1. Distinct() function is used to remove the duplicate rows(all columns).
2. dropDuplicates() is used to drop rows based on selected (one or multiple) columns.

from pyspark.sql import \*  
from pyspark.sql.functions import \*  
  
spark = SparkSession.builder.appName('distinct\_dropDuplicates').getOrCreate()  
  
  
myData = [(1,'Harsha','male',2000),  
 (2,'Mokshit','male',4000),  
 (2,'Mokshit','male',4000),  
 (3,'Harika','female',3000)]  
  
mySchema = ['id','name','gender','salary']  
  
df = spark.createDataFrame(myData,mySchema)  
  
df.distinct().show()  
  
df.dropDuplicates().show()  
  
df.dropDuplicates(['gender']).show()  
  
df.dropDuplicates(['gender','salary']).show()

**Output:**

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

| id| name|gender|salary|

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

| 1| Harsha| male| 2000|

| 2|Mokshit| male| 4000|

| 3| Harika|female| 3000|

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

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

| id| name|gender|salary|

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

| 1| Harsha| male| 2000|

| 2|Mokshit| male| 4000|

| 3| Harika|female| 3000|

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

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

| id| name|gender|salary|

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

| 3|Harika|female| 3000|

| 1|Harsha| male| 2000|

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

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

| id| name|gender|salary|

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

| 3| Harika|female| 3000|

| 1| Harsha| male| 2000|

| 2|Mokshit| male| 4000|

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