Similar to SQL Group By clause, Pyspark groupBy() function is used to collect the identical data into groups on DataFrame and perform count, sum, avg, min, max functions on the grouped Data.

from pyspark.sql import \*  
from pyspark.sql.functions import \*  
  
spark = SparkSession.builder.appName('groupBy').getOrCreate()  
  
myData = [(1,'Harsha','male',2000,'IT'),  
 (2,'Mokshit','male',4000,'HR'),  
 (3, 'Harika', 'female', 2000, 'payroll'),  
 (4, 'Kiran', 'male', 4000, 'HR'),  
 (5, 'venkat', 'male', 2000, 'IT'),  
 (6, 'Rama', 'female', 4000, 'payroll')  
 ]  
  
mySchema = ['id','name','gender','salary','dept']  
  
df = spark.createDataFrame(myData,mySchema)  
  
df.groupBy(df.dept).count().show()  
df.groupBy(df.dept).min('salary').show()  
df.groupBy(df.dept).max('salary').show()  
df.groupBy(df.dept,df.gender).count().show()

**Output:**

+-------+-----+

| dept|count|

+-------+-----+

| IT| 2|

| HR| 2|

|payroll| 2|

+-------+-----+

+-------+-----------+

| dept|min(salary)|

+-------+-----------+

| IT| 2000|

| HR| 4000|

|payroll| 2000|

+-------+-----------+

+-------+-----------+

| dept|max(salary)|

+-------+-----------+

| IT| 2000|

| HR| 4000|

|payroll| 4000|

+-------+-----------+

+-------+------+-----+

| dept|gender|count|

+-------+------+-----+

| IT| male| 2|

| HR| male| 2|

|payroll|female| 2|

+-------+------+-----+