```
# in Python
from pyspark.sql.functions import skewness, kurtosis
df.select(skewness("Quantity"), kurtosis("Quantity")).show()

-- in SQL
SELECT skewness(Quantity), kurtosis(Quantity) FROM dfTable
```

```
+------------------+------------------+
| skewness(Quantity)|kurtosis(Quantity)|
+------------------+------------------+
|-0.2640755761052562|119768.05495536952|
+------------------+------------------+
```

协方差和相关性

我们讨论了单列聚合，不过有的函数是去比较两个不同列的值之间的相互关系。其中两个函数就是cov和corr，它们分别用于计算协方差和相关性。相关性采用Pearson相关系数来衡量，范围是-1~+1。协方差的范围由数据中的输入决定。

跟var函数一样，协方差又分为样本协方差和总体协方差，因此在使用的时候需要指定，这一点很重要。相关性没有这个概念，因此没有总体或样本的相关性之分。以下是它们的使用方式：

```
// in Scala
import org.apache.spark.sql.functions.{corr, covar_pop, covar_samp}
df.select(corr("InvoiceNo", "Quantity"), covar_samp("InvoiceNo", "Quantity"),
    covar_pop("InvoiceNo", "Quantity")).show()

# in Python
from pyspark.sql.functions import corr, covar_pop, covar_samp
df.select(corr("InvoiceNo", "Quantity"), covar_samp("InvoiceNo", "Quantity"),
    covar_pop("InvoiceNo", "Quantity")).show()

-- in SQL
SELECT corr(InvoiceNo, Quantity), covar_samp(InvoiceNo, Quantity),
  covar_pop(InvoiceNo, Quantity)
FROM dfTable
```

```
+-------------------------+-----------------------------+---------------------+
|corr(InvoiceNo, Quantity)|covar_samp(InvoiceNo, Quantity)|covar_pop(InvoiceN...|
+-------------------------+-----------------------------+---------------------+
|     4.912186085635685E-4|            1052.7280543902734|            1052.7...|
+-------------------------+-----------------------------+---------------------+
```

聚合输出复杂类型

在Spark中，不仅可以在数值型上执行聚合操作，还能在复杂类型上执行聚合操作。例如，可以收集某列上的值到一个list列表里，或者将unique唯一值收集到一个set集合里。