUSTOMERNAME", "AGE_GROUP", "INCOME_LEVEL",␣
 ↪"LOYALTY_STATUS"]


customer_demographics_df = spark.createDataFrame(customer_demographics_data,␣
 ↪schema=customer_demographics_columns)

customer_demographics_df.show()
```

```
+--------------------+---------+------------+--------------+
|        CUSTOMERNAME|AGE_GROUP|INCOME_LEVEL|LOYALTY_STATUS|
+--------------------+---------+------------+--------------+
|   Land of Toys Inc.|    25-34|        High|           VIP|
|  Reims Collectables|    35-44|      Medium|       Regular|
|     Lyon Souveniers|    45-54|      Medium|       Regular|
|   Toys4GrownUps.com|    25-34|        High|           VIP|
|Corporate Gift Ideas|    35-44|        High|           VIP|
|Technics Stores Inc.|    45-54|      Medium|       Regular|
|Daedalus Designs …|    55-64|         Low|           New|
|        Herkku Gifts|      65+|      Norway|           New|
|     Mini Wheels Co.|    25-34|      Medium|       Regular|
|    Auto Canal Petit|    35-44|      Medium|       Regular|
+--------------------+---------+------------+--------------+
```

[102]:
```
from pyspark.sql.functions import sum


joined_df = sales_df.join(customer_demographics_df, on="CUSTOMERNAME")


total_sales_per_demographic_df = joined_df.groupBy("AGE_GROUP","CUSTOMERNAME",␣
 ↪"INCOME_LEVEL", "LOYALTY_STATUS").agg(
    sum("SALES").alias("TOTAL_SALES")
)
```

```
total_sales_per_demographic_df.show()
```

```
+---------+--------------------+------------+--------------+------------------+
|AGE_GROUP|        CUSTOMERNAME|INCOME_LEVEL|LOYALTY_STATUS|       TOTAL_SALES|
+---------+--------------------+------------+--------------+------------------+
|    35-44|    Auto Canal Petit|      Medium|       Regular| 93170.66000000002|
|    45-54|     Lyon Souveniers|      Medium|       Regular|          78570.34|
|    35-44|  Reims Collectables|      Medium|       Regular|         135042.94|
|    25-34|   Toys4GrownUps.com|        High|           VIP|104561.95999999998|
|    25-34|     Mini Wheels Co.|      Medium|       Regular|          74476.18|
|      65+|        Herkku Gifts|      Norway|           New|         111640.28|
|    45-54|Technics Stores Inc.|      Medium|       Regular|120783.06999999999|
|    25-34|   Land of Toys Inc.|        High|           VIP|164069.43999999994|
+---------+--------------------+------------+--------------+------------------+
```

n) Can you implement a cumulative distribution function (CDF) over the SALES value for each CUSTOMERNAME? What insights can you gather from analyzing the CDF distribution for each customer?

[138]:
```
window_spec = Window.partitionBy('CUSTOMERNAME').orderBy('SALES')

rank_df = sales_df1.withColumn('rank', row_number().over(window_spec))
count_df = rank_df.withColumn('total_count', count('SALES').over(Window.
 ↪partitionBy('CUSTOMERNAME')))
cdf_df = count_df.withColumn('CDF', col('rank') / col('total_count'))
cdf_df1 = cdf_df.select('CUSTOMERNAME', 'SALES', 'CDF').show(50)
```

```
+------------------+-------+-------------------+
|       CUSTOMERNAME|  SALES|                CDF|
+------------------+-------+-------------------+
|Suominen Souveniers| 891.03| 0.03333333333333333|
|Suominen Souveniers| 1086.6| 0.06666666666666667|
|Suominen Souveniers|1103.76|                0.1|
|Suominen Souveniers|1629.04| 0.13333333333333333|
|Suominen Souveniers| 1988.4| 0.16666666666666666|
|Suominen Souveniers|2140.11|                0.2|
|Suominen Souveniers|2447.76| 0.23333333333333334|
|Suominen Souveniers|2632.89| 0.26666666666666666|
|Suominen Souveniers| 2773.8|                0.3|
|Suominen Souveniers|2775.08|  0.3333333333333333|
|Suominen Souveniers|2817.87| 0.36666666666666664|
|Suominen Souveniers|2851.84|                0.4|
|Suominen Souveniers|2931.98| 0.43333333333333335|
|Suominen Souveniers|3128.65|  0.4666666666666667|
|Suominen Souveniers|3288.82|                0.5|
```

```
|Suominen Souveniers|3595.62|  0.5333333333333333|
|Suominen Souveniers|3686.54|  0.5666666666666667|
|Suominen Souveniers| 3784.8|                 0.6|
|Suominen Souveniers| 4068.7|  0.6333333333333333|
|Suominen Souveniers|4142.64|  0.6666666666666666|
|Suominen Souveniers|4157.73|                 0.7|
|Suominen Souveniers|4381.25|  0.7333333333333333|
|Suominen Souveniers| 4836.5|  0.7666666666666667|
|Suominen Souveniers|5154.41|                 0.8|
|Suominen Souveniers|5500.44|  0.8333333333333334|
|Suominen Souveniers|5938.53|  0.8666666666666667|
|Suominen Souveniers|6287.66|                 0.9|
|Suominen Souveniers| 6576.5|  0.9333333333333333|
|Suominen Souveniers| 6756.0|  0.9666666666666667|
|Suominen Souveniers|10606.2|                 1.0|
| Amica Models & Co.|  577.6|0.038461538461538464|
| Amica Models & Co.|1381.05| 0.07692307692307693|
| Amica Models & Co.|1557.36| 0.11538461538461539|
| Amica Models & Co.| 1574.0| 0.15384615384615385|
| Amica Models & Co.|1656.69| 0.19230769230769232|
| Amica Models & Co.|1921.92| 0.23076923076923078|
| Amica Models & Co.|2084.81|  0.2692307692307692|
| Amica Models & Co.|2137.05|  0.3076923076923077|
| Amica Models & Co.|2418.24| 0.34615384615384615|
| Amica Models & Co.|2800.08| 0.38461538461538464|
| Amica Models & Co.|2819.28|  0.4230769230769231|
| Amica Models & Co.|2941.89| 0.46153846153846156|
| Amica Models & Co.|2954.53|                 0.5|
| Amica Models & Co.|3006.43|  0.5384615384615384|
| Amica Models & Co.|3474.46|  0.5769230769230769|
| Amica Models & Co.| 3668.6|  0.6153846153846154|
| Amica Models & Co.|3704.05|  0.6538461538461539|
| Amica Models & Co.|4242.24|  0.6923076923076923|
| Amica Models & Co.| 4455.0|  0.7307692307692307|
| Amica Models & Co.| 4750.8|  0.7692307692307693|
+-------------------+-------+--------------------+
only showing top 50 rows
```

Insights

- CDF values show how sales are distributed within each customer group.
- Sales Concentration for a customer like "Suominen Souveniers," we can see that a significant portion of the total sales is concentrated at higher values.
- For example, 0.7 CDF corresponds to a sales value of 4381.25, meaning that 70% of the sales data falls below this amount.

o) Write spark dataframe code to rank products by total revenue within each country (COUN-

TRY)?

[20]:
```
from pyspark.sql.functions import col
from pyspark.sql.window import Window
from pyspark.sql.functions import rank

window_spec = Window.partitionBy("COUNTRY").orderBy(col("Total Revenue").desc())


ranked_df = sales_df2.withColumn("RANK", rank().over(window_spec))


ranked_df.select(col('PRODUCTCODE'),col('COUNTRY'),col('Total␣
 ↪revenue'),col('RANK')).show()
```

