Universal College of Engineering, Kaman
Department of Computer Engineering
Subject: Big Data Analytics Laboratory

Experiment No. 08

Roll No: 65	**Name: Hari Jha**	**Div: A**	**Batch: A3**

Aim:

Exploratory Data Analysis using Spark/PySpark.

Theory:

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, particularly when using large datasets where traditional tools might struggle. When working with Apache Spark and PySpark, EDA can leverage the distributed computing capabilities of Spark to efficiently analyze and visualize large volumes of data. Here’s a theoretical overview of EDA in the context of Spark/PySpark:

1. Introduction to Exploratory Data Analysis (EDA)

EDA is a set of techniques used to summarize the main characteristics of a dataset, often using visual methods. It helps in understanding the underlying structure, detecting anomalies, and testing hypotheses.

Key Goals of EDA:

- Identify patterns and trends.
- Detects outliers and anomalies.
- Gain insights into the data distribution.
- Formulate hypotheses for further analysis.

2. Spark and PySpark Overview

Apache Spark is an open-source, distributed computing system that provides an interface for programming entire clusters with implicit data parallelism and fault tolerance. **PySpark** is the Python API for Spark, allowing Python developers to harness the simplicity and power of Python while utilizing Spark's capabilities.

3. Key Concepts in EDA with PySpark

3.1 Data Loading and Transformation

- Loading Data: Use `SparkSession` to load data from various sources (CSV, JSON, Parquet, etc.).
- DataFrames: Spark's primary data structure, similar to pandas DataFrames, allows for distributed data manipulation.

3.2 Data Cleaning and Preprocessing

- Handling Missing Values: Use functions like `fillna()`, `dropna()`, and `replace()`.
- Data Types: Convert data types using `cast()` for accurate analysis.
- Removing Duplicates: Use `dropDuplicates()` to clean up the data.

3.3 Descriptive Statistics

- Use methods such as `describe()`, `summary()`, and `countDistinct()` to get insights into numerical and categorical features.
- Calculate aggregates using `groupBy()` and `agg()` to understand distributions across different categories.

3.4 Data Visualization

While PySpark does not have built-in visualization tools, you can convert DataFrames to pandas DataFrames and use libraries like Matplotlib or Seaborn for visual analysis.

- Histograms: Understand the distribution of numerical features.
- Box Plots: Identify outliers and understand data spread.
- Scatter Plots: Explore relationships between features.

4. Advanced EDA Techniques

4.1 Correlation Analysis

- Use `corr()` to compute correlation coefficients between numerical columns, helping to identify relationships between features.

4.2 Feature Engineering

- Create new features based on existing ones to enhance model performance (e.g., extracting date components).

4.3 Clustering

- Implement clustering algorithms (like K-means) to discover natural groupings in the data.

Conclusion

EDA is a vital part of the data analysis pipeline, especially when using tools like Spark/PySpark. It enables data scientists and analysts to uncover insights, prepare the data for modeling, and ultimately make data-driven decisions. By leveraging the scalability of Spark, one can efficiently perform EDA on large datasets that wouldn't fit into memory with traditional tools.

Code:

Installing PySpark

```
!sudo apt update
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
#Check this site for the latest download link
https://www.apache.org/dyn/closer.lua/spark/spark3.2.1/spark-3.2.1-bin-hadoop3.2.tgz
!wget -q https://dlcdn.apache.org/spark/spark-3.2.1/spark-3.2.1-bin-hadoop3.2.tgz
!tar xf spark-3.2.1-bin-hadoop3.2.tgz
!pip install -q findspark
!pip install pyspark
!pip install py4j
import os import
sys
# os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64" #
os.environ["SPARK_HOME"] = "/content/spark-3.2.1-bin-hadoop3.2"
import findspark findspark.init() findspark.find() import pyspark
from pyspark.sql import DataFrame, SparkSession
from typing import List import pyspark.sql.types
as T import pyspark.sql.functions as F spark=
SparkSession \
        .builder \
        .appName("Our First Spark Example") \
        .getOrCreate() spark
```

