7PAM2000 Applied Data Science 1

Assignment 2: Statistics and trends.

This second assignment will focus on exploring statistics and trends in more detail. You are expected to produce a two page report conforming to the guidelines set out below. Think of this as a report for your team leader.

This time you will be exploring public data from the World Bank, and specifically country-by-country indicators related to climate change: `https://data.worldbank.org/topic/climate-change`. There are a range of indicators relevant to climate change, for example access to electricity, agricultural activity, urban population, etc.

Your goal is to:

- Ingest and manipulate the data using pandas dataframes. Your program should include a function which takes a filename as argument, reads a dataframe in Worldbank format and returns two dataframes: one with years as columns and one with countries as columns. Do not forget to clean the transposed dataframe.
- Explore the statistical properties of a few indicators, that are of interest to you, and cross-compare between individual countries (you do not have to do all the countries, just a select few will do) and produce appropriate summary statistics. You can also use aggregated data for regions and other categories. You are expected to use the `.describe()` method to explore your data and two other statistical methods.
- Explore and understand any correlations (or lack of) between indicators (e.g. population growth and energy consumption). Does this vary between country, have any correlations or trends changed with time?
- You are expected to use your initiative and "tell a story" with the data. You should use appropriate visualisation (hint: time series could be useful) and provide a text narrative to communicate and explain your findings. Details of the implementation and the coding do not belong in such a report. Your boss wants to see results and interpretation. What are the key findings?
- You will be assessed on the overall quality of the report, good use of visualisation tools and good use of the methods and tools available for dataframes. See mark scheme for details. Good reports often combine information from graphs to draw conclusions or follow up on insights/questions from one graph with another graph.
- You do not have to document all your data exploration. Usually one tries different plots and graphs and select the most meaningful ones for the report.

Coding quality marks are given for

- Adherence to the PEP-8 guidelines.
- Well structured and commented program following the good programming style guide. Good use of functions. No spaghetti code please.
- Good use of your repository with repeat commits.

This assignment does intentionally not specify which data sets to choose. Some ideas, definitely not exhaustive. You may find more interesting combinations.

- CO_2 production vs. GDP (energy efficiency)
- Arable land vs. land covered by forests (deforestation)
- Electric power consumption, access to electricity, overall energy use and CO_2 emission.
- Agricultural and non-agricultural methane production. How does it link to other parameters like poverty, headcount, energy consumption, access to electricity?
- How does this look for countries in different phases of development? Countries in different parts of the world?
- Numbers per capita (e.g., GDP/head) are often useful.

Format guidelines for the report.

- It should be no more than two A4 pages plus a title page (PDF format please) with 1.5 cm margins all around.
- The minimum font size should be 11 pt, with the exception of figure labels, footnotes and references. Text in graphs should still be large enough to be readable.
- The report should have a clear title and short abstract and an introductory paragraph explaining the topic.
- The title page should contain:
  - The title.
  - Your name.
  - A short abstract.
  - Link to your github repository. (Please insert it selecting `Link` from the `Insert` tab. That makes it clickable.)
  - Links to non-Worldbank data you were using, if applicable.
  - **No** table of content. No need for a short report.

How to make the most out of two pages?

- The default margin size in Word is larger than 1.5 cm. It can be changed in the `Layout` tab, select `Margins`.
- You can divide a page into columns in Word: go to the `Layout` tab, select `Columns` and the number of columns.
- The line spacing can be reduced to 4 pt, but not to less. `Design` tab, select `Paragraph spacing` and `Compact`.
- Some fonts use less space than others – but do not overdo it.
- You can use portrait or landscape format `Layout` ⟶ `Orientation`. Whatever works best.
- One can make text flow around figures.

What data can I use?

Your report needs to use Worldbank data. Additional files can be used (e.g. sales of electric vehicles not included in the Worldbank data). You can make use of Worldbank data not included in the collection, of course.

What modules can I use?

- Use of a minimum number of functions from pyplot and statistical tools from the lecture (numpy, scipy) is expected (see mark sheet).
- Functions from other modules can be used in addition. The lecture material provides sufficient tools for the assignments, but sometimes you will find functions in other modules doing something special.

What to submit?

- Word or PDF file of your report. PDFs are preferred because they avoid potential format problems (different defaults on different systems). The report should be uploaded as is and not be wrapped into a zip file or similar.
- Your program as python file. Notebooks are depreciated. We expect a link to your repo, but this way we can do most of the marking if there is a problem with the link.
- Do not upload data files. Worldbank files do not need to be referenced. Links to other data suffice. You can include links in the report or include them in a comment.
- *Repositories:* Repositories are a tool to different store versions of programming code. We expect repeat commits for full marks. Other material can also be committed, but that is optional.

What to submit?

- PDF file of report. PDFs are preferred because they avoid potential format problems (different defauls on different systems).
- Your program as python file or notebook file.
- No need to upload files. Worldbank files do not need to be referenced. Links to other data suffice.
- PDF files should be uploaded as is and not be wrapped into a zip file or similar other files can be uploaded in a zip file, but that will usually not be necessary.
- *Repo:* Repos are mainly to store version of programming code. We expect repeat commits for full marks. Other material can also be comitted, but that is optional.

What data can I use?

Your report needs to use Worldbank data. Additional files can be used (e.g. sales of electric vehicles not included in the Worldbank data).

What modules can I use?

- Use of a minimum number of functions from pyplot and statistical tools from the lecture (numpy, scipy) is expected (see rubrics).
- Functions from other modules can be used in addition. The lecture material provides sufficient tools for the assignments, but sometimes you will find functions in other modules doing something special. Use of additional modules is ok.