```
+-----------+-------+-----------------+----+
|PRODUCTCODE|COUNTRY|     Total revenue|RANK|
+-----------+-------+-----------------+----+
|   S18_4600| Sweden|           4900.0|   1|
|   S18_4600| Sweden|           4800.0|   2|
|   S24_2300| Sweden|           4800.0|   2|
|   S12_4675| Sweden|           4700.0|   4|
|   S18_2949| Sweden|           4700.0|   4|
|   S18_2319| Sweden|           4600.0|   6|
|   S24_1578| Sweden|           4500.0|   7|
|   S10_4757| Sweden|           4400.0|   8|
|   S18_4522| Sweden|           4300.5|   9|
|   S18_1662| Sweden|           4300.0|  10|
|   S24_2300| Sweden|           4200.0|  11|
|   S12_1666| Sweden|           4100.0|  12|
|   S18_1097| Sweden|           4100.0|  12|
|   S24_2011| Sweden|           4100.0|  12|
|   S24_2000| Sweden|3988.6000000000004|  15|
|   S18_2625| Sweden|           3900.0|  16|
|   S18_1889| Sweden|3881.7799999999997|  17|
|   S10_1949| Sweden|           3700.0|  18|
|   S24_3151| Sweden|          3519.85|  19|
|   S12_3380| Sweden|           3500.0|  20|
+-----------+-------+-----------------+----+
only showing top 20 rows
```

p) Calculate a running total of SALES for each customer and show the top 5 customers by this cumulative total?

[93]:
```
window_spec = Window.partitionBy("CUSTOMERNAME").orderBy("ORDERDATE").
 ↪rowsBetween(Window.unboundedPreceding, Window.currentRow)
```

```
running_total_df = sales_df1.withColumn("RUNNING_TOTAL", sum("SALES").
 ↪over(window_spec))
per_customer_df = running_total_df.groupBy("CUSTOMERNAME").agg(
    sum("RUNNING_TOTAL").alias("TOTAL_CUMULATIVE_SALES")

top_customers_df = per_customer_df.orderBy(col("TOTAL_CUMULATIVE_SALES").
 ↪desc()).limit(5)
top_customers_df.show()
```

```
+--------------------+----------------------+
|        CUSTOMERNAME|TOTAL_CUMULATIVE_SALES|
+--------------------+----------------------+
|Euro Shopping Cha…|  1.1636376556000003E8|
|Mini Gifts Distri…|   5.858921097000004E7|
|Australian Collec…|     5550391.160000001|
|  Muscle Machine Inc|     5175832.140000001|
|   La Rochelle Gifts|     4766253.330000002|
+--------------------+----------------------+
```

q) Identify and handle Outliers in DataFrame.

[21]:
```
from pyspark.sql.functions import col

quantiles = sales_df.approxQuantile("SALES", [0.25, 0.75], 0.001)

Q1 = quantiles[0]
Q3 = quantiles[1]
IQR = Q3 - Q1



lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR


outliers_df = sales_df1.filter((col("SALES") < lower_bound) | (col("SALES") >␣
 ↪upper_bound))
outliers_df.show()


handled_df = sales_df1.filter((col("SALES") >= lower_bound) & (col("SALES") <=␣
 ↪upper_bound))


handled_df.show()
```

```
+----------+--------------+---------+--------------+-------+-----------------
```

```
--+----------+------+--------+-------+-----------+----+----------+-----------
--------+-------------+--------------------+-----------+-------------+-------
---+----------+---------+---------+--------------+---------------+--------+---
-+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES|
ORDERDATE|    STATUS|QTR_ID|MONTH_ID|YEAR_ID| PRODUCTLINE|MSRP|PRODUCTCODE|
CUSTOMERNAME|         PHONE|         ADDRESSLINE1|ADDRESSLINE2|           CITY|
STATE|POSTALCODE|
COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTFIRSTNAME|DEALSIZE|YEAR|
+-----------+--------------+---------+--------------+-------+----------------
--+----------+------+--------+-------+-----------+----+----------+-----------
--------+-------------+--------------------+-----------+-------------+-------
---+----------+---------+---------+--------------+---------------+--------+---
-+
|      10150|             45|    100.0|             8|10993.5|2003-09-19
00:00:00|   Shipped|     3|       9|   2003|Classic Cars| 214|   S10_1949|Dragon
Souveniers…|  +65 221 7555|Bronz Sok., Bronz…|        Nill|     Singapore|
Nill|     79903|Singapore|    Japan|      Natividad|            Eric|
Large|2003|
|      10174|             34|    100.0|             4|8014.82|2003-11-06
00:00:00|   Shipped|     4|      11|   2003|Classic Cars| 214|
S10_1949|Australian Gift N…|61-7-3844-6555|31 Duncan St. Wes…|
Nill|South Brisbane|Queensland|      4101|Australia|     APAC|       Calaghan|
Tony|   Large|2003|
|      10206|             47|    100.0|             6|9064.89|2003-12-05
00:00:00|   Shipped|     4|      12|   2003|Classic Cars| 214|
S10_1949|Canadian Gift Exc…|(604) 555-3392|        1900 Oak St.|        Nill|
Vancouver|        BC|   V3F 2K1|   Canada|       NA|      Tannamuri|
Yoshi|   Large|2003|
|      10280|             34|    100.0|             2|8014.82|2004-08-17
00:00:00|   Shipped|     3|       8|   2004|Classic Cars| 214|   S10_1949|
Amica Models & Co.|   011-4988555| Via Monte Bianco 34|        Nill|
Torino|      Nill|     10100|    Italy|     EMEA|        Accorti|
Paolo|   Large|2004|
|      10304|             47|    100.0|             6|10172.7|2004-10-11
00:00:00|   Shipped|     4|      10|   2004|Classic Cars| 214|   S10_1949|  Auto
Assoc. & Cie.|    30.59.8555|67, avenue de l'E…|        Nill|    Versailles|
Nill|     78000|   France|     EMEA|         Tonini|          Daniel|
Large|2004|
|      10312|             48|    100.0|             3|11623.7|2004-10-21
00:00:00|   Shipped|     4|      10|   2004|Classic Cars| 214|   S10_1949|Mini
Gifts Distri…|    4155551450|     5677 Strong St.|        Nill|    San Rafael|
CA|     97562|      USA|       NA|         Nelson|         Valarie|
Large|2004|
|      10381|             36|    100.0|             3| 8254.8|2005-02-17
00:00:00|   Shipped|     1|       2|   2005|Classic Cars| 214|
S10_1949|Corporate Gift Id…|    6505551386|     7734 Strong St.|        Nill|
San Francisco|        CA|      Nill|      USA|       NA|          Brown|
```