Reading Data

```
import requests
```

```
path = "https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owid-covid-data.csv"
req = requests.get(path) url_content = req.content csv_file_name = 'owid-covid-data.csv'
csv_file = open(csv_file_name, 'wb')

csv_file.write(url_content) csv_file.close()
df = spark.read.csv('/content/'+csv_file_name, header=True, inferSchema=True)
```

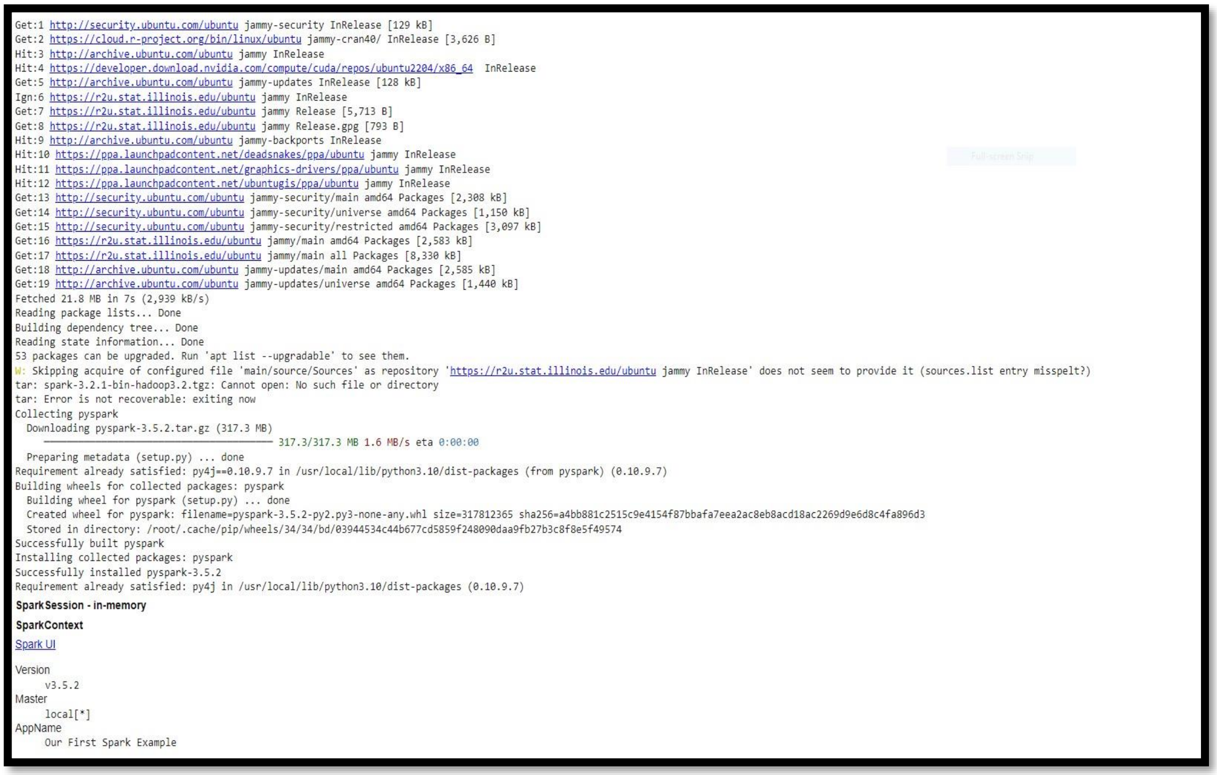
PySpark DataFrames

```
#Viewing the dataframe schema
df.printSchema() #Converting a
date column
df.select(F.to_date(df.date).alias('date'))
#Summary stats
df.describe().show() #DataFrame
Filtering
df.filter(df.location == "United States").orderBy(F.desc("date")).show()
#Simple Group by Function
df.groupBy("location").sum("new_cases").orderBy(F.desc("sum(new_cases)")).show(truncate=False)
```

Spark SQL

```
#Creating a table from the dataframe
df.createOrReplaceTempView("covid_data") #temporary view
# df.saveAsTable("covid_data") #Save as a table
# df.write.mode("overwrite").saveAsTable("covid_data") #Save as table and overwrite table if exits
df2 = spark.sql("SELECT * from covid_data")
df2.printSchema() df2.show()
groupDF = spark.sql("SELECT location, count(*) from covid_data group by location")
groupDF.show()
```

