Project 1

Information

Instructions

- To answer the following question you need to combine your knowledge from different topics of the course. But also it requires aspects not explicitly covered so far. Thus, you may need to do your own research and try out new things.
- There may be multiple ways to approach the questions, possibly even resulting in different sets of results. If you are in such a situation, then explain your thought process: What is your goal? What are your steps to reach it? Why do you choose these steps? If your choices are reasonable, then your answer will be counted as correct.
- Submit a single, well structured and cleaned up Jupyter notebook that contains only (!) the code, output and explanations that are relevant for the questions asked. Your results must be fully reproducible. I recommend that you proceed in two steps: (1) Create a notebook for experimentation, in which you try out a lot of different things. (2) Create a notebook for the submission, in which you revise your first notebook and make it clean and clear.
- For the grading, I will consider in particular the completeness, correctness and quality of solutions. However, for an excellent grade I also expect a clean and well structured notebook, pythonic code and well formulated answers.

Data

You are provided with a number of data files related to olympic games:

- `results_2006.csv`, `results_2008.csv`, etc. contains information about individual results of athletes in the olympic games of the respective year (2006, 2008, etc.): Which athlete participated in which olympic games, sports and events, achieving which position?
- The file `metadata.xlsx` contains two tabs: one tab with biographic data on athletes such as their names and date of birth, and another tab with data about the olympic games such as their location and opening dates.
- If you want to get a better idea of what these information mean and where they come from, see: athletes, games, results. Note that some minor processing steps have been carried out. So don't expect your data to be 100% identical to the data on the website.

Excercises

Exercise 1

- Combine all provided data adequately into a single, tidy data set (e.g. by merging, concatenating or reshaping the data) such that you are able to answer the subsequent questions.
- Explain your steps in a few sentences. How do you combine the data? What are the main challenges? How do you deal with them and why in this way?
- Sort the data permanently by year, sport, event and position. Then display the first 3 rows and the last 3 rows of the data set and print the number of rows and columns.

Notes:

- Further data processing steps will be required to answer the subsequent questions: e.g. cleaning data, creating new columns, or others. You can carry out these further processing steps either also as part of exercise 1, or as part of the later exercises.
- Good code avoids redundancy. If you find yourself writing the same code multiple times, then there is probably a better way to do it.

Exercise 2

- Which woman won the Gold medal in the 100 meters race of the 2012 olympic games?
- Which athlete has won the most gold medals, considering only athletes from the following countries: "Jamaica", "Trinidad and Tobago", "Barbados", "Grenada", "Saint Kitts and Nevis"?
- What was the best position, worst position, and average position achieved by athletes from "Nepal"? (Hint: you can ignore missing values in the position column)

Exercise 3

Visualize how the number of participants has changed over the years. Create a single figure with one subplot for the summer games and one subplot for the winter games.

Note: If a single athlete participates at multiple events at a given olympic game, then count this athlete only once. For instance, Usain Bolt partcipated at three events (100 meters, 200 meters, and 4x100 meters relay) in the 2016 olympic games. He should be counted only as one participant, not as three participants.

Exercise 4

Which lastname is the most common among the athletes of the 2016 olympic games, and how often does this name occur?

Exercise 5

Which are the 5 sports with the highest average age of gold medal winners? Provide the sport and the average age.

Note: We define age as the number of years between the athlete's birth and the opening date of the olympic games.

Exercise 6

Calculate the medal table for the olympic games 2016. See the official medal table as a reference. Your medal table should have the same structure (same column names and same sorting) as the official medal table. Display the top 10 countries.

Hint: In team events such as Basketball, all team members of the winning team receive a gold medal, but for the medal table it should only count as one gold medal. The same logic also applies to silver and bronze medals. You can recognize team events either via the column `team` (which is then non-missing) or by the fact that multiple athletes of a country have won the same medal in a single competition.