from pyspark.sql import SparkSession  
from pyspark.sql.functions import col, avg  
  
customer=[  
 ("Smith",23,5.3),  
 ("Rashmi",27,5.8),  
 ("Smith",23,5.3),  
 ("Payal",27,5.8),  
 ("Megha",27,5.4)  
]  
  
columns=["Name","Age","Height"]  
spark = SparkSession.builder.appName("calculating").getOrCreate()  
df=spark.createDataFrame(customer,columns)  
#df.show()  
  
#data cleaning -removing all duplicate record  
  
df\_remove\_dups = df.dropDuplicates(['Name', 'age','Height'])  
#df\_remove\_dups.show()  
  
#remove duplicate on basis of height and age  
df\_remove\_dups1 = df\_remove\_dups.dropDuplicates(['age','Height'])  
df\_remove\_dups1.show()