Exercise - Product Analysis Sales

# 1.Join the product\_df and sales\_df DataFrames on ProductID to create a combined DataFrame with product and sales data.

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

print("Joined dataframes: ")

combined\_df.show()

Output:

+---------+-----------+-----------+-----+------+--------+

|ProductID|ProductName| Category|Price|SaleID|Quantity|

+---------+-----------+-----------+-----+------+--------+

| 1| Laptop|Electronics|50000| 1| 2|

| 1| Laptop|Electronics|50000| 4| 1|

| 2| Smartphone|Electronics|30000| 2| 1|

| 2| Smartphone|Electronics|30000| 6| 2|

| 3| Table| Furniture|15000| 3| 3|

| 3| Table| Furniture|15000| 8| 1|

| 4| Chair| Furniture| 5000| 5| 5|

| 5| Headphones|Electronics| 2000| 7| 10|

+---------+-----------+-----------+-----+------+--------+

# 2. For each product, calculate the total sales value by multiplying the price by the quantity sold.

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

print("Total sales value: ")

sales\_value\_df.show()

Output:

Total sales value:

+---------+-----------+-----------+-----+------+--------+---------------+

|ProductID|ProductName| Category|Price|SaleID|Quantity|TotalSalesValue|

+---------+-----------+-----------+-----+------+--------+---------------+

| 1| Laptop|Electronics|50000| 1| 2| 100000|

| 1| Laptop|Electronics|50000| 4| 1| 50000|

| 2| Smartphone|Electronics|30000| 2| 1| 30000|

| 2| Smartphone|Electronics|30000| 6| 2| 60000|

| 3| Table| Furniture|15000| 3| 3| 45000|

| 3| Table| Furniture|15000| 8| 1| 15000|

| 4| Chair| Furniture| 5000| 5| 5| 25000|

| 5| Headphones|Electronics| 2000| 7| 10| 20000|

+---------+-----------+-----------+-----+------+--------+---------------+

# 3.Group the data by the Category column and calculate the total sales value for each product category

category\_sales\_value\_df = sales\_value\_df.groupBy("Category").agg({"TotalSalesValue": "sum"})

print("Category wise sales value: ")

category\_sales\_value\_df.show()

Output:

Category wise sales value:

+-----------+--------------------+

| Category|sum(TotalSalesValue)|

+-----------+--------------------+

|Electronics| 260000|

| Furniture| 85000|

+-----------+--------------------+

# 4.Find the product that generated highest sales value

highest\_sales\_product = sales\_value\_df.orderBy(col("TotalSalesValue").desc()).limit(1)

print("Product with highest sales value: ")

highest\_sales\_product.show()

Output:

Product with highest sales value:

+---------+-----------+-----------+-----+------+--------+---------------+

|ProductID|ProductName| Category|Price|SaleID|Quantity|TotalSalesValue|

+---------+-----------+-----------+-----+------+--------+---------------+

| 1| Laptop|Electronics|50000| 1| 2| 100000|

+---------+-----------+-----------+-----+------+--------+---------------+

# 5. Sort the products by total sales value in descending order

sorted\_products\_df = sales\_value\_df.orderBy(col("TotalSalesValue").desc())

print("Sorted products by total sales value: ")

sorted\_products\_df.show()

Output:

Sorted products by total sales value:

+---------+-----------+-----------+-----+------+--------+---------------+

|ProductID|ProductName| Category|Price|SaleID|Quantity|TotalSalesValue|

+---------+-----------+-----------+-----+------+--------+---------------+

| 1| Laptop|Electronics|50000| 1| 2| 100000|

| 2| Smartphone|Electronics|30000| 6| 2| 60000|

| 1| Laptop|Electronics|50000| 4| 1| 50000|

| 3| Table| Furniture|15000| 3| 3| 45000|

| 2| Smartphone|Electronics|30000| 2| 1| 30000|

| 4| Chair| Furniture| 5000| 5| 5| 25000|

| 5| Headphones|Electronics| 2000| 7| 10| 20000|

| 3| Table| Furniture|15000| 8| 1| 15000|

+---------+-----------+-----------+-----+------+--------+---------------+

# 6. Count the number of sales for each product

sales\_count\_df = sales\_value\_df.groupBy("ProductID","ProductName").agg({"Quantity": "sum"})

print("Number of sales for each product: ")

sales\_count\_df.show()

Output:

Number of sales for each product:

+---------+-----------+-------------+

|ProductID|ProductName|sum(Quantity)|

+---------+-----------+-------------+

| 1| Laptop| 3|

| 2| Smartphone| 3|

| 3| Table| 4|

| 4| Chair| 5|

| 5| Headphones| 10|

+---------+-----------+-------------+

#7. Filter out the products that have a total sales value greater than ₹50,000

high\_value\_products\_df = sales\_value\_df.filter(col("TotalSalesValue") > 50000)

print("Products with total sales value greater than ₹50,000: ")

high\_value\_products\_df.show()

Output:

Products with total sales value greater than ₹50,000:

+---------+-----------+-----------+-----+------+--------+---------------+

|ProductID|ProductName| Category|Price|SaleID|Quantity|TotalSalesValue|

+---------+-----------+-----------+-----+------+--------+---------------+

| 1| Laptop|Electronics|50000| 1| 2| 100000|

| 2| Smartphone|Electronics|30000| 6| 2| 60000|

+---------+-----------+-----------+-----+------+--------+---------------+