Output:



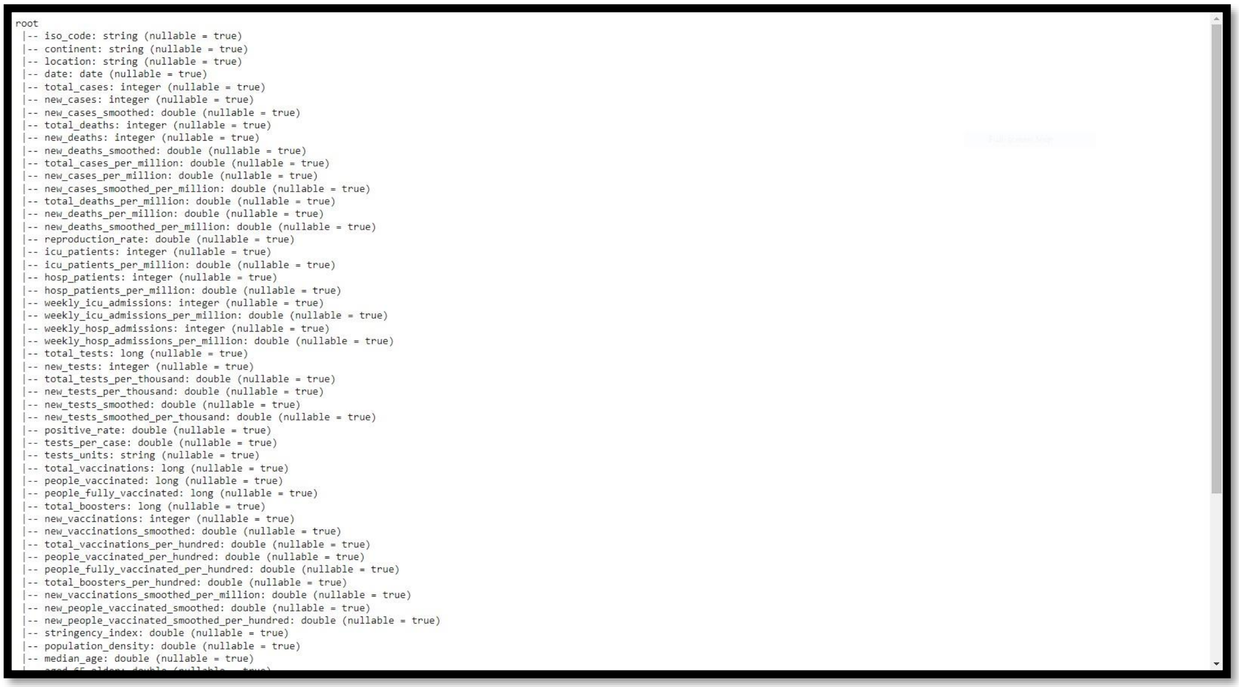
```
Get:1 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Get:2 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease [3,626 B]
Hit:3 http://archive.ubuntu.com/ubuntu jammy InRelease
Hit:4 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease
Get:5 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Ign:6 https://r2u.stat.illinois.edu/ubuntu jammy InRelease
Get:7 https://r2u.stat.illinois.edu/ubuntu jammy Release [5,713 B]
Get:8 https://r2u.stat.illinois.edu/ubuntu jammy Release.gpg [793 B]
Hit:9 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:10 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:11 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
Hit:12 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:13 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages [2,308 kB]
Get:14 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 Packages [1,150 kB]
Get:15 http://security.ubuntu.com/ubuntu jammy-security/restricted amd64 Packages [3,097 kB]
Get:16 https://r2u.stat.illinois.edu/ubuntu jammy/main amd64 Packages [2,583 kB]
Get:17 https://r2u.stat.illinois.edu/ubuntu jammy/main all Packages [8,330 kB]
Get:18 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [2,585 kB]
Get:19 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1,440 kB]
Fetched 21.8 MB in 7s (2,939 kB/s)
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
53 packages can be upgraded. Run 'apt list --upgradable' to see them.
W: Skipping acquire of configured file 'main/source/Sources' as repository 'https://r2u.stat.illinois.edu/ubuntu jammy InRelease' does not seem to provide it (sources.list entry misspelt?)
tar: spark-3.2.1-bin-hadoop3.2.tgz: Cannot open: No such file or directory
tar: Error is not recoverable: exiting now
Collecting pyspark
  Downloading pyspark-3.5.2.tar.gz (317.3 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 317.3/317.3 MB 1.6 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.5.2-py2.py3-none-any.whl size=317812365 sha256=a4bb881c2515c9e4154f87bbafa7eea2ac8eb8acd18ac2269d9e6d8c4fa896d3
  Stored in directory: /root/.cache/pip/wheels/34/34/bd/03944534c44b677cd5859f248090daa9fb27b3c8f8e5f49574
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.5.2
Requirement already satisfied: py4j in /usr/local/lib/python3.10/dist-packages (0.10.9.7)
```

