111-1 Database - HW1

HW1 requests:

1. Launch a database service with PostgreSQL engines
2. Create a “covid19” database (10pts)
3. In the covid19 database, create tables for the following three .csv files with public schema, and set a suitable primary key set for each table.
   a. Source 1: (creating the new tables, setting data type, and keys, 5pts each) https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/10-11-2022.csv
   b. Source 2: (creating the new tables, setting the data type, and keys) https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/10-01-2022.csv
   c. Source 3: (creating the new tables, setting the data type, and keys, 5pts each)
      Country code and continents mapping table
      https://gist.github.com/stevewithington/20a69c0b6d2ff846ea5d35e5fc47f26c#file-country-and-continent-codes-list-csv-csv
   d. Coding book
   e. Hint
4. Try to write queries in Query Tool to
   a. extract the total case number (Confirmed) in California, US on 2022-10-11 **or 2022-10-12** (data update date, you can use either date for answering the question, and if you haven’t started/or submitted your answers, please use both dates to choose the date) (10 pts)
   b. extract the total case number (Confirmed) in California, US on 2022-10-01 **or 2022-10-02** (data update date) (10 pts)
   c. extract the diagnosed case number (Confirmed) in California, US on 2022-10-11 **or 2022-10-12**, compared with 2022-10-01 **or 2022-10-02** (data update date), and return the difference between them (=newly diagnosed case number on 2022-10-11 or 2022-10-12), in one SQL statement (please don’t just do 4a-4b in excel or other software) (10 pts)
   d. extract the country names (return Country_Region column) and total confirmed COVID cases (return Confirmed column) with more than 20,000,000 total COVID-19 cases on 2022-10-11 **or 2022-10-12** (data update date) (10 pts)
   e. extract the country names (return Country_Region column) and total confirmed COVID cases (return Confirmed column) with more than 20,000,000 total COVID-19 cases on 2022-10-11 **or 2022-10-12** (data update date). Try to join the Country code and continents mapping table, and return only the data from countries in Asia. (10 pts)
   f. extract the country names (return Country_Region column) and newly diagnosed case number of countries with a newly diagnosed case number (calculate the number by yourself) > 100,000 on 2022-10-11 **or 2022-10-12**

(compare with 2022-10-01 **or 2022-10-02**). In descending order of newly diagnosed case numbers. (10 pts)

5. Please **describe** the detailed process, including launching the DB, creating the new tables, setting the attributes, data type, and keys, and the query process with query results in a PDF file, following the format we provided. (extra 10pts for redesigning the table schema and detailed description)
6. submit the PDF file to the E3 system HW1 section by **2022/10/27 23:59**

HW1 format:

1. The process of creating the "covid19" databases (can be screenshot and/or SQL/non-SQL statements with text explanation) (10pts)
   Ans:

2. The process of importing three required .csv files into covid19 database (can be screenshot and/or SQL/non-SQL statements with text explanation). Please included/described the data type and keys of the imported table in your screenshot, SQL statements, and explanations (30pts)
   Ans:

3. The **SQL statements** and **output results** of 4a (10pt). If the SQL statements or output results are not provided, you will not get the points.
   Ans:

4. The SQL statements and output results of 4b (10pt)
   Ans:

5. The SQL statements and output results of 4c (10pt)
   Ans:

6. The SQL statements and output results of 4d (10pt)
   Ans:

7. The SQL statements and output results of 4e (10pt)
   Ans:

8. The SQL statements and output results of 4f (10pt)
   Ans:

HW1 Q&A:

Q: The files on GitHub are changed over time, which versions should I use for the HW?
A:
Feel free to use the one you downloaded. Don’t worry if you are not using the most updated one. You will get the score if the SQL statement and logic are correct.

Q: How to write a partial match using SQL statement? Should we have to use partial match when we deal with country names?
A:
Usually, we will split the string first and then do the partial match. For example, split_part, left, right, substring, or other functions that can split the string.
In this HW, you may not need to do the partial match.

Q: In the data I downloaded, there is no data update date “2022-10-01” and “2022-10-11”. What happened?
A:
Due to the data update issues, we have modified the questions. In all the questions using “2022-10-01”, you can use “2022-10-02” in these questions. In all the questions using “2022-10-11”, you can use “2022-10-12” in these questions.

Q: Some countries have different names in “country-and-continent-codes-list-csv.csv”, for example, Korea, South or Korea, Republic of. Do we need to keep all the names or have some modifications? Or can we merge or edit some data?
A:
You should keep the name consistent throughout the databases, so please modify the country name in **“country-and-continent-codes-list-csv.csv”**. And please describe what you have done in the answer sheet. You do not need to use only SQL statements to do the modification part. All the strategies are welcome and acceptable.

Q: Some countries have different names in “country-and-continent-codes-list-csv.csv”. Can I use another dataset that I found on the internet to map the country?
A:
If you can use the other dataset to do the same work, you can use another country list to fulfill the requirements of this HW. Please describe in detail what data you use and your data process steps.

Q: Is “diamond princess” a country?
A:
No. “Diamond princess” is a really special case in epidemiology. That’s the reason why these cases were listed separately, and not included in the cases of Japan.

Q: Do we have to use pgAdmin 4 to access PostgreSQL?
A:

No. You do not need to use pgAdmin 4 to access PostgreSQL. You can use other tools to access your database with PostgreSQL.

Q: I want to create an attribute with a double data type, but I get an error message saying: " type double doesn't exist". Why?
A:
You can check the official manual for PostgreSQL. Maybe double precision is the data type you need in the PostgreSQL environment?

Q: I get an error message saying that the "Big5" encoding is not supported in the PostgreSQL environment. What can I do?
A:
Maybe you can try to convert the encoding, from "Big5" to "utf8" by using the other software. For example, VS code, Notepad++, etc. After encoding conversion, you may try to import the file again.