Python Programming

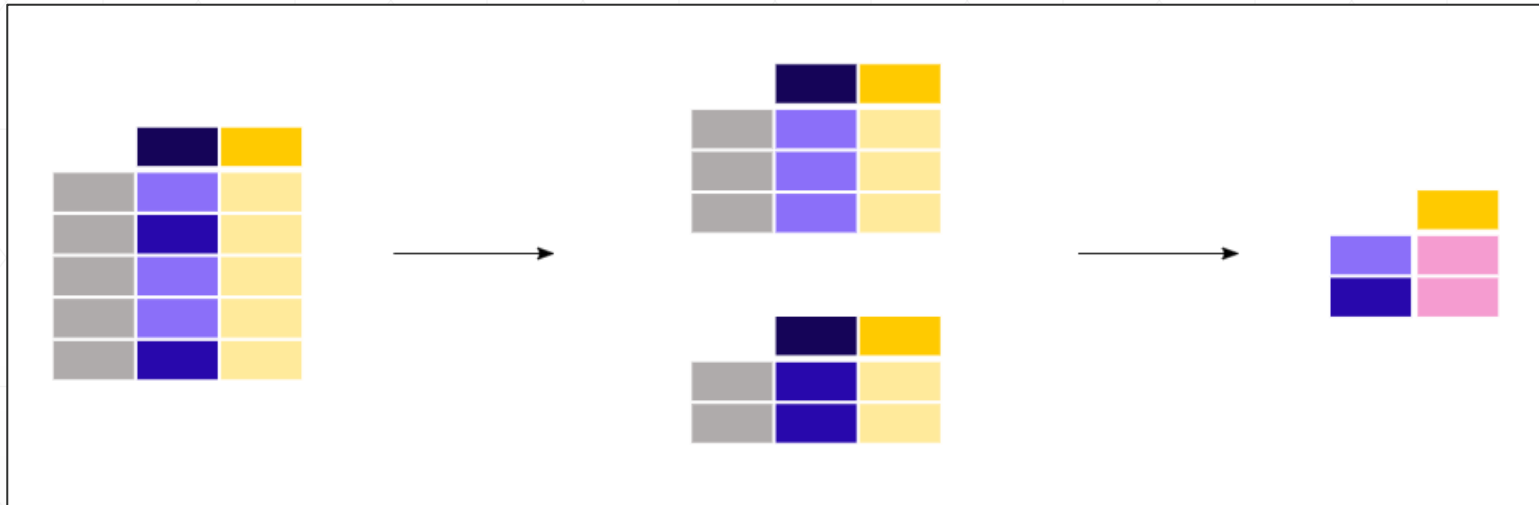Lesson 9 – Data Visualization with Graph Plotting

Lesson 9 - Outline

- More case studies with Pandas

- Data Visualization with Graph Plotting
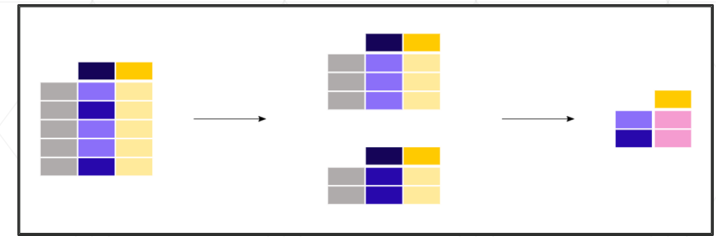
More case studies with Pandas

More about Data Analysis

- **Aggregating statistics grouped by category**

1. **Split** the data into groups
2. **Apply** the function to each group individually
3. **Combine** the results into a data structure

More about Data Analysis
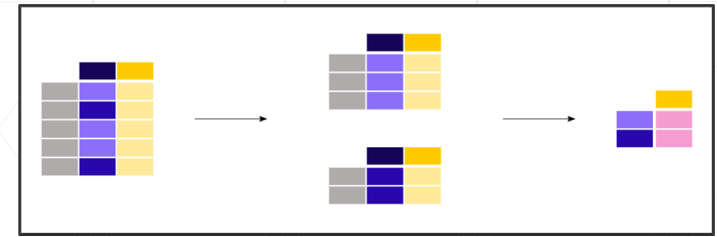
- **Aggregating statistics grouped by category**

```
df[["col1", "col2"]].groupby("col1").mean()
```

