Mid-course review

In today's lesson we are going to stop for a moment and review the content in an interactive way. We are going to use the Etherpad at:

- https://yopad.eu/p/ch701-midcourse-review

And then fill the form at:

- https://forms.office.com/e/DGRE6FmWeP **DONE**

Exercises

Today we're reviewing — or *refreshing* if you will — what we've seeing so far in the course. And an exercise combining different elements of the course so far.

Pandas

Solve the Pandas exercises marked as `Medium` difficulty from page:

- https://github.com/ajcr/100-pandas-puzzles

Data Pipeline

Using "this year's top-1000 Github Python-based repositories", we're going to transform, structure, and visualise summary data. In other words, a data pipeline — from data collection up to exploration and visualisation — must be the result of this exercise.

The Github data we're collecting here is ****similar**** to the previous exercise: we're still collecting the repositories metadata but NOT the ***readme*** (file) content NOR are we writing the JSON files in disk. Instead, we will scrape the repositories top-level page for the following information (make them attributes/columns of a table named `context` :

- Number of `Releases` ***and* latest release date**
- Number of `Contributors`
- Number of `Used by` repositories
- List of `Languages` ***and*** their **percentages** in the code base

💡 Regarding the “URL” attributes in the repositories metadata, discard all ***but*** the `html_url` and `url`.

Questions & Tasks

1. Create a summary of each of the tables in the database (`repos` , `owner` , `topics` , `license` , `context`) with (at least) the following quantities:

    1. number of records;
    2. number of unique values per column;
    3. If numeric: maximum, minimum, and average values.

2. Which *owner* have more than *one* repository, and which are the repositories?
3. List all the licenses used, in decreasing order, and their (use) percentage overall.
4. Create a word-cloud from `topics` .
5. Make a (bar) plot of the top-5 languages used (x axis) and their average (percent) participation.
6. What are the top-ten projects with most contributors?
7. What are the top-ten projects most used by other projects?
8. Make a (scatter) plot of `contributors` **vs** `used-by` counts.
9. ***Make a questions on your own, and answer it.***

Key points and guidelines

- Structure the pipeline across **two** Jupyter Notebooks:

    - one for the ETL part,

        💡 The first — ETL — notebook ends by either setting up an SQLite database **or** a set of CSV files. It will depend how you want to retrieve the data in the second part.

    - another one for summarising and visualising the data.

        💡 The second — summary/visual — notebook starts from the SQLite or CSV files database to answer the questions.

- The very first cell of each notebook is an explanation about the content in it, with a *table of contents* (aka, *index*) for the different sections/questions on the notebook, and a paragraph stating “where we are” and “where we are going”.