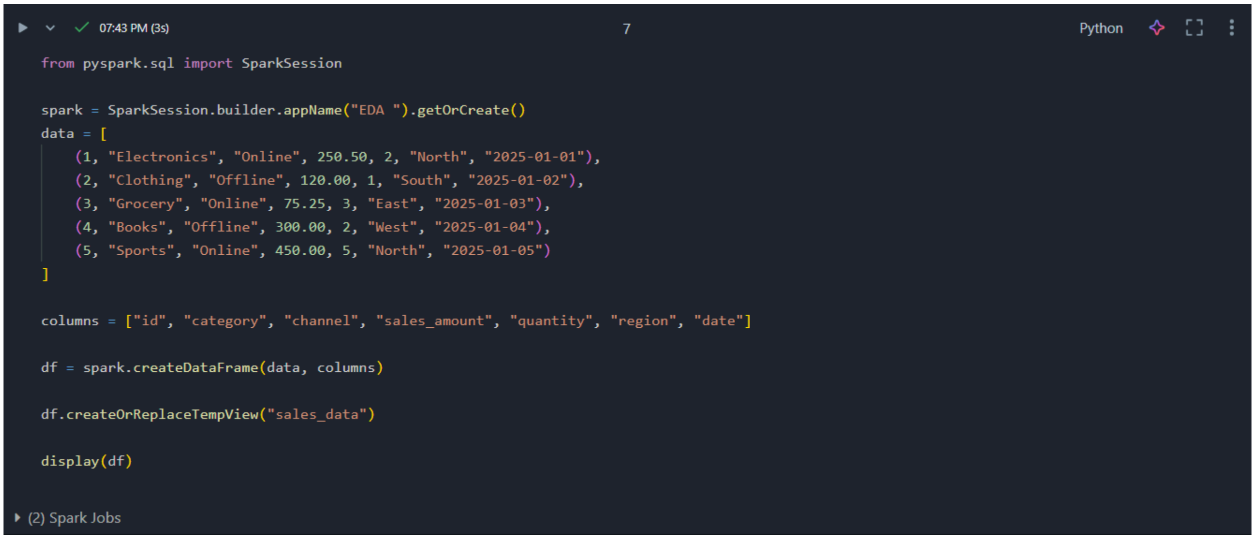
Tejashree G
tejashreeg.ai2021@dce.edu.in

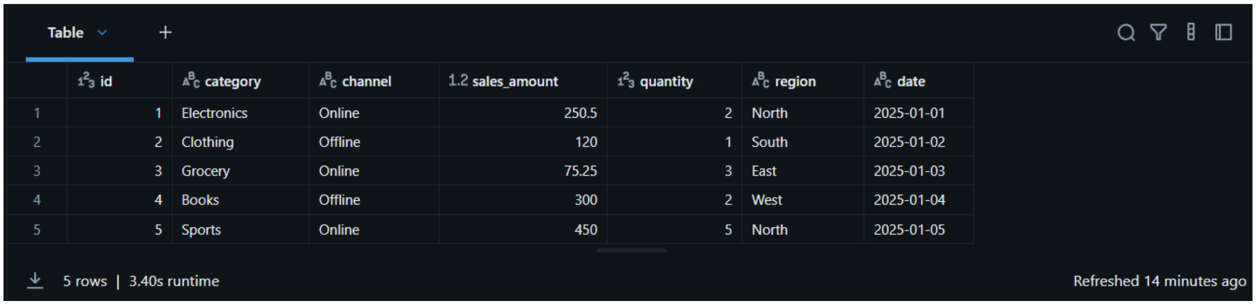
EDA Assignment

Create Sample Dataset



```python
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("EDA ").getOrCreate()
data = [
    (1, "Electronics", "Online", 250.50, 2, "North", "2025-01-01"),
    (2, "Clothing", "Offline", 120.00, 1, "South", "2025-01-02"),
    (3, "Grocery", "Online", 75.25, 3, "East", "2025-01-03"),
    (4, "Books", "Offline", 300.00, 2, "West", "2025-01-04"),
    (5, "Sports", "Online", 450.00, 5, "North", "2025-01-05")
]

columns = ["id", "category", "channel", "sales_amount", "quantity", "region", "date"]

df = spark.createDataFrame(data, columns)

df.createOrReplaceTempView("sales_data")

display(df)
```

