CIA Country Analysis SQL Project

Use SparkSQL and SQL to analyze important information on several countries based on CIA data.

```
!pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.11/dist-packages (3.5.5)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.11/dist-packages (from pyspark) (0.10.9.7)
```

```
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("CIA_SQL_Project").getOrCreate()
```

```
file_path = "/content/CIA_Country_Facts.csv"
df = spark.read.csv(file_path, header=True, inferSchema=True)
```

```
df.createOrReplaceTempView("country_facts")
```

View the column names

```
column_names = spark.sql(
    """
    SELECT *
    FROM country_facts
    """
)

column_names.show()
```

```
+----------------+--------------------+----------+--------------+-------------------------+----------------------------+-------------+
|         Country|              Region|Population|Area (sq. mi.)|Pop. Density (per sq. mi.)|Coastline (coast/area ratio)|Net migration|
+----------------+--------------------+----------+--------------+-------------------------+----------------------------+-------------+
|     Afghanistan|ASIA (EX. NEAR EA...|  31056997|        647500|                     48.0|                         0.0|        23.06|
|         Albania|EASTERN EUROPE   ...|   3581655|         28748|                    124.6|                        1.26|        -4.93|
|         Algeria|NORTHERN AFRICA  ...|  32930091|       2381740|                     13.8|                        0.04|        -0.39|
|  American Samoa|OCEANIA          ...|     57794|           199|                    290.4|                       58.29|       -20.71|
|         Andorra|WESTERN EUROPE   ...|     71201|           468|                    152.1|                         0.0|         6.6|
|          Angola|SUB-SAHARAN AFRIC...|  12127071|       1246700|                      9.7|                        0.13|         0.0|
|        Anguilla|LATIN AMER. & CAR...|     13477|           102|                    132.1|                        59.8|        10.76|
|Antigua & Barbuda|LATIN AMER. & CAR...|     69108|           443|                    156.0|                       34.54|        -6.15|
|       Argentina|LATIN AMER. & CAR...|  39921833|       2766890|                     14.4|                        0.18|         0.61|
|         Armenia|C.W. OF IND. STATES |   2976372|         29800|                     99.9|                         0.0|        -6.47|
|           Aruba|LATIN AMER. & CAR...|     71891|           193|                    372.5|                       35.49|         0.0|
|       Australia|OCEANIA          ...|  20264082|       7686850|                      2.6|                        0.34|         3.98|
|         Austria|WESTERN EUROPE   ...|   8192880|         83870|                     97.7|                         0.0|         2.0|
|      Azerbaijan|C.W. OF IND. STATES |   7961619|         86600|                     91.9|                         0.0|         -4.9|
|    Bahamas, The|LATIN AMER. & CAR...|    303770|         13940|                     21.8|                       25.41|         -2.2|
|         Bahrain|NEAR EAST        ...|    698585|           665|                   1050.5|                       24.21|         1.05|
|      Bangladesh|ASIA (EX. NEAR EA...| 147365352|        144000|                   1023.4|                         0.4|        -0.71|
|        Barbados|LATIN AMER. & CAR...|    279912|           431|                    649.5|                       22.51|        -0.31|
|         Belarus|C.W. OF IND. STATES |  10293011|        207600|                     49.6|                         0.0|         2.54|
|         Belgium|WESTERN EUROPE   ...|  10379067|         30528|                    340.0|                        0.22|         1.23|
+----------------+--------------------+----------+--------------+-------------------------+----------------------------+-------------+
only showing top 20 rows
```

Calculate total number of rows.

The row count returned indicates the number of non-null rows.

```
row_count = spark.sql(
    """
    SELECT COUNT(*)
    FROM country_facts
    """
)

row_count.show()
```

```
+--------+
|count(1)|
+--------+
|     227|
+--------+
```

Number of countries in CIA data.

```
total_countries = spark.sql(
    """
    SELECT COUNT(Country)
    FROM country_facts
    """
)

total_countries.show()
```

```
+--------------+
|count(Country)|
+--------------+
|           227|
+--------------+
```

Calculate the population of a region.

Note the total population of a region is based on the combined population of the countries within the given region.

Here, the most populated region is Asia at 3,687,982,236 and the least populated is the Baltics at 7,184,974 people.

```
region_population = spark.sql(
    """
    SELECT DISTINCT Region, SUM(Population) AS Total_Population
    FROM country_facts AS cf
    GROUP BY Region
    ORDER BY Total_Population DESC
    """
)

region_population.show()
```

```
+--------------------+----------------+
|              Region|Total_Population|
+--------------------+----------------+
|ASIA (EX. NEAR EA...|      3687982236|
|SUB-SAHARAN AFRIC...|       749437000|
|LATIN AMER. & CAR...|       561824599|
|WESTERN EUROPE   ...|       396339998|
|NORTHERN AMERICA ...|       331672307|
|C.W. OF IND. STATES |       280081548|
|NEAR EAST        ...|       195068377|
|NORTHERN AFRICA  ...|       161407133|
|EASTERN EUROPE   ...|       119914717|
|OCEANIA          ...|        33131662|
|BALTICS          ...|         7184974|
+--------------------+----------------+
```

List the most and least densely populated regions.

Though ASIA and BALTICS are listed as the most and least populated regions respectively, that is not neccessarily the same as being the most or least dense regions. To determine which region is the most dense, the population density per square mile needs to be considered.

The query below confirms that the column for population density per square mile was calculated by dividing the population of each country with its corresponding area. The original Pop. Density column rounds the values up to the 10th as shown with the first 5 most dense countries.

```
country_density = spark.sql(
    """
    SELECT Country, Population, `Area (sq. mi.)`, `Pop. Density (per sq. mi.)`, Population / `Area (sq. mi.)` AS Calc_pop_density
    FROM country_facts AS cf
```

```
    ORDER BY Calc_pop_density DESC
    LIMIT 5
    """
)

country_density.show()
```

```
+---------+----------+--------------+-------------------------+------------------+
|  Country|Population|Area (sq. mi.)|Pop. Density (per sq. mi.)|  Calc_pop_density|
+---------+----------+--------------+-------------------------+------------------+
|   Monaco|     32543|             2|                  16271.5|           16271.5|
|    Macau|    453125|            28|                  16183.0|16183.035714285714|
|Singapore|   4492150|           693|                   6482.2| 6482.178932178932|
|Hong Kong|   6940432|          1092|                   6355.7| 6355.706959706959|
|Gibraltar|     27928|             7|                   3989.7| 3989.714285714286|
+---------+----------+--------------+-------------------------+------------------+
```

The total population density per square mile for all countries within a region will determine the regional density.

If the average density is being considered, ASIA is both the most populated and dense region. However, though BALTICS is the least populated on average, it is not the least dense. The least dense population on average is NORTHERN AFRICA.

Similarly, regional population and density do not go hand in hand for most other regions listed.

```
regional_density = spark.sql(
    """
    SELECT DISTINCT Region, AVG(`Pop. Density (per sq. mi.)`) AS Avg_Regional_Density
    FROM country_facts AS cf
    GROUP BY Region
    ORDER BY AVG_Regional_Density DESC
    """
)

regional_density.show()
```

```
+--------------------+--------------------+
|              Region|Avg_Regional_Density|
+--------------------+--------------------+
|ASIA (EX. NEAR EA...|            1264.825|
|WESTERN EUROPE   ...|   952.0428571428571|
|NEAR EAST        ...|           427.08125|
|NORTHERN AMERICA ...|              260.86|
|LATIN AMER. & CAR...|  136.20222222222222|
|OCEANIA          ...|   131.1809523809524|
|EASTERN EUROPE   ...|  100.89999999999999|
|SUB-SAHARAN AFRIC...|   92.26470588235293|
|C.W. OF IND. STATES |  56.708333333333336|
|BALTICS          ...|  39.833333333333336|
|NORTHERN AFRICA  ...|   38.93333333333334|
+--------------------+--------------------+
```

Determine correlation between literacy and birth/death rates for countries.

- Calculate the average literacy rate for each country and determine if any trends exist with the average birth and death rates.
- Utilize the query from the first point as a CTE to calculate the number of countries with 100%, NULL, or low literacy rates.

```
country_100_literacy_rate = spark.sql(
    """
    SELECT DISTINCT Country,
    AVG(`Literacy (%)`) AS Avg_Literacy_Rate,
    AVG(Birthrate) AS Avg_Birth_Rate, AVG(Deathrate) AS Avg_Death_Rate,
    AVG(`Infant mortality (per 1000 births)`) AS Avg_Infant_Mortality
    FROM country_facts AS cf
    GROUP BY Country
    HAVING AVG(`Literacy (%)`) = 100
    ORDER BY Avg_Literacy_Rate DESC
    """
)

country_100_literacy_rate.show()

country_100_literacy_rate_count = spark.sql(
    """
```

```
    WITH country_100_literacy AS (
      SELECT DISTINCT Country,
      AVG(`Literacy (%)`) AS Avg_Literacy_Rate,
      AVG(Birthrate) AS Avg_Birth_Rate, AVG(Deathrate) AS Avg_Death_Rate,
      AVG(`Infant mortality (per 1000 births)`) AS Avg_Infant_Mortality
      FROM country_facts AS cf
      GROUP BY Country
      HAVING AVG(`Literacy (%)`) = 100
      ORDER BY Avg_Literacy_Rate DESC
      )

    SELECT COUNT(Country)
    FROM country_100_literacy
    """
)

country_100_literacy_rate_count.show()
```

```
+-------------+-----------------+--------------+--------------+--------------------+
|      Country|Avg_Literacy_Rate|Avg_Birth_Rate|Avg_Death_Rate|Avg_Infant_Mortality|
+-------------+-----------------+--------------+--------------+--------------------+
|      Finland|            100.0|         10.45|          9.86|                3.57|
|      Andorra|            100.0|          8.71|          6.25|                4.05|
|       Norway|            100.0|         11.46|           9.4|                 3.7|
|      Denmark|            100.0|         11.13|         10.36|                4.56|
|Liechtenstein|            100.0|         10.21|          7.18|                 4.7|
|   Luxembourg|            100.0|         11.94|          8.41|                4.81|
|    Australia|            100.0|         12.14|          7.51|                4.69|
+-------------+-----------------+--------------+--------------+--------------------+

+--------------+
|count(Country)|
+--------------+
|             7|
+--------------+
```

```
country_null_literacy_rate = spark.sql(
    """
    SELECT DISTINCT Country,
    AVG(`Literacy (%)`) AS Avg_Literacy_Rate,
    AVG(Birthrate) AS Avg_Birth_Rate, AVG(Deathrate) AS Avg_Death_Rate,
    AVG(`Infant mortality (per 1000 births)`) AS Avg_Infant_Mortality
    FROM country_facts AS cf
    GROUP BY Country
    HAVING AVG(`Literacy (%)`) IS NULL
    ORDER BY 2
    """
)

country_null_literacy_rate.show()

country_null_literacy_rate_count = spark.sql(
    """
    WITH country_null_literacy AS (
      SELECT DISTINCT Country,
      AVG(`Literacy (%)`) AS Avg_Literacy_Rate,
      AVG(Birthrate) AS Avg_Birth_Rate, AVG(Deathrate) AS Avg_Death_Rate,
      AVG(`Infant mortality (per 1000 births)`) AS Avg_Infant_Mortality
      FROM country_facts AS cf
      GROUP BY Country
      HAVING AVG(`Literacy (%)`) IS NULL
      ORDER BY 2
      )

    SELECT COUNT(Country)
    FROM country_null_literacy
    """
)

country_null_literacy_rate_count.show()
```

```
+--------------------+-----------------+--------------+--------------+--------------------+
|             Country|Avg_Literacy_Rate|Avg_Birth_Rate|Avg_Death_Rate|Avg_Infant_Mortality|
+--------------------+-----------------+--------------+--------------+--------------------+
|            Kiribati|             NULL|         30.65|          8.26|               48.52|
|              Jersey|             NULL|           9.3|          9.28|                5.24|
|      Western Sahara|             NULL|          NULL|          NULL|                NULL|
|            Slovakia|             NULL|         10.65|          9.45|                7.41|
```

```
|          West Bank|             NULL|         31.67|          3.92|               19.62|
|           Guernsey|             NULL|          8.81|         10.01|                4.71|
|             Tuvalu|             NULL|         22.18|          7.11|               20.03|
|         Isle of Man|             NULL|         11.05|         11.19|                5.93|
|      Virgin Islands|             NULL|         13.96|          6.43|                8.03|
|          Gaza Strip|             NULL|         39.45|           3.8|               22.93|
|               Nauru|             NULL|         24.76|           6.7|                9.95|
|           Gibraltar|             NULL|         10.74|          9.31|                5.13|
|             Mayotte|             NULL|         40.95|           7.7|                62.4|
|       Faroe Islands|             NULL|         14.05|           8.7|                6.24|
|           Macedonia|             NULL|         12.02|          8.77|               10.09|
|           Greenland|             NULL|         15.93|          7.84|               15.82|
|Bosnia & Herzegovina|             NULL|          8.77|          8.27|               21.05|
|     Solomon Islands|             NULL|         30.01|          3.92|               21.29|
+--------------------+-----------------+--------------+--------------+--------------------+

+--------------+
|count(Country)|
+--------------+
|            18|
+--------------+
```

```
country_lowest_literacy_rate = spark.sql(
    """
    SELECT DISTINCT Country,
    AVG(`Literacy (%)`) AS Avg_Literacy_Rate,
    AVG(Birthrate) AS Avg_Birth_Rate, AVG(Deathrate) AS Avg_Death_Rate,
    AVG(`Infant mortality (per 1000 births)`) AS Avg_Infant_Mortality
    FROM country_facts AS cf
    GROUP BY Country
    HAVING AVG(`Literacy (%)`) < 50
    ORDER BY 2
    """
)

country_lowest_literacy_rate.show()

country_lowest_literacy_rate_count = spark.sql(
    """
    WITH country_lowest_literacy AS (
      SELECT DISTINCT Country,
      AVG(`Literacy (%)`) AS Avg_Literacy_Rate,
      AVG(Birthrate) AS Avg_Birth_Rate, AVG(Deathrate) AS Avg_Death_Rate,
      AVG(`Infant mortality (per 1000 births)`) AS Avg_Infant_Mortality
      FROM country_facts AS cf
      GROUP BY Country
      HAVING AVG(`Literacy (%)`) < 50
      ORDER BY 2
      )

    SELECT COUNT(Country)
    FROM country_lowest_literacy
    """
)

country_lowest_literacy_rate_count.show()
```

```
+-------------+-----------------+--------------+--------------+--------------------+
|      Country|Avg_Literacy_Rate|Avg_Birth_Rate|Avg_Death_Rate|Avg_Infant_Mortality|
+-------------+-----------------+--------------+--------------+--------------------+
|        Niger|             17.6|         50.73|         20.91|              121.69|
| Burkina Faso|             26.6|         45.62|          15.6|               97.57|
| Sierra Leone|             31.4|         45.76|         23.03|              143.64|
|       Guinea|             35.9|         41.76|         15.48|               90.37|
|  Afghanistan|             36.0|          46.6|         20.34|              163.07|
|      Somalia|             37.8|         45.13|         16.63|               116.7|
|  Gambia, The|             40.1|         39.37|         12.25|               72.02|
|      Senegal|             40.2|         32.78|          9.42|               55.51|
|         Iraq|             40.4|         31.98|          5.37|               50.25|
|        Benin|             40.9|         38.85|         12.22|                85.0|
|   Mauritania|             41.7|         40.99|         12.16|               70.89|
|       Angola|             42.0|         45.11|          24.2|              191.19|
|       Bhutan|             42.2|         33.65|          12.7|              100.44|
|Guinea-Bissau|             42.4|         37.22|         16.53|              107.17|
|     Ethiopia|             42.7|         37.98|         14.86|               95.32|
|   Bangladesh|             43.1|          29.8|          8.27|                62.6|
|        Nepal|             45.2|         30.98|          9.31|               66.98|
|     Pakistan|             45.7|         29.74|          8.23|               72.44|
|         Mali|             46.4|         49.82|         16.89|              116.79|
```

```
|         Chad|             47.5|         45.73|         16.38|               93.82|
+-------------+-----------------+--------------+--------------+---------------------+
only showing top 20 rows

+--------------+
|count(Country)|
+--------------+
|            21|
+--------------+
```

