**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

*>>> Covid19DF=spark.read.option("header",True).option("inferschema",True).csv ("country\_wise\_latest.csv")*

*>>> Covid19DF.printSchema()*

*>>> Covid19DF.show()*

*>>> Covid19DF.createOrReplaceTempView("Covid19")*

A screenshot of a computer

Description automatically generated

A close-up of a document

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

*>>> Covid19DF.withColumnRenamed("Deathsper100 recovered", "Deathsper100\_Recovered").printSchema()*

A screenshot of a computer program

Description automatically generated

*>>> Covid19DF.withColumnRenamed("Deathsper100 Recovered","Deathsper100\_Recovered").show(truncate=False)*

A screenshot of a computer

Description automatically generated

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

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

*>>> Covid19DF\_agg = Covid19DF.agg(\*[F.count(F.when(F.isnull(c), c)).alias(c) for c in Covid19DF.columns])*

*>>> Covid19DF\_agg.show()*

A screenshot of a computer

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