SparkSession - in-memory

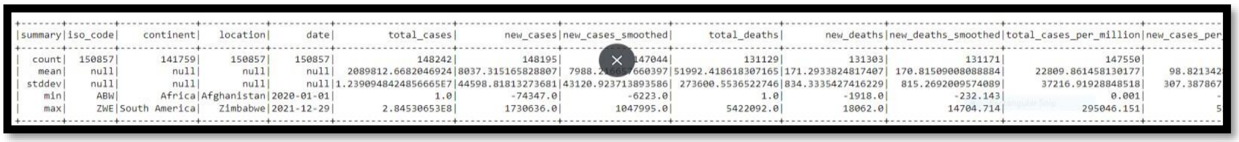
SparkContext

Spark UI

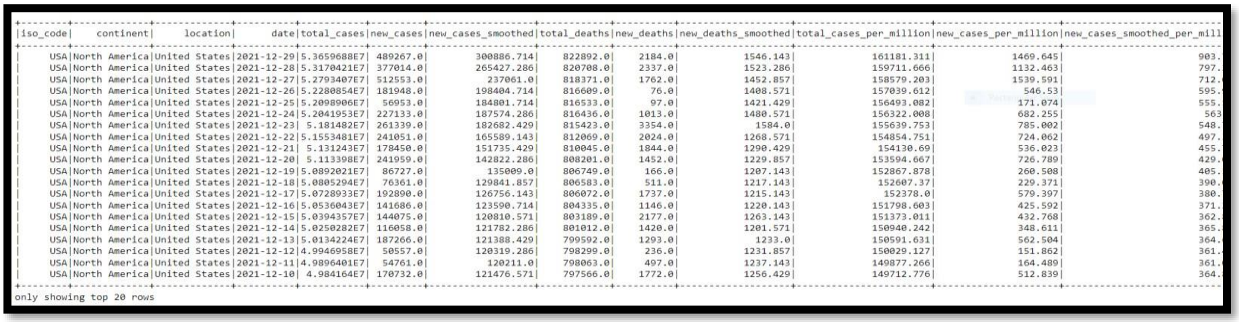
Version
    v3.5.2
Master
    local[*]
AppName
    Our First Spark Example



