DST2 COVID-19 mini-project

Roll number: 9057

Please note that the structure of the documentation is not 100% based on the guidance. Part 3 and 4 are integrated into part 1b. Part 3 is replaced by "details on the python script", which briefly explains how the python code works. Part 4 is the instructions for building the database.

1. The design of the software

a. User interface description

After running the code, please make the user interface into full screen. The three buttons at the left-hand side are for query 1, while the two buttons at the right are for query 2. Once you click one of the buttons of query 1 or 2, you enter this query. The buttons are quite similar to the demo's. After you type in all the input, you need to press enter to get the result. If you want to change a query, you can simply click one of the buttons of that query and the results of the previous query will be cleared. If your input is wrong, the warning massages will be displayed on the screen.

b. The provided queries

Query 1: What is the number of confirmed cases in a certain country or province in the whole year or a certain month?

Users can choose to query either country or province. If you want to query country, you do not need to type in anything in the province field. However, if you want to query province, you should input both country and province, and they should be matched. Users can also choose to query either a specific month or the whole year of 2020. The table below shows how the outputs related to different inputs.

	Country only	Country + province
Year (2020)	A table which contains the number of	Similar to the Year + country query,
	confirmed cases in the last day of	but the output will be the confirmed
	every month of the input country (e.g.:	cases of the input province.
	1.31, 2.29).	
	A plot which shows the trend of the	
	confirmed cases.	
Month (an	A table which contains the confirmed	Similar to the Month + country query,
integer	cases in every day of the input month	but the output will be the confirmed
between 1 and	and country. (e.g.: input month = 1,	cases of the input province.
12)	output: 1.22-1.31)	
	A plot which shows the trend of the	
	confirmed cases.	

Note that once the query occurs, the plot is saved in the same directory where the python script is. Tables cannot be saved.

Example input & output:

Input	Table	Plot
Country:	(2020-02-01,599) (2020-02-02,661) (2020-02-03,724)	Comfirmed COVID-19 cases in Zhejiang in 2
China	(2020-02-04,829) (2020-02-05,895) (2020-02-06,954)	1200 -
	(2020-02-07,1006) (2020-02-08,1048) (2020-02-09,1075)	1100
Province:	(2020-02-10,1092) (2020-02-11,1117) (2020-02-12,1131)	§ 1000 -
Zhejiang	(2020-02-13,1145) (2020-02-14,1155) (2020-02-15,1162)	900 -
	(2020-02-16,1167) (2020-02-17,1171) (2020-02-18,1172)	S 800 - S 1000 - S 10
Year/month	(2020-02-19,1174) (2020-02-20,1175) (2020-02-21,1203) (2020-02-22,1205) (2020-02-23,1205) (2020-02-24,1205)	700 -
: 2	(2020-02-25,1205) (2020-02-23,1205) (2020-02-24,1205)	600 -
. 2	(2020-02-28,1205) (2020-02-29,1205)	
	(,	
		\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
Country:	(2020-01-31,9802) (2020-02-29,79356) (2020-03-31,82279)	Comfirmed COVID-19 cases in US in 2020
US	(2020-04-30,83956) (2020-05-31,84146) (2020-06-30,84785)	2000000 -
	(2020-07-31,87655) (2020-08-31,89914) (2020-09-30,90545)	15000000 -
Province:	(2020-10-31,91366) (2020-11-30,92902) (2020-12-29,95797)	% 12500000 -
Blank		9 10000000 - E 7500000 -
Year/month		5000000 - 2500000 -
		0 -
2020		\$\\\ \delta_1 \delta_
2020		to the to

Why do we design this query in this way? Since we have a large amount of data, our query needs to fetch the data that we want quickly and conveniently. If we want to know how many confirmed cases in China in 2020-3-12, we can input "China" and "3". Then, we can easily find 2020-3-12 in the result. A large amount of data is also difficult to interpret, that's why include a plot. We can get a rough idea of the trend from the plot easily.

Query 2: Which country has the fastest (or second or third ... to the tenth place) COVID-19 diagnosis in a certain mouth (users can input the mouth for themselves)? The rate of COVID-19 diagnosis is defined as (confirmed cases this mouth- last mouth)/last month. According to this definition, users can input month (2-12) and n th place (1-10). The output will be the country name and the rate. If a country's last month confirmed cases is 0, then this country does not have the diagnosis rate in this month.

Example input & output:

Input	Output	
Month: 12	Cyprus	
N th place: 10	1.0175106483672502	

2. Database

a. The database used

Two databases are used: countries-aggregated, time-series-19-covid-combined.

"countries-aggregated" contains date, country, confirmed, recovered and deaths.

"time-series-19-covid-combined" contains the information of provinces of some countries (but not all countries).

List of countries which have the province information: Australia, Canada, China, Denmark, France, Netherlands, United Kingdom. Although there is province information of US in us_confirmed.csv, the data is for the specific counties in a certain state. It is difficult to get the data of whole states. So I decided not to include province information of US.

I only focused on the confirmed cases in this project.

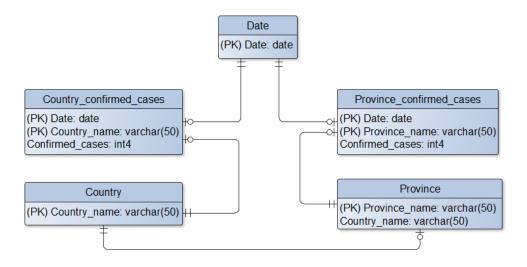
Note that all the databases have the date from 2020-1-22 to 2020-12-29.

b. The design of the database

Entities represent various real-word notion. The entities in the provided dataset can be country, province and date. The attribute of the country entity is just the country name (also serves as primary key). The attribute of date is a list of the date of 2020 (from 2020-1-1 to 2020-12-31) (also serves as primary key). Since not all countries have province information, it is appropriate to set the province name as the primary key in the province entity. The second attribute will be the country that the province belongs to. It links to the country entity (as a foreign key). There is no need to add an id to the province table since there are no two countries that have the provinces with the same name. Another possible way to build the province entity is to set the country name as the primary key. So the province name will be a multivalued attribute. However, I am not sure how to create a multivalued attribute in pgAdmin.