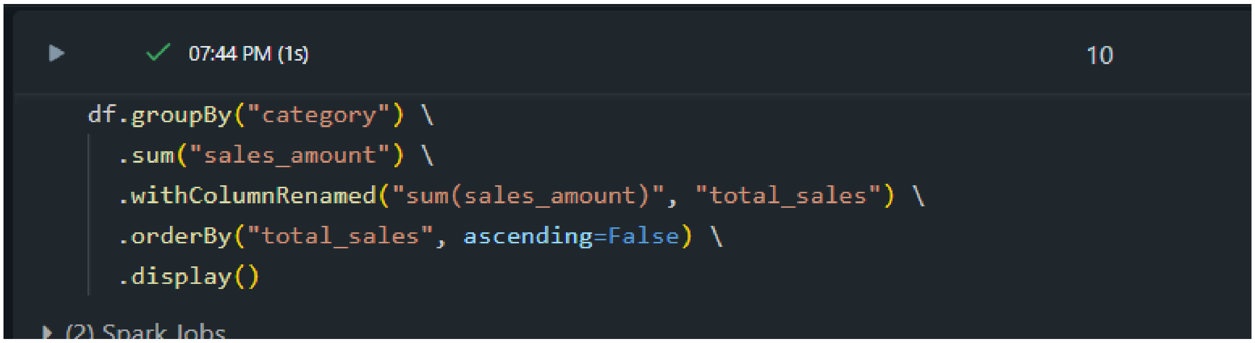
Output



	id	category	channel	sales_amount	quantity	region	date
1	1	Electronics	Online	250.5	2	North	2025-01-01
2	2	Clothing	Offline	120	1	South	2025-01-02
3	3	Grocery	Online	75.25	3	East	2025-01-03
4	4	Books	Offline	300	2	West	2025-01-04
5	5	Sports	Online	450	5	North	2025-01-05

