import pyspark  
from pyspark.sql import SparkSession  
import pyspark.sql.functions as F  
from pyspark.sql.functions import col, sum, max, avg  
  
spark = SparkSession.builder.master(**'local'**).appName(**'Covid\_data'**).getOrCreate()

print(“**first Question Answer**”)  
df= spark.read.format(**'csv'**).option(**'header'**,**'true'**).option(**'inferSchema'**,**'true'**) .load(**'C:**\\**Users**\\**zakeerm**\\**Downloads**\\**PysparkLearning**\\**resources**\\**country\_wise\_latest.csv'**)  
  
df.printSchema()  
  
df.show(2)  
  
print(**"second Question Answer"**)  
df1 = df.withColumnRenamed(**"Country/Region"**,**"country"**) \  
 .withColumnRenamed(**"New cases"**,**"New\_cases"**) \  
 .withColumnRenamed(**"New deaths"**,**"New\_deaths"**) \  
 .withColumnRenamed(**"New recovered"**,**"New\_recovered"**) \  
 .withColumnRenamed(**"Deaths / 100 Cases"**,**"Deaths\_100\_Cases"**) \  
 .withColumnRenamed(**"Recovered / 100 Cases"**,**"Recovered\_100\_Cases"**) \  
 .withColumnRenamed(**"Deaths / 100 Recovered"**,**"Deaths\_100\_Recovered"**) \  
 .withColumnRenamed(**"Confirmed last week"**,**"Confirmed\_last\_week"**) \  
 .withColumnRenamed(**"1 week change"**,**"1\_week\_change"**) \  
 .withColumnRenamed(**"1 week % increase"**,**"1\_week\_%\_increase"**) \  
 .withColumnRenamed(**"WHO Region"**,**"WHO\_Region"**)  
  
df1.printSchema()  
df1.show(2)  
df1.createOrReplaceTempView(**"covid"**)  
  
print(**"Count and check if there any null values in any of the columns"**)  
null\_counts = df1.agg(\*[F.count(F.when(F.isnull(c), c)).alias(c) for c in df1.columns])  
null\_counts.show()  
  
print(**"4. What are the top 10 countries under the WHO region with covid 19 Confirmed cases"**)  
fourth = spark.sql( **"SELECT country,WHO\_Region,sum(Confirmed) FROM covid group by country, WHO\_Region order by 3 desc"**)  
fourth.show(10)  
  
print(**"5. What are the bottom 10 countries under the WHO region with covid 19 Confirmed cases"**)  
fifth = spark.sql( **"SELECT Country,WHO\_Region,sum(Confirmed) FROM covid GROUP BY Country, WHO\_Region order by 3 asc"**)  
fifth.show(10)  
  
print(**"6. What are the total number of countries/ total no. of WHO regions and also list the various WHO regions"**)  
sixth = spark.sql(**"SELECT count(distinct Country), count(distinct WHO\_Region),WHO\_Region from covid group by WHO\_Region"**)

sixth.show()

Output =>

C:\Users\zakeerm\IdeaProjects\Spark-Scala\venv\Scripts\python.exe C:/Users/zakeerm/Downloads/PysparkLearning/pysparkprograms/Covid\_19\_Data\_analysis.py

23/11/11 21:51:46 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: java.io.FileNotFoundException: HADOOP\_HOME and hadoop.home.dir are unset. -see https://wiki.apache.org/hadoop/WindowsProblems

Setting default log level to "WARN".

To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).

23/11/11 21:51:47 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

root

|-- Country/Region: string (nullable = true)

|-- Confirmed: integer (nullable = true)

|-- Deaths: integer (nullable = true)

|-- Recovered: integer (nullable = true)

|-- Active: integer (nullable = true)

|-- New cases: integer (nullable = true)

|-- New deaths: integer (nullable = true)

|-- New recovered: integer (nullable = true)

|-- Deaths / 100 Cases: double (nullable = true)

|-- Recovered / 100 Cases: double (nullable = true)

|-- Deaths / 100 Recovered: string (nullable = true)

|-- Confirmed last week: integer (nullable = true)

|-- 1 week change: integer (nullable = true)

|-- 1 week % increase: double (nullable = true)

|-- WHO Region: string (nullable = true)

