Python Científico

Andre Nepomuceno

January 10, 2022

Exercícios - Lista 02 - Pandas

Os arquivos de dados estão disponíveis no github (pasta pcientifico_dados)

2.1 A tabela *Sun and Solar System*, disponível no link
`https://en.wikipedia.org/wiki/Abundances_of_the_elements_(data_page)`,
fornece uma lista da abundância de elementos no Sistema Solar (coluna **Y1**). Use o método `read_html` para ler os dados da tabela e plotar um gráfico de barras que demostre a regra de Oddo–Harkins: os elementos com número atômico par são mais abundantes que os elementos vizinhos com número atômico ímpar. Para melhor visualização, utilize escala logarítmica no eixo y.

2.2 O diagrama de Hertzsprung–Russell é um gráfico de dispersão onde cada estrela é representada como um ponto sendo a abcissa x a temperatura efetiva e o eixo y a luminosidade. Utilize o Pandas para ler o arquivo `hygdata_v3-file.csv` (dados de 119614 estrelas) e plot o diagrama de Hertzsprung–Russell. No arquivo em questão, a luminosidade está na coluna **lum**, e a temperatura (em Kelvin) pode ser calculada usando a fórmula de Ballesteros:

$$T = 4600\left(\frac{1}{0.92(B-V)+1.7}+\frac{1}{0.92(B-V)+0.62}\right),$$

onde o fator $(B-V)$ é chamado índice de cor (disponível no arquivo na coluna **ci**). Utilize escala logarítmica para a luminosidade e plot o eixo da temperatura no sentido reverso.

2.3 O grupo de tecnologia digital da Universidade de Cambridge coleta dados meteorológicos locais desde 1995. Os dados estão disponíveis em `https://www.cl.cam.ac.uk/research/dtg/weather/`. Leia o arquivo referente a todo o dataset (*single CSV file*) e determine: a) a direção do vento mais frequente; b) a maior velocidade de vento medida; c) o ano que teve o mês de Junho mais ensolarado; d) o dia mais chuvoso ; e) a menor temperatura medida.

2.4 Os dados de covid-19 de todo o mundo são reunidos na plataforma da universidade John Hopkins `https://coronavirus.jhu.edu/map.html`. No link do github deste minicurso, encontramos o arquivo `time_series_covid19_confirmed_global.csv` que contém o **acumulado** diário de casos desde o começo da pandemia, em todos os países. Utilize o Pandas para ler esse arquivo e plotar o número **diário** de casos confirmados no Brasil em 2021, bem como a média de casos a cada sete dias. Sugestão: Utilize o nome dos países como índices e as datas como colunas. Como cada coluna terá o acumulado de casos até aquela data, use o método diff() para obter o número diário de casos.

2.5 Use Pandas para ler os quatro arquivos do exercício 1.2 (lista 01), e junte-os num mesmo DataFrame. Use o método `pd.concat()`, com as opções `axis=1` e `sort=True`, e o método `dropna()` para eliminar os valores inválidos.

2.6 No site `https://www.kaggle.com/` encontramos mais de cinquenta mil datasets de domínio público, dos mais diversos interesses. Como um exercício adicional, procure o dataset 'Daily Sun Spot Data', que contém o número diário de manchas solares observados desde 1818, e faça um gráfico do número mensal de manchas solares observadas entre 1985 e 2019. Você deve observar os três últimos ciclos solares (um ciclo solar tem aproximadamente 11 anos).