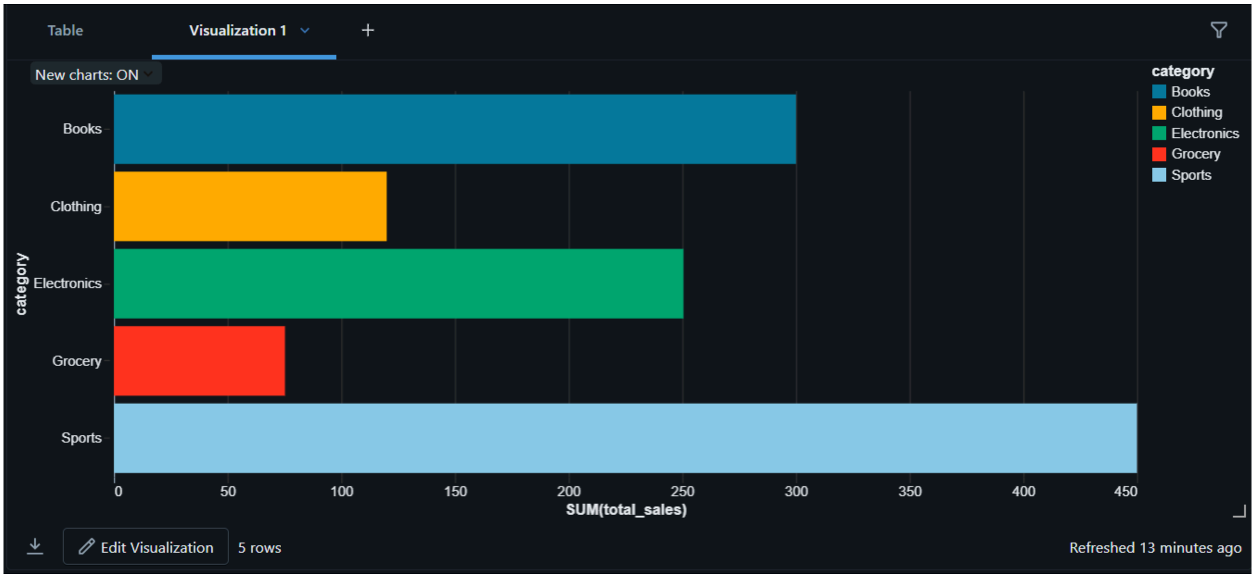
Visualizations

1. Bar Chart: Total Sales by Category



```python
df.groupBy("category") \
    .sum("sales_amount") \
    .withColumnRenamed("sum(sales_amount)", "total_sales") \
    .orderBy("total_sales", ascending=False) \
    .display()
```

