#### \*\*Exercise: Product Sales Analysis\*\*

#### \*\*Step 1: Create DataFrames\*\*

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", "Name", "Category", "Price"]

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

#Create DataFrames

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

transaction\_columns = ["TransactionID", "CustomerID", "Amount"]

#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()

#### \*\*Step 2: Perform the Following Tasks\*\*

#1. \*\*Join the DataFrames:\*\*

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

print("Product Sales DataFrame:")

product\_sales\_df.show()

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

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

print("Combined 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)", "TotalCategorySales")

print("Total Sales for Each Product Category:")

category\_sales\_df.show()

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

product\_sales\_df = combined\_df.groupBy("ProductID", "Name").agg({"TotalSalesValue": "sum"}).withColumnRenamed("sum(TotalSalesValue)", "TotalProductSales")

top\_selling\_product\_df = product\_sales\_df.orderBy(col("TotalProductSales").desc()).limit(1)

print("Top-Selling Product:")

top\_selling\_product\_df.show()

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

sorted\_products\_df = combined\_df.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 = sales\_df.groupBy("ProductID").count().withColumnRenamed("count", "NumberOfSales")

print("Number of Sales for Each Product:")

sales\_count\_df.show()

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

product\_sales\_df = product\_sales\_df.withColumnRenamed("sum(TotalSalesValue)", "TotalProductSales")

filtered\_products\_df = product\_sales\_df.filter(col("TotalProductSales") > 50000)

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

filtered\_products\_df.show()