Generally, the higher the literacy rate is, the lower the infant mortality rate and vice versa.

For the 7 countries with 100% literacy rate (i.e., Finland, Andorra, Norway, rtc.), the mortality rate for babies is 3 to 5 babies per 1000 births. On the other hand, the literacy rate for a country like Niger is only 17% and their infant mortality rate is very high at 121.69 babies for every 1000 births.

Sometimes, certain countries like Andola will contradict this trend of lower literacy equals higher mortality rate. That is, their literacy is at 42% (higher than Niger), but their infant mortality is higher at 191.19 babies per 1000 births instead of lower (than 121.69).

This might be due to a difference in population density. A hypothesis could be that less density populated regions have less financing for staffing or that it's a rural area with less resoutces.

Determine correlation between literacy and birth/death rates for regions.

```
region_literacy_rates = spark.sql(
    """
    SELECT DISTINCT Region,
    AVG(`Literacy (%)`) AS Avg_Literacy_Rate,
    AVG(Birthrate) AS Avg_Birth_Rate, AVG(Deathrate) AS Avg_Death_Rate,
    AVG(`Infant mortality (per 1000 births)`) AS Avg_Infant_Mortality
    FROM country_facts AS cf
    GROUP BY Region
    ORDER BY Avg_Literacy_Rate
    """
)

region_literacy_rates.show()
```

```
+--------------------+-----------------+------------------+------------------+--------------------+
|              Region|Avg_Literacy_Rate|    Avg_Birth_Rate|    Avg_Death_Rate|Avg_Infant_Mortality|
+--------------------+-----------------+------------------+------------------+--------------------+
|SUB-SAHARAN AFRIC...|            62.51| 36.04392156862746|15.160000000000002|   80.03921568627453|
|NORTHERN AFRICA  ...|            67.24|            20.814| 4.805999999999999|  30.916000000000004|
|NEAR EAST        ...|79.52142857142857|         25.031875|          4.809375|  23.377499999999998|
|ASIA (EX. NEAR EA...|79.55357142857143|21.157857142857146| 7.637142857142856|   41.78000000000001|
|OCEANIA          ...|88.83529411764707|22.108000000000004| 5.810526315789474|  20.203684210526315|
|LATIN AMER. & CAR...|90.65454545454544| 19.08111111111111|             6.376|   20.09266666666667|
|EASTERN EUROPE   ...|97.08888888888889|10.370909090909091|10.284545454545453|  12.686666666666666|
|NORTHERN AMERICA ...|            97.75|            13.154|             7.694|               8.628|
|WESTERN EUROPE   ...| 98.3913043478261|10.553571428571429| 9.354642857142858|   4.730357142857144|
|C.W. OF IND. STATES |98.72500000000001|17.855833333333333|10.341666666666667|               44.41|
|BALTICS          ...|99.73333333333333| 9.343333333333334|             12.63|   8.103333333333333|
+--------------------+-----------------+------------------+------------------+--------------------+
```

The average literacy rate does seem to play a significant role in the corresponding average birth, death, and infant mortality rates. As the literacy rate for the regions increase, the mortality rate seems to decrease for the most part.

The SUB-SAHARAN AFRICA region has the lowest literacy rate among all the regions and the highest infant mortality rate too. The BALTICS region does have the highest literacy rate and the lowest infant mortality rate compared to other regions despite having the lowest average birth rate.

