**Covid 19 Data analysis :**

**Task1:**

1. **Load the covid 19 data in to a spark dataframe**

**(country\_wise\_latest.csv) with the correct schema definition**

[mavricbdhoct07@ip-10-1-1-204 ~]$ hdfs dfs -put country\_wise\_latest.csv

[mavricbdhoct07@ip-10-1-1-204 ~]$ hdfs dfs -ls

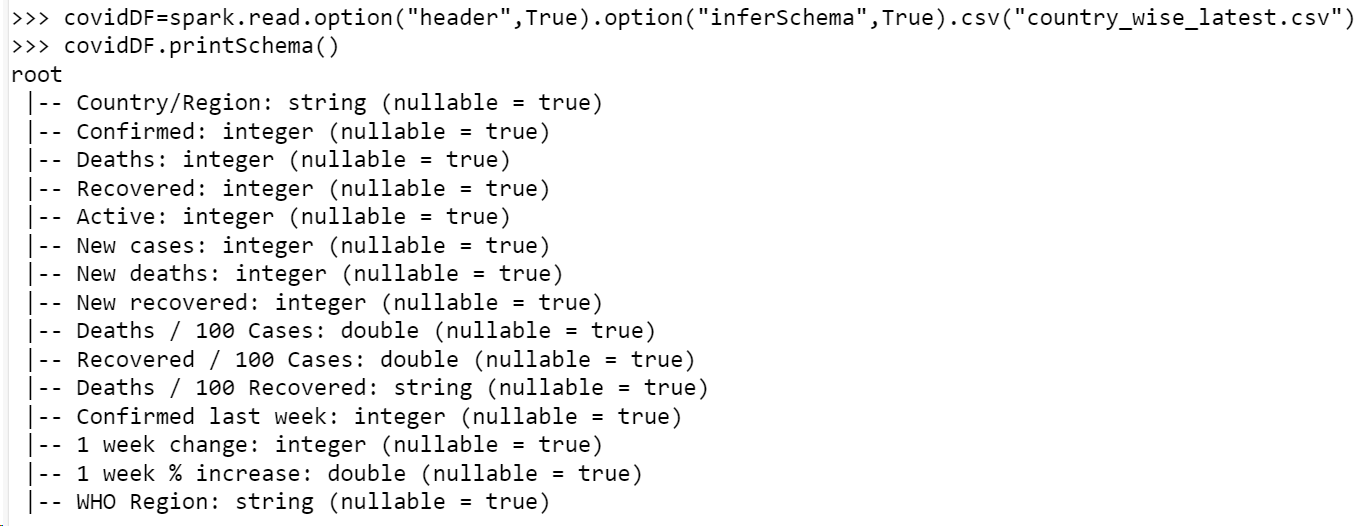
A black and white text

Description automatically generated

**>>>covidDF=spark.read.option("header",True).option("inferSchema",True)**

**.csv("country\_wise\_latest.csv”)**

**>>>covidDF.printSchema()**



**>>>covidDF.show()**

**>>>covidDF.createOrReplaceTempView("covid")**

A screenshot of a computer

Description automatically generated

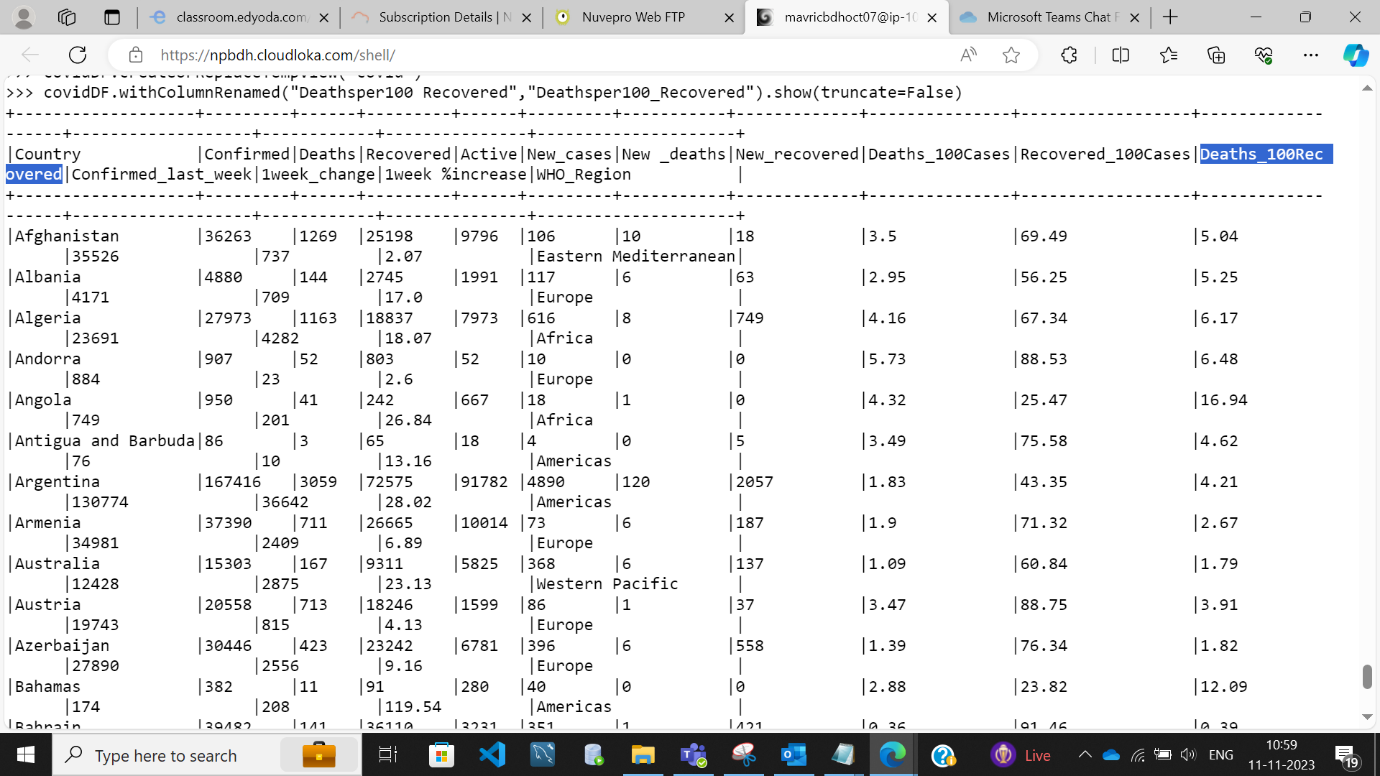
A close-up of a computer code

Description automatically generated

1. **The are some column names which are long, contains specia**

**characters, spaces etc. Rename all such column names accordingly. Example Country/Region → country New cases → New\_cases etc**

covidDF.withColumnRenamed("Deathsper100 Recovered","Deathsper100\_Recovered").show(truncate=False)



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

**import pyspark.sql.functions as F**

**covidDF\_agg = covidDF.agg(\*[F.count(F.when(F.isnull(c), c)).alias(c) for c in covidDF.columns])**

**covidDF\_agg.show()**

A screenshot of a computer

Description automatically generated

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

**Confirmed cases**

spark.sql( "SELECT Country,WHO\_Region,Confirmed FROM covid").show()

spark.sql( "SELECT Country,WHO\_Region,sum(Confirmed) FROM covid GROUP BY Country, WHO\_Region order by 3 desc").show(10)

A screenshot of a computer

Description automatically generated

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

**Confirmed cases**

spark.sql( "SELECT Country,WHO\_Region,Confirmed FROM covid").show()

spark.sql( "SELECT Country,WHO\_Region,sum(Confirmed) FROM covid GROUP BY Country, WHO\_Region order by 3 asc").show(10)