- **Example – the average age for male versus female Titanic passengers**

```
sex_age_mean =
titanic[["Sex","Age"]].groupby("Sex").mean()

print(sex_age_mean)
```

```
                Age
Sex
female   27.915709
male     30.726645
```

More about Data Analysis
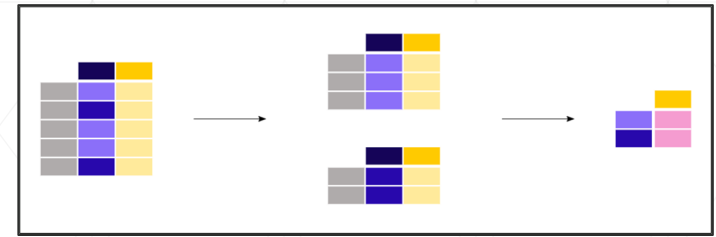
- **Aggregating statistics grouped by category**

```
df[["col1", "col2"]].groupby("col1").mean()
```

- **Example – the average of all columns for male versus female Titanic passengers**

```
sex_mean = titanic.groupby("Sex").mean()

print(sex_mean)
```

|	PassengerId	Survived	Pclass	...	SibSp	Parch	Fare
Sex				...			
female	431.028662	0.742038	2.159236	...	0.694268	0.649682	44.479818
male	454.147314	0.188908	2.389948	...	0.429809	0.235702	25.523893

More about Data Analysis

- **Aggregating statistics grouped by category**
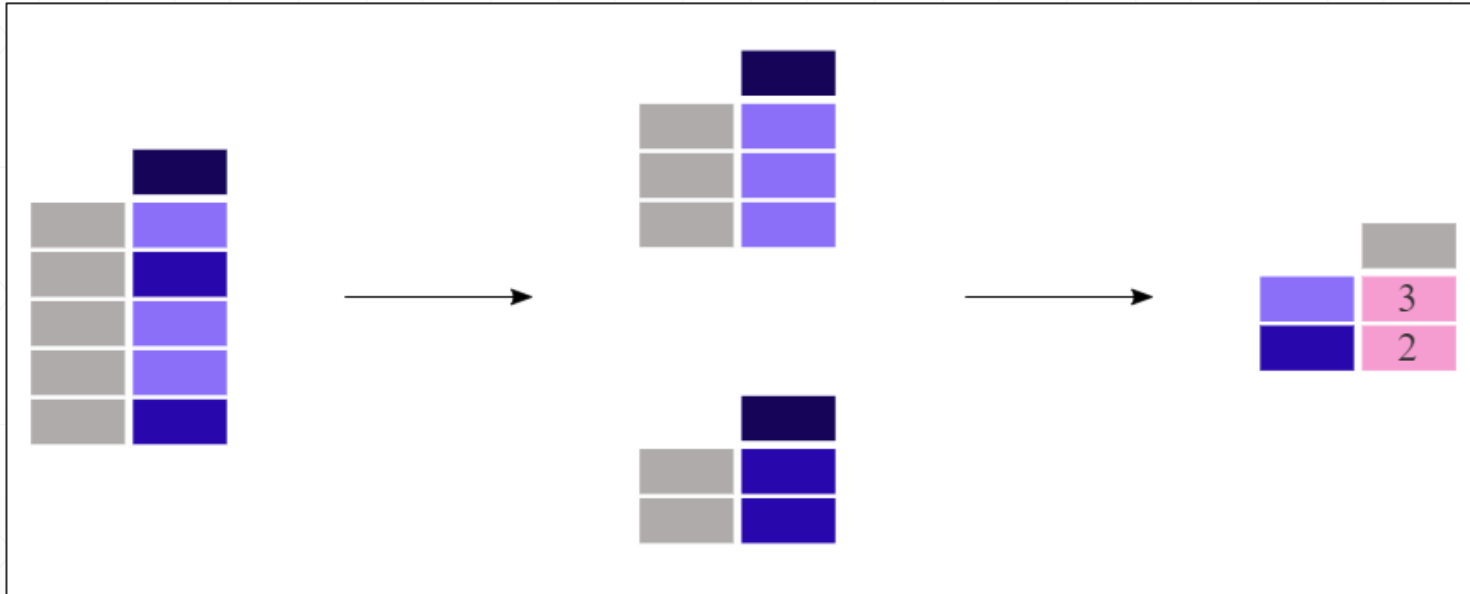
```
df[["col1", "col2"]].groupby("col1").mean()
```

