Fundamentals of Data Science and Engineering

Databases • Practical Assignment • 28-11-2021

Each group will deliver its project inside a **zip** containing, at least, a README file. This file should contain a **brief** description of what was done and the contribution of each group member.

Words highlighted in **yellow** represent the name of the files expected in the delivered zip file.

1) Download Files.

a) **Download** the *world-happiness* dataset from here. Save the five files (one per year as "wh-2015.csv, wh-2016.csv, wh-2017.csv, wh-2018.csv, wh-2019.csv"). This dataset contains results from a world annual survey about the state of global happiness.
b) **Download** the *countries-of-the-world* ("countries of the world.csv") dataset from here. This dataset has static data about the **population**, **area**, coastline, migration, **infant mortality**, **GDP,** and **literacy**.

2) Design the Database.

a) **Draw** a UML diagram of a database capable of holding data about each country's **population**, **area**, **infant mortality**, **GDP,** and **literacy**, and the **score** column of the *world-happiness* (uml.png).
b) **Convert** this diagram into the relational model (relational.txt).
c) **Write** a SQL script that creates the database (happiness.sql).
d) **Create** the corresponding tables in your PostgreSQL database.

3) Prepare Data.

a) You noticed that some countries do not have the same name in both datasets. **Write** a Python script (compare-countries.py) that reads the *world-happiness* dataset files and compares each country's name against the *countries-of-the-world* dataset. If a country in the *world-happiness* dataset does not exist in the *countries-of-the-world* dataset, write that country to the console to be corrected/removed (manually or using another Python script).

4) Load Data.

a) Create a Python script (load_countries_of_the_world.py) that:
   i) removes all data from the database (using the DELETE command),

    ii) reads the countries of the world.csv file, and
    iii) populates the database with new data (using the INSERT command).

b) Create a Python script (load_happiness.py) that:
    i) removes all happiness data from the database for a certain year (passed as an argument)
    ii) reads data from one of the world-happiness datasets files (passed as an argument), and
    iii) populates the database with data from that file and year.

Note: To receive arguments in Python, you can use the following code (example.py):

```
import sys

print (sys.argv[1])
print (sys.argv[2])
```

If called like "example.py 2015 wh-2015.csv" this would print:

```
2015
wh-2015.csv
```

5) Ask Questions

Ask the following questions using SQL (question.sql):

a) What was the happiest country each year (year, country)?
b) What was the average happiness each year (year, happiness)?
c) In which position was Portugal in the happiness index each year (year, position)?
d) What is the average happiness of the 10 countries with the higher GDP in each year (year, happiness)? What about the lower GDP, or higher infant mortality or literacy?
e) What are the three countries with a greater improvement in the happiness index during the years (notice that we can have more years than those in the dataset) present in the database (country, improvement)? What about those that had the larger regression in the index?

Extra points: Think of other interesting questions to ask!

6) Extra

Use other libraries (panda, matplotlib, sci-kit, …) to extract meaningful information from the database. For example, output charts that show the relationship between different variables in the datasets (e.g., GDP vs literacy, happiness vs GDP, …) or histograms for each variable.