The confirmed cases will only be meaningful if both date and country/province name are provided. Therefore, it is necessary to create two tables that contain country/province name, date and confirmed cases. In these two tables, name and date will be the primary keys. It is a bit like the composite unique attributes. None of the three attributes is unique, but the combination of date and country/province name is unique. We link the tables with the entities and obtain the following ER diagram.

c. Database's ERD



Some explanation to the ERD:

Not all countries have province information, but a province must belong to a country. Thus, the cardinality constraint of country to province is optional one. The opposite cardinality constraint should be mandatory one.

The cardinality constraints of date to country_confirmed_cases and province_confirmed_cases are optional one, since we don't have records for every day in 2020. Similarly, the cardinality constraints of country/province to country_confirmed_cases and province_confirmed_cases are also optional one. It seems that we have data for all countries/province. However, data missing should be allowed. It is possible that we do not have data for a certain country/province in a future date. The date, country and province which appears in country_confirmed_cases and province_confirmed_cases should be found in the date, country and province entities. So the

opposite cardinality constraints should be mandatory one.

One may ask why it is necessary to create an entity with just one attribute (country & date). This is because both country_confirmed_cases and province_confirmed_cases have two primary keys. Two primary keys should be linked to other entities (form foreign keys). Otherwise, pgAdmin will give you an error when you are creating the tables. The right-hand side is an example from the dvdental database. Another reason is that the country and date are real-world notions. They agree with the definition of the entity. Moreover, the country name is the only information we can get from the provided data. We can add more information to the country entity, such as population, continent etc, if we have that information. We can also add holidays etc to date entity. Then, we can do other interesting queries such as which continent has the highest confirmed cases or whether holidays increase the spread of COVID-19 and so on.



3. Details on the python script

The python script can be divided into two parts. The first part defines all variables and functions. The second part builds up the user interface, possesses the data derived from SQL guery and displays the data in the user interface.

a. The first part (focused on functions)connect_postgres: to connect Python with PostgreSQL.

For query 1:

- 1. return_country: return the query result of the input country. Query from "Country_confirmed_cases" using the pre-defined SQL function "return_country" (code can be found in "return_country.sql" file). The SQL function basically contains if statement to judge the input month/year and find the country using "where" clause.
- check_country_province: check whether the input country and province match. If not, the
 function returns []. Query from Country_province (using "SELECT Country_name from
 Country_province where country_name = country (input) and province_name = province
 (input)"). If country and province do not match, the return will be [].
- 3. return_province: return the query result of the input province. Query from "province_confirmed_cases" using pre-defined SQL function "return_province" (similar to

```
"return_country").
```

4. plot: gather necessary information for plotting the data. Since the ('(2020-12-01,13791945)' ('(2020-12-02,13992765)' query return is a list of string (date and confirmed cases are combined into one string, see figure), we need to extract date and confirmed case into two lists respectively (using regex expression).

```
In [2]: query1_return
Out[2]:
 ('(2020-12-03,14212649)',),
 ('(2020-12-04,14442788)
 ('(2020-12-05,14658875)'
 ('(2020-12-06,14840269)',),
 ('(2020-12-07,15030195)
 ('(2020-12-08,15249204)
```

For query 2:

1. return rate: return the necessary data for the calculation of the rate of diagnosis. For example, if the input month is 2, the function will return a list of confirmed cases of Jan. 31st and Feb. 29th of each country using pre-defined SQL function "return rate" (similar to "return country"). The rate calculation is done in the pygame part.

b. The second part

1. How we display the data? (for query 1)

The data we get from the SQL query is quite dirty. If we simply display the data on the screen, some words will overflow the screen. This is because pyGame will print all records line by line. For instance, if the input month is 3, the return has 31 lines. The screen is not big enough to display 31 lines. Therefore, in line 328-338 and 349-359, we are trying to combine 3 records into one line.

2. How we calculate the diagnosis rate? (for query 2, in line 496-515)

It is easy to calculate the fastest diagnosis rate. We calculate the rate country by country and record the rate and country if a bigger rate appears. For 2nd, 3rd, 4th to 10th place, we use while-loop. In every-loop, we record the highest rate and pass the highest rate to the previous rate. In the next loop, only of the rate that is higher than the current rate and lower than the previous rate will be recorded. That is, if we want to know the n th place, we need to loop for n times.

4. Instructions for building the database

The following steps will help you build the same database as mine. Please follow them if you want to run my code.

- 1. Create a database named "DST2final_project" using pgAdmin. Change the user_pass to your database password in the python script.
- 2. Create tables using the SQL code in the create table.sql file using pgAdmin. The following website can help you import the csv files into the tables. https://www.postgresqltutorial.com/import-csv-file-into-posgresql-table/
- 3. Import all three SQL functions using pgAdmin.
- 4. Hurray!!! Now you can run the code. If you encounter any problem, feel free to contact 1742673512@qq.com. It does not contain information that can show my identity.

When running the code, you may find it a bit difficult to type in words if there is a figure displayed in the user interface. This is because showing the figure slow down the pygame loop. Thus, the computer may not catch every keypress. So please be patient! Please do not open the plot when the code is running.