Data Science II - Data Visualization

Lab 1 - Summer Semester 2024
https://moodle.haw-landshut.de/course/view.php?id=10969

1. Datasources

Try to find high quality data sources for the following types of data:

- Sports
- Geography
- World – Health and Development Indicators
- Medical Image Data
- Traffic Data
- Government and Politics
- COVID-19 (Germany, EU, World)
- Weather
- Financial and Economic Data
- Real Estate
- any other categories you are eager to learn something about and present later on in this course

As a starting point you can use general data search engines like

- `https://datasetsearch.research.google.com/`,
- `https://ckan.org/`
- `https://opendatamonitor.eu`

and the data sources in data_sources.html (provided by the previous students of this course).

2. HTML & CSS

Create a small web page with HTML and CSS showcasing the datasources you have found in the previous exercise. Include images, tables and links (if it makes sense for the content and makes it easier to understand what is behind the linked data source). Upload the HTML+CSS to the Moodle course as a zip-file.

3. Converting HTML tables into pandas dataframes

(a) Implement a function `get_extreme_min_temperature(html_table)` that extracts from the table at

`https://en.wikipedia.org/wiki/List_of_extreme_temperatures_in_Germany`

the columns `State` and, from the `Extreme Minimum` part of this table, the columns `Temperature`, `Location` and `Date`. The function must return the extracted part of the table as a pandas dataframe.

(b) Generate two different bar plots from the dataframe returned by `get_extreme_min_temperature(html_table)`. The first bar plot should use the column `State` for the x-axis and the column `Temperature` for the y-axis. The second bar plot should use the column `Date` for the x-axis, the column `Temperature` for the y-axis and should be ordered by date on the x-axis.

4. Web scraping box office revenue in Germany from boxofficemojo.com

Write a Python script that scrapes the information from the table at

`https://www.boxofficemojo.com/weekend/by-year/2023/?area=DE`,

converts it into a pandas dataframe and stores it locally as a CSV file. What do you need to change to get the same information for year 2022 or 2021?