from pyspark.sql import SparkSession  
from pyspark.sql.functions import col, sum  
  
  
saledata =[  
 (1001, 2001, 3001, 5, 20.0, '2024-09-01'),  
 (1002, 2002, 3002, 2, 50.0, '2024-09-01'),  
 (1003, 2003, 3003, 1, 120.0, '2024-09-02'),  
 (1004, 2001, 3002, 3, 40.0, '2024-09-03'),  
 (1005, 2004, 3001, 10, 15.0, '2024-09-03'),  
 (1006, 2005, 3004, None, 30.0, '2024-09-03')  
]  
columns=["transaction\_id", "customer\_id", "product\_id", "quantity", "price", "date"]  
spark = SparkSession.builder.appName("Sales Transactions Analysis").getOrCreate()  
df = spark.createDataFrame(data=saledata, schema=columns)  
#df.show()  
  
#1 filter  
#df\_filter=df.filter((df.quantity\*df.price)>100)  
#df\_filter.show()  
  
#2 Handle\_null  
#df\_filled = df.na.fill({'quantity': 0})  
#df\_filled.show()  
  
#3 drop duplicate  
#df\_no\_duplicates = df.dropDuplicates(['transaction\_id'])  
#df\_no\_duplicates.show()  
  
#4 select specific column  
#Select 'customer\_id', 'product\_id', and 'quantity'.  
  
#df\_selected = df.select('customer\_id', 'product\_id', 'quantity')  
#df\_selected.show()  
  
#5 grouping and aggregate  
#calculate total sale per product  
  
#df\_grouped = df.groupBy('product\_id').agg({'quantity': 'sum'})  
#df\_grouped.show()  
  
#Joining DataFrames  
  
products=[  
 (3001,"laptop"),  
 (3002,"mouse"),  
 (3003,"keyboard"),  
 (3004,"monitor")  
 ]  
 #Create a movie details DataFrame  
columns2= ["product\_id","item\_name"]  
df2 = spark.createDataFrame(data = products, schema = columns2)  
  
#df\_joined = df.join(df2, on='product\_id', how='inner')  
#df\_joined.show()  
  
#\*\*\*\*\*\*\*\* ERROR\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*8  
df\_new\_transaction =[  
 (1007, 2003, 3006, 5, 20.0, '2024-09-01')  
 ]  
columns3= ["transaction\_id", "customer\_id", "product\_id", "quantity", "price", "date"]  
df3 = spark.createDataFrame(data = df\_new\_transaction, schema = columns2)  
  
df\_union = df.union(df3)  
df\_union.show()  
  
  
#df.createOrReplaceTempView('transactions')  
#sql\_result = spark.sql('SELECT customer\_id, SUM(quantity \* price) as total\_spent FROM transactions GROUP BY customer\_id')  
#sql\_result.show()  
  
  
if \_\_name\_\_ == '\_\_main\_\_':  
 ()