Designing a Relational Database: Part III

PROJECT: Modelling the Vaccine Distribution in Finland

- In the virtual server, you can start Sqlite with the command `sqlite3`.
- Open your existing database using `.open NAME OF DATABASE`
- View all the tables by typing in `.tables`
- View the structure of a table by typing in `.schema NAME OF TABLE`

Your final task is to now perform data analysis. You and your group members are expected to uncover trends for side-effects of different vaccine types. More specifically, you are to use pandas and SQL to do the following:

Part III

Your final task is to now perform data analysis. You and your group members are expected to uncover trends in the given data. You will be given data analysis tasks which you should solve by using pandas library and SQL queries.

The tasks (2), (6), (7), and (10) are slightly more difficult tasks, so reserve enough time to work on them.

(1/4) Requirements: Analysing the data

1. Create a dataframe for patients and symptoms containing the following columns: (1) ssNO, (2) gender, (3) dateOfBirth, (4) symptom, (5) diagnosisDate. Create a table named "PatientSymptoms" using the command `to_sql` with options `index = True`, `if exists = "replace"`.
2. Create a dataframe for patients and vaccines containing the following columns: (1) patientssNO, (2) date1, (3) vaccinetype1, (4) date2, (5) vaccinetype2. The attribute "date1" and "date2" refer to the date when the first and/or second dose were given to a patient respectively. Similarly, "vaccinetype1" and "vaccinetype2" are the type of vaccine used for the first and/or second dose. The value of the attribute should be NULL if the patient has not received some dose. Create a table named "PatientVaccineInfo" using the dataframe as in Task 1.

(2/4) Requirements: Analysing the data

3. Create a dataframe using the table ”PatientSymptoms” and separate it into two dataframes, one for males and one for females. What are the top three most common symptoms for males and females?
4. Create a dataframe using table ”Patient” and add the ”ageGroup” column for each patient. The age groups are ”0-10”, ”10-20”, ”20-40”, ”40-60”, ”60+”
5. Using the same dataframe as in the previous step, add a column describing each patient's vaccination status. The statuses are defined as ”0” for not vaccinated, ”1” for vaccinated once, and ”2” for fully-vaccinated.
6. For each age group, calculate the percentage of people who have received zero, one, or two doses of vaccines. Show the results in a dataframe, where the index is the vaccination status from task (5) and the columns are the age groups. The sum over each age group column should be 100%.
   EXTRA: Solve this task using pivoting.

(3/4) Requirements: Analysing the data

7. Create a dataframe for symptoms with three additional columns: 'V01', 'V02', and 'V03'. The columns should tell the relative frequency of the symptom with the following values:

≥ 0.1	”very common”
≥ 0.05	”common”
> 0.0	”rare”
0.0	”-”

8. Estimate the amount of vaccines (as a percentage) that should be reserved for each vaccination to minimize waste. Do this by first finding the expected percentage of patients that will attend and increase the number by STD of the percentage of attending patients.

(4/4) Requirements: Analysing the data

9. Plot the total number of vaccinated patients with respect to date (Hint: functions `cumsum()` and `strftime()`).
   EXTRA: Plot the number of patients who have gotten two doses to the same figure.
10. Suppose that we found out that the nurse with ssNo "19740919-7140" has been tested positive for corona on 15.5.2021. You should find the social security numbers and names of the patients and staff members that the nurse may have met in vaccination events in the past 10 days? (You are allowed to solve this task using multiple steps and queries).