**Spark Assignment**

**Covid 19 Data analysis**

1. Load the covid 19 data in to a spark dataframe (country\_wise\_latest.csv) with the correct schema definition

A screenshot of a computer program

Description automatically generated

A screenshot of a computer screen

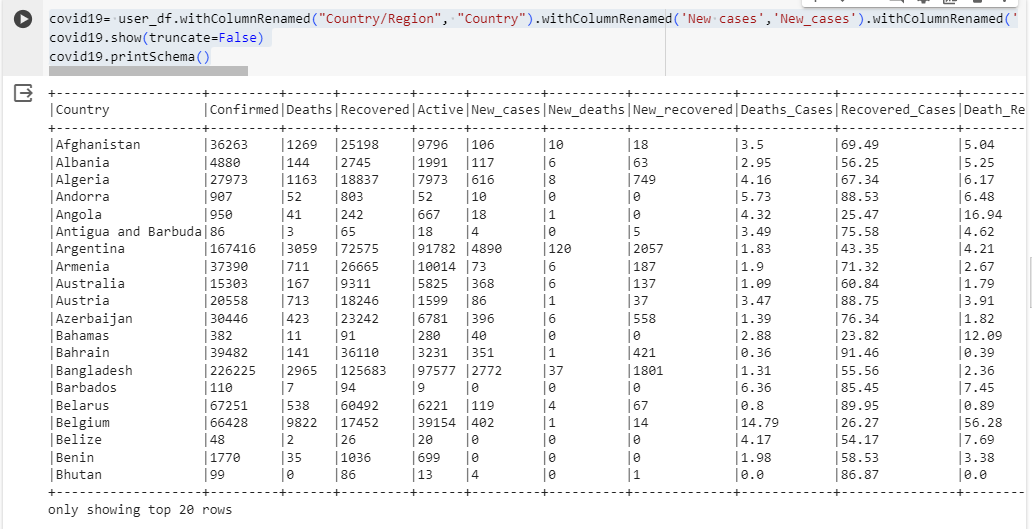
Description automatically generated

1. The are some column names which are long, contains special characters, spaces etc. Rename all such column names accordingly. Example Country/Region → country New cases → New\_cases etc

covid19= user\_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\_Cases').withColumnRenamed('Recovered / 100 Cases','Recovered\_Cases').withColumnRenamed('Deaths / 100 Recovered','Death\_Recovered').withColumnRenamed('Confirmed last week','Confimred\_last\_week').withColumnRenamed('1 week change','1\_week\_change').withColumnRenamed('1 week % increase ','1\_week\_%Increase').withColumnRenamed('WHO Region','WHO\_Region')

covid19.show(truncate=False)

covid19.printSchema()



A screenshot of a computer

Description automatically generated

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

A screenshot of a computer program

Description automatically generated

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

***Note: not used limit function instead filtered top 10 records in show ()***

*>>> spark.sql( "SELECT Country,WHORegion,sum(Confirmed) FROM Covid19 GROUP BY Country, WHORegion order by 3 desc").show(10)*

***Note: Used limit function to get top 10 rows***

*>>> spark.sql( "SELECT Country,WHORegion,sum(Confirmed) FROM Covid19 GROUP BY Country, WHORegion order by 3 desc limit 10 ").show()*

A white background with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

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

*>>> spark.sql( "SELECT Country,WHORegion,sum(Confirmed) FROM Covid19 GROUP BY Country, WHORegion order by 3 asc limit 10").show()*

A screenshot of a computer

Description automatically generated

1. 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 WHORegion),WHORegion from Covid19 group by WHORegion").show()*

A close-up of a white screen

Description automatically generated

*>>> spark.sql("SELECT count(distinct Country), count(distinct WHORegion) from Covid19").show()*

A computer screen shot of a computer code

Description automatically generated