Project 2

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 zip folder that contains both the data file and a Jupyter Notebook. Your results must be fully reproducible. Therefore make sure that you are using relative paths that will also work if the Notebook is executed on some other computer**.
- The Jupyter Notebook that you submit should be well structured and contain only (!) the code, output and explanations that are relevant for the questions asked. Therefore 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.

Overview

This project consists of three parts:

- In **exercise 1** you retrieve demographic and economic data for countries from the Worldbank API.
- In **exercise 2** you combine the data from the previous exercise with data on countries performances at Olympic Games and prepare the combined data for exercise 3.
- In **exercise 3** you train a linear regression model to predict the performance of countries at Olympic Games based on demographic and economic features.
- **Note**: Since machine learning is not a focus topic of this course, you do not need to optimize the model. Just demonstrate that you are able to apply the steps we discussed in the course and correctly interpret the results.

Exercise 1

Write a Python function that can be used to query data from the Worldbank Indicator API. Your function should:

- take the following input parameters: `indicators`, `countries`, and `years`.
- return a Pandas DataFrame of the queried data
- have a docstring that explains what the function does, what the input parameters are, and what the output is
- minimize the number of API calls necessary to retrieve the data

Demonstrate that your function works by querying the following data (codes are provided in parentheses):

a) The total population (SP.POP.TOTL) of Germany (DE) and France (FR) between 2015 and 2020.
b) The total population (SP.POP.TOTL), GDP in current US$ (NY.GDP.MKTP.CD), and life expectancy in years at birth (SP.DYN.LE00.IN) of all countries (all) in 2012. Print the shape of the resulting DataFrame and display its first 10 rows.
c) State how many API calls your function makes for a) and b) respectively.

Notes:

- To solve the exercises study the documentation of the basic call structures. Most of the information you need is provided there.
- If needed, additional information about the API is available here. For instance, you will find links to the list of available indicators and countries, and explanations on error codes.
- Note that by default the API returns only the first 50 items. Check the documentation on how to retrieve more items.

Exercise 2

The file `medal_table.csv` contains information about the number of medals won by each country at the Olympic Games 2012.

a) Preprocess both the medal table data and the Worldbank data retrieved in exercise 1 b) and combine the two datasets suitably into one tidy dataset. The final dataset should be such that it allows you to answer the following exercises (2b and 3). Explain your actions and decisions in a few sentences. **Notes**: 1. If there are missing values in the Worldbank data set (e.g. if no population data is available for Germany), then you do NOT need to impute these values. 2. Exercises 2b and 3 may require different handling of missing values. Therefore, it is fine if you create slightly different versions of the combined dataset for these exercises.

b) Create an alternative medal table for the 2012 Olympic Games by calculating the number of Gold, Silver, and Bronze medals won per 10 million inhabitants. Display the 10 most successful countries according to this alternative medal table.

Exercise 3

Carry out a simple supervised machine learning experiment, in which you train a model to predict the number of medals a country wins at the Olympic Games based on demographic and economic features. **Note**: Since machine learning is not a focus topic of this course, you do not need to optimize the model. Just demonstrate that you are able to apply the steps we discussed in the course and correctly interpret the results.

a) Train and evaluate a linear regression model: 1. Split your data into a training and a test set. 2. Train a linear regression model using population, life expectancy and the GDP per capita of a country as features. 3. Evaluate the model using the root mean squared error as the performance metric.
b) Briefly discuss the results: How do you judge the performance? What are possible reasons for this performance? How could the model be improved?
c) Predict the number of medals a hypothetical country with a population of 10 million, life expectancy of 70 years, and a GDP per capita of 20.000 US$ would win.