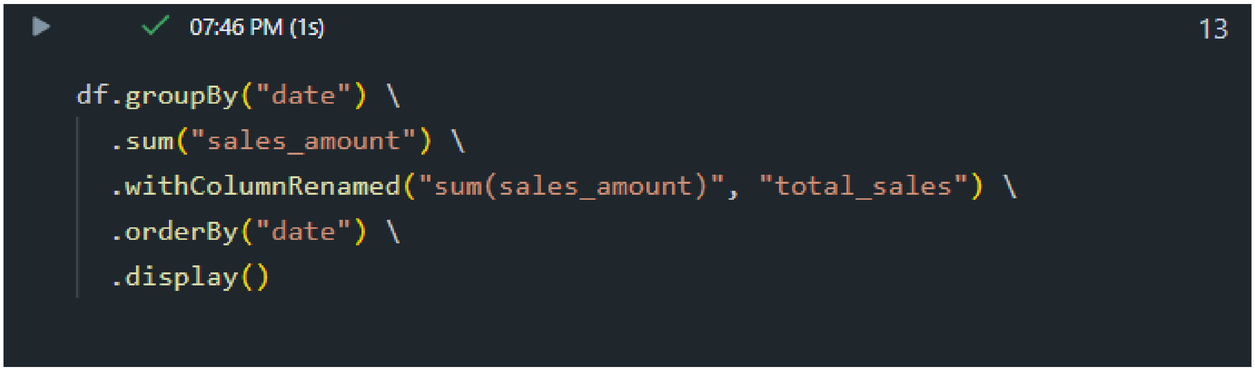
Output



5 rows

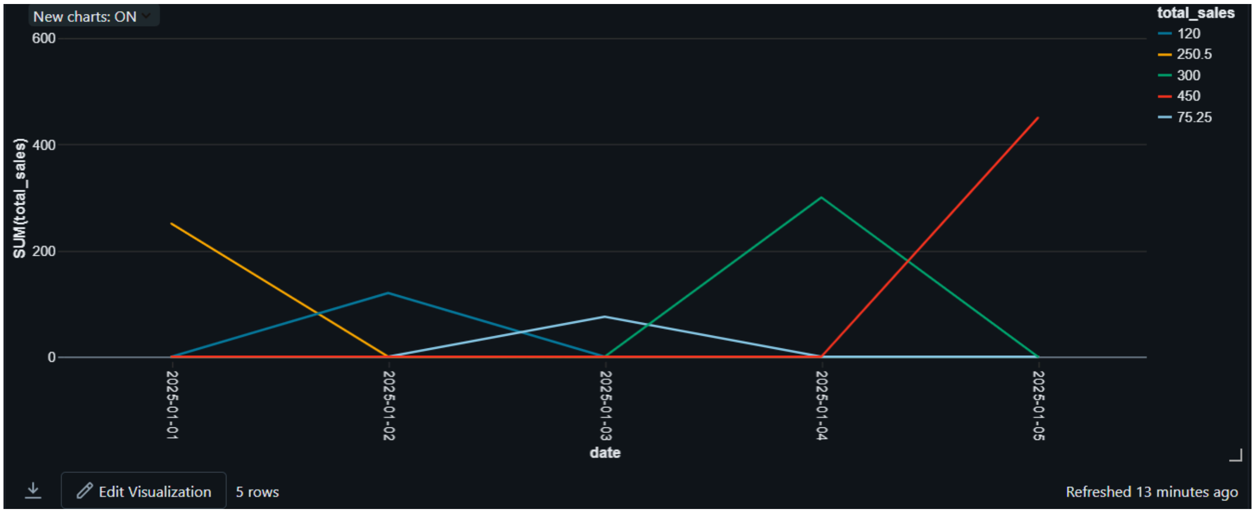
Refreshed 13 minutes ago

2. Line Chart: Sales Trend Over Time



```
df.groupBy("date") \
   .sum("sales_amount") \
   .withColumnRenamed("sum(sales_amount)", "total_sales") \
   .orderBy("date") \
   .display()
```