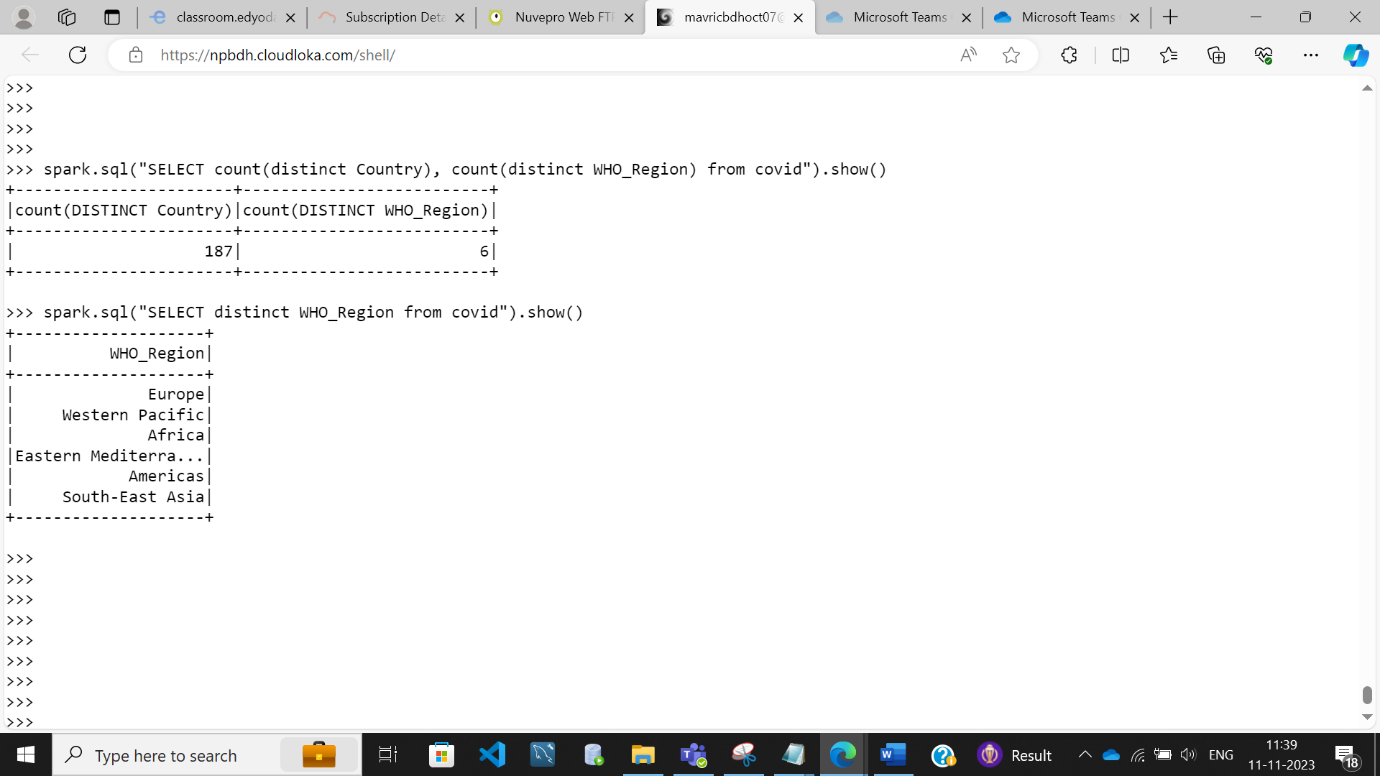
A screenshot of a computer

Description automatically generated

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

spark.sql("SELECT count(distinct Country), count(distinct WHO\_Region) from covid").show()

spark.sql("SELECT distinct WHO\_Region from covid").show()



**spark.sql("SELECT count(distinct Country), count(distinct WHO\_Region),WHO\_Region from covid group by WHO\_Region").show()**

A screenshot of a computer

Description automatically generated