```
Julie|   Large|2005|
|      10424|             50|    100.0|              6|12001.0|2005-05-31
00:00:00|In Process|     2|       5|   2005|Classic Cars| 214|   S10_1949|Euro
Shopping Cha…|(91) 555 94 44|  C/ Moralzarzal, 86|        Nill|        Madrid|
Nill|     28034|    Spain|     EMEA|         Freyre|           Diego|
Large|2005|
|      10120|             46|    100.0|              2|9264.86|2003-04-29
00:00:00|   Shipped|     2|       4|   2003| Motorcycles| 193|
S10_4698|Australian Collec…|  03 9520 4555|   636 St Kilda Road|     Level 3|
Melbourne|  Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Large|2003|
|      10180|             41|    100.0|             11| 8892.9|2003-11-11
00:00:00|   Shipped|     4|      11|   2003| Motorcycles| 193|
S10_4698|Daedalus Designs …|    20.16.1555|184, chausse de T…|        Nill|
Lille|      Nill|     59000|   France|     EMEA|          Rance|
Martine|   Large|2003|
|      10188|             45|    100.0|              3| 8714.7|2003-11-18
00:00:00|   Shipped|     4|      11|   2003| Motorcycles| 193|   S10_4698|
Herkku Gifts| +47 2267 3215|Drammen 121, PR 7…|        Nill|        Bergen|
Nill|    N 5804|   Norway|     EMEA|         Oeztan|          Veysel|
Large|2003|
|      10201|             49|    100.0|              4|8065.89|2003-12-01
00:00:00|   Shipped|     4|      12|   2003| Motorcycles| 193|   S10_4698|
Mini Wheels Co.|    6505555787|5557 North Pendal…|        Nill| San Francisco|
CA|      Nill|      USA|       NA|         Murphy|           Julie|
Large|2003|
|      10223|             49|    100.0|              3|9774.03|2004-02-20
00:00:00|   Shipped|     1|       2|   2004| Motorcycles| 193|
S10_4698|Australian Collec…|  03 9520 4555|   636 St Kilda Road|     Level 3|
Melbourne|  Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Large|2004|
|      10263|             41|    100.0|              4|8336.94|2004-06-28
00:00:00|   Shipped|     2|       6|   2004| Motorcycles| 193|   S10_4698|
Gift Depot Inc.|    2035552570| 25593 South Bay Ln.|        Nill|   Bridgewater|
CT|     97562|      USA|       NA|           King|           Julie|
Large|2004|
|      10403|             66|    100.0|              9|11886.6|2005-04-08
00:00:00|   Shipped|     2|       4|   2005| Motorcycles| 193|   S10_4698|UK
Collectables, …|(171) 555-2282|Berkeley Gardens …|        Nill|
Liverpool|      Nill|   WX1 6LT|       UK|     EMEA|          Devon|
Elizabeth|   Large|2005|
|      10417|             56|    100.0|              4|9218.16|2005-05-13
00:00:00|  Disputed|     2|       5|   2005| Motorcycles| 193|   S10_4698|Euro
Shopping Cha…|(91) 555 94 44|  C/ Moralzarzal, 86|        Nill|        Madrid|
Nill|     28034|    Spain|     EMEA|         Freyre|           Diego|
Large|2005|
|      10400|             64|    100.0|              9|9661.44|2005-04-01
00:00:00|   Shipped|     2|       4|   2005|Classic Cars| 136|   S10_4757|The
```

```
Sharp Gifts W…|    4085553659|      3086 Ingle Ln.|        Nill|      San
Jose|        CA|     94217|      USA|       NA|          Frick|             Sue|
Large|2005|
|      10135|             42|    100.0|             7|8008.56|2003-07-02
00:00:00|   Shipped|     3|       7|   2003|Classic Cars| 194|   S12_1099|Mini
Gifts Distri…|    4155551450|     5677 Strong St.|        Nill|    San Rafael|
CA|     97562|      USA|       NA|         Nelson|         Valarie|
Large|2003|
|      10147|             48|    100.0|             7|9245.76|2003-09-05
00:00:00|   Shipped|     3|       9|   2003|Classic Cars| 194|
S12_1099|Collectables For …|    6175558555|    7825 Douglas Av.|        Nill|
Brickhaven|        MA|     58339|      USA|       NA|         Nelson|
Allen|   Large|2003|
|      10159|             41|    100.0|             2|8296.35|2003-10-10
00:00:00|   Shipped|     4|      10|   2003|Classic Cars| 194|
S12_1099|Corporate Gift Id…|    6505551386|     7734 Strong St.|        Nill|
San Francisco|        CA|      Nill|      USA|       NA|          Brown|
Julie|   Large|2003|
+-----------+---------------+---------+---------------+-------+-----------------
--+----------+------+--------+-------+------------+----+----------+-----------
--------+-------------+--------------------+-----------+-------------+-------
---+----------+---------+---------+---------------+---------------+--------+---
-+
only showing top 20 rows

+-----------+---------------+---------+---------------+-------+-----------------
--+-------+------+--------+-------+----------+----+----------+---------------
----+----------------+--------------------+-----------+-------------+--------+-
---------+---------+---------+---------------+---------------+--------+----+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES|
ORDERDATE| STATUS|QTR_ID|MONTH_ID|YEAR_ID|PRODUCTLINE|MSRP|PRODUCTCODE|
CUSTOMERNAME|           PHONE|        ADDRESSLINE1|ADDRESSLINE2|         CITY|
STATE|POSTALCODE|
COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTFIRSTNAME|DEALSIZE|YEAR|
+-----------+---------------+---------+---------------+-------+-----------------
--+-------+------+--------+-------+----------+----+----------+---------------
----+----------------+--------------------+-----------+-------------+--------+-
---------+---------+---------+---------------+---------------+--------+----+
|      10107|             30|     95.7|              2| 2871.0|2003-02-24
00:00:00|Shipped|     1|       2|   2003|Motorcycles|  95|   S10_1678|   Land of
Toys Inc.|      2125557818|897 Long Airport …|        Nill|          NYC|
NY|     10022|      USA|       NA|             Yu|            Kwai|
Small|2003|
|      10121|             34|    81.35|              5| 2765.9|2003-05-07
00:00:00|Shipped|     2|       5|   2003|Motorcycles|  95|   S10_1678|  Reims
Collectables|      26.47.1555|  59 rue de l'Abbaye|        Nill|        Reims|
Nill|     51100|   France|     EMEA|        Henriot|            Paul|
Small|2003|
```

```
|      10134|             41|    94.74|             2|3884.34|2003-07-01
00:00:00|Shipped|     3|       7|   2003|Motorcycles|  95|   S10_1678|     Lyon
Souveniers|+33 1 46 62 7555|27 rue du Colonel…|        Nill|        Paris|
Nill|     75508|   France|     EMEA|       Da Cunha|          Daniel|
Medium|2003|
|      10145|             45|    83.26|             6| 3746.7|2003-08-25
00:00:00|Shipped|     3|       8|   2003|Motorcycles|  95|   S10_1678|
Toys4GrownUps.com|      6265557265|  78934 Hillside Dr.|        Nill|
Pasadena|      CA|     90003|      USA|       NA|          Young|
Julie|  Medium|2003|
|      10159|             49|    100.0|            14|5205.27|2003-10-10
00:00:00|Shipped|     4|      10|   2003|Motorcycles|  95|   S10_1678|Corporate
Gift Id…|      6505551386|     7734 Strong St.|        Nill|San Francisco|
CA|      Nill|      USA|       NA|          Brown|           Julie|
Medium|2003|
|      10168|             36|    96.66|             1|3479.76|2003-10-28
00:00:00|Shipped|     4|      10|   2003|Motorcycles|  95|   S10_1678|Technics
Stores Inc.|      6505556809|   9408 Furth Circle|        Nill|   Burlingame|
CA|     94217|      USA|       NA|         Hirano|            Juri|
Medium|2003|
|      10180|             29|    86.13|             9|2497.77|2003-11-11
00:00:00|Shipped|     4|      11|   2003|Motorcycles|  95|   S10_1678|Daedalus
Designs …|      20.16.1555|184, chausse de T…|        Nill|        Lille|
Nill|     59000|   France|     EMEA|          Rance|         Martine|
Small|2003|
|      10188|             48|    100.0|             1|5512.32|2003-11-18
00:00:00|Shipped|     4|      11|   2003|Motorcycles|  95|   S10_1678|
Herkku Gifts|   +47 2267 3215|Drammen 121, PR 7…|        Nill|       Bergen|
Nill|    N 5804|   Norway|     EMEA|         Oeztan|          Veysel|
Medium|2003|
|      10201|             22|    98.57|             2|2168.54|2003-12-01
00:00:00|Shipped|     4|      12|   2003|Motorcycles|  95|   S10_1678|     Mini
Wheels Co.|      6505555787|5557 North Pendal…|        Nill|San Francisco|
CA|      Nill|      USA|       NA|         Murphy|           Julie|
Small|2003|
|      10211|             41|    100.0|            14|4708.44|2004-01-15
00:00:00|Shipped|     1|       1|   2004|Motorcycles|  95|   S10_1678|    Auto
Canal Petit|  (1) 47.55.6555|   25, rue Lauriston|        Nill|        Paris|
Nill|     75016|   France|     EMEA|        Perrier|       Dominique|
Medium|2004|
|      10223|             37|    100.0|             1|3965.66|2004-02-20
00:00:00|Shipped|     1|       2|   2004|Motorcycles|  95|   S10_1678|Australian
Collec…|    03 9520 4555|   636 St Kilda Road|     Level 3|
Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|2004|
|      10237|             23|    100.0|             7|2333.12|2004-04-05
00:00:00|Shipped|     2|       4|   2004|Motorcycles|  95|   S10_1678|
Vitachrome Inc.|      2125551500|   2678 Kingston Rd.|   Suite 101|
```

