Top Three Densely Populated Countries: Analyze Factbook Data with SQL

In this project, I will work with data from the CIA World Factbook, a compendium of statistics about all of the countries on Earth to find the top three countries that are the most densely populated. The Factbook contains demographic information like the following:

- `population` — the global population.
- `population_growth` — the annual population growth rate, as a percentage.
- `area` — the total land and water area.

In [52]:
```
%%capture
%load_ext sql
%sql sqlite:///factbook.db
```

Out[52]:
```
'Connected: None@factbook.db'
```

Overview of the Data

I will begin by exploring the data.

In [53]:
```
%%sql
SELECT *
  FROM facts
 LIMIT 5;
```

```
Done.
```

Out[53]:

id	code	name	area	area_land	area_water	population	population_growth	birth_rate	death_rate	migration_rate
1	af	Afghanistan	652230	652230	0	32564342	2.32	38.57	13.89	1.51
2	al	Albania	28748	27398	1350	3029278	0.3	12.92	6.58	3.3
3	ag	Algeria	2381741	2381741	0	39542166	1.84	23.67	4.31	0.92
4	an	Andorra	468	468	0	85580	0.12	8.13	6.96	0.0
5	ao	Angola	1246700	1246700	0	19625353	2.78	38.78	11.49	0.46

Here are the descriptions for some of the columns:

`name` — the name of the country.

`area` — the total land and sea area of the country.

`population` — the country's population.

`population_growth` — the country's population growth as a percentage.

`birth_rate` — the country's birth rate, or the number of births a year per 1,000 people.

`death_rate` — the country's death rate, or the number of death a year per 1,000 people.

`area` — the country's total area (both land and water).

`area_land` — the country's land area in square kilometers.

`area_water` — the country's water area in square kilometers.

I will start by calculating some summary statistics to see what information they provide.

Summary Statistics

In [54]:
```
%%sql
SELECT MIN(population) AS min_pop,
       MAX(population) AS max_pop,
       MIN(population_growth) AS min_pop_growth,
       MAX(population_growth) AS max_pop_growth
FROM facts;
```

```
Done.
```

Out[54]:

min_pop	max_pop	min_pop_growth	max_pop_growth
0	7256490011	0.0	4.02

Exploring Outliers

The last query showed some interesting results:

- There's a country with a population of 0
- There's a country with a population of 7256490011 (or more than 7.2 billion people)

I will now zoom in on these countries to investigate this.

In [55]:
```
%%sql
SELECT name, MIN(population) AS min_population
FROM facts
GROUP BY population
LIMIT 5;
```

```
Done.
```

Out[55]:

name	min_population
Southern Ocean	None
Antarctica	0
Pitcairn Islands	48
Cocos (Keeling) Islands	596
Holy See (Vatican City)	842

In [56]:
```
%%sql
SELECT name, MAX(population) AS max_population
FROM facts
GROUP BY name
ORDER BY max_population DESC
LIMIT 5;
```

```
Done.
```

Out[56]:

name	max_population
World	7256490011
China	1367485388
India	1251695584
European Union	513949445
United States	321368864

It seems like the table contains a row for Antarctica, which explains the population of 0. This matches with the CIA Factbook page for Antarctica.

I also see that the table contains a row for the whole world, which explains the maximum population of over 7.2 billion that was found earlier.

Now I will recalculate the summary statistics that were calculated earlier, while excluding the row for the whole world.

Summary Statistics Revisited

In [57]:
```
%%sql
SELECT MIN(population) AS min_pop,
       MAX(population) AS max_pop,
       MIN(population_growth) AS min_pop_growth,
       MAX(population_growth) AS max_pop_growth
FROM facts
WHERE name <> 'World';
```

```
Done.
```

Out[57]:

min_pop	max_pop	min_pop_growth	max_pop_growth
0	1367485388	0.0	4.02

From the summary statitics that exclude the `world` entry, I can see that the country with the highest population in this database is close to 1.4 billion.

In [58]:
```
%%sql
SELECT AVG(population) AS 'Average Population', AVG(area) AS 'Average Area'
FROM facts
WHERE name <> 'World';
```

```
Done.
```

Out[58]:

Average Population	Average Area
32242666.56846473	555093.546184739

The average population in the facts table is around 32 million and the average area is 555 thousand square kilometers.

Finding Densely Populated Countries

To finish, I will build on the query above to find countries that are densely populated. I will identify countries that have the following:

- Above-average values for population.
- Below-average values for area.

In [59]:
```
%%sql
SELECT name, population, area
FROM facts
WHERE name <> 'World'
AND population > (SELECT AVG(population) FROM facts WHERE name <> 'World')
AND area < (SELECT AVG(area) FROM facts WHERE name <> 'World')
LIMIT 3;
```

```
Done.
```

Out[59]:

name	population	area
Bangladesh	168957745	148460
Germany	80854408	357022
Iraq	37056169	438317

The top three most densely populated countries are:

1. Bangladesh
2. Germany
3. Iraq

I have confidence in these results because these countries are commonly known to be densely populated.