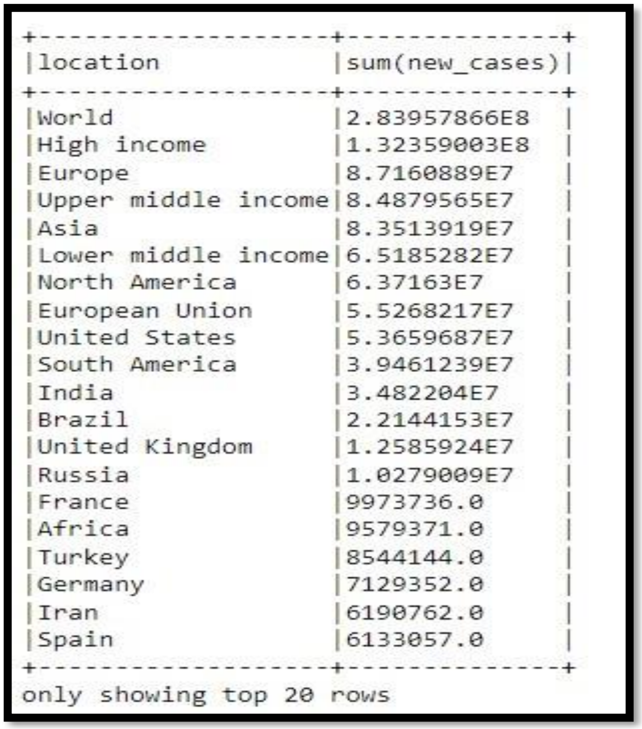
```
root
 |-- iso_code: string (nullable = true)
 |-- continent: string (nullable = true)
 |-- location: string (nullable = true)
 |-- date: date (nullable = true)
 |-- total_cases: integer (nullable = true)
 |-- new_cases: integer (nullable = true)
 |-- new_cases_smoothed: double (nullable = true)
 |-- total_deaths: integer (nullable = true)
 |-- new_deaths: integer (nullable = true)
 |-- new_deaths_smoothed: double (nullable = true)
 |-- total_cases_per_million: double (nullable = true)
 |-- new_cases_per_million: double (nullable = true)
 |-- new_cases_smoothed_per_million: double (nullable = true)
 |-- total_deaths_per_million: double (nullable = true)
 |-- new_deaths_per_million: double (nullable = true)
 |-- new_deaths_smoothed_per_million: double (nullable = true)
 |-- reproduction_rate: double (nullable = true)
 |-- icu_patients: integer (nullable = true)
 |-- icu_patients_per_million: double (nullable = true)
 |-- hosp_patients: integer (nullable = true)
 |-- hosp_patients_per_million: double (nullable = true)
 |-- weekly_icu_admissions: integer (nullable = true)
 |-- weekly_icu_admissions_per_million: double (nullable = true)
 |-- weekly_hosp_admissions: integer (nullable = true)
 |-- weekly_hosp_admissions_per_million: double (nullable = true)
 |-- total_tests: long (nullable = true)
 |-- new_tests: integer (nullable = true)
 |-- total_tests_per_thousand: double (nullable = true)
 |-- new_tests_per_thousand: double (nullable = true)
 |-- new_tests_smoothed: double (nullable = true)
 |-- new_tests_smoothed_per_thousand: double (nullable = true)
 |-- positive_rate: double (nullable = true)
 |-- tests_per_case: double (nullable = true)
 |-- tests_units: string (nullable = true)
 |-- total_vaccinations: long (nullable = true)
 |-- people_vaccinated: long (nullable = true)
 |-- people_fully_vaccinated: long (nullable = true)
 |-- total_boosters: long (nullable = true)
 |-- new_vaccinations: integer (nullable = true)
 |-- new_vaccinations_smoothed: double (nullable = true)
 |-- total_vaccinations_per_hundred: double (nullable = true)
 |-- people_vaccinated_per_hundred: double (nullable = true)
 |-- people_fully_vaccinated_per_hundred: double (nullable = true)
 |-- total_boosters_per_hundred: double (nullable = true)
 |-- new_vaccinations_smoothed_per_million: double (nullable = true)
 |-- new_people_vaccinated_smoothed: double (nullable = true)
 |-- new_people_vaccinated_smoothed_per_hundred: double (nullable = true)
 |-- stringency_index: double (nullable = true)
 |-- population_density: double (nullable = true)
 |-- median_age: double (nullable = true)
 |-- aged_65_older: double (nullable = true)
```



