task2-spark-analysis

July 10, 2025

1 Task 2: Data Analysis Using Apache Spark

Objective:
Analyze the `SampleSuperstore.csv` dataset using Apache Spark: - Perform filtering, grouping, and aggregation - Gain insights into sales, profit, and product distribution

Concept	Function	Example
Filter	`filter()` / `where()`	Select high-profit orders
Group By	`groupBy()`	Group sales by region/category
Aggregate	`sum()`, `avg()`, `count()`	Calculate totals, averages
Sort	`orderBy()`	Show top-performing categories or products

[ ]:
```
from pyspark.sql import SparkSession
from pyspark.sql.functions import sum, avg, count

spark = SparkSession.builder.appName("SuperstoreAnalysis").getOrCreate()
```

1.1 Step 1: Load the Cleaned Dataset

[ ]:
```
import os
print(os.path.exists("/content/Sample - Superstore.csv"))
```

```
True
```

[ ]:
```
df_clean = spark.read.csv("/content/Sample - Superstore.csv", header=True,␣
↪inferSchema=True)
df_clean = df_clean.dropna().dropDuplicates().withColumnRenamed("Postal Code",␣
↪"Postal_Code")
df_clean.show(5)
```

```
+------+--------------+----------+---------+--------------+-----------+---------
-------+-----------+-------------+-------------+-------------+-----------+-------
+--------------+---------------+------------+--------------------+-------+-----
---+--------+-------+
|Row ID|      Order ID|Order Date|Ship Date|     Ship Mode|Customer ID|
Customer Name|    Segment|      Country|         City|       State|Postal_Code|
```

```
Region|     Product ID|       Category|Sub-Category|        Product Name|
Sales|Quantity|Discount| Profit|
+------+--------------+---------+---------+--------------+-----------+--------
-------+-----------+-------------+-------------+------------+-----------+-------
+---------------+---------------+------------+--------------------+-------+-----
---+--------+-------+
|   188|CA-2016-157000| 7/16/2016|7/22/2016|Standard Class|   AM-10360|  Alice
McCarthy|  Corporate|United States|Grand Prairie|       Texas|
75051|Central|OFF-ST-10001328|Office Supplies|     Storage|Personal Filing T…|
37.224|       3|     0.2| 3.7224|
|   303|CA-2016-142545|10/28/2016|11/3/2016|Standard Class|   JD-15895|Jonathan
Doherty|  Corporate|United States|   Belleville|  New Jersey|       7109|
East|OFF-BI-10002706|Office Supplies|     Binders|Avery Premier Hea…|  14.28|
1|       0| 6.5688|
|   392|US-2014-135972| 9/21/2014|9/23/2014|  Second Class|   JG-15115|
Jack Garza|   Consumer|United States|   Des Moines|  Washington|      98198|
West|TEC-PH-10003012|     Technology|      Phones|Nortel Meridian M…|246.384|
2|     0.2|27.7182|
|   466|CA-2016-109869| 4/22/2016|4/29/2016|Standard Class|   TN-21040|   Tanja
Norvell|Home Office|United States|      Phoenix|     Arizona|      85023|
West|OFF-SU-10003505|Office Supplies|    Supplies|Premier Electric …|185.376|
2|     0.2|-34.758|
|   698|CA-2015-119291| 5/14/2015|5/17/2015|   First Class|   JO-15550|    Jesus
Ocampo|Home Office|United States|      Chester|Pennsylvania|      19013|
East|OFF-LA-10003510|Office Supplies|      Labels|Avery 4027 File F…| 97.696|
4|     0.2|31.7512|
+------+--------------+---------+---------+--------------+-----------+--------
-------+-----------+-------------+-------------+------------+-----------+-------
+---------------+---------------+------------+--------------------+-------+-----
---+--------+-------+
only showing top 5 rows
```

