In [4]:

```
from pyspark.sql import SparkSession
spark_2 = SparkSession.builder.appName('aggs').getOrCreate()
```

In [5]:

```
df = spark_2.read.csv(r'D:/Python/PySpark/Python-and-Spark-for-Big-Data-master/Spark_DataFrames/sales_info.csv', inferSchema= True, header= True)
```

In [6]:

```
df.show()
```

```
+-------+-------+-----+
|Company| Person|Sales|
+-------+-------+-----+
|   GOOG|    Sam|200.0|
|   GOOG|Charlie|120.0|
|   GOOG|  Frank|340.0|
|   MSFT|   Tina|600.0|
|   MSFT|    Amy|124.0|
|   MSFT|Vanessa|243.0|
|     FB|   Carl|870.0|
|     FB|  Sarah|350.0|
|   APPL|   John|250.0|
|   APPL|  Linda|130.0|
|   APPL|   Mike|750.0|
|   APPL|  Chris|350.0|
+-------+-------+-----+
```

In [9]:

```
print((df.count(),len(df.columns)))
```

```
(12, 3)
```

In [10]:

```
df.printSchema()
```

```
root
 |-- Company: string (nullable = true)
 |-- Person: string (nullable = true)
 |-- Sales: double (nullable = true)
```

Group By :

In [15]:

```
df.groupBy("Company")
```

Out[15]:

```
<pyspark.sql.group.GroupedData at 0x24b70329f28>
```

In [17]:

```
df.groupBy("Company").mean().show()
```

```
+-------+-----------------+
|Company|       avg(Sales)|
+-------+-----------------+
|   APPL|            370.0|
|   GOOG|            220.0|
|     FB|            610.0|
|   MSFT|322.3333333333333|
+-------+-----------------+
```

In [20]:

```
df.groupBy("Company").sum().show()
```

```
+-------+----------+
|Company|sum(Sales)|
+-------+----------+
|   APPL|    1480.0|
|   GOOG|     660.0|
|     FB|    1220.0|
|   MSFT|     967.0|
+-------+----------+
```

In [21]:

```
df.groupBy("Company").min().show()
```

```
+-------+----------+
|Company|min(Sales)|
+-------+----------+
|   APPL|     130.0|
|   GOOG|     120.0|
|     FB|     350.0|
|   MSFT|     124.0|
+-------+----------+
```

In [28]:
```
df.groupBy("Company").count().show()
```

```
+-------+-----+
|Company|count|
+-------+-----+
|   APPL|    4|
|   GOOG|    3|
|     FB|    2|
|   MSFT|    3|
+-------+-----+
```

Aggregate :

In [29]:
```
df.agg({'Sales':'sum'}).show()
```

```
+----------+
|sum(Sales)|
+----------+
|    4327.0|
+----------+
```

In [42]:
```
# Another way to do agg function
group_data = df.groupBy()
group_data.agg({'Sales':'max'}).show()
print('='*12)
group_data.agg({'Sales':'sum'}).show()
```

```
+----------+
|max(Sales)|
+----------+
|     870.0|
+----------+

============
+----------+
|sum(Sales)|
+----------+
|    4327.0|
+----------+
```

In [43]:
```
from pyspark.sql.functions import countDistinct,avg,stddev
```

In [45]:
```
df.select(countDistinct('Sales')).show()
```

```
+---------------------+
|count(DISTINCT Sales)|
+---------------------+
|                   11|
+---------------------+
```

In [47]:
```
df.select(countDistinct('Company')).show()
```

```
+-----------------------+
|count(DISTINCT Company)|
+-----------------------+
|                      4|
+-----------------------+
```

In [50]:
```
# column with alias name
df.select(avg('Sales').alias('Avg. Sales')).show()
```

```
+-----------------+
|       Avg. Sales|
+-----------------+
|360.5833333333333|
+-----------------+
```

Orded By :

In [52]:
```
df.orderBy("Sales").show()
```

```
+-------+-------+-----+
|Company| Person|Sales|
+-------+-------+-----+
|   GOOG|Charlie|120.0|
|   MSFT|    Amy|124.0|
|   APPL|  Linda|130.0|
|   GOOG|    Sam|200.0|
|   MSFT|Vanessa|243.0|
|   APPL|   John|250.0|
|   GOOG|  Frank|340.0|
|     FB|  Sarah|350.0|
|   APPL|  Chris|350.0|
|   MSFT|   Tina|600.0|
|   APPL|   Mike|750.0|
|     FB|   Carl|870.0|
+-------+-------+-----+
```

In [54]:
```
df.orderBy("Sales", ascending = False).collect()
```

Out[54]:
```
[Row(Company='FB', Person='Carl', Sales=870.0),
 Row(Company='APPL', Person='Mike', Sales=750.0),
 Row(Company='MSFT', Person='Tina', Sales=600.0),
 Row(Company='FB', Person='Sarah', Sales=350.0),
 Row(Company='APPL', Person=' Chris', Sales=350.0),
 Row(Company='GOOG', Person='Frank', Sales=340.0),
 Row(Company='APPL', Person='John', Sales=250.0),
 Row(Company='MSFT', Person='Vanessa', Sales=243.0),
 Row(Company='GOOG', Person='Sam', Sales=200.0),
 Row(Company='APPL', Person='Linda', Sales=130.0),
 Row(Company='MSFT', Person='Amy', Sales=124.0),
 Row(Company='GOOG', Person='Charlie', Sales=120.0)]
```

In [57]:
```
df.orderBy("Sales", ascending = True).show()
```

```
+-------+-------+-----+
|Company| Person|Sales|
+-------+-------+-----+
|   GOOG|Charlie|120.0|
|   MSFT|    Amy|124.0|
|   APPL|  Linda|130.0|
|   GOOG|    Sam|200.0|
|   MSFT|Vanessa|243.0|
|   APPL|   John|250.0|
|   GOOG|  Frank|340.0|
|     FB|  Sarah|350.0|
|   APPL|  Chris|350.0|
|   MSFT|   Tina|600.0|
|   APPL|   Mike|750.0|
|     FB|   Carl|870.0|
+-------+-------+-----+
```

In [ ]: