```
# Databricks notebook source
from pyspark.sql import SparkSession
from pyspark.sql.functions import col,sum,avg,max

spark = SparkSession.builder \
                    .appName('SparkByExamples.com') \
                    .getOrCreate()

simpleData = [("James","Sales","NY",90000,34,10000),
    ("Michael","Sales","NV",86000,56,20000),
    ("Robert","Sales","CA",81000,30,23000),
    ("Maria","Finance","CA",90000,24,23000),
    ("Raman","Finance","DE",99000,40,24000),
    ("Scott","Finance","NY",83000,36,19000),
    ("Jen","Finance","NY",79000,53,15000),
    ("Jeff","Marketing","NV",80000,25,18000),
    ("Kumar","Marketing","NJ",91000,50,21000)
  ]

schema = ["employee_name","department","state","salary","age","bonus"]
df = spark.createDataFrame(data=simpleData, schema = schema)
df.printSchema()
df.show(truncate=False)

df.groupBy("state").sum("salary").show()

dfGroup=df.groupBy("state") \
          .agg(sum("salary").alias("sum_salary"))

dfGroup.show(truncate=False)

dfFilter=dfGroup.filter(dfGroup.sum_salary > 100000)
dfFilter.show()

from pyspark.sql.functions import asc
dfFilter.sort("sum_salary").show()

from pyspark.sql.functions import desc
dfFilter.sort(desc("sum_salary")).show()

df.groupBy("state") \
  .agg(sum("salary").alias("sum_salary")) \
  .filter(col("sum_salary") > 100000)  \
  .sort(desc("sum_salary")) \
  .show()

df.createOrReplaceTempView("EMP")
spark.sql("select state, sum(salary) as sum_salary from EMP " +
          "group by state having sum_salary > 100000 " +
          "order by sum_salary desc").show()
```

```
df.groupBy("state") \
  .sum("salary") \
  .withColumnRenamed("sum(salary)", "sum_salary") \
  .show()

df.groupBy("state") \
  .sum("salary") \
  .select(col("state"),col("sum(salary)").alias("sum_salary")) \
  .show()




# COMMAND ----------
```