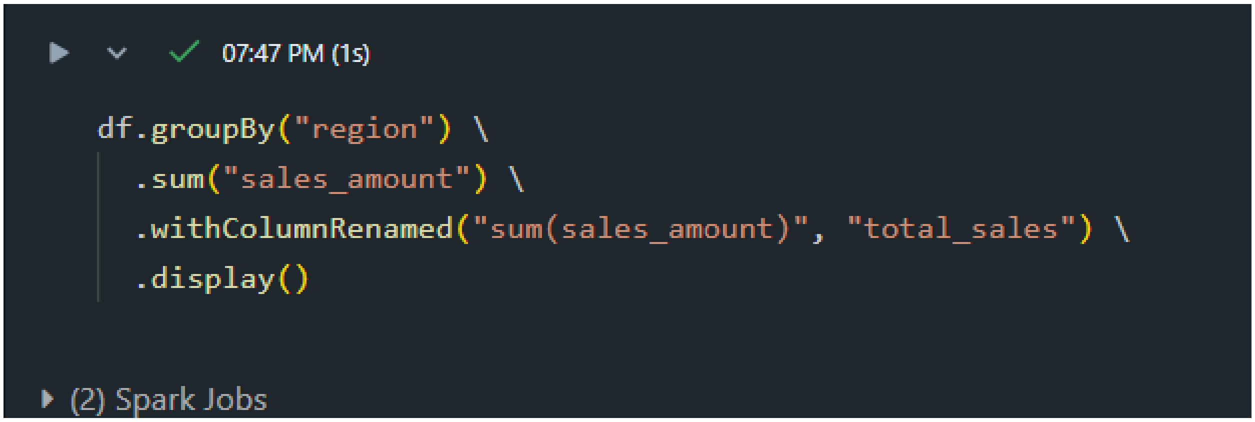
Output



5 rows

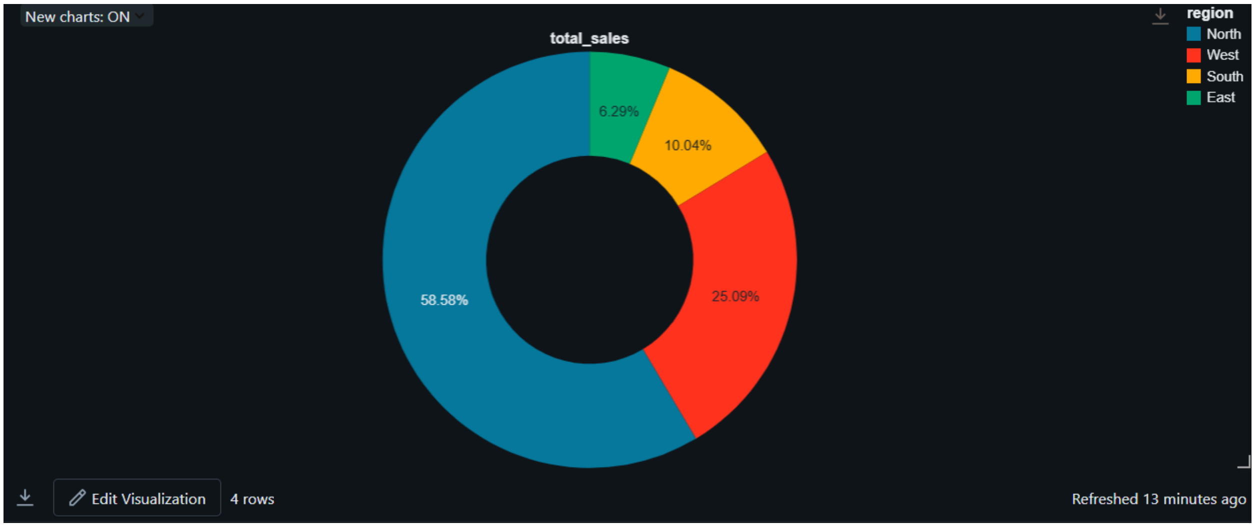
Refreshed 13 minutes ago

3. Pie Chart: Sales Distribution by Region

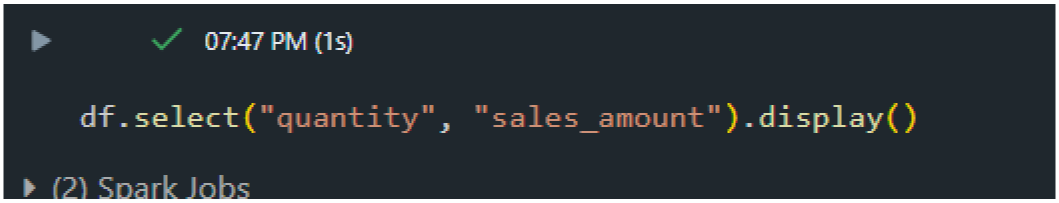


```
df.groupBy("region") \
  .sum("sales_amount") \
  .withColumnRenamed("sum(sales_amount)", "total_sales") \
  .display()
```