Determine if the average GDP in dollars per capita and average literacy rate have a correlation.

Higher literacy rate does not neccessarily correlate to a higher GDP. The BALTICS have the highest average literacy rate, but its average GDP of $11,300 falls in the middle range in comparison to the other regions. Though the first few regions for SUB-SAHARAN and NORTHERN AFRICA and the NEAR EAST do follow a positive correlation.

```
region_gdp_literacy = spark.sql(
    """
```

```
    SELECT DISTINCT Region,
    AVG(`Literacy (%)`) AS Avg_Literacy_Rate,
    AVG(`GDP ($ per capita)`) AS Avg_GDP
    FROM country_facts AS cf
    GROUP BY Region
    ORDER BY Avg_Literacy_Rate
    """
)

region_gdp_literacy.show()
```

```
+--------------------+-----------------+------------------+
|              Region|Avg_Literacy_Rate|           Avg_GDP|
+--------------------+-----------------+------------------+
|SUB-SAHARAN AFRIC...|            62.51| 2323.529411764706|
|NORTHERN AFRICA  ...|            67.24|            5460.0|
|NEAR EAST        ...|79.52142857142857|          10456.25|
|ASIA (EX. NEAR EA...|79.55357142857143| 8053.571428571428|
|OCEANIA          ...|88.83529411764707| 8247.619047619048|
|LATIN AMER. & CAR...|90.65454545454544| 8682.222222222223|
|EASTERN EUROPE   ...|97.08888888888889| 9808.333333333334|
|NORTHERN AMERICA ...|            97.75|           26100.0|
|WESTERN EUROPE   ...| 98.3913043478261|27046.428571428572|
|C.W. OF IND. STATES |98.72500000000001|            4000.0|
|BALTICS          ...|99.73333333333333|           11300.0|
+--------------------+-----------------+------------------+
```

Determine if crops, climate, and agriculture have a relationship with the GDP.

No obvious correlation or trend appears.

```
region_gdp_crop = spark.sql(
    """
    SELECT DISTINCT Region,
    AVG(`Crops (%)`) AS Avg_crop_percentage,
    AVG(Climate) AS Avg_climate,
    AVG(Agriculture) AS Avg_agriculture,
    AVG(`GDP ($ per capita)`) AS Avg_GDP
    FROM country_facts AS cf
    GROUP BY Region
    ORDER BY Avg_GDP
    """
)

region_gdp_crop.show()
```

```
+--------------------+-------------------+------------------+--------------------+------------------+
|              Region|Avg_crop_percentage|       Avg_climate|     Avg_agriculture|           Avg_GDP|
+--------------------+-------------------+------------------+--------------------+------------------+
|SUB-SAHARAN AFRIC...| 3.7888000000000006|1.8854166666666667|  0.2835510204081633| 2323.529411764706|
|C.W. OF IND. STATES | 2.0224999999999995|              2.55| 0.19200000000000003|            4000.0|
|NORTHERN AFRICA  ...| 2.8049999999999997|               1.5|               0.135|            5460.0|
|ASIA (EX. NEAR EA...|  3.848928571428572| 1.962962962962963| 0.17764285714285713| 8053.571428571428|
|OCEANIA          ...|  14.71952380952381|               2.0| 0.17512499999999998| 8247.619047619048|
|LATIN AMER. & CAR...|  4.913555555555555| 2.033333333333333| 0.09102325581395349| 8682.222222222223|
|EASTERN EUROPE   ...|  2.430833333333333| 3.111111111111111| 0.09216666666666666| 9808.333333333334|
|NEAR EAST        ...|           5.105625|1.6666666666666667|           0.0638125|          10456.25|
|BALTICS          ...|               0.61|               3.0|0.045000000000000005|           11300.0|
|NORTHERN AMERICA ...|              0.048|               2.0|               0.014|           26100.0|
|WESTERN EUROPE   ...| 1.6848148148148143|3.0952380952380953|0.044480000000000006|27046.428571428572|
+--------------------+-------------------+------------------+--------------------+------------------+
```