- **Example – the average of all columns for male versus female Titanic passengers**

```
sex_mean = titanic.groupby("Sex").mean()

print(sex_mean)
```
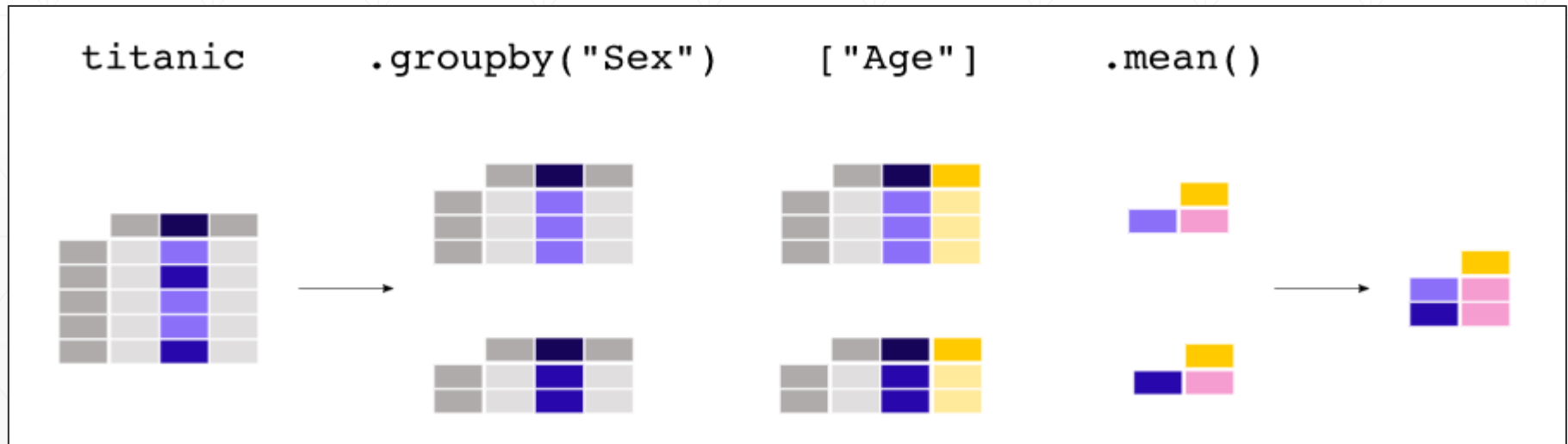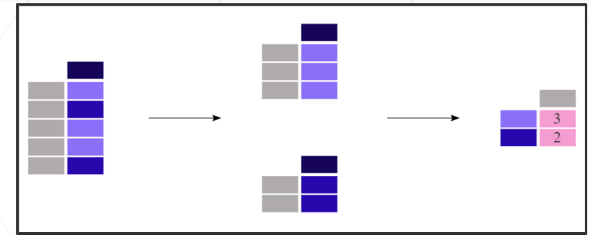
More about Data Analysis

▪ <u>Count number of records by category</u>

1. **Split** the data into groups
2. **Apply** the function to each group individually (i.e. Count)
3. **Combine** the results into a data structure