1.2 Step 2: Total Sales by Region

[ ]:
```
df_clean.groupBy("Region").agg(sum("Sales").alias("Total_Sales")).show()
```

```
+-------+------------------+
| Region|       Total_Sales|
+-------+------------------+
|  South|388983.58499999996|
|Central|497800.87279999984|
|   East| 672194.0540000002|
|   West|  713471.344500001|
+-------+------------------+
```

1.3 Step 3: Average Profit by Category

```
[ ]: df_clean.groupBy("Category").agg(avg("Profit").alias("Avg_Profit")).show()
```

```
+---------------+-----------------+
|       Category|       Avg_Profit|
+---------------+-----------------+
|Office Supplies|20.01873189512116|
|      Furniture|9.281672418670434|
|     Technology|78.71591586356256|
+---------------+-----------------+
```

1.4 Step 4: Total Orders by Ship Mode

```
[ ]: df_clean.groupBy("Ship Mode").agg(count("*").alias("Order_Count")).show()
```

```
+--------------+-----------+
|     Ship Mode|Order_Count|
+--------------+-----------+
|   First Class|       1538|
|      Same Day|        543|
|  Second Class|       1945|
|Standard Class|       5968|
+--------------+-----------+
```

1.5 Step 5: Top 5 Profitable Orders

```
[ ]: df_clean.orderBy(df_clean["Profit"].desc()).select("Order ID", "Product Name",␣
     ↪"Profit").show(5)
```

```
+--------------+--------------------+---------+
|      Order ID|        Product Name|   Profit|
+--------------+--------------------+---------+
|CA-2016-118689|Canon imageCLASS …| 8399.976|
|CA-2017-140151|Canon imageCLASS …|6719.9808|
|CA-2017-166709|Canon imageCLASS …|5039.9856|
|CA-2016-117121|GBC Ibimaster 500…|  4946.37|
|CA-2014-116904|Ibico EPK-21 Elec…|4630.4755|
+--------------+--------------------+---------+
only showing top 5 rows
```

1.6 Step 6: Orders with High Discount (> 30%)

```
[ ]: *df_clean.filter(df_clean["Discount"] > 0.3).select("Order ID", "Product Name",␣
      ↪"Discount").show(5)
```

```
+--------------+--------------------+--------+
|      Order ID|        Product Name|Discount|
+--------------+--------------------+--------+
|CA-2014-169775|Balt Solid Wood R…|    0.45|
|US-2017-131583|"REDIFORM Incomin…|       6|
|CA-2014-169775|GBC DocuBind 200 …|     0.7|
|CA-2017-133067|"Black Avery Memo…|     0.7|
|CA-2016-130393|Acco 6 Outlet Gua…|     0.8|
+--------------+--------------------+--------+
only showing top 5 rows
```

1.7 Step 7: Total Profit by Region and Category

```
[ ]: df_clean.groupBy("Region", "Category").agg(sum("Profit").alias("Total_Profit")).
      ↪orderBy("Region").show()
```

```
+-------+---------------+------------------+
| Region|       Category|      Total_Profit|
+-------+---------------+------------------+
|Central|Office Supplies| 9038.715400000012|
|Central|     Technology|33693.441399999996|
|Central|      Furniture|-2581.653800000001|
|   East|Office Supplies| 40786.45890000001|
|   East|      Furniture|3376.6402000000035|
|   East|     Technology| 47439.95760000001|
|  South|Office Supplies| 19595.75349999999|
|  South|      Furniture| 7071.571900000001|
|  South|     Technology|19983.015600000002|
|   West|Office Supplies| 51211.95060000004|
|   West|     Technology|         44271.882|
|   West|      Furniture|11819.868899999998|
+-------+---------------+------------------+
```

1.8 Conclusion:

- **West** and **Central** regions have the highest total sales.
- **Technology** products give better profit margins.
- **Standard Class** shipping has the highest number of orders.
- High discounts often correlate with lower profits.