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

|Country/Region|Confirmed|Deaths|Recovered|Active|New cases|New deaths|New recovered|Deaths / 100 Cases|Recovered / 100 Cases|Deaths / 100 Recovered|Confirmed last week|1 week change|1 week % increase| WHO Region|

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

| Afghanistan| 36263| 1269| 25198| 9796| 106| 10| 18| 3.5| 69.49| 5.04| 35526| 737| 2.07|Eastern Mediterra...|

| Albania| 4880| 144| 2745| 1991| 117| 6| 63| 2.95| 56.25| 5.25| 4171| 709| 17.0| Europe|

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

only showing top 2 rows

second Question Answer

root

|-- country: string (nullable = true)

|-- Confirmed: integer (nullable = true)

|-- Deaths: integer (nullable = true)

|-- Recovered: integer (nullable = true)

|-- Active: integer (nullable = true)

|-- New\_cases: integer (nullable = true)

|-- New\_deaths: integer (nullable = true)

|-- New\_recovered: integer (nullable = true)

|-- Deaths\_100\_Cases: double (nullable = true)

|-- Recovered\_100\_Cases: double (nullable = true)

|-- Deaths\_100\_Recovered: string (nullable = true)

|-- Confirmed\_last\_week: integer (nullable = true)

|-- 1\_week\_change: integer (nullable = true)

|-- 1\_week\_%\_increase: double (nullable = true)

|-- WHO\_Region: string (nullable = true)

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

| country|Confirmed|Deaths|Recovered|Active|New\_cases|New\_deaths|New\_recovered|Deaths\_100\_Cases|Recovered\_100\_Cases|Deaths\_100\_Recovered|Confirmed\_last\_week|1\_week\_change|1\_week\_%\_increase| WHO\_Region|

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

|Afghanistan| 36263| 1269| 25198| 9796| 106| 10| 18| 3.5| 69.49| 5.04| 35526| 737| 2.07|Eastern Mediterra...|

| Albania| 4880| 144| 2745| 1991| 117| 6| 63| 2.95| 56.25| 5.25| 4171| 709| 17.0| Europe|

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

only showing top 2 rows

Count and check if there any null values in any of the columns

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

|country|Confirmed|Deaths|Recovered|Active|New\_cases|New\_deaths|New\_recovered|Deaths\_100\_Cases|Recovered\_100\_Cases|Deaths\_100\_Recovered|Confirmed\_last\_week|1\_week\_change|1\_week\_%\_increase|WHO\_Region|

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

| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0| 0|

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

4. What are the top 10 countries under the WHO region with covid 19 Confirmed cases

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

| country| WHO\_Region|sum(Confirmed)|

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

| US| Americas| 4290259|

| Brazil| Americas| 2442375|

| India| South-East Asia| 1480073|

| Russia| Europe| 816680|

| South Africa| Africa| 452529|

| Mexico| Americas| 395489|

| Peru| Americas| 389717|

| Chile| Americas| 347923|

|United Kingdom| Europe| 301708|

| Iran|Eastern Mediterra...| 293606|

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

only showing top 10 rows

5. What are the bottom 10 countries under the WHO region with covid 19 Confirmed cases

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

| Country| WHO\_Region|sum(Confirmed)|

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

| Western Sahara| Africa| 10|

| Holy See| Europe| 12|

| Greenland| Europe| 14|

|Saint Kitts and N...| Americas| 17|

| Dominica| Americas| 18|

| Laos|Western Pacific| 20|

| Grenada| Americas| 23|

| Saint Lucia| Americas| 24|

| Timor-Leste|South-East Asia| 24|

| Fiji|Western Pacific| 27|

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

only showing top 10 rows

6. What are the total number of countries/ total no. of WHO regions and also list the various WHO regions

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

|count(DISTINCT Country)|count(DISTINCT WHO\_Region)| WHO\_Region|

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

| 56| 1| Europe|

| 16| 1| Western Pacific|

| 48| 1| Africa|

| 22| 1|Eastern Mediterra...|

| 35| 1| Americas|

| 10| 1| South-East Asia|

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

Process finished with exit code 0