More about Data Analysis

- **<u>Aggregating statistics grouped by category</u>**
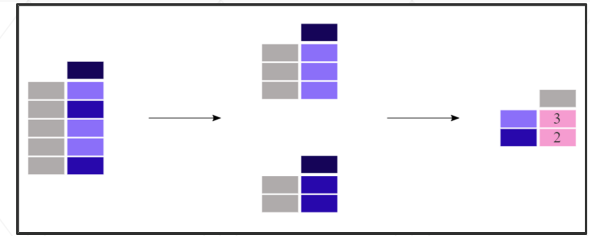
```
df["col1"].value_counts()
```

- **<u>Example – the number of passengers in each of the cabin classes</u>**

```
class_count = titanic["Pclass"].value_counts()

print(class_count)
```

```
3     491
1     216
2     184
Name: Pclass, dtype: int64
```

How to sort the result in order?

More about Data Analysis

- **<u>Aggregating statistics grouped by category (in order)</u>**

```
df["col1"].value_counts().sort_index()
```

- **<u>Example – the number of passengers in each of the cabin classes (in order)</u>**

```
class_count =
titanic["Pclass"].value_counts().sort_index()

print(class_count)
```

```
1     216
2     184
3     491
Name: Pclass, dtype: int64
```
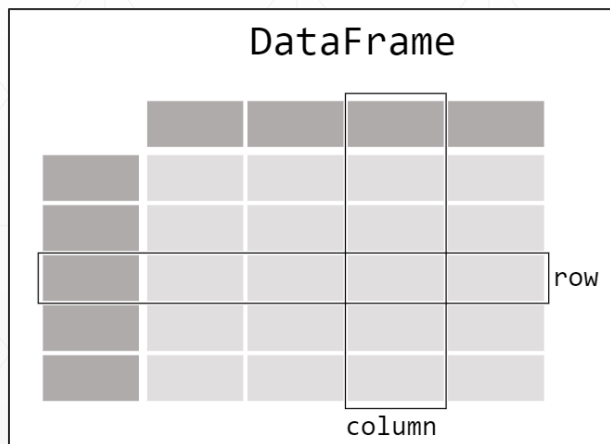
Recap the basic of Pandas

- **<u>Install of Pandas module</u>**
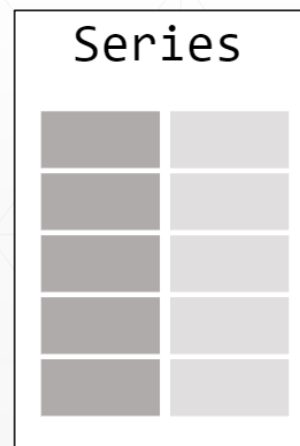
  ```
  pip install pandas
  ```

- **<u>Import Pandas module</u>**

  ```
  import pandas as pd
  ```

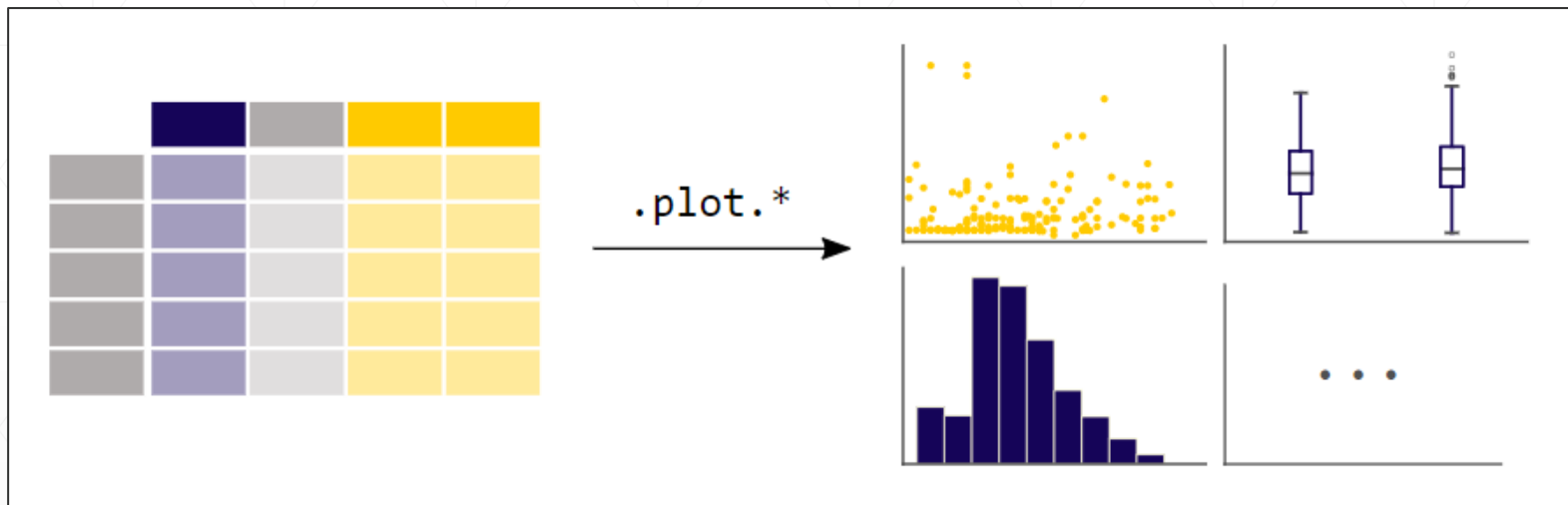| Pandas data table representation	Each column in a DataFrame is a Series

