

Activity A20

1. The **air_quality_no2.csv** data set contains data on NO2 values, in parts per billion (ppb), measured in Antwerp, Paris and London.

The 1st field reports the date in which the air quality value was recorded.

The 2nd 3rd and 4th fields, the recorded values in Antwerp, Paris, and London.

In a file called **A20-air-quality.py**, do the following:

Q1. Import pandas, Matplotlib, and NumPy

Q2. Read in data by using `read_csv()` and convert the dates to the datetime object. Use the `parse_dates` parameter. Store the DataFrame in variable `df`.

Q3. Display the first 4 lines of the DataFrame.

Q4. Get info about your data.

Q5. How many rows and columns There are? Use an attribute to answer the question.

Q6. Check if there are duplicate rows, and in case delete them.

Q7. Calculate the number of missing values in each column of the dataset.

Q8. Remove columns containing missing values.

Q9. Display the 1D array of the column labels.

Q10. Rename the columns to: paris, london. Use list comprehension to remove the 'station_' from each column label.

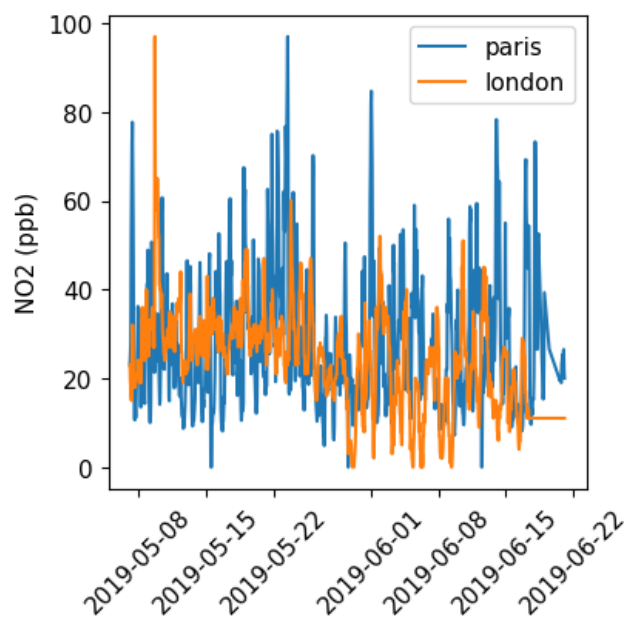
Q11. Calculate the maximum value of NO2 recorded in the Paris, and in which date occurred. Print the results to screen.

Q12. Calculate the mean and std for both the Paris and London station, and print results to screen with 2 decimal precisions:

```
London mean=24.30 and std=11.27 (ppb)
Paris mean=27.70 and std=15.11 (ppb)
```

Q13. Make this plot. Use rcParams to set the Font size to 11, and figure size 4,4
To rotate the x labels use tick_params()

https://matplotlib.org/stable/api/_as_gen/matplotlib.axes.Axes.tick_params.html



Activity: make a bar graph

2. The **plants_data.txt** dataset provides information about various plant species, including their characteristics and occurrences in a specific area. Each row in the dataset represents a different plant species, and the columns contain the following information:

species: The name of the plant species.

height: The average height of the plant species, measured in centimeters.

occurrences: The number of times the plant species has been observed or recorded within a specific surface area.

Surface area (km²): The surface area in square kilometers where the plant species was observed.

In a script **A20-plants.py** do the following:

Q1. Read the provided dataset `plant_data_updated.csv` and store it in a DataFrame using `read_csv()`. Set the first field as row labels. Display the first 3 rows of the DataFrame.

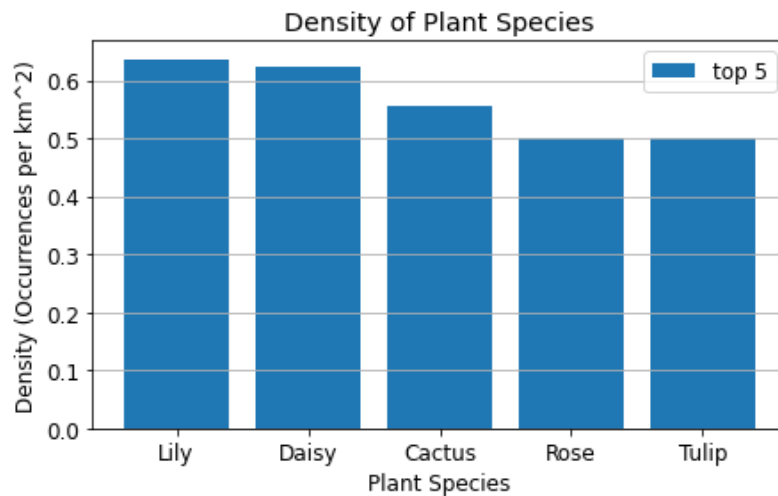
	height	occurrences	surface_area_km2
species			
Rose	30	25	50
Tulip	20	30	60
Sunflower	40	20	40

Q2. Calculate the density for each plant species by dividing the occurrence by the surface area and add a new column 'density' to the DataFrame.

	height	occurrences	surface_area_km2	density
species				
Rose	30	25	50	0.500000
Tulip	20	30	60	0.500000
Sunflower	40	20	40	0.500000
Lily	25	35	55	0.636364
Daisy	15	28	45	0.622222
Orchid	35	22	70	0.314286
Dandelion	10	40	80	0.500000
Cactus	5	50	90	0.555556
Hydrangea	18	33	75	0.440000

Q3. Sort the DataFrame values by the density column in descending order.

Q4 Select the top 5 most dense plants and make this bar graph. Set the font size 12, and figure size (7, 4).



Submit to A20:

- A20-air-quality.py
- A20-plants.py