```
NYC|      NY|     10022|      USA|       NA|          Frick|         Michael|
Small|2004|
|      10251|             28|    100.0|             2|3188.64|2004-05-18
00:00:00|Shipped|     2|       5|   2004|Motorcycles|  95|   S10_1678|Tekni
Collectable…|      2015559350|       7476 Moss Rd.|        Nill|       Newark|
NJ|     94019|      USA|       NA|          Brown|         William|
Medium|2004|
|      10263|             34|    100.0|             2|3676.76|2004-06-28
00:00:00|Shipped|     2|       6|   2004|Motorcycles|  95|   S10_1678|     Gift
Depot Inc.|      2035552570| 25593 South Bay Ln.|        Nill|  Bridgewater|
CT|     97562|      USA|       NA|           King|           Julie|
Medium|2004|
|      10275|             45|    92.83|             1|4177.35|2004-07-23
00:00:00|Shipped|     3|       7|   2004|Motorcycles|  95|   S10_1678|   La
Rochelle Gifts|      40.67.8555|67, rue des Cinqu…|        Nill|       Nantes|
Nill|     44000|   France|     EMEA|        Labrune|          Janine|
Medium|2004|
|      10285|             36|    100.0|             6|4099.68|2004-08-27
00:00:00|Shipped|     3|       8|   2004|Motorcycles|  95|   S10_1678|Marta's
Replicas Co.|      6175558555| 39323 Spinnaker Dr.|        Nill|    Cambridge|
MA|     51247|      USA|       NA|      Hernandez|           Marta|
Medium|2004|
|      10299|             23|    100.0|             9|2597.39|2004-09-30
00:00:00|Shipped|     3|       9|   2004|Motorcycles|  95|   S10_1678|Toys of
Finland, Co.|     90-224 8555|       Keskuskatu 45|        Nill|     Helsinki|
Nill|     21240|  Finland|     EMEA|      Karttunen|           Matti|
Small|2004|
|      10309|             41|    100.0|             5|4394.38|2004-10-15
00:00:00|Shipped|     4|      10|   2004|Motorcycles|  95|   S10_1678|  Baane
Mini Imports|      07-98 9555|Erling Skakkes ga…|        Nill|      Stavern|
Nill|      4110|   Norway|     EMEA|     Bergulfsen|           Jonas|
Medium|2004|
|      10318|             46|    94.74|             1|4358.04|2004-11-02
00:00:00|Shipped|     4|      11|   2004|Motorcycles|  95|   S10_1678|Diecast
Classics …|      2155551555|    7586 Pompton St.|        Nill|    Allentown|
PA|     70267|      USA|       NA|             Yu|           Kyung|
Medium|2004|
|      10329|             42|    100.0|             1|4396.14|2004-11-15
00:00:00|Shipped|     4|      11|   2004|Motorcycles|  95|   S10_1678|   Land of
Toys Inc.|      2125557818|897 Long Airport …|        Nill|          NYC|
NY|     10022|      USA|       NA|             Yu|           Kwai|
Medium|2004|
+-----------+--------------+---------+--------------+-------+----------------
--+-------+------+--------+-------+-----------+----+-----------+---------------
----+----------------+--------------------+------------+-------------+--------+-
---------+---------+---------+--------------+---------------+--------+----+
only showing top 20 rows
```

r) How would you cache a DataFrame containing sales data from the top 10 countries by sales to avoid recomputation in subsequent transformations? What persistence level (e.g. MEMORY_ONLY, MEMORY_AND_DISK) would you choose and why?

[106]:
```
from pyspark.storagelevel import StorageLevel
sales_country = sales_df1.groupBy("COUNTRY").agg(sum("SALES")
                                        .alias("TOTAL_SALES")).
 ↪orderBy("TOTAL_SALES", ascending=False).limit(10)
sales_country.show(5)
sales_country.cache()
sales_country.persist(StorageLevel.MEMORY_AND_DISK)
```

```
+---------+------------------+
|  COUNTRY|       TOTAL_SALES|
+---------+------------------+
|      USA|        3627982.83|
|    Spain|1215686.9200000009|
|   France|1110916.5199999993|
|Australia| 630623.1000000001|
|       UK| 478880.4600000001|
+---------+------------------+
only showing top 5 rows
```

[106]:
```
DataFrame[COUNTRY: string, TOTAL_SALES: double]
```

Recommended Persistence Level For caching the DataFrame of sales data from the top 10 countries, I would recommend using MEMORY_AND_DISK. This choice is suitable because:

```
Data Size: The DataFrame might be relatively small after filtering to the top 10 countries, so
Resilience: If the DataFrame size grows or if you face memory constraints, it will spill to dis
Performance Balance: Provides a good balance between performance (memory access) and reliabilit
```

s) How would you pivot the data to show PRODUCTLINE as columns and the total SALES for each ORDERDATE as the values? What are the implications of pivoting large datasets in Spark?

[141]:
```
pivot_df = sales_df1.groupBy("ORDERDATE", "PRODUCTLINE").agg(sum("SALES").
 ↪alias("TOTAL_SALES"))
pivot_df.groupBy("ORDERDATE").pivot("PRODUCTLINE").sum("TOTAL_SALES").show()
```

```
+-------------------+-----------------+-----------------+------------------+--
---------------+-------+---------------+------------+
|          ORDERDATE|      Classic Cars|       Motorcycles|            Planes|
Ships| Trains|Trucks and Buses|Vintage Cars|
+-------------------+-----------------+-----------------+------------------+--
---------------+-------+---------------+------------+
|2005-03-02 00:00:00|             null|           4175.6|              null|
null|   null|           null|        null|
```

