

Python & pandas

A one day course

Uwe Ziegenhagen

`github.com/UweZiegenhagen/OneDayPythonPandasCourse`

Cologne, 24. August 2022

Introduction

Why Python/pandas?

- ▶ You *have* a CSV-file with semicolon as column separator and comma as decimal separator
- ▶ You *need* a CSV-file with comma as column separator and dot as decimal separator

```
1 import pandas as pd
2
3 df = pd.read_csv('myfile.csv', sep=';', decimal = ',')
4 df.to_csv('myfile_new.csv', sep=',', decimal = '.')
```

- ▶ Or: You need an Excel-file:

```
1 import pandas as pd
2
3 df = pd.read_csv('myfile.csv', sep=';', decimal = ',')
4 df.to_excel('myfile_new.xlsx', index=FALSE)
```

Limits of this Course

- ▶ Basis for this tutorial is a course I held at the FOM („Fachhochschule für Ökonomie und Management“) in Cologne
- ▶ It is not a *full* Python & pandas course, we would need a whole week or more. . .
- ▶ Goal: give you an overview of Python and to teach you enough Python to a) read and b) understand Python-Code and c) write smaller programs relevant for your job
- ▶ We will skip many interesting things (that you do not necessarily need for your job)

Introduction

Python

- Datatypes

- Functions

- Flow control

- File-Operations

- Ranges and Strings

- Sequential Datatypes

pandas

Links

Python

- ▶ Invented by Guido van Rossum at the „Centrum Wiskunde & Informatica“ in Amsterdam as successor for the teaching language ABC
- ▶ Published in 1991, so it is quite a bit older than Java (1995)
- ▶ Current version is 3.11
- ▶ For a long time, Python 3.x and Python 2.7 existed together
- ▶ Python 2.7 support expired in 2019:
- ▶ How to spot 2.7 code: → `print 'hello'` instead of `print('hello')`

Python versus Java & C

Python code is often much slower than C or Java but:

- ▶ the implementation time for Python is way faster
- ▶ speed only matters sometimes, not always
- ▶ many computing-intensive Python modules use C/C++ modules „under the hood“

Python

Datatypes

Functions

Flow control

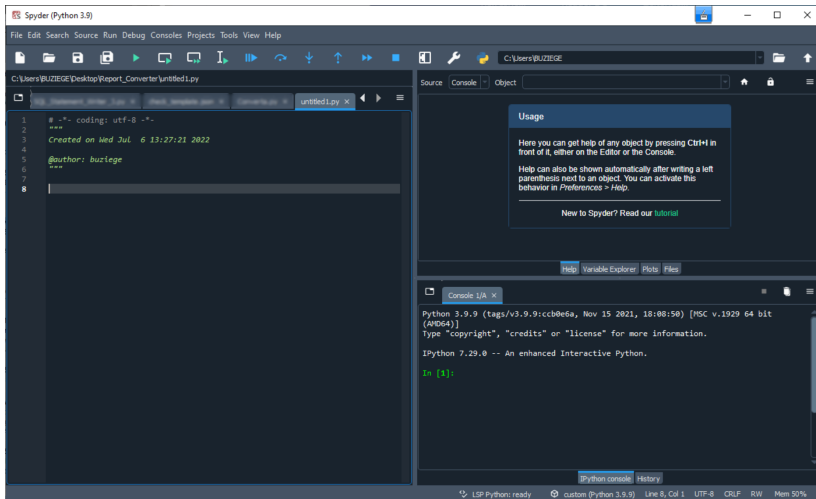
File-Operations

Ranges and Strings

Sequential Datatypes

Spyder

- ▶ We will use the Spyder5 IDE with Python 3.9.9, make sure it is installed



Python as a Calculator

- ▶ Spyder5 runs an IPython kernel, this runs our programs
- ▶ We can also use it as a calculator

```
1 In [1]: 4*5.4
2 Out[1]: 21.6
3
4 In [2]: 4/12
5 Out[2]: 0.3333333333333333
6
7 In [3]: _*3
8 Out[3]: 1.0
9
10 'hello'
11 Out[4]: 'hello'
```

Python as a Calculator

```
1
2 In [1]: 4%2
3 Out[1]: 0
4
5 In [2]: 5%2 # Modulo
6 Out[2]: 1
7
8 In [3]: 3**3 # Power
9 Out[3]: 27
10
11 In [4]: 5//2
12 Out[4]: 2
```

By the way: # indicates a comment

Exercise 1

- ▶ Start Spyder
- ▶ Run some basic calculations in the Console window
- ▶ Put them also in a Python file and run them from there using F5 and F9
- ▶ What is the difference between using F5 or F9?

Priority of Operators

- ▶ Round brackets have highest priority
- ▶ followed by Power
- ▶ followed by multiplication and division
- ▶ followed by addition and subtraction

Basic Input & Output

```
1 print('Hello World')
2
3 yourName = input('Tell me your name: ')
4
5 print('Hello ' + yourName + ', welcome to this class')
6
7 print('Hello ', yourName, ', welcome to this class', sep='')
8
9 print(f'Hello {yourName}, welcome to this class')
```

- ▶ Strings use either single or double quotation marks
- ▶ `input()` only reads strings
- ▶ If you need to process a number, you need to convert it
- ▶ there are better ways than `print()` for logging, but it works...
- ▶ f-Strings (last row) are recommended for mixed output!

Exercise 2

- ▶ The formula to convert degrees Celsius to degrees Fahrenheit is

$$(c \times 1.8) + 32$$

with c as the value in Celsius

- ▶ Write some code that
 - ▶ asks a celsius-value from the user
 - ▶ converts it into Fahrenheit
 - ▶ stores the result in a variable
 - ▶ prints the result using f-Strings
 - ▶ Hint: to convert the string into a floating-point number, use `float(<string>)`

Basic Input & Output - Raw Strings

- ▶ Certain characters need to be escaped
- ▶ A list is e. g. here: https://www.w3schools.com/python/gloss_python_escape_characters.asp
- ▶ Raw strings can be used to prevent (most) processing, simply put an „r“ before the string¹

```
1 print('\\') # for a backslash
2
3 print('a\nb') # for a line-break
4
5 print(r'c:\windows') # to prevent most processing
```

¹A raw string must not end with a backslash

Rules for Variables

- ▶ must start with a letter or `_`
- ▶ Case-sensitivity: `'A'` is not `'a'`
- ▶ For naming conventions see <https://realpython.com/python-pep8/>
- ▶ Most important hint: let them speak for themselves: `'diameter'` is good, `'d'` is bad

Reserved Keywords

The following keywords are reserved and must not be used to name variables:

| | | | | |
|----------|-------|----------|-------|-------|
| and | as | assert | break | class |
| continue | def | del | elif | else |
| except | False | finally | for | from |
| global | if | import | in | is |
| lambda | None | nonlocal | not | or |
| pass | raise | return | True | try |
| while | with | yield | | |

Datatypes

- ▶ Integer (integer numbers)
- ▶ Float (Floating point number)
- ▶ Strings
- ▶ Booleans
- ▶ Complex numbers

Integer

- ▶ unlimited length
- ▶ must not start with 0 if they shall represent decimal numbers
- ▶ Leading 0 by representation in hex-, binary- or Octal system:
 - `0b/0B` binary
 - `0x/0X` hex
 - `0o/0O` octal
- ▶ Functions `hex()`, `bin()`, `oct()` for conversions into string, are internally represented as decimal numbers

Float

- ▶ floating point numbers
- ▶ 3.1415927
- ▶ 3.1e8
- ▶ Hint: Not every floating point number can be represented *exactly* („Floating-Point Arithmetic“)
- ▶ This can get tricky, if you compare numbers for equality
- ▶ docs.python.org/3/tutorial/floatingpoint.html

Strings

- ▶ Single or double quotes
- ▶ For multiline strings:
 - ▶ Triple double or single quotes
 - ▶ Alternatively backslash at the end of the line
- ▶ Python has numerous functions for strings, more on this later
- ▶ Comments start with a hash #

```
1 a = "I am a string"
2
3 b = 'me too'
4
5 c = """I am
6 as string as well """
7
8 # I am a comment
```

Booleans

- ▶ Named after George Boole
- ▶ 1854: „An investigation into the Laws of Thought“
- ▶ Essence of modern computing
- ▶ Boolean operators or (\cup) , and (\cap) , not

```
1 a = True
2 b = False
3
4 a == b #False
5 a or b # True
6 a and b # False
7 a and not b # True
8 not a and b # False
```

Type Conversions

- ▶ Mixing strings and floats/integers requires explicit type conversion using the `str()` function
- ▶ Use `int()` and `float()` to convert from string to number

```
1 >>> a + str(b)
2 'abc123'
3 >>> a+str(c)
4 'abc3.141'
5 >>> a*str(b)
6 Traceback (most recent call last):
7   File "<stdin>", line 1, in <module>
8   TypeError: can't multiply sequence by non-int of type 'str'
9
10 >>> 'a' * 3
11 'aaa'
12
13 >>> 3 * 'a'
14 'aaa'
```


Functions

- ▶ Functions: named sequence of commands
- ▶ Purpose: encapsule code for multiple calls
- ▶ *Can* take arguments as input, *can* have return values
- ▶ Define a new function as
 - ▶ Determine name
 - ▶ Determine arguments
 - ▶ Define function code

Built-in Functions

| | | | | |
|-------------|-----------|------------|------------|--------------|
| abs | delattr | hash | memoryview | set |
| all | dict | help | min | setattr |
| any | dir | hex | next | slice |
| ascii | divmod | id | object | sorted |
| bin | enumerate | input | oct | staticmethod |
| bool | eval | int | open | str |
| breakpoint | exec | isinstance | ord | sum |
| bytearray | filter | issubclass | pow | super |
| bytes | float | iter | print | tuple |
| callable | format | len | property | type |
| chr | frozenset | list | range | vars |
| classmethod | getattr | locals | repr | zip |
| compile | globals | map | reversed | __import__ |
| complex | hasattr | max | round | |

Simple Functions with any parameter

- ▶ def starts function definition
- ▶ do not forget : at the end of the def line
- ▶ indent the code inside the function using tabulator²
- ▶ No input parameter in round brackets
- ▶ No return value (void)

```
1 def print_hello():  
2     print('Hello')  
3  
4  
5 print_hello()
```

Listing 1: funktion-01.py Code

```
1 Hello
```

²Spyder expands it to four space characters

Simple Functions

Empty Functions

- ▶ `pass` is often used in the development process
- ▶ Useful, when parts of the code are not ready, yet
- ▶ without `pass` one gets `IndentationError: expected an indented block` error

```
1 def unfinished_function():  
2     pass  
3  
4  
5 unfinished_function()
```

Listing 2: funktion-12.py **Code**

Functions with Arguments

- ▶ Function with one argument text
- ▶ Error message, when argument is missing

```
1 def print_text(text):  
2     print(text)  
3  
4  
5 print_text('Hello World!')
```

Listing 3: funktion-02.py **Code**

```
1 Hello World!
```

Functions with Arguments

- ▶ Two arguments, text and count

```
1 def print_text_multiple(text, anzahl):  
2     print(text*count)  
3  
4  
5 print_text_multiple('Hello!', 0)  
6  
7 print_text_multiple('Hello!', 1)  
8  
9 print_text_multiple('Hello!', 3)  
10  
11 print_text_multiple('Hello!', -1)
```

Listing 4: funktion-03.py Code

```
1  
2 Hello!  
3 Hello!Hello!Hello!
```

Functions with Arguments

- ▶ Setting standard values for parameters
- ▶ Allows calling the function without parameters

```
1 def print_text_multiple(text='Hello CGN', count=2):
2     print(text*count)
3
4
5 print_text_multiple()
6
7 print_text_multiple('Hello!')
8
9 print_text_multiple('Hello!',3)
```

Listing 5: funktion-04.py Code

```
1 Hello CGNHello CGN
2 Hello!Hello!
3 Hello!Hello!Hello!
```

Functions with return values

- Functions can return values for further processing

```
1 def clone_text(text, count=2):  
2     return text*count  
3  
4  
5 a = clone_text('Huhu')  
6  
7 print(a)
```

Listing 6: funktion-05.py Code

```
1 HuhuHuhu
```


Exercise 3

- ▶ Use the code from the temperature conversion
- ▶ create a function for the conversion from Celsius to Fahrenheit
- ▶ Call the function multiple times

Solution for Exercise 3

```
1 def c2f(celsius):  
2     f = (celsius * 1.8) + 32  
3     return f  
4  
5  
6 c = 0  
7 f = c2f(c)  
8 print(f'{c} degrees Celsius are {f} degrees Fahrenheit')  
9  
10 c = 10  
11 f = c2f(c)  
12 print(f'{c} degrees Celsius are {f} degrees Fahrenheit')  
13  
14 c = 20  
15 f = c2f(c)  
16 print(f'{c} degrees Celsius are {f} degrees Fahrenheit')
```

Listing 7: c2f-function Code

Info: Functions with multiple return values

- ▶ Functions can have more than one return value
- ▶ Function then return tuples, an unmutable list of values³
- ▶ Dealing with the tuple is called „Unpacking“
- ▶ Remark: Parameter sep in the example is a parameter of the `print()` function

```
1 def clone_text(text, count=2):  
2     return count, text*count  
3  
4  
5 i, a = clone_text('Huhu')  
6  
7 print(i, a, sep='>')
```

Listing 8: funktion-06.py Code

```
1 2>HuhuHuhu
```

³more on this later

Info: Functions with variable count of arguments

- ▶ Parameter with *: variable count of arguments
- ▶ Parameter with **: variable count of key-value arguments

```
1 def addthem(*args):  
2     result = 0  
3     for number in args:  
4         result += number  
5     return result  
6  
7  
8 print(addthem(1, 2, 3, 4, 5))
```

Listing 9: funktion-08.py Code

```
1 def give(**args):  
2     for key, value in args.items():  
3         print(key, value, sep=': ')  
4  
5  
6 print(give(Vorname='Uwe', Nachname='Ziegenhagen'))
```

Listing 10: funktion-09.py Code

Flow control

- ▶ Input & Output ✓
 - ▶ input
 - ▶ print
- ▶ Branching
 - ▶ if else elif
- ▶ Loops
 - ▶ for
 - ▶ while

Branching

► `if <condition>:`

```
1 def check_length(text):  
2     if len(text)>8:  
3         print('Longer than 8 characters!')  
4  
5 check_length('Köln')  
6 check_length('Düsseldorf')
```

Listing 11: if-01.py Code

```
1 Longer than 8 characters!
```

Branching

- ▶ There is no `switch()` in Python < 3.10⁴
- ▶ `if` statements can be used multiple times
- ▶ maybe not the most pythonic approach

```
1 def check_length(text):
2     if len(text)>8:
3         print(text, 'Longer than 8 characters!', sep=': ')
4     if len(text)<=8:
5         print(text, 'Shorter or equal to 8 characters!', sep=': ')
6
7 check_length('Köln')
8 check_length('Düsseldorf')
```

Listing 12: if-02.py Code

```
1 Köln: Shorter or equal to 8 characters!
2 Düsseldorf: Longer than 8 characters!
```

⁴From 3.10 there is `match/case`, see
stackoverflow.com/questions/11479816/

Branching

► more „pythonic“: `else:`

```
1 def check_length(text):
2     if len(text)>8:
3         print(text, 'Longer than 8 characters!', sep=': ')
4     else:
5         print(text, 'Shorter or equal to 8 characters!', sep=': ')
6
7 check_length('Köln')
8 check_length('Düsseldorf')
```

Listing 13: if-03.py Code

```
1 Köln: Shorter or equal to 8 characters!
2 Düsseldorf: Longer than 8 characters!
```


Branching

- Nesting von `if <condition>`: `else`: leads to confusing code

```
1 def check_length(text):
2     if len(text)>8:
3         print(text, 'Longer than 8 chars!', sep=': ')
4     else:
5         if len(text)<=5:
6             print(text, 'Shorter or equal 5 chars!', sep=': ')
7         else:
8             print(text, 'Longer than 5, shorter than 8', sep=': ')
9
10 check_length('Köln')
11 check_length('Berlin')
12 check_length('Düsseldorf')
```

Listing 14: if-04.py Code

```
1 Köln: Shorter or equal 5 chars!
2 Berlin: Longer than 5, shorter than 8
3 Düsseldorf: Longer than 8 chars!
```

Branching

► `if <condition>: else:` can be shortened to `elif`:

```
1 def check_length(text):
2     if len(text)>8:
3         print(text, 'Longer than 8 chars!', sep=': ')
4     elif len(text)<=5:
5         print(text, 'Shorter or equal 5 chars!', sep=': ')
6     else:
7         print(text, 'Longer than 5, shorter than 8', sep=': ')
8
9 check_length('Köln')
10 check_length('Berlin')
11 check_length('Düsseldorf')
```

Listing 15: if-05.py Code

```
1 Köln: Shorter or equal 5 chars!
2 Berlin: Longer than 5, shorter than 8
3 Düsseldorf: Longer than 8 chars!
```

Loops

for

- ▶ `for` loops iterate over a sequence
- ▶ sequence can be a string, a list, etc.
- ▶ `range(start, end, stepsize=1)` creates numerical sequence from start until below (!) end with step size stepsize

```
1 for char in 'Hallo Welt':  
2     print(char)  
3  
4 for j in range(1, 10):  
5     print(j) # 1 bis 9  
6  
7 for j in range(1, 10, 3):  
8     print(j) # 1, 4 und 7
```

Listing 16: for-01.py Code

Loops

while

- ▶ `while` loop runs, until condition is not fulfilled anymore

```
1 s = 'Hallo Welt'
2 l = len(s)
3
4 while (l>0):
5     print(s[l-1])
6     l-=1
7
8 i = 1
9 while (i<100):
10     i += 1
11
12 print(i)
```

Listing 17: while-01.py **Code**

Loops

break and continue

- ▶ `break` and `continue` influence loops
- ▶ `break` can e. g. be used to exit a loop

```
1 s = 'Hallo Welt'
2 l = len(s)
3
4 while (l>0):
5     temp = s[l-1]
6     if temp == 'W':
7         break
8     print(temp)
9     l-=1
```

Listing 18: while-02.py **Code**

File-Operations

- ▶ `open()` opens file for read/write access
- ▶ two parameters:

Path Path to the file

Mode Read, Write, Binary, Text

```
1 f = open('u:/hello.txt', 'w')
2 f.write("Hello World!")
3 f.close()
```

Listing 19: Simple example for `write()`

Not optimal, in case of errors the file-handle might remain open!
The file is not usable by other applications.

File-Operations

Improved version, uses „Context manager“, closes file handle in each situation

```
1 with open('r:/hello.txt', 'w') as f:  
2     f.write("Hello World!")
```

Listing 20: Improved example for write()

Hint: Context managers are also very useful when dealing with SQL databases.

Exercise 4: Files and Branching

- ▶ Ask the user to input a number
- ▶ If the number < 0 write „Hello“ to a text file
- ▶ If the number > 0 write „World“ to a text file
- ▶ If the number $= 0$ write „Foobar“ to a text file
- ▶ Open the file and printout the file on the screen
- ▶ Delete the created file afterwards

File-Operations

Read-/Write- parameters

- `r` Read; Error, when file is not present
- `r+` Read and Write
- `a+` Read and append, file is created if not existent
- `x` Creates file, error if file exists
- `a` Append; creates file if not existent, appends at the end
- `w` Write; creates file if not existent; overwrite, if file exists

Content format

- `t` for text files
- `b` for binary files (images, zip, etc)

File-Operations

```
1 with open('r:/hello.txt', 'rt') as file:  
2     print(file.read())
```

Listing 21: Read a complete file

```
1 with open('r:/hello.txt', 'rt') as file:  
2     for line in file:  
3         print(x)
```

Listing 22: Row-wise reading a file

File-Operations

Deleting files

```
1 import os
2
3 if os.path.exists('r:/hello.txt'):
4     os.remove('r:/hello.txt')
5 else:
6     print("File does not exist!")
```

Listing 23: Deleting a file

The `range()` Function

- ▶ `range()` function creates a sequence of numbers via generator \Rightarrow see next slide
- ▶ Three parameters maximum : start, stop and step
- ▶ `range(<stop>)` with one parameter, `range(0,<stop>,1)` implicitly
- ▶ `range(<start>, <stop>)` with two parameters, `<start>,<stop>,1` implicitly
- ▶ `range(<start>,<stop>,<step>)` with three parameters
- ▶ Important: `<start>` is inclusive, `<stop>` not!!!
- ▶ `range(0,10)` runs from 0 to 9

Examples for range()-Function

```
1  # -*- coding: utf-8 -*-
2
3  for i in range(10):
4      print('1', i)
5
6  print('\n')
7  for i in range(2, 10):
8      print('2', i)
9
10 print('\n')
11 for i in range(2, 10, 2):
12     print('3', i)
13
14 print('\n')
15 for i in range(10, -10, -2):
16     print('4', i)
```

Listing 24: range_beispiel.py Code

```
1  1: 0  1  1  1  2  1  3  1  4  1  5  1  6  1  7  1  8  1  9
2  2: 2  2  3  2  4  2  5  2  6  2  7  2  8  2  9
3  3: 2  3  4  3  6  3  8
4  4: 10 4  8  4  6  4  4  4  2  4  0  4  -2  4  -4  4  -6  4  -8
```

Sequential Datatypes aka: for i in whatever

- ▶ Sequential Datatypes = Datatypes that store elements sequentially
 - Strings contain characters only
 - Lists different objects possible, mutable (changeable)
 - Tuple different objects possible, not mutable (not changeable)
- ▶ Identical methods for the access: `object[<number>]` to access a specific element, `len()` for the length, Slicing-Notation

String Functions

- ▶ Numerous string-functions are available, see `docs.python.org/3/library/stdtypes.html#text-sequence-type-str`
- ▶ Length of a string using `len(<String>)`
- ▶ Upper- or lowercase a string with `upper(<String>)` and `lower(<String>)`
- ▶ `index(<String>)`, `find(<String>)` and `replace(<String>)`
- ▶ `startswith(<String>)` and `endwith(<String>)`
- ▶ `split(<String>)` and `strip(<String>)`

Lists

- ▶ Lists contain arbitrary objects
- ▶ Square brackets, elements separated by comma
- ▶ first element is `listname[0]`
- ▶ Can be nested
- ▶ Can be changed at runtime \Rightarrow „mutable“

```
1 abc = ['a', 'b', 'c', 3.1234]
2 efg = [1, 2, [1, 2, 3], 3, 4]
```


Tuples

- ▶ Round brackets, can be left out
- ▶ Recommendation: do not leave them out
- ▶ Immutable, objects cannot be changed after creation
- ▶ can be unpacked via multi-assignment: `a, b, c = (1, 2, 3)`
- ▶ hint: switch two numbers by `a, b = b, a`

Indexing Sequential Datatypes

- ▶ Two ways of indexing: from 0 to $n - 1$ and $-n$ to -1

| | | | | | | | | | | | |
|-------|-----|-----|----|----|----|----|----|----|----|----|----|
| Index | -11 | -10 | -9 | -8 | -7 | -6 | -5 | -4 | -3 | -2 | -1 |
| | H | e | l | l | o | | W | o | r | l | d |
| Index | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |

Slicing

- ▶ Slicing = very powerful, especially for strings
- ▶ two parameters, start and stop, separated by :, both are optional
- ▶ third parameter: step size

```
1 >>> a = 'Hello World'
2 >>> a[:]
3 'Hello World'
4 >>> a[1:-1]
5 'ello Worl'
6 >>> a[1:-5]
7 'ello '
8 >>> a[::1]
9 'Hello World'
10 >>> a[::2]
11 'HloWrld'
12 >>> a[1:-1:3]
13 'eoo'
14 >>> a[::-1]
```

Exercise 5: String manipulation

- ▶ Ask the user to input a file name, or alternatively: use the filenames in a directory of your choice
- ▶ print the filename without the file extension as well as the file extension separately

Dictionaries

- ▶ Dictionaries = Associative fields, maps, hashes
- ▶ consist of Key-Value pairs, for each („Key“) a („Value“) is assigned
- ▶ each key must only exist once in each dict
- ▶ can arbitrarily grow and shrink
- ▶ all immutable objects can be keys: strings, floats, integers, tuples, but no lists or dictionaries
- ▶ Dictionaries can be nested (well, if you ever need this, think again...)

Dictionaries

```
1 de_en = {'Glücklich': 'Happy', 'Baum': 'Tree'}
2 de_en['Freunde'] = 'Friends'
3
4 print(de_en['Baum'])
5 # print(de_en['Tree']) # => KeyError
6 'Baum' in de_en
7
8 monatsnamen = {1: 'Januar', 2: 'Februar', 3: 'März'}
9 print(monatsnamen[1])
```

Listing 25: dict-01.py Code

```
1 Tree
2 Januar
```

Accessing a Dictionary

```
1 d = {1: 'one', 2: 'two', 3: 'three'}
2
3 print(d.keys())
4 print(d.values())
5 print(d.items())
6
7 print(d.get(4))
8 print(d.get(3))
9
10 for k, v in d.items():
11     print(k, v)
```

Listing 26: dict-04.py Code

```
1 dict_keys([1, 2, 3])
2 dict_values(['one', 'two', 'three'])
3 dict_items([(1, 'one'), (2, 'two'), (3, 'three')])
4 None
5 three
6 1 one
7 2 two
8 3 three
```

Functions for Dictionaries (Selection)

`clear()` deletes all entries

`copy()` creates a flat copy

`keys()` set of all keys

`pop()` removes key and its value from dict

`update()` adds dict2 to dict, overwrites values eventually

`popitem()` removes arbitrary key-value combination from dict, `KeyError` if dict is empty. Important: „arbitrary“ \neq random!

pandas

- ▶ A Python library for data wrangling and management
- ▶ Invented by Wes McKinney during his time at AQR Capital Management
- ▶ In his own words: „I tell them that it enables people to analyze and work with data who are not expert computer scientists,” he says. „You still have to write code, but it’s making the code intuitive and accessible. It helps people move beyond just using Excel for data analysis.”⁵

⁵qz.com/1126615/

Das SciPy Framework

pandas is just a piece among many:

NumPy Matrices, vectors, algorithms

IPython Matlab/Mathematica-like environment

Matplotlib Scientific Plotting, Basis for seaborn library

SymPy Symbolic mathematics

... etc, etc

We only focus on pandas today:

```
1 import pandas as pd
```

Series and DataFrames

- ▶ central data structures in pandas: series and dataframes
- ▶ quite similar to dataframes in R
- ▶ Definition „Series“: a vector with data of the same type and an index
- ▶ Definition „Dataframe“: matrix from various series, the series can have different data types haben but they share the same index

Series und DataFrames

The diagram illustrates a DataFrame structure. A red arrow on the left points downwards, labeled "Row Index". A red arrow at the top points to the right, labeled "Column Index". The DataFrame consists of 7 rows and 7 columns. The first column contains row indices from 0 to 6. The subsequent columns contain numerical values, currency codes, and ellipses. The first row of data (row index 0) has values 0.2, 'USD', and three ellipses, followed by three empty cells. The second row (index 1) has values 0.4, 'EUR', and two ellipses, followed by three empty cells. The third row (index 2) has values 0.1, 'USD', and two ellipses, followed by three empty cells. The fourth row (index 3) has values 0.7, 'EUR', and two ellipses, followed by three empty cells. The fifth row (index 4) has values 0.5, 'YEN', and two ellipses, followed by three empty cells. The sixth row (index 5) has values 0.5, 'USD', and two ellipses, followed by three empty cells. The seventh row (index 6) has values 0.0, 'AUD', and two ellipses, followed by three empty cells.

| | | 'var 0' | 'var 1' | 'var 2' | 'var 3' | 'var 4' | 'var 5' | 'var 6' |
|---|-----|---------|---------|---------|---------|---------|---------|---------|
| 0 | 0.2 | 'USD' | ... | | | | | |
| 1 | 0.4 | 'EUR' | ... | | | | | |
| 2 | 0.1 | 'USD' | ... | | | | | |
| 3 | 0.7 | 'EUR' | ... | | | | | |
| 4 | 0.5 | 'YEN' | ... | | | | | |
| 5 | 0.5 | 'USD' | ... | | | | | |
| 6 | 0.0 | 'AUD' | ... | | | | | |

Manual Creation of pandas Objects

- ▶ pandas-objects *can* be created manually
- ▶ normally not used, data is usually loaded from files/databases

```
1 import pandas as pd
2
3 d = pd.DataFrame({'A': ['A0', 'A1', 'A2', 'A3'],
4   'Key': ['K0', 'K1', 'K2', 'K4']})
5 a = pd.Series([1,2,3,4,5,6,7,8,9,10])
6 b = pd.Series(['A', 'C', 'D', 'B', 'F', 'G', 'I', 'K', 'L', 'P'])
7 df = pd.concat([a,b], axis=1)
8 # alternatively
9 df = pd.DataFrame({'a': a, 'b':b})
10 df = a.to_frame().join(b.to_frame())
11 df = pd.DataFrame(data=dict(a=a, b=b))
```

Daten einlesen

- ▶ various functions to read files

| Befehl | Beschreibung |
|-----------------------------|------------------------|
| <code>read_pickle</code> | reads Pickle objects |
| <code>read_table</code> | table-like formats |
| <code>read_csv</code> | Comma-Separated Values |
| <code>read_fwf</code> | fixed-width formats |
| <code>read_clipboard</code> | clipboard |
| <code>read_excel</code> | Excel-files |

other commands for HTML, JSON, HDF5, ...

Reading CSV

- ▶ „CSV“: Comma-Separated Value
- ▶ CSV is not a unique format
 - ▶ Column-separator
 - ▶ Decimal-separator
 - ▶ Text Encoding
- ▶ Specifications: http://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html
 - `sep` Column-separator
 - `thousands` separator for thousands
 - `encoding` Encoding
 - `decimal` Decimal-separator
 - `converters` `converters={'A': str}` for explicit conversion to a format

Reading Excel

- ▶ `pd.read_excel()` to read XLSX-files (!)
- ▶ Documentation:
http://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_excel.html
- ▶ Export to Excel using `pd.to_excel()`
- ▶ Remarks:
 - ▶ Excel-Export is way slower than CSV-Export
 - ▶ Export of well-formatted Excel is possible but takes effort
 - ▶ One can even control Excel via COM (Common Object Model)

DataFrames bearbeiten

Exploratory Data Analysis

- ▶ We use Northwind data as an example
- ▶ First task after loading: check data consistency

```
1 customers = pd.read_excel('Northwind.xlsx', sheetname = '
    Customers')
2
3 print(len(customers)) # number of rows
4 print(customers.head()) # first five rows
5 print(customers.tail()) # last five rows
6 print(list(customers)) # list of columns
```

Aufgabe

- ▶ Read the files `Bestandsdaten.csv`, `Bestandsdaten_Vortrag.csv` and `Bewertung.xlsx`
- ▶ Check the data using the functions from last slide
- ▶ Export the data to HTML, which columns could be used to connect the data sets with eachother?

Pandas Dataframe Operations

Selection and Filtering I

- ▶ pandas has advanced methods for selecting, filtering and transforming data

- ▶ Select only specific columns

```
df = df[['colA', 'colB']]
```

- ▶ Select the top two rows (Index starting at 0)

```
df.iloc[:1]
```

- ▶ Select only those rows where one row > 50

```
df[df['colA'] > 50]
```

Pandas Dataframe Operations

Selection and Filtering II

- ▶ Select only those rows where value is between two values

```
df[(df['colA'] > 50) | (df['colA'] < 500)]
```

- ▶ Select rows that **do not** have a certain value

```
df[~(df['colA'] == 'HelloWorld')]
```

- ▶ Select rows, where column b is either value 'A' or 'B'

```
df = df[(df['b'] == 'A') | (df['b'] == 'I')]
```

- ▶ Use alternatively `isin()`

```
df = df[df['b'].isin(['A', 'I'])]
```

- ▶ or the opposite

```
df = df[~df['b'].isin(['A', 'I'])]
```

- ▶ More here: http://chrisalbon.com/python/pandas_indexing_selecting.html

Apply functions to pandas columns

```
1 import pandas as pd
2
3 def capitalizeColumn(text):
4     return text.capitalize()
5
6 df = pd.DataFrame({'Key': ['K0', 'K1', 'K2', 'K4'],
7                     'Name': ['anna', 'bernd', 'cesar', 'dana']})
8
9 print(df)
10
11 df['Nachname'] = df['Name'].apply(capitalizeColumn)
12
13 print(df)
```

| | Key | Name | Nachname |
|---|-----|-------|----------|
| 0 | K0 | anna | Anna |
| 1 | K1 | bernd | Bernd |
| 2 | K2 | cesar | Cesar |
| 3 | K4 | dana | Dana |

Mapping I

- ▶ Similar to Excel's `vlookup()` function
- ▶ Example: `countries` is a key-value dictionary
- ▶ `country` column in the dataframe is used as key in the dictionary, a new column is created

```
1 import pandas as pd
2
3 df = pd.DataFrame({'A': ['A0', 'A1', 'A2', 'A3'],
4   'Country': ['DEU', 'USA', 'ARE', 'ESP']})
5
6 countries = { 'DEU': 'Germany',
7               'USA': 'United States',
8               'ARE': 'Arabic Emirates',
9               'ESP': 'Spain'}
10
11 df['Country'] = df['Country'].map(countries)
12
13 print(df)
```

Mapping II

- ▶ `map()` can also be used to make simple calculations
- ▶ keyword here is „lambda“ \Rightarrow anonymous function

```
1 import pandas as pd
2
3 df = pd.DataFrame({'A': ['A0', 'A1', 'A2', 'A3'],
4   'Net': [100, 200, 300, 400]})
5
6 df['Gross'] = df['Net'].map(lambda x: x * 1.19)
7
8 print(df)
```


Pandas Dataframe Operations

Merge and Join

- ▶ `merge()` SQL-like merging of datasets
- ▶ useful to combine data
- ▶ Supports the following join types:
 - ▶ Left
 - ▶ Right
 - ▶ Inner
 - ▶ Full Outer
- ▶ `join()` is a special alias for `merge()`, works on the index, not the columns

Pandas Dataframe Operations

Merge and Join

► Standard-command for merge()

```
1 leftDataFrame.merge(rightDataFrame, how='inner',  
2 on=None, left_on=None, right_on=None, left_index=False,  
3 right_index=False, sort=False, suffixes=('_x', '_y'),  
4 copy=True, indicator=False)
```

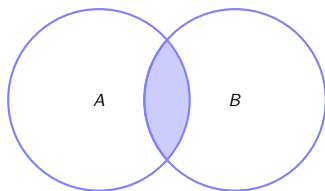
Workflow

1. define other dataset
2. define type of merge
3. define keys for the merger

Merging

Inner Join

- Select all data, that is in A **and** B



| left | | |
|------|----|-----|
| | A | Key |
| 0 | A0 | K0 |
| 1 | A1 | K1 |
| 2 | A2 | K2 |
| 3 | A3 | K4 |

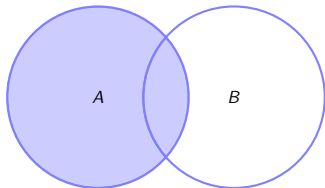
| right | | |
|-------|----|-----|
| | B | Key |
| 0 | B0 | K0 |
| 1 | B1 | K1 |
| 2 | B2 | K2 |
| 3 | C3 | K5 |

| merged | | | |
|--------|----|----|-----|
| | A | B | Key |
| 0 | A0 | B0 | K0 |
| 1 | A1 | B1 | K1 |
| 2 | A2 | B2 | K2 |

Merging

Left Join

- Select all data in A, get data from B if available



| left | | |
|------|----|-----|
| | A | Key |
| 0 | A0 | K0 |
| 1 | A1 | K1 |
| 2 | A2 | K2 |
| 3 | A3 | K4 |

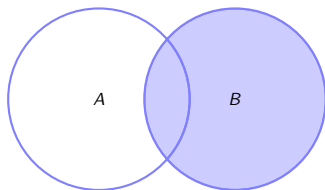
| right | | |
|-------|----|-----|
| | B | Key |
| 0 | B0 | K0 |
| 1 | B1 | K1 |
| 2 | B2 | K2 |
| 3 | C3 | K5 |

| merged | | | |
|--------|----|-----|-----|
| | A | B | Key |
| 0 | A0 | B0 | K0 |
| 1 | A1 | B1 | K1 |
| 2 | A2 | B2 | K2 |
| 3 | A3 | NaN | K4 |

Merging

Right Join

- Select all data in B, get data from A if available



| left | | |
|------|----|-----|
| | A | Key |
| 0 | A0 | K0 |
| 1 | A1 | K1 |
| 2 | A2 | K2 |
| 3 | A3 | K4 |

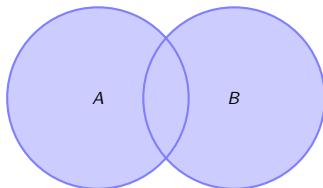
| right | | |
|-------|----|-----|
| | B | Key |
| 0 | B0 | K0 |
| 1 | B1 | K1 |
| 2 | B2 | K2 |
| 3 | C3 | K5 |

| merged | | | |
|--------|-----|----|-----|
| | A | B | Key |
| 0 | A0 | B0 | K0 |
| 1 | A1 | B1 | K1 |
| 2 | A2 | B2 | K2 |
| 3 | NaN | B3 | K5 |

Merging

Full Outer Join

- Select all data which is in A **or** B



| left | | |
|------|----|-----|
| | A | Key |
| 0 | A0 | K0 |
| 1 | A1 | K1 |
| 2 | A2 | K2 |
| 3 | A3 | K4 |

| right | | |
|-------|----|-----|
| | B | Key |
| 0 | B0 | K0 |
| 1 | B1 | K1 |
| 2 | B2 | K2 |
| 3 | C3 | K5 |

| merged | | | |
|--------|-----|-----|-----|
| | A | B | Key |
| 0 | A0 | B0 | K0 |
| 1 | A1 | B1 | K1 |
| 2 | A2 | B2 | K2 |
| 3 | A3 | NaN | K4 |
| 4 | NaN | B3 | K5 |

Loop through DataFrames

```
1 import pandas as pd
2
3 df = pd.DataFrame({'Key': ['K0', 'K1', 'K2', 'K4'],
4                     'Name': ['Anna', 'Bernd', 'Cesar', 'Dana']})
5
6 for index, row in df.iterrows():
7     print(row['Key'], 'belongs to', row['Name'])
```

```
1 K0 belongs to Anna
2 K1 belongs to Bernd
3 K2 belongs to Cesar
4 K4 belongs to Dana
```

Links

Links

- ▶ https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html
- ▶ <https://pandas.pydata.org>
- ▶
- ▶
- ▶
- ▶