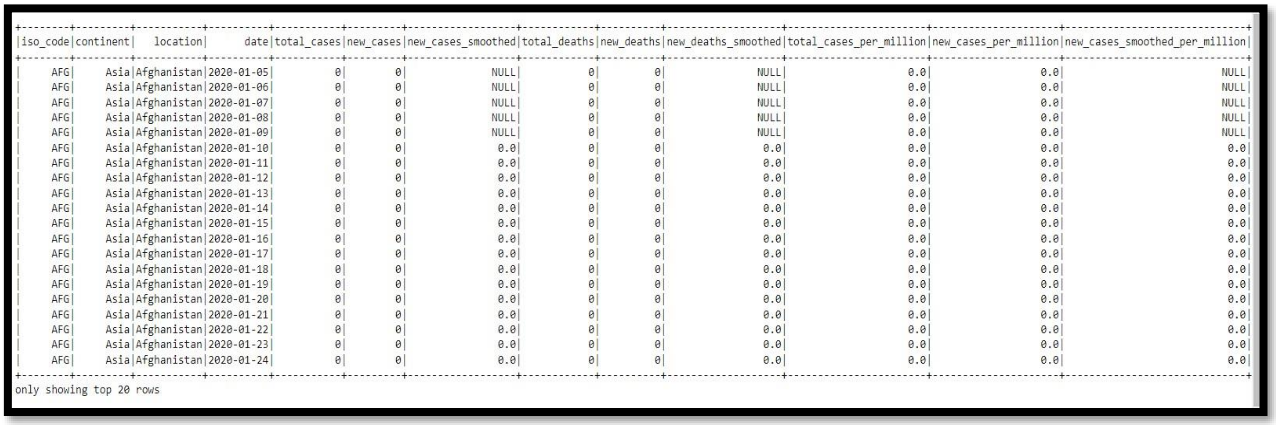
```
+-------+--------+-------------+-----------+----------+--------------------+-----------------+--------------------+------------------+-----------------+------------------+-----------------------+-----------
|summary|iso_code|    continent|   location|      date|         total_cases|        new_cases|new_cases_smoothed|      total_deaths|       new_deaths|new_deaths_smoothed|total_cases_per_million|new_cases_per
+-------+--------+-------------+-----------+----------+--------------------+-----------------+--------------------+------------------+-----------------+------------------+-----------------------+-----------
|  count|  150857|       141759|     150857|    150857|              148242|           148195|              147044|            131129|           131303|            131171|                 147550|
|   mean|    null|         null|       null|      null|  2089812.6682046924| 8037.315165828807| 7988.216657660397| 51992.418618307165|171.2933824817407| 170.81509008088884|     22809.861458130177|    98.821342
| stddev|    null|         null|       null|      null|1.2390948424856665E7|44598.81813273681|43120.923713893586| 273600.5536522746|834.3335427416229|  815.2692009574089|     37216.9192884851 8|   307.387867
|    min|     ABW|       Africa|Afghanistan|2020-01-01|                 1.0|         -74347.0|           -6223.0|               1.0|          -1918.0|           -232.143|                  0.001|            -
|    max|     ZWE|South America|   Zimbabwe|2021-12-29|        2.84530653E8|        1730636.0|         1047995.0|         5422092.0|          18062.0|          14704.714|             295046.151|            5
+-------+--------+-------------+-----------+----------+--------------------+-----------------+--------------------+------------------+-----------------+------------------+-----------------------+-----------
```