```
|2004-11-09 00:00:00|              null|              null|              null|
6673.29|3807.68|            null|     9665.35|
|2005-05-03 00:00:00|25040.629999999997|              null|              null|
null|   null|        27247.11|        null|
|2003-09-11 00:00:00| 43593.54000000001|              null|              null|
null|   null|            null|     3598.22|
|2005-04-01 00:00:00|           9661.44|              null|           9036.06|
6284.0|   null|            null|    12545.34|
|2005-05-10 00:00:00|              null|            7567.8|30429.010000000002|
null|   null|            null|     3513.13|
|2004-04-26 00:00:00|              null|              null|              null|
null|   null|            null|      7129.0|
|2003-10-17 00:00:00| 40321.60999999999|              null|              null|
null|   null|            null|        null|
|2004-09-27 00:00:00|              null|5307.9800000000005|              null|
null|   null|            null|        null|
|2003-10-20 00:00:00|           4860.24|              null|              null|
null|   null|            null|    20424.51|
|2004-04-09 00:00:00|          31329.56|              null|              null|
null|   null|            null|        null|
|2005-05-06 00:00:00|           2764.88|              null|
null|23664.609999999997|5808.48|            null|    16078.92|
|2003-02-17 00:00:00|              null|              null|39205.310000000005|
6598.34|   null|            null|    10377.67|
|2004-12-07 00:00:00|          19489.57|10394.560000000001|              null|
null|   null|            null|        null|
|2005-04-22 00:00:00|          40207.85|              null|              null|
null|   null|            null|    18229.19|
|2003-12-01 00:00:00|              null|25431.879999999997|           7120.96|
null|   null|            null|      1113.6|
|2003-01-06 00:00:00|              null|              null|              null|
null|   null|            null|    12133.25|
|2004-05-18 00:00:00|              null|          27987.07|              null|
null|   null|            null|        null|
|2005-01-05 00:00:00|              null|          13529.57|              null|
null|   null|            null|        null|
|2004-11-17 00:00:00| 47423.57000000001|12786.920000000002|              null|
null|   null|            null|    37747.89|
+-------------------+------------------+------------------+------------------+--
---------------+-------+---------------+------------+
only showing top 20 rows
```

Implications of Pivoting Large Datasets

* Memory Usage: Pivoting reshapes data, creating many new columns for each unique pivot value, which can lead to high memory consumption.

* Performance: The operation can be slow with large datasets or many unique pivot values due to increased computational complexity.

* Column Explosion: A large number of unique pivot values can result in a DataFrame with many c making it unwieldy and hard to manage.

* Data Skew: Uneven distribution of data (e.g., some dates having much more sales) can cause performance issues due to uneven partitioning.

* Shuffling: Pivoting involves shuffling data across nodes, which can be expensive in terms of time and resources.

* Disk I/O: If the data exceeds memory capacity, Spark will spill data to disk, potentially slowing down the process.

t) How would you calculate the percentage growth of total sales month over month for each PRODUCTLINE using Spark DataFrame?

[110]:
```
from pyspark.sql.functions import *
from pyspark.sql.window import Window


salesData = sales_df1.withColumn("year", year("ORDERDATE"))
salesData = salesData.withColumn("month", month("ORDERDATE"))


monthlySales = salesData.groupBy("year", "month", "PRODUCTLINE").
 ↪agg(sum("SALES").alias("total_sales"))

windowSpec = Window.partitionBy("PRODUCTLINE").orderBy("year", "month")

monthlySales = monthlySales.withColumn("previous_month_sales",␣
 ↪lag("total_sales").over(windowSpec))

monthlySales = monthlySales.withColumn(
    "percentage_growth",
    when(
        col("previous_month_sales").isNotNull(),
        (col("total_sales") - col("previous_month_sales")) /␣
 ↪col("previous_month_sales") * 100
    ).otherwise(lit(None))
)

monthlySales.show()
```

```
+----+-----+-----------+-----------------+-------------------+---------------
---+
|year|month|PRODUCTLINE|      total_sales|previous_month_sales|
```

```
percentage_growth|
+----+-----+-----------+-----------------+-------------------+---------------
---+
|2003|    2|Motorcycles|25783.760000000002|               null|
null|
|2003|    3|Motorcycles|          12639.15|  25783.760000000002|
-50.98019063162239|
|2003|    4|Motorcycles|23475.590000000004|            12639.15|
85.7370946622202|
|2003|    5|Motorcycles|          22097.32|
23475.590000000004|-5.8710771486467594|
|2003|    6|Motorcycles|           2642.01|            22097.32|
-88.04375372217082|
|2003|    7|Motorcycles| 37924.23000000001|             2642.01|
1335.430978686682|
|2003|    8|Motorcycles|44164.909999999996|   37924.23000000001|
16.455653812878953|
|2003|    9|Motorcycles|           3155.58|  44164.909999999996|
-92.85500638402749|
|2003|   10|Motorcycles| 64235.65000000001|             3155.58|
1935.6210268793695|
|2003|   11|Motorcycles|          109345.5|   64235.65000000001|
70.22556788948191|
|2003|   12|Motorcycles|25431.879999999997|            109345.5|
-76.74172233882508|
|2004|    1|Motorcycles|          41200.52|  25431.879999999997|
62.00343820433252|
|2004|    2|Motorcycles|           49066.5|            41200.52|
19.091943499742246|
|2004|    4|Motorcycles| 36269.07000000001|             49066.5|
-26.08180734309558|
|2004|    5|Motorcycles|46848.950000000004|   36269.07000000001|
29.17053015144859|
|2004|    6|Motorcycles|          47237.41|  46848.950000000004|
0.8291754671129217|
|2004|    7|Motorcycles|           22774.0|
47237.41|-51.788211927791984|
|2004|    8|Motorcycles|          62704.93|             22774.0|
175.3356020022833|
|2004|    9|Motorcycles| 42471.04999999999|            62704.93|
-32.26840377622623|
|2004|   10|Motorcycles|          39413.96|
42471.04999999999|-7.1980560876173065|
+----+-----+-----------+-----------------+-------------------+---------------
---+
only showing top 20 rows
```

u) How can you rebalance the data by portioning based on the COUNTRY column to ensure that large data partitions are avoided?

[126]:
```
unique_country_count = sales_df1.select("COUNTRY").distinct().count()


repartitioned_df = sales_df1.repartition(unique_country_count, "COUNTRY")


print(f"No of partitions: {repartitioned_df.rdd.getNumPartitions()}")
```

```
No of partitions: 19
```

v) Suppose you have a smaller lookup table with customer details. How would you perform a broadcast join with the large sales_data_sample dataset to improve join performance? What are the key considerations when using broadcast joins?

[121]:
```
# Example data for customer details
customer_data = [
    (1, "Land of Toys Inc.", "USA"),
    (2, "Reims Collectables", "Canada"),
    (3, "Lyon Souveniers", "Mexico")
]

customer_columns = ["customer_id", "name", "country"]

customer_details = spark.createDataFrame(customer_data, customer_columns)

from pyspark.sql.functions import broadcast


broadcast_customer_details = broadcast(customer_details)


joined_df = sales_df1.join(broadcast_customer_details,
                                    sales_df1["CUSTOMERNAME"] ==␣
 ↪broadcast_customer_details["name"])


joined_df.show()
```

```
+----------+--------------+---------+--------------+-------+----------------
--+---------+------+--------+-------+-----------+----+----------+-------------
-----+----------------+--------------------+-----------+-----+-----+----------+
-------+---------+---------------+---------------+--------+----+--------+------
-----+-----------------+-------+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES|
ORDERDATE|   STATUS|QTR_ID|MONTH_ID|YEAR_ID| PRODUCTLINE|MSRP|PRODUCTCODE|
CUSTOMERNAME|           PHONE|        ADDRESSLINE1|ADDRESSLINE2| CITY|STATE|POST
```

