Project: Part 1

Information

Data

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

- The file `athletes.csv` contains biographic data on athletes such as their names and date of birth
- The excel file `results.xlsx` contains information about individual results of athletes in the olympic games of different years. Each tab holds information for one year. For instance, the data shows at which olympic games, sports and events the athletes participated and which position or medal they achieved.
- The file `games.csv` contains information about the olympic games, e.g. their locations 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

1. Read in all data sets and inspect their basic properties.
2. Combine all provided data suitably 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.
3. Describe and justify your data processing steps in a few sentences: How do you combine the data? What are the main challenges? How do you deal with these challenges and why?
4. Sort the data permanently by year, sport, event and position. Then display the first 3 rows of the data set, and print the number of rows and columns.

Notes:

- If you find yourself writing the same or similar code multiple times, then there is probably a better way to do it. Write your code to avoid redundancies as much as possible.
- 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.

Exercise 2

1. Which lastname is the most common among the athletes of the 2012 olympic games, and how often does this name occur?
2. Which woman won the Gold medal in the 100 meters race of the 2012 olympic games?
3. Which athlete has won the largest number of gold medals (in all years covered by the dataset), considering only athletes from the following countries: "Jamaica", "Trinidad and Tobago", "Barbados", "Grenada", "Saint Kitts and Nevis"?
4. What was the best position, worst position, median position achieved by athletes from "Mongolia"?

Exercise 3

1. 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 completed years between the athlete's birth and the opening date of the olympic games.)
2. The Body-Mass-Index (BMI) is a rule-of-thumb measure to classify people into groups such as "underweight", "normal", "overweight" etc. The BMI is defined as the body mass (in kilograms) divided by the square of the body height (in metres), and is expressed in units of kg/m2. Derive a new column `bmi` from the provided data. Display the 5 sports with the lowest average BMI of male athletes.

Exercise 4

Calculate the medal table for the olympic games 2012. 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 by the fact that multiple athletes of a country have won the same medal in a single competition.

Exercise 5

1. Identify an interesting pattern, trend, or relationship in the data. This could relate to countries, sports, gender, time periods, medal counts, or any other aspect. Create a single visualization that clearly and effectively communicates the insight you found. You are free to choose the type of chart, but make sure it suits the message you want to convey.
2. Briefly describe what the visualization shows and explain why the pattern is interesting/surprising.