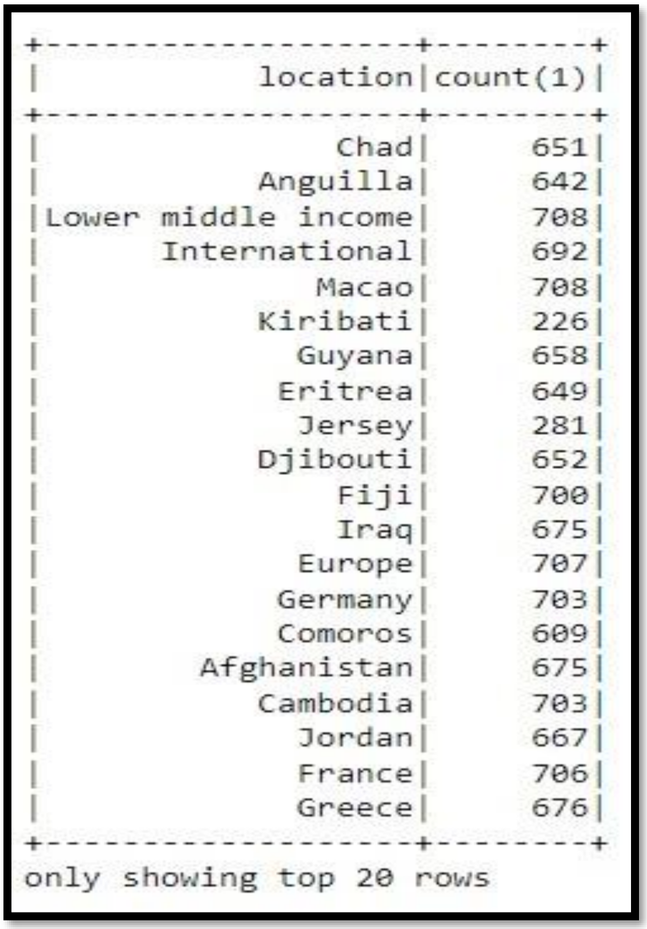
```
+--------+-------------+-------------+----------+------------+----------+------------------+------------+----------+-------------------+-----------------------+---------------------+--------------------------------
|iso_code|    continent|     location|      date| total_cases| new_cases|new_cases_smoothed|total_deaths|new_deaths|new_deaths_smoothed|total_cases_per_million|new_cases_per_million|new_cases_smoothed_per_mill
+--------+-------------+-------------+----------+------------+----------+------------------+------------+----------+-------------------+-----------------------+---------------------+--------------------------------
|     USA|North America|United States|2021-12-29|5.3659688E7|  489267.0|        300886.714|    822892.0|    2184.0|           1546.143|             161181.311|             1469.645|                          903.
|     USA|North America|United States|2021-12-28|5.3170421E7|  377014.0|        265427.286|    820708.0|    2337.0|           1523.286|             159711.666|             1132.463|                          797.
|     USA|North America|United States|2021-12-27|5.2793407E7|  512553.0|          237061.0|    818371.0|    1762.0|           1452.857|             158579.203|             1539.591|                          712.
|     USA|North America|United States|2021-12-26|5.2280854E7|  181948.0|        198404.714|    816609.0|      76.0|           1408.571|             157039.612|               546.53|                          595.
|     USA|North America|United States|2021-12-25|5.2098906E7|   56953.0|        184801.714|    816533.0|      97.0|           1421.429|             156493.082|              171.074|                          555.
|     USA|North America|United States|2021-12-24|5.2041953E7|  227133.0|        187574.286|    816436.0|    1013.0|           1480.571|             156322.008|              682.255|                          563
|     USA|North America|United States|2021-12-23| 5.181482E7|  261339.0|        182682.429|    815423.0|    3354.0|             1584.0|             155639.753|              785.002|                          548.
|     USA|North America|United States|2021-12-22|5.1553481E7|  241051.0|        165589.143|    812069.0|    2024.0|           1268.571|             154854.751|              724.062|                          497.
|     USA|North America|United States|2021-12-21| 5.131243E7|  178450.0|        151735.429|    810045.0|    1844.0|           1290.429|              154130.69|              536.023|                          455.
|     USA|North America|United States|2021-12-20| 5.113398E7|  241959.0|        142822.286|    808201.0|    1452.0|           1229.857|             153594.667|              726.789|                          429.
|     USA|North America|United States|2021-12-19|5.0892021E7|   86727.0|          135009.0|    806749.0|     166.0|           1207.143|             152867.878|              260.508|                          405.
|     USA|North America|United States|2021-12-18|5.0805294E7|   76361.0|        129841.857|    806583.0|     511.0|           1217.143|              152607.37|              229.371|                          390.
|     USA|North America|United States|2021-12-17|5.0728933E7|  192890.0|        126756.143|    806072.0|    1737.0|           1215.143|               152378.0|              579.397|                          380.
|     USA|North America|United States|2021-12-16|5.0536043E7|  141686.0|        123590.714|    804335.0|    1146.0|           1220.143|             151798.603|              425.592|                          371.
|     USA|North America|United States|2021-12-15|5.0394357E7|  144075.0|        120810.571|    803189.0|    2177.0|           1263.143|             151373.011|              432.768|                          362.
|     USA|North America|United States|2021-12-14|5.0250282E7|  116058.0|        121782.286|    801012.0|    1420.0|           1201.571|             150940.242|              348.611|                          365.
|     USA|North America|United States|2021-12-13|5.0134224E7|  187266.0|        121388.429|    799592.0|    1293.0|             1233.0|             150591.631|              562.504|                          364.
|     USA|North America|United States|2021-12-12|4.9946958E7|   50557.0|        120319.286|    798299.0|     236.0|           1231.857|             150029.127|              151.862|                          361.
|     USA|North America|United States|2021-12-11|4.9896401E7|   54761.0|          120211.0|    798063.0|     497.0|           1237.143|             149877.266|              164.489|                          361.
|     USA|North America|United States|2021-12-10| 4.984164E7|  170732.0|        121476.571|    797566.0|    1772.0|           1256.429|             149712.776|              512.839|                          364.