```
ALCODE|COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTFIRSTNAME|DEALSIZE|YEAR|CATEGORY
|customer_id|              name|country|
+----------+--------------+---------+--------------+-------+----------------
--+---------+------+--------+-------+------------+----+-----------+-------------
-----+----------------+--------------------+------------+-----+-----+----------+
-------+---------+--------------+---------------+--------+----+--------+------
-----+-----------------+-------+
|      10107|             30|     95.7|             2| 2871.0|2003-02-24
00:00:00|  Shipped|     1|       2|   2003| Motorcycles|  95|   S10_1678| Land
of Toys Inc.|      2125557818|897 Long Airport …|        Nill|  NYC|   NY|
10022|    USA|       NA|             Yu|            Kwai|   Small|2003|  Medium|
1| Land of Toys Inc.|    USA|
|      10121|             34|    81.35|             5| 2765.9|2003-05-07
00:00:00|  Shipped|     2|       5|   2003| Motorcycles|  95|   S10_1678|Reims
Collectables|      26.47.1555|  59 rue de l'Abbaye|        Nill|Reims| Nill|
51100| France|     EMEA|        Henriot|            Paul|   Small|2003|  Medium|
2|Reims Collectables| Canada|
|      10134|             41|    94.74|             2|3884.34|2003-07-01
00:00:00|  Shipped|     3|       7|   2003| Motorcycles|  95|   S10_1678|   Lyon
Souveniers|+33 1 46 62 7555|27 rue du Colonel…|        Nill|Paris| Nill|
75508| France|     EMEA|       Da Cunha|          Daniel|  Medium|2003|  Medium|
3|   Lyon Souveniers| Mexico|
|      10329|             42|    100.0|             1|4396.14|2004-11-15
00:00:00|  Shipped|     4|      11|   2004| Motorcycles|  95|   S10_1678| Land
of Toys Inc.|      2125557818|897 Long Airport …|        Nill|  NYC|   NY|
10022|    USA|       NA|             Yu|            Kwai|  Medium|2004|    High|
1| Land of Toys Inc.|    USA|
|      10107|             39|    99.91|             5|3896.49|2003-02-24
00:00:00|  Shipped|     1|       2|   2003| Motorcycles| 118|   S10_2016| Land
of Toys Inc.|      2125557818|897 Long Airport …|        Nill|  NYC|   NY|
10022|    USA|       NA|             Yu|            Kwai|  Medium|2003|  Medium|
1| Land of Toys Inc.|    USA|
|      10134|             27|    100.0|             5|3307.77|2003-07-01
00:00:00|  Shipped|     3|       7|   2003| Motorcycles| 118|   S10_2016|   Lyon
Souveniers|+33 1 46 62 7555|27 rue du Colonel…|        Nill|Paris| Nill|
75508| France|     EMEA|       Da Cunha|          Daniel|  Medium|2003|  Medium|
3|   Lyon Souveniers| Mexico|
|      10329|             20|    100.0|             2| 3176.0|2004-11-15
00:00:00|  Shipped|     4|      11|   2004| Motorcycles| 118|   S10_2016| Land
of Toys Inc.|      2125557818|897 Long Airport …|        Nill|  NYC|   NY|
10022|    USA|       NA|             Yu|            Kwai|  Medium|2004|  Medium|
1| Land of Toys Inc.|    USA|
|      10107|             27|    100.0|             4|6065.55|2003-02-24
00:00:00|  Shipped|     1|       2|   2003| Motorcycles| 193|   S10_4698| Land
of Toys Inc.|      2125557818|897 Long Airport …|        Nill|  NYC|   NY|
10022|    USA|       NA|             Yu|            Kwai|  Medium|2003|    High|
1| Land of Toys Inc.|    USA|
|      10134|             31|    100.0|             4|7023.98|2003-07-01
```

00:00:00| Shipped| 3| 7| 2003| Motorcycles| 193| S10_4698| Lyon Souveniers|+33 1 46 62 7555|27 rue du Colonel…| Nill|Paris| Nill| 75508| France| EMEA| Da Cunha| Daniel| Large|2003| High| 3| Lyon Souveniers| Mexico|
10329	26	100.0	3	5868.2	2004-11-15 00:00:00	Shipped	4	11	2004	Motorcycles	193	S10_4698	Land of Toys Inc.	2125557818	897 Long Airport …	Nill	NYC	NY	10022	USA	NA	Yu	Kwai	Medium	2004	High	1	Land of Toys Inc.	USA
10248	20	100.0	3	2910.4	2004-05-07 00:00:00	Cancelled	2	5	2004	Classic Cars	136	S10_4757	Land of Toys Inc.	2125557818	897 Long Airport …	Nill	NYC	NY	10022	USA	NA	Yu	Kwai	Small	2004	Medium	1	Land of Toys Inc.	USA
10359	48	54.68	6	2624.64	2004-12-15 00:00:00	Shipped	4	12	2004	Classic Cars	136	S10_4757	Reims Collectables	26.47.1555	59 rue de l'Abbaye	Nill	Reims	Nill	51100	France	EMEA	Henriot	Paul	Small	2004	Medium	2	Reims Collectables	Canada
10395	32	100.0	2	3370.56	2005-03-17 00:00:00	Shipped	1	3	2005	Classic Cars	136	S10_4757	Lyon Souveniers	+33 1 46 62 7555	27 rue du Colonel…	Nill	Paris	Nill	75508	France	EMEA	Da Cunha	Daniel	Medium	2005	Medium	3	Lyon Souveniers	Mexico
10329	41	71.47	5	2930.27	2004-11-15 00:00:00	Shipped	4	11	2004	Classic Cars	194	S12_1099	Land of Toys Inc.	2125557818	897 Long Airport …	Nill	NYC	NY	10022	USA	NA	Yu	Kwai	Small	2004	Medium	1	Land of Toys Inc.	USA
10359	42	100.0	8	4764.48	2004-12-15 00:00:00	Shipped	4	12	2004	Classic Cars	207	S12_1108	Reims Collectables	26.47.1555	59 rue de l'Abbaye	Nill	Reims	Nill	51100	France	EMEA	Henriot	Paul	Medium	2004	High	2	Reims Collectables	Canada
10395	33	69.12	1	2280.96	2005-03-17 00:00:00	Shipped	1	3	2005	Classic Cars	207	S12_1108	Lyon Souveniers	+33 1 46 62 7555	27 rue du Colonel…	Nill	Paris	Nill	75508	France	EMEA	Da Cunha	Daniel	Small	2005	Low	3	Lyon Souveniers	Mexico
10107	21	100.0	1	3036.6	2003-02-24 00:00:00	Shipped	1	2	2003	Motorcycles	150	S12_2823	Land of Toys Inc.	2125557818	897 Long Airport …	Nill	NYC	NY	10022	USA	NA	Yu	Kwai	Medium	2003	Medium	1	Land of Toys Inc.	USA
10121	50	100.0	4	8284.0	2003-05-07 00:00:00	Shipped	2	5	2003	Motorcycles	150	S12_2823	Reims Collectables	26.47.1555	59 rue de l'Abbaye	Nill	Reims	Nill	51100	France	EMEA	Henriot	Paul	Large	2003	High