Data Visualization with Graph Plotting

Data Visualization with Graph Plotting

▪ **How to create plots in pandas?**

Data Visualization with Graph Plotting

- **<u>Install of Pandas module</u>**

  ```
  pip install pandas
  ```

- **<u>Import Pandas module</u>**

  ```
  import pandas as pd
  ```

- **<u>Install of matplotlib module</u>**

  ```
  pip install matplotlib
  ```

- **<u>Import matplotlib module</u>**

  ```
  import matplotlib.pyplot as plt
  ```
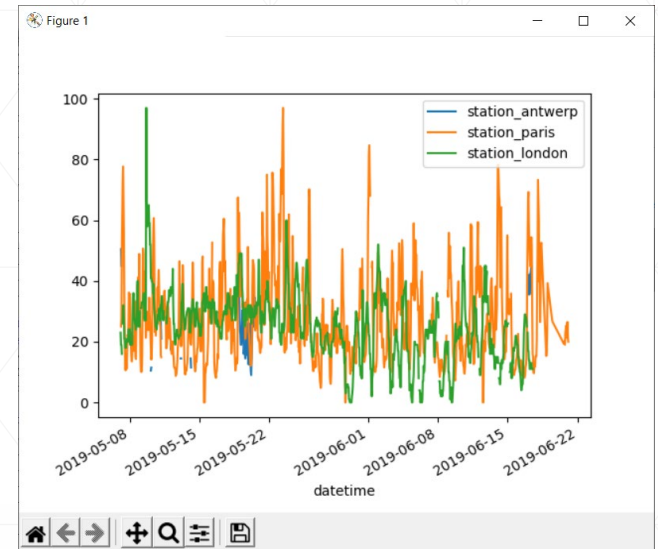
Data Visualization with Graph Plotting

- **Quick visual check of the data**

```
df.plot()

plt.show()
```

- **Example of plotting all data**

```
import pandas as pd

import matplotlib.pyplot as plt

air_quality = pd.read_csv("air_quality_no2.csv",
index_col=0, parse_dates=True)

print(air_quality)

print(air_quality.plot())

plt.show()
```
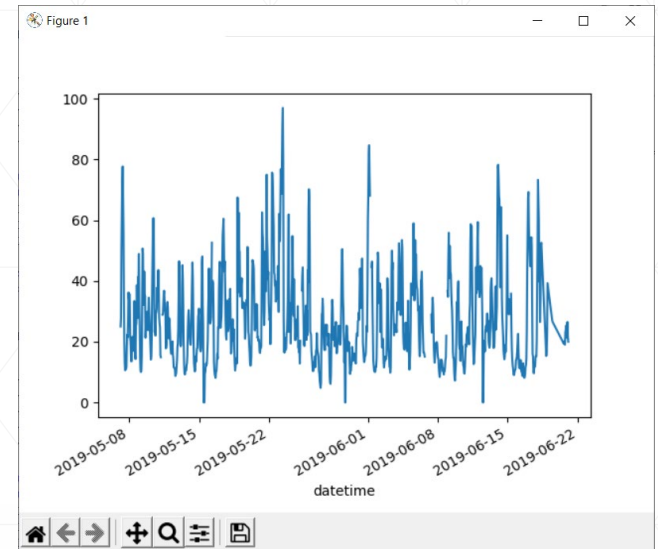
Data Visualization with Graph Plotting

