Lab 1: Why Databases?
Part 2

Due date: Tuesday, January 24, 7:00pm

Note: Lab 2 will be assigned on Tuesday in class, so please be ready to start working on Lab 2 during the January 24 lab period.

Lab Assignment

Assignment Preparation

This is an individual assignment designed to accomplish two goals. First, it demonstrates the kinds of tasks that are typically accomplished by the database management systems, the roles DBMS play in data processing and delivery of results to end users. Second, it also tests your knowledge of Python, and potentially mildly encourages you to learn Python's declarative programming techniques for working with data.

The Task

This is the second part of your Lab 1 assignment. This part consists of another Jupyter Python notebook shared with you on the course web page. This time, it comes with **two** accompanying data files (a CSV file).

Data. You will be working with two data files that come from two different Kaggle datasets. The first file, `world_population.csv` records information about world population at different times in the 20th and 21st century for every country in the world. The file has two columns identifying the country (a three-letter country code and a full name of the country), several columns for population counts at different years, and several other columns for information like population density, land area, etc. Column names are largely self-explanatory.

The second file, corruption_data.csv comes from another Kaggle dataset. This file documents a metric called "corruption index" for about 170 different countries. Corrution index is a measure of perceived corruption in a country. It is a value on a scale from 0 (*total corruption*) to 100 (*no corruption whatsover*), and it is assigned on an annual basis to each country on the list. The file contains a column with country names (most if not all should match the names of the countries in the `world_population.csv` file), and one column per year for corruption indexes for the years 2012 – 2021.

Assignment. The Jupyter notebook shared with you loads both CSV files, renders them in three different ways (as a pandas data frame, as a numPy array, and as a Python list (of lists)). You can choose to work with any of the three data representations, on create your own (there are other ways to store 2-dimensional data in Python). The notebook contains some instructions that you need to read before working on the assignment. The assignment itself is 10 questions about the data contained in the either one of the CSV files, or - for some questions - in both of them. Each question needs to be answered by performing some computations over the contents of the CSV files - searching for information, and possibly using some of the found information to perform additional computations.

For each question you are given a Jupyter code cell to put your code in. The result of running the cell should be clear and consise output containing the information requested (either display the variable used to store the results, or pretty print the results). All questions shall be answered in isolation: that is, no variables used for answering one question shall be used to answer another. You can, however (and are encouraged to where appropriate), copy-paste the code.

Submission. When you are done, make sure that your notebook is properly saved. Put your name and email address at the top of the notebook. Submit using the `handin` command:

```
$ handin dekhtyar 365-lab01-2 <FILE>
```

The notebook contains submission instructions that allow you to submit directly from your Jupyter notebooks environment, without having to ssh into unix1.

Good Luck!