```
2|Reims Collectables| Canada|
|      10134|             20|    100.0|              1| 2711.2|2003-07-01
00:00:00|  Shipped|     3|       7|   2003| Motorcycles| 150|   S12_2823|   Lyon
Souveniers|+33 1 46 62 7555|27 rue du Colonel…|        Nill|Paris| Nill|
75508| France|     EMEA|       Da Cunha|          Daniel|   Small|2003|  Medium|
3|   Lyon Souveniers| Mexico|
|      10329|             24|    100.0|              6|3542.64|2004-11-15
00:00:00|  Shipped|     4|      11|   2004| Motorcycles| 150|   S12_2823| Land
of Toys Inc.|      2125557818|897 Long Airport …|        Nill|  NYC|   NY|
10022|    USA|       NA|             Yu|            Kwai|  Medium|2004|  Medium|
1| Land of Toys Inc.|    USA|
+-----------+---------------+---------+---------------+-------+-----------------
--+---------+------+--------+-------+-----------+----+----------+-------------
-----+----------------+--------------------+-----------+-----+-----+----------+
-------+---------+---------------+---------------+--------+----+--------+------
-----+------------------+--------+
only showing top 20 rows
```

w) Create a UDF that categorizes the sales values (SALES) into custom buckets like “Low”, “Medium”, “High”. Apply this UDF to the DataFrame and calculate the count of orders in each category per COUNTRY.

[122]:
```
from pyspark.sql.functions import udf
from pyspark.sql.types import StringType

def categorize_sales(ntile):
    if ntile == 1:
        return "Low"
    elif ntile == 2:
        return "Medium"
    else:
        return "High"

categorize_sales_udf = udf(categorize_sales, StringType())
window_spec = Window.orderBy(col("SALES"))
sales_df2 =sales_df2.withColumn("NTILE", ntile(3).over(window_spec))
df_categorized = sales_df2.withColumn('SALES_CATEGORY',␣
 ↪categorize_sales_udf(col('NTILE')))
df_categorized.groupBy('COUNTRY', 'SALES_CATEGORY').count().show()
```

```
+-----------+--------------+-----+
|    COUNTRY|SALES_CATEGORY|count|
+-----------+--------------+-----+
|Philippines|           Low|    7|
|     Norway|        Medium|   22|
|        USA|           Low|  319|
|    Austria|           Low|   17|
```

```
|    Denmark|           Low|   18|
|     Canada|          High|   18|
|     Canada|        Medium|   24|
|     Canada|           Low|   28|
|      Italy|          High|   32|
|Switzerland|        Medium|   12|
|    Ireland|          High|    8|
|    Ireland|           Low|    7|
|Philippines|          High|    9|
|  Australia|          High|   57|
|  Singapore|           Low|   28|
|    Finland|        Medium|   34|
|     Norway|          High|   33|
|        USA|        Medium|  338|
|      Italy|           Low|   38|
|        USA|          High|  347|
+-----------+--------------+-----+
only showing top 20 rows
```

x) Create a Python UDF to calculate discounts for specific product lines. For example, give a 10% discount for Classic Cars and 5% for Motorcycles. Apply this UDF to derive new discounted sales values.

[124]:
```python
from pyspark.sql.functions import udf, col
from pyspark.sql.types import FloatType


def calculate_discounted_sales(product_line, sales):
    if product_line == 'Classic Cars':
        discount = 0.10
    elif product_line == 'Motorcycles':
        discount = 0.05
    else:
        discount = 0.0
    return float(sales * (1 - discount))


calculate_discounted_sales_udf = udf(calculate_discounted_sales, FloatType())


sales_df2 = sales_df2.withColumn(
    'DISCOUNT_SALES',
    calculate_discounted_sales_udf(col('PRODUCTLINE'), col('SALES'))
)


sales_df2.select(col('PRODUCTLINE'), col('SALES'), col('DISCOUNT_SALES')).show()
```

```
+-----------+-------+--------------+
|PRODUCTLINE|  SALES|DISCOUNT_SALES|
+-----------+-------+--------------+
|Motorcycles| 2871.0|        2727.45|
|Motorcycles| 2765.9|       2627.605|
|Motorcycles|3884.34|       3690.123|
|Motorcycles| 3746.7|       3559.365|
|Motorcycles|5205.27|      4945.0063|
|Motorcycles|3479.76|       3305.772|
|Motorcycles|2497.77|      2372.8816|
|Motorcycles|5512.32|       5236.704|
|Motorcycles|2168.54|       2060.113|
|Motorcycles|4708.44|       4473.018|
|Motorcycles|3965.66|       3767.377|
|Motorcycles|2333.12|       2216.464|
|Motorcycles|3188.64|       3029.208|
|Motorcycles|3676.76|       3492.922|
|Motorcycles|4177.35|      3968.4824|
|Motorcycles|4099.68|       3894.696|
|Motorcycles|2597.39|      2467.5205|
|Motorcycles|4394.38|       4174.661|
|Motorcycles|4358.04|       4140.138|
|Motorcycles|4396.14|       4176.333|
+-----------+-------+--------------+
only showing top 20 rows
```

z) How do you implement a cumulative distribution function (CDF) over the SALES value for each CUSTOMERNAME? What insights can you gather from analyzing the CDF distribution for each customer?

[142]:
```
window_spec = Window.partitionBy('CUSTOMERNAME').orderBy('SALES')
df_with_rank = sales_df1.withColumn('rank', row_number().over(window_spec))
df_with_count = df_with_rank.withColumn('total_count', count('SALES').
  ↪over(Window.partitionBy('CUSTOMERNAME')))
df_cdf = df_with_count.withColumn('CDF', col('rank') / col('total_count'))
df_cdf = df_cdf.select('CUSTOMERNAME', 'SALES', 'CDF').show(50)
```

```
+------------------+-------+-------------------+
|      CUSTOMERNAME|  SALES|                CDF|
+------------------+-------+-------------------+
|Suominen Souveniers| 891.03| 0.03333333333333333|
|Suominen Souveniers| 1086.6| 0.06666666666666667|
|Suominen Souveniers|1103.76|                0.1|
|Suominen Souveniers|1629.04| 0.13333333333333333|
|Suominen Souveniers| 1988.4| 0.16666666666666666|
|Suominen Souveniers|2140.11|                0.2|
|Suominen Souveniers|2447.76| 0.23333333333333334|
```

```
|Suominen Souveniers|2632.89| 0.26666666666666666|
|Suominen Souveniers| 2773.8|                   0.3|
|Suominen Souveniers|2775.08|  0.3333333333333333|
|Suominen Souveniers|2817.87| 0.36666666666666664|
|Suominen Souveniers|2851.84|                   0.4|
|Suominen Souveniers|2931.98| 0.43333333333333335|
|Suominen Souveniers|3128.65|  0.4666666666666667|
|Suominen Souveniers|3288.82|                   0.5|
|Suominen Souveniers|3595.62|  0.5333333333333333|
|Suominen Souveniers|3686.54|  0.5666666666666667|
|Suominen Souveniers| 3784.8|                   0.6|
|Suominen Souveniers| 4068.7|  0.6333333333333333|
|Suominen Souveniers|4142.64|  0.6666666666666666|
|Suominen Souveniers|4157.73|                   0.7|
|Suominen Souveniers|4381.25|  0.7333333333333333|
|Suominen Souveniers| 4836.5|  0.7666666666666667|
|Suominen Souveniers|5154.41|                   0.8|
|Suominen Souveniers|5500.44|  0.8333333333333334|
|Suominen Souveniers|5938.53|  0.8666666666666667|
|Suominen Souveniers|6287.66|                   0.9|
|Suominen Souveniers| 6576.5|  0.9333333333333333|
|Suominen Souveniers| 6756.0|  0.9666666666666667|
|Suominen Souveniers|10606.2|                   1.0|
| Amica Models & Co.|  577.6|0.038461538461538464|
| Amica Models & Co.|1381.05| 0.07692307692307693|
| Amica Models & Co.|1557.36| 0.11538461538461539|
| Amica Models & Co.| 1574.0| 0.15384615384615385|
| Amica Models & Co.|1656.69| 0.19230769230769232|
| Amica Models & Co.|1921.92| 0.23076923076923078|
| Amica Models & Co.|2084.81|  0.2692307692307692|
| Amica Models & Co.|2137.05|  0.3076923076923077|
| Amica Models & Co.|2418.24| 0.34615384615384615|
| Amica Models & Co.|2800.08| 0.38461538461538464|
| Amica Models & Co.|2819.28|  0.4230769230769231|
| Amica Models & Co.|2941.89| 0.46153846153846156|
| Amica Models & Co.|2954.53|                   0.5|
| Amica Models & Co.|3006.43|  0.5384615384615384|
| Amica Models & Co.|3474.46|  0.5769230769230769|
| Amica Models & Co.| 3668.6|  0.6153846153846154|
| Amica Models & Co.|3704.05|  0.6538461538461539|
| Amica Models & Co.|4242.24|  0.6923076923076923|
| Amica Models & Co.| 4455.0|  0.7307692307692307|
| Amica Models & Co.| 4750.8|  0.7692307692307693|
+-------------------+-------+--------------------+
only showing top 50 rows
```

Insights

* CDF values show how sales are distributed within each customer group.
* Sales Concentration for a customer like "Suominen Souveniers," we can see that a significant
* For example, 0.7 CDF corresponds to a sales value of 4381.25, meaning that 70% of the sales

[95]:
```
sales_df1.printSchema()
```

