**CODE:**

from pyspark.sql import SparkSession

from pyspark.sql.functions import col

# Initialize SparkSession

spark = SparkSession.builder \

.appName("Product Sales Analysis") \

.getOrCreate()

# Sample data for products

products = [

(1, "Laptop", "Electronics", 50000),

(2, "Smartphone", "Electronics", 30000),

(3, "Table", "Furniture", 15000),

(4, "Chair", "Furniture", 5000),

(5, "Headphones", "Electronics", 2000),

]

# Sample data for sales transactions

sales = [

(1, 1, 2),

(2, 2, 1),

(3, 3, 3),

(4, 1, 1),

(5, 4, 5),

(6, 2, 2),

(7, 5, 10),

(8, 3, 1),

]

# Define schema for DataFrames

product\_columns = ["ProductID", "ProductName", "Category", "Price"]

sales\_columns = ["SaleID", "ProductID", "Quantity"]

# Create DataFrames

product\_df = spark.createDataFrame(products, schema=product\_columns)

sales\_df = spark.createDataFrame(sales, schema=sales\_columns)

# Show the DataFrames

print("Products DataFrame:")

product\_df.show()

print("Sales DataFrame:")

sales\_df.show()

#1. \*Join the DataFrames:\*

combined\_df = product\_df.join(sales\_df, on="ProductID")

print("Combined DataFrame:")

combined\_df.show()

#2. \*Calculate Total Sales Value:\*

from pyspark.sql.functions import expr

combined\_df = combined\_df.withColumn("TotalSalesValue", col("Price") \* col("Quantity"))

print("DataFrame with Total Sales Value:")

combined\_df.show()

#3. \*Find the Total Sales for Each Product Category:\*

category\_sales\_df = combined\_df.groupBy("Category").agg({"TotalSalesValue": "sum"}).withColumnRenamed("sum(TotalSalesValue)", "CategoryTotalSales")

print("Total Sales Value by Category:")

category\_sales\_df.show()

#4. \*Identify the Top-Selling Product:\*

top\_selling\_product = combined\_df.groupBy("ProductName").agg({"TotalSalesValue": "sum"}).withColumnRenamed("sum(TotalSalesValue)", "TotalSalesValue").orderBy(col("TotalSalesValue").desc()).limit(1)

print("Top-Selling Product:")

top\_selling\_product.show()

#5. \*Sort the Products by Total Sales Value:\*

sorted\_products\_df = combined\_df.groupBy("ProductName").agg({"TotalSalesValue": "sum"}).withColumnRenamed("sum(TotalSalesValue)", "TotalSalesValue").orderBy(col("TotalSalesValue").desc())

print("Products Sorted by Total Sales Value:")

sorted\_products\_df.show()

#6. \*Count the Number of Sales for Each Product:\*

sales\_count\_df = combined\_df.groupBy("ProductName").count().withColumnRenamed("count", "NumberOfSales")

print("Number of Sales per Product:")

sales\_count\_df.show()

#7. \*Filter the Products with Total Sales Value Greater Than ₹50,000:\*

filtered\_products\_df = combined\_df.groupBy("ProductName").agg({"TotalSalesValue": "sum"}).withColumnRenamed("sum(TotalSalesValue)", "TotalSalesValue").filter(col("TotalSalesValue") > 50000)

print("Products with Total Sales Value Greater Than ₹50,000:")

filtered\_products\_df.show()

**OUTPUT:**