Output

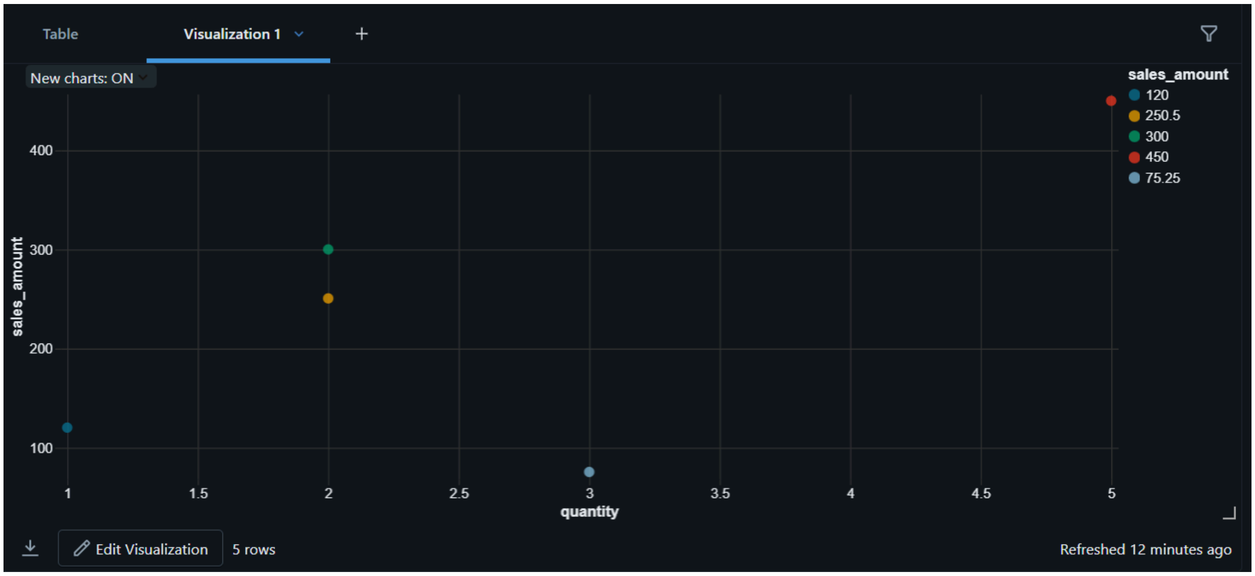


4. Scatter Plot: Quantity vs. Sales Amount



```
df.select("quantity", "sales_amount").display()
```

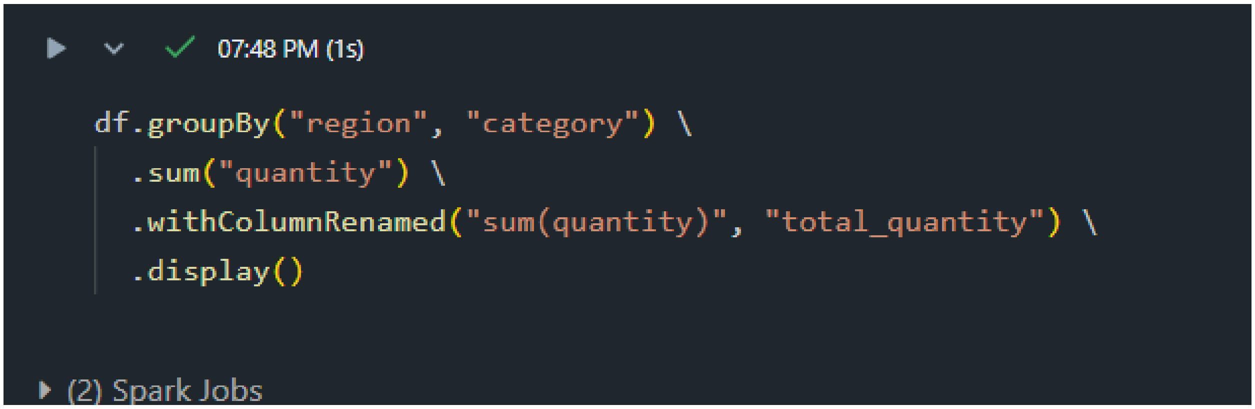
Output



5 rows

Refreshed 12 minutes ago

5. Heatmap: Quantity by Region and Category

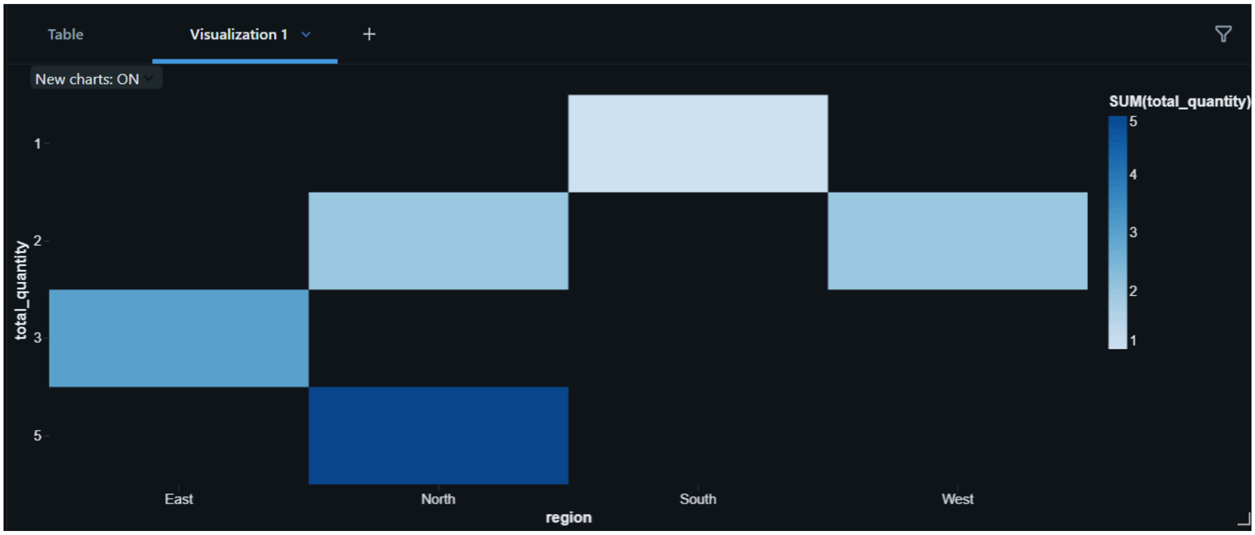


07:48 PM (1s)