- **Plot of the data of one column**

```
df["col1"].plot()

plt.show()
```

- **Example of plotting one column**

```
import pandas as pd

import matplotlib.pyplot as plt

air_quality = pd.read_csv("air_quality_no2.csv",
index_col=0, parse_dates=True)

print(air_quality)

print(air_quality["station_paris"].plot())

plt.show()
```
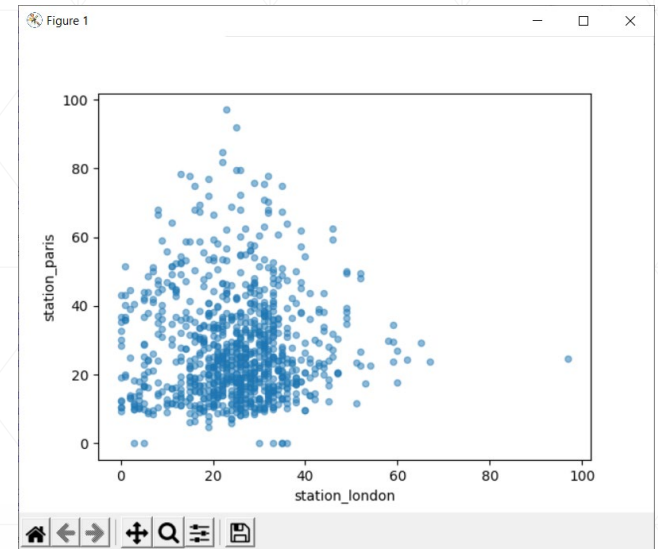
Data Visualization with Graph Plotting

- **Plot scatter chart**

```
df.plot.scatter()

plt.show()
```

- **Example**

```
import pandas as pd

import matplotlib.pyplot as plt

air_quality = pd.read_csv("air_quality_no2.csv",
index_col=0, parse_dates=True)

air_quality.plot.scatter(x="station_london",
y="station_paris", alpha=0.5)

plt.show()
```
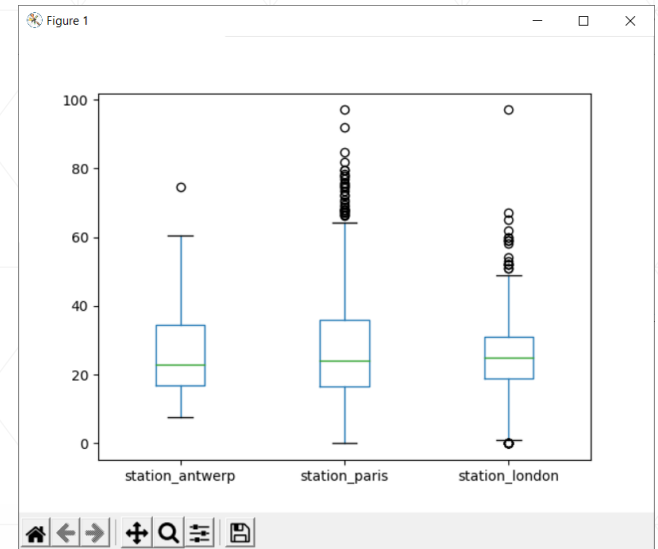
Data Visualization with Graph Plotting

- **Plot box chart**

```
df.plot.box()

plt.show()
```

- **Example**

```
import pandas as pd

import matplotlib.pyplot as plt

air_quality = pd.read_csv("air_quality_no2.csv",
index_col=0, parse_dates=True)

air_quality.plot.box()

plt.show()
```

Data Visualization with Graph Plotting

- **<u>Each column in a separate subplot</u>**

```
df.plot.area(figsize=(x_inch, y_inch),
subplots=True)

plt.show()
```
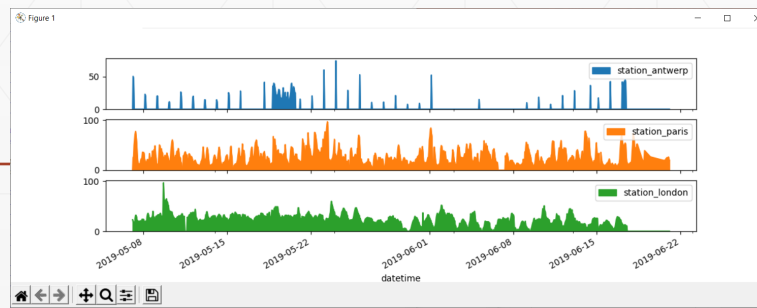
- **<u>Example</u>**

```
import pandas as pd

import matplotlib.pyplot as plt

air_quality = pd.read_csv("air_quality_no2.csv",
index_col=0, parse_dates=True)

air_quality.plot.area(figsize=(12, 4),
subplots=True)

plt.show()
```

Data Visualization with Graph Plotting

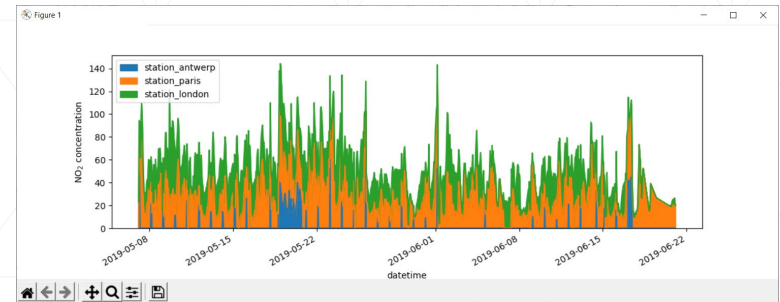- **More customization of subplot**

- **Example**

```python
import pandas as pd

import matplotlib.pyplot as plt

air_quality = pd.read_csv("air_quality_no2.csv",
index_col=0, parse_dates=True)

fig, axs = plt.subplots(figsize=(12, 4))

air_quality.plot.area(ax=axs)

axs.set_ylabel("NO$_2$ concentration")

fig.savefig("no2_concentrations.png")

plt.show()
```

Explanation on next slide

Data Visualization with Graph Plotting

- **More customization of subplot**

- **Explanation**

- **Create an empty matplotlib Figure and Axes**

- `fig, axs = plt.subplots(figsize=(12, 4))`

- **Use pandas to put the area plot on the prepared Figure/Axes**

- `air_quality.plot.area(ax=axs)`

- **Do any matplotlib customization you like**

- `axs.set_ylabel("NO$_2$ concentration")`

- **Save the Figure/Axes using the existing matplotlib method**

- `fig.savefig("no2_concentrations.png")`
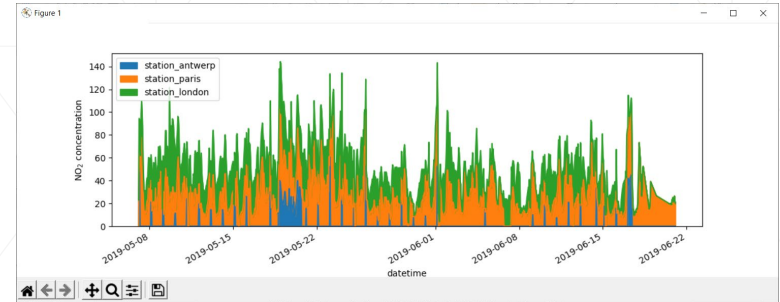
References

- References/Examples of Pandas
  - https://pandas.pydata.org/

- References/Examples of matplotlib
  - https://matplotlib.org/3.1.1/api/pyplot_summary.html

Thank you