NOVA IMS Information Management School

Programming for Data Science

Pandas

Flávio Pinheiro
fpinheiro@novaims.unl.pt
Liah Rosenfeld
lrosenfeld@novaims.unl.pt
Maria Almeida
malmeida@novaims.unl.pt
Niclas Sturm
nsturm@novaims.unl.pt

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

What's next?

Class topics

- Pandas

Python libraries: Pandas

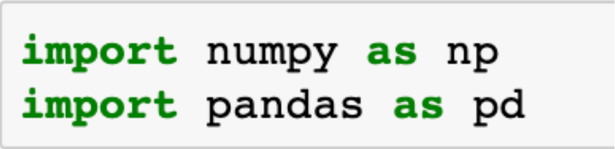
Libraries!

- In a general sense, a library in Python is a **piece of reusable code**.
- There are many of them and they help us solve problems, store data, represent data, perform statistic tests, run algorithms, etc...

Pandas

Pandas!

- **Pandas** is a widely-used Python library built on top of NumPy.



```
import numpy as np
import pandas as pd
```

- **Pandas** was designed to work with 2-dimensional data (like Excel spreadsheets/csv files).

- It is known for its very useful data structure called the **DataFrame** (and for pandas Series).

Pandas Series

Pandas Series

- **Series** are a special type of data structure available in the pandas Python library.

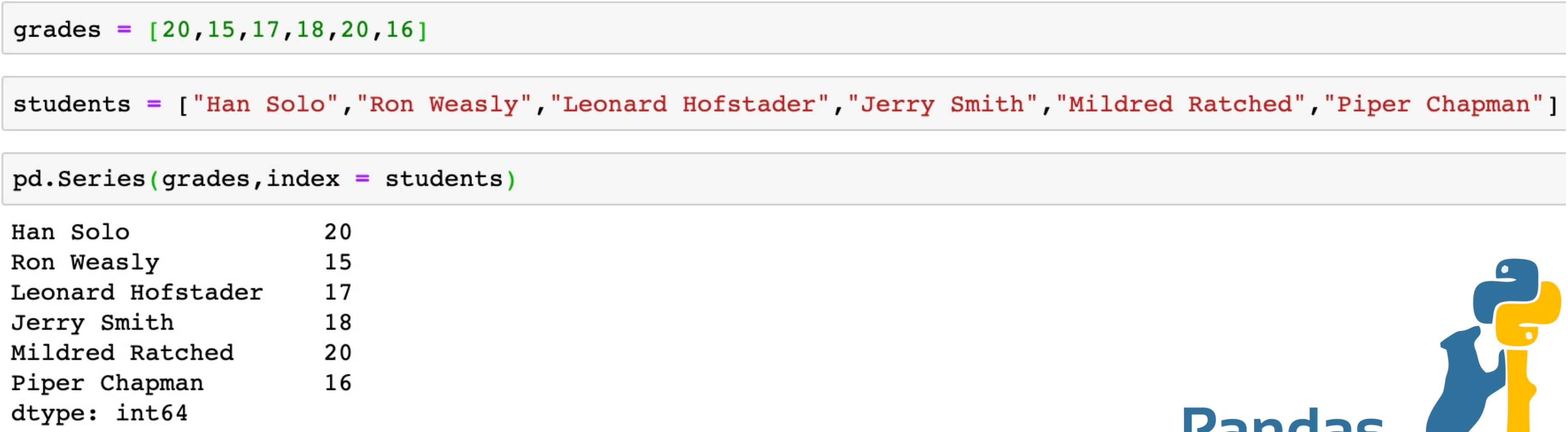
- Pandas Series are like NumPy arrays, except that we can give them a **named or datetime index** instead of just a numerical index.

- You will learn about NumPy soon, but for now we can think of NumPy arrays as being very similar to lists.

Pandas Series

- We can think of a Pandas **Series as a List** where each value can be indexed with a name and not only by a number!

- They help giving meaning to our data:



```
grades = [20,15,17,18,20,16]
```

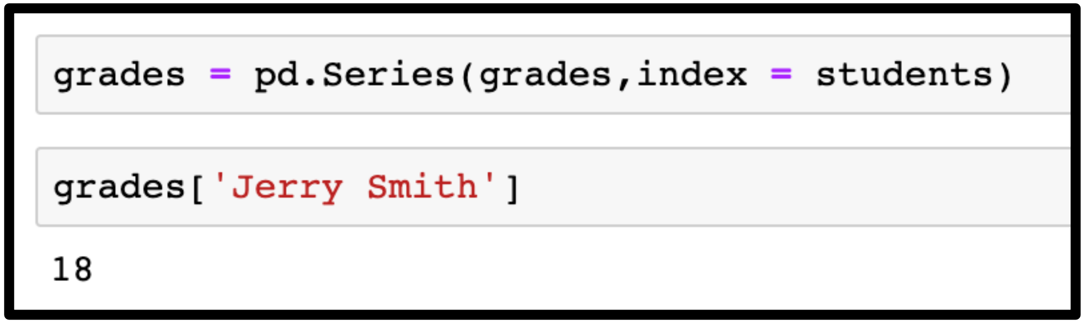
```
students = ["Han Solo","Ron Weasly","Leonard Hofstader","Jerry Smith","Mildred Ratched","Piper Chapman"]
```

```
pd.Series(grades,index = students)
```

```
Han Solo             20
Ron Weasly           15
Leonard Hofstader    17
Jerry Smith          18
Mildred Ratched      20
Piper Chapman        16
dtype: int64
```

Pandas Series

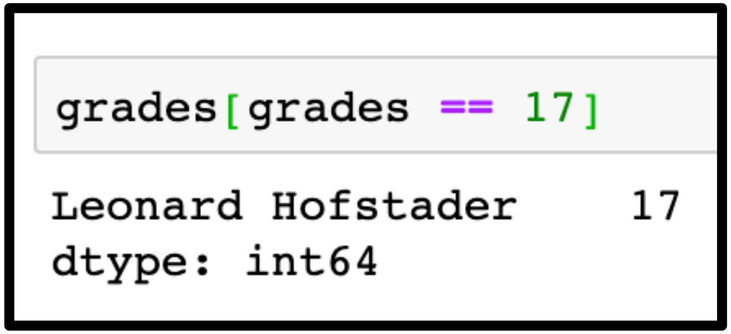
Pandas Series:



```
grades = pd.Series(grades,index = students)
```

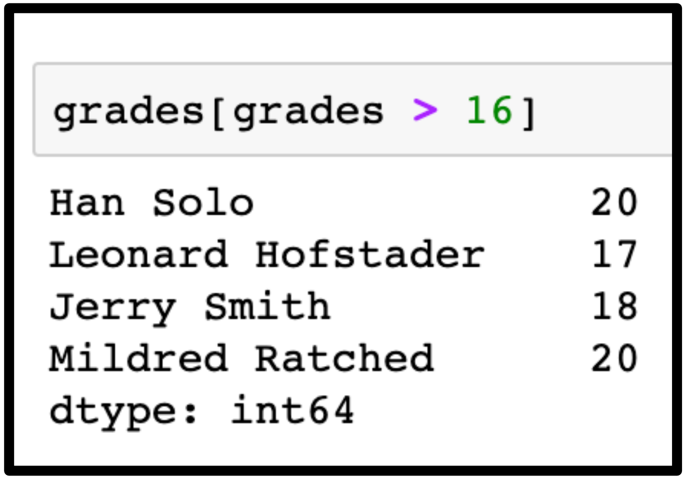
```
grades['Jerry Smith']
```

```
18
```



```
grades[grades == 17]
```

```
Leonard Hofstader     17
dtype: int64
```



```
grades[grades > 16]
```

```
Han Solo              20
Leonard Hofstader     17
Jerry Smith           18
Mildred Ratched       20
dtype: int64
```

Pandas

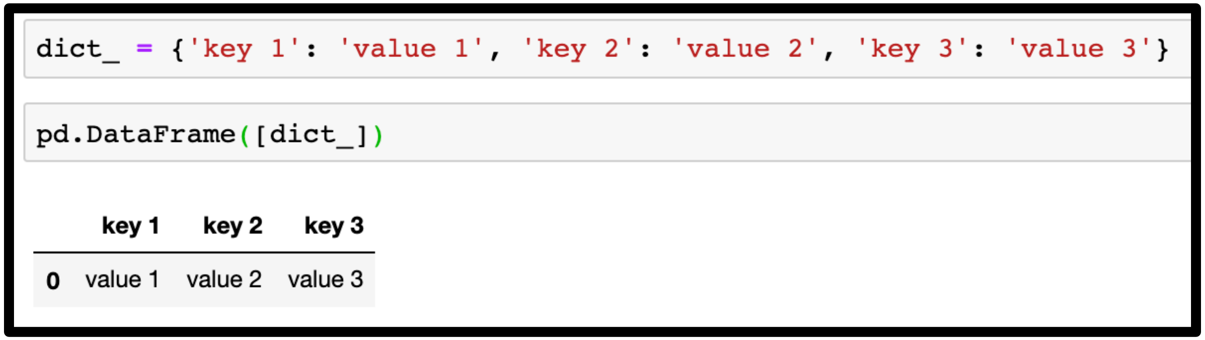
Pandas DataFrames

- **DataFrames** are the most important data structure in the Pandas library.

- A pandas DataFrame is a **2-dimensional data structure** that has labels for both its rows and columns.

- A DataFrame can be created in many ways. The most common by loading a .csv file (or an Excel sheet).

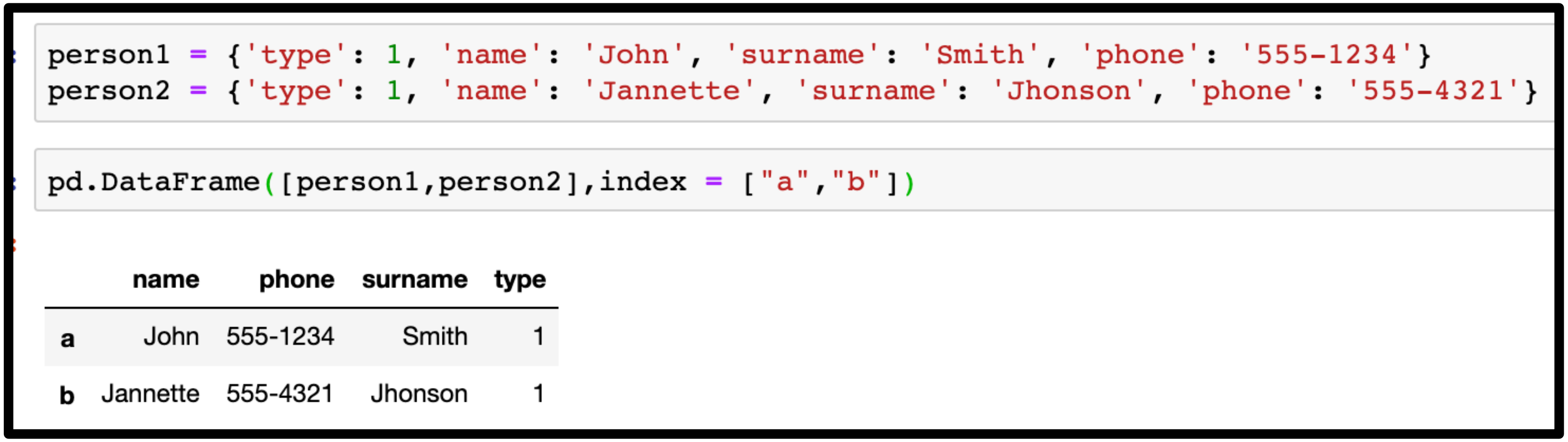
Pandas DataFrames



```
dict_ = {'key 1': 'value 1', 'key 2': 'value 2', 'key 3': 'value 3'}
```

```
pd.DataFrame([dict_])
```

	key 1	key 2	key 3
0	value 1	value 2	value 3



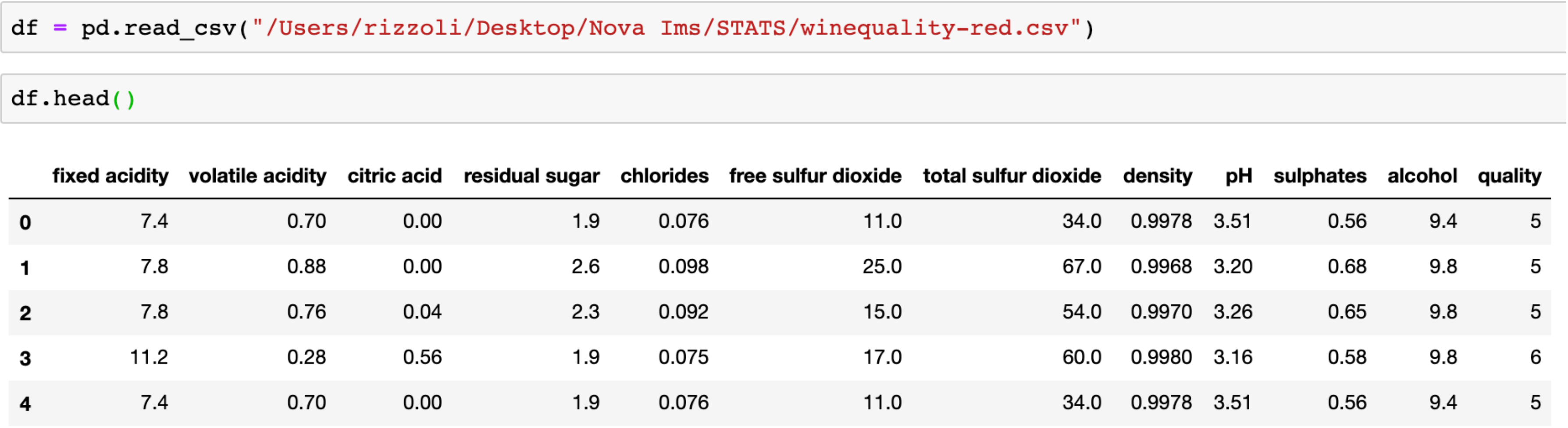
```
person1 = {'type': 1, 'name': 'John', 'surname': 'Smith', 'phone': '555-1234'}
person2 = {'type': 1, 'name': 'Jannette', 'surname': 'Jhonson', 'phone': '555-4321'}
```

```
pd.DataFrame([person1,person2],index = ["a","b"])
```

	name	phone	surname	type
a	John	555-1234	Smith	1
b	Jannette	555-4321	Jhonson	1

Pandas

Pandas DataFrames



```
df = pd.read_csv("/Users/rizzoli/Desktop/Nova Ims/STATS/winequality-red.csv")
```

```
df.head()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Pandas

End