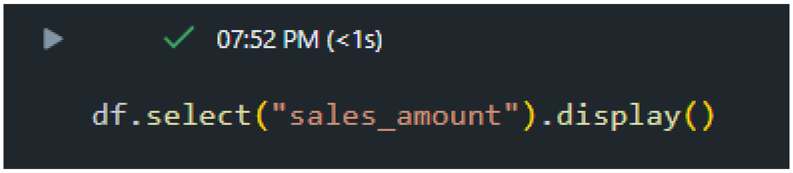
```
df.groupBy("region", "category") \
  .sum("quantity") \
  .withColumnRenamed("sum(quantity)", "total_quantity") \
  .display()
```

(2) Spark Jobs

Output

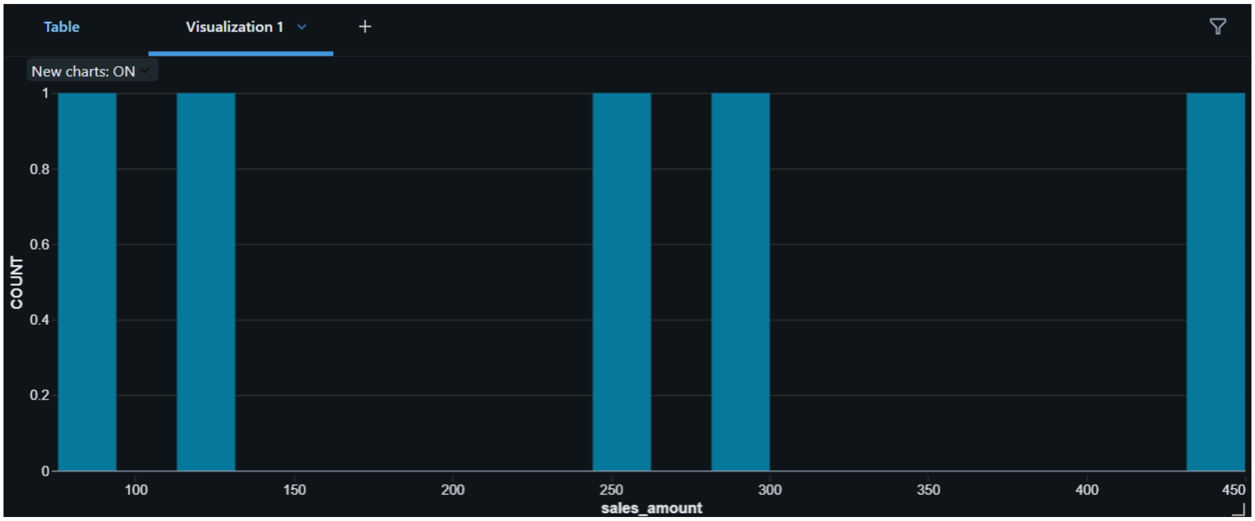


6. Histogram: Distribution of Sales Amount

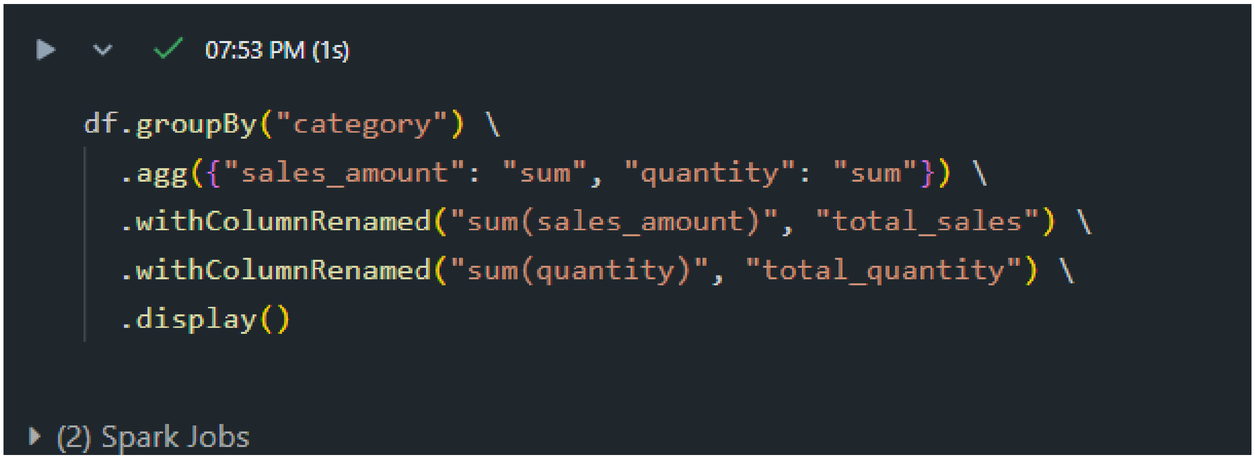


```
df.select("sales_amount").display()
```

Output

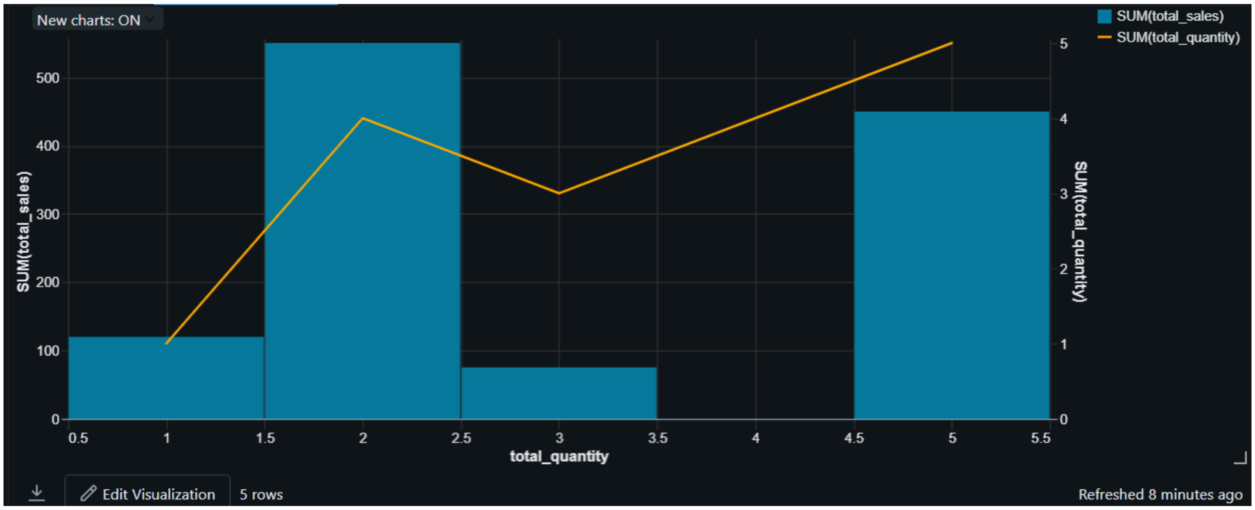


7. Combo Chart: Sales & Quantity by Category



```
df.groupBy("category") \
    .agg({"sales_amount": "sum", "quantity": "sum"}) \
    .withColumnRenamed("sum(sales_amount)", "total_sales") \
    .withColumnRenamed("sum(quantity)", "total_quantity") \
    .display()
```

Output


New charts: ON
SUM(total_sales)
SUM(total_quantity)
500
400
300
200
100
0
5
4
3
2
1
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
total_quantity
Edit Visualization
5 rows
Refreshed 8 minutes ago