+--------+-------------+-------------+----------+------------+----------+------------------+------------+----------+-------------------+-----------------------+---------------------+--------------------------------
only showing top 20 rows
```



```
+-------------------+--------------+
|location           |sum(new_cases)|
+-------------------+--------------+
|World              |2.83957866E8  |
|High income        |1.32359003E8  |
|Europe             |8.7160889E7   |
|Upper middle income|8.4879565E7   |
|Asia               |8.3513919E7   |
|Lower middle income|6.5185282E7   |
|North America      |6.37163E7     |
|European Union     |5.5268217E7   |
|United States      |5.3659687E7   |
|South America      |3.9461239E7   |
|India              |3.482204E7    |
|Brazil             |2.2144153E7   |
|United Kingdom     |1.2585924E7   |
|Russia             |1.0279009E7   |
|France             |9973736.0     |
|Africa             |9579371.0     |
|Turkey             |8544144.0     |
|Germany            |7129352.0     |
|Iran               |6190762.0     |
|Spain              |6133057.0     |
+-------------------+--------------+
only showing top 20 rows
```



```
+--------+---------+-----------+----------+-----------+---------+------------------+------------+----------+-------------------+-----------------------+---------------------+------------------------------+
|iso_code|continent|   location|      date|total_cases|new_cases|new_cases_smoothed|total_deaths|new_deaths|new_deaths_smoothed|total_cases_per_million|new_cases_per_million|new_cases_smoothed_per_million|
+--------+---------+-----------+----------+-----------+---------+------------------+------------+----------+-------------------+-----------------------+---------------------+------------------------------+
|     AFG|     Asia|Afghanistan|2020-01-05|          0|        0|              NULL|           0|         0|               NULL|                    0.0|                  0.0|                          NULL|
|     AFG|     Asia|Afghanistan|2020-01-06|          0|        0|              NULL|           0|         0|               NULL|                    0.0|                  0.0|                          NULL|
|     AFG|     Asia|Afghanistan|2020-01-07|          0|        0|              NULL|           0|         0|               NULL|                    0.0|                  0.0|                          NULL|
|     AFG|     Asia|Afghanistan|2020-01-08|          0|        0|              NULL|           0|         0|               NULL|                    0.0|                  0.0|                          NULL|
|     AFG|     Asia|Afghanistan|2020-01-09|          0|        0|              NULL|           0|         0|               NULL|                    0.0|                  0.0|                          NULL|
|     AFG|     Asia|Afghanistan|2020-01-10|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-11|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-12|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-13|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-14|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-15|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-16|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-17|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-18|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-19|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-20|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-21|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-22|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-23|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
|     AFG|     Asia|Afghanistan|2020-01-24|          0|        0|               0.0|           0|         0|                0.0|                    0.0|                  0.0|                           0.0|
+--------+---------+-----------+----------+-----------+---------+------------------+------------+----------+-------------------+-----------------------+---------------------+------------------------------+
only showing top 20 rows
```



```
+-------------------+--------+
|           location|count(1)|
+-------------------+--------+
|               Chad|     651|
|           Anguilla|     642|
|Lower middle income|     708|
|      International|     692|
|              Macao|     708|
|           Kiribati|     226|
|             Guyana|     658|
|            Eritrea|     649|
|             Jersey|     281|
|           Djibouti|     652|
|               Fiji|     700|
|               Iraq|     675|
|             Europe|     707|
|            Germany|     703|
|            Comoros|     609|
|        Afghanistan|     675|
|           Cambodia|     703|
|             Jordan|     667|
|             France|     706|
|             Greece|     676|
+-------------------+--------+
only showing top 20 rows
```

Conclusion:

Hence, we have performed Exploratory Data Analysis using PySpark in Google Colab and generated the required output.