```
root
 |-- ORDERNUMBER: integer (nullable = true)
 |-- QUANTITYORDERED: integer (nullable = true)
 |-- PRICEEACH: double (nullable = true)
 |-- ORDERLINENUMBER: integer (nullable = true)
 |-- SALES: double (nullable = true)
 |-- ORDERDATE: timestamp (nullable = true)
 |-- STATUS: string (nullable = false)
 |-- QTR_ID: integer (nullable = true)
 |-- MONTH_ID: integer (nullable = true)
 |-- YEAR_ID: integer (nullable = true)
 |-- PRODUCTLINE: string (nullable = false)
 |-- MSRP: integer (nullable = true)
 |-- PRODUCTCODE: string (nullable = false)
 |-- CUSTOMERNAME: string (nullable = false)
 |-- PHONE: string (nullable = false)
 |-- ADDRESSLINE1: string (nullable = false)
 |-- ADDRESSLINE2: string (nullable = false)
 |-- CITY: string (nullable = false)
 |-- STATE: string (nullable = false)
 |-- POSTALCODE: string (nullable = false)
 |-- COUNTRY: string (nullable = false)
 |-- TERRITORY: string (nullable = false)
 |-- CONTACTLASTNAME: string (nullable = false)
 |-- CONTACTFIRSTNAME: string (nullable = false)
 |-- DEALSIZE: string (nullable = false)
 |-- YEAR: integer (nullable = true)
 |-- CATEGORY: string (nullable = false)
```

[96]:
```
sales_df1.show(10)
```

```
+-----------+---------------+---------+---------------+-------+-----------------
--+-------+------+--------+-------+-----------+----+-----------+----------------
----+---------------+-------------------+-----------+-------------+-----+----
------+-------+---------+---------------+---------------+--------+----+--------
+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES|
ORDERDATE| STATUS|QTR_ID|MONTH_ID|YEAR_ID|PRODUCTLINE|MSRP|PRODUCTCODE|
CUSTOMERNAME|           PHONE|        ADDRESSLINE1|ADDRESSLINE2|         CITY|ST
ATE|POSTALCODE|COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTFIRSTNAME|DEALSIZE|YEAR|
CATEGORY|
```

```
+----------+--------------+--------+--------------+-------+----------------
--+-------+------+--------+-------+----------+----+----------+---------------
----+---------------+-------------------+-----------+------------+-----+----
------+-------+---------+--------------+---------------+--------+----+--------
+
|      10107|             30|     95.7|             2| 2871.0|2003-02-24
00:00:00|Shipped|     1|       2|   2003|Motorcycles|  95|   S10_1678|   Land of
Toys Inc.|      2125557818|897 Long Airport …|        Nill|          NYC|
NY|     10022|    USA|       NA|             Yu|            Kwai|   Small|2003|
Medium|
|      10121|             34|    81.35|             5| 2765.9|2003-05-07
00:00:00|Shipped|     2|       5|   2003|Motorcycles|  95|   S10_1678|  Reims
Collectables|      26.47.1555|  59 rue de l'Abbaye|        Nill|        Reims|
Nill|     51100| France|     EMEA|        Henriot|            Paul|
Small|2003|  Medium|
|      10134|             41|    94.74|             2|3884.34|2003-07-01
00:00:00|Shipped|     3|       7|   2003|Motorcycles|  95|   S10_1678|     Lyon
Souveniers|+33 1 46 62 7555|27 rue du Colonel…|        Nill|        Paris|
Nill|     75508| France|     EMEA|       Da Cunha|          Daniel|
Medium|2003|  Medium|
|      10145|             45|    83.26|             6| 3746.7|2003-08-25
00:00:00|Shipped|     3|       8|   2003|Motorcycles|  95|   S10_1678|
Toys4GrownUps.com|      6265557265|  78934 Hillside Dr.|        Nill|
Pasadena|   CA|     90003|    USA|       NA|          Young|           Julie|
Medium|2003|  Medium|
|      10159|             49|    100.0|            14|5205.27|2003-10-10
00:00:00|Shipped|     4|      10|   2003|Motorcycles|  95|   S10_1678|Corporate
Gift Id…|      6505551386|     7734 Strong St.|        Nill|San Francisco|
CA|      Nill|    USA|       NA|          Brown|           Julie|  Medium|2003|
High|
|      10168|             36|    96.66|             1|3479.76|2003-10-28
00:00:00|Shipped|     4|      10|   2003|Motorcycles|  95|   S10_1678|Technics
Stores Inc.|      6505556809|   9408 Furth Circle|        Nill|   Burlingame|
CA|     94217|    USA|       NA|         Hirano|            Juri|  Medium|2003|
Medium|
|      10180|             29|    86.13|             9|2497.77|2003-11-11
00:00:00|Shipped|     4|      11|   2003|Motorcycles|  95|   S10_1678|Daedalus
Designs …|      20.16.1555|184, chausse de T…|        Nill|        Lille|
Nill|     59000| France|     EMEA|          Rance|         Martine|
Small|2003|     Low|
|      10188|             48|    100.0|             1|5512.32|2003-11-18
00:00:00|Shipped|     4|      11|   2003|Motorcycles|  95|   S10_1678|
Herkku Gifts|   +47 2267 3215|Drammen 121, PR 7…|        Nill|       Bergen|
Nill|    N 5804| Norway|     EMEA|         Oeztan|          Veysel|
Medium|2003|    High|
|      10201|             22|    98.57|             2|2168.54|2003-12-01
00:00:00|Shipped|     4|      12|   2003|Motorcycles|  95|   S10_1678|     Mini
Wheels Co.|      6505555787|5557 North Pendal…|        Nill|San Francisco|
```

```
CA|      Nill|    USA|       NA|         Murphy|           Julie|   Small|2003|
Low|
|      10211|             41|    100.0|             14|4708.44|2004-01-15
00:00:00|Shipped|     1|       1|   2004|Motorcycles|  95|   S10_1678|    Auto
Canal Petit|  (1) 47.55.6555|   25, rue Lauriston|        Nill|        Paris|
Nill|     75016| France|     EMEA|        Perrier|       Dominique|
Medium|2004|    High|
+-----------+--------------+---------+--------------+-------+----------------
--+-------+------+--------+-------+-----------+----+-----------+---------------
----+----------------+--------------------+------------+-------------+-----+----
------+-------+---------+--------------+---------------+--------+----+--------
+
only showing top 10 rows
```

[ ]: