Data Analysis with Python

Course Introduction

Martin Uray

Josef Ressel Center for Intelligent and Secure Industrial Automation
Department for Information Technologies and Digitalisation

Salzburg University of Applied Sciences, Austria

September 13, 2023

Salzburg University
of Applied Sciences

Course Admin I

Course Team

Martin Uray
Lecture ITS,
Lecture and Lab ITS-B

Maximilian Schirl
Lab ITS

Course Admin II

Lecturer

Martin Uray

- Lecturer / Researcher
- Office U414
- Open Student Hours:
  - Friday, 08.15 am - 9.00 am,
  - and before / after class,
  - online and in person
- for further information see `hhttps://www.fh-salzburg.ac.at/personen/martin-uray`

Course Admin III

Moodle LMS

- Course: *WF: Datenanalyse mit Python (ID: 229953)*
- direct link here

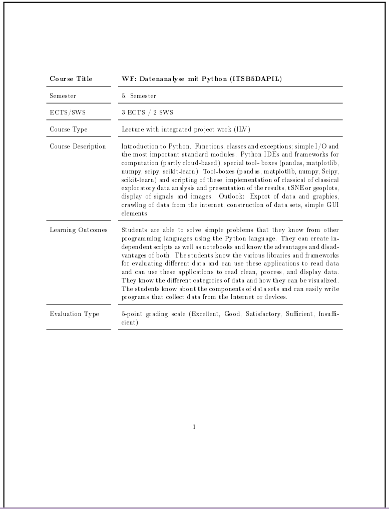
Course Admin IV

Syllabus

See Moodle: `https://elearn.fh-salzburg.ac.at/pluginfile.php/19606/mod_resource/content/3/syllabus.pdf`

Attention

Course is of *immanent character*!



Course Title	**WF: Datenanalyse mit Python (ITSB5DAPIL)**
Semester	5. Semester
ECTS/SWS	3 ECTS / 2 SWS
Course Type	Lecture with integrated project work (ILV)
Course Description	Introduction to Python. Functions, classes and exceptions; simple I/O and the most important standard modules. Python IDEs and frameworks for computation (partly cloud-based), special tool- boxes (pandas, matplotlib, numpy, scipy, scikit-learn). Tool-boxes (pandas, matplotlib, numpy, Scipy, scikit-learn) and scripting of these, implementation of classical of classical exploratory data analysis and presentation of the results, tSNEor geoplots, display of signals and images. Outlook: Export of data and graphics, crawling of data from the internet, construction of data sets, simple GUI elements
Learning Outcomes	Students are able to solve simple problems that they know from other programming languages using the Python language. They can create independent scripts as well as notebooks and know the advantages and disadvantages of both. The students know the various libraries and frameworks for evaluating different data and can use these applications to read data and can use these applications to read clean, process, and display data. They know the different categories of data and how they can be visualized. The students know about the components of data sets and can easily write programs that collect data from the Internet or devices.
Evaluation Type	5-point grading scale (Excellent, Good, Satisfactory, Sufficient, Insufficient)

1

Course Admin

Format	Exam modality	Flexible	Points	Must be positive in itself	Minimum attendance
Readings	Quizzes	no	25	no	50%[1]
Lab Assignments	Jupyter Notebooks	no	30	no	50%[1]
Project	Presentation and Deliverables	no	45	no	33.3%[2]

[1] Submissions have to be turned in, irrespective of attendance.

[2] Attendance only mandatory for final presentation. All other units are optional, but highly recommended.

Course Admin V

Plagiarism - Intentional or unintentional

Be referred to the *Examination Regulations* (ER), §30f, (see this link):

“Plagiarism occurs when someone[a]

1. Uses words, ideas, or work products
2. Attributable to another identifiable person or source
3. Without attributing the work to the source from which it was obtained
4. In a situation in which there is a legitimate expectation of original authorship
5. In order to obtain some benefit, credit, or gain which need not be monetary”

- Plagiarism also exists if it “happens unintentionally” and has the benefit described in the ER (“grade”, “degree”)
- Instructors are instructed to guarantee fair conditions for all, and in particular to strictly prosecute all forms of plagiarism.

If in any doubt, please consolidate the instructor(s).

[a] Teddi Fishman. “We know it when we see it is not good enough: Toward a standard definition of plagiarism that transcends theft, fraud, and copyright”. In: *Educational Integrity: Creating an Inclusive Approach. Proceedings of the 4th Asia Pacific Conference on*

Course Admin VI

Schedule

Lecture	
1	Course Introduction
2	Introduction to Python
3	Data Handling
4	Data Visualization
5	Project Introduction
6	Machine Learning
7	Dashboards

Lab	
1	Categegorical Data
2	Geographical Data
3	Continous Data
4	Consulting for the Project (optional)
5	Time Series Data
6	Classification (ML)
7	Project Presentation

Course Admin VII

Applied Data Science Talks

- Talks on different topics in the field of Data Analysis, AI and their application
- speaker from academia and industry
- regional and international

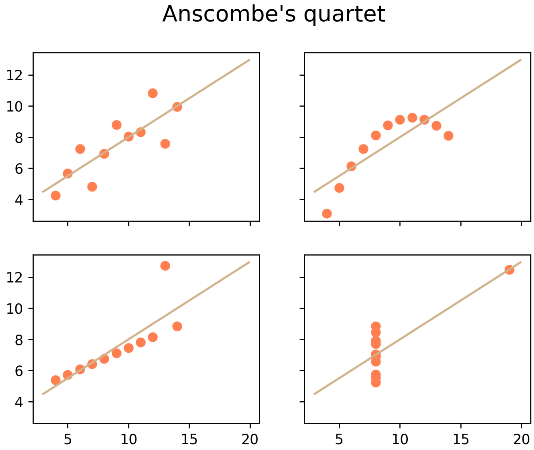
More information and registration here.

Course

Data Analysis with Python

Motivation I

Ancombe’s Quartet

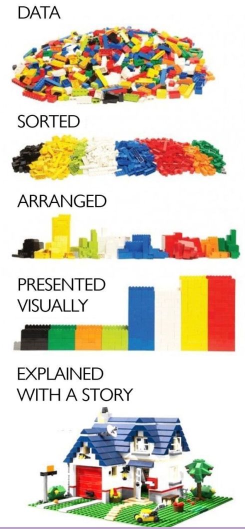


Properties of all four datasets[a]:

Property	Value
Mean of x	9
Sample variance of x: s_x^2	11
Mean of y	7.50
Sample variance of y:s_y^2	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$

[a] F. J. Anscombe. “Graphs in Statistical Analysis”. In: *The American Statistician* 27.1 (Feb. 1973), pp. 17–21. ISSN: 0003-1305, 1537-2731. DOI: 10.1080/00031305.1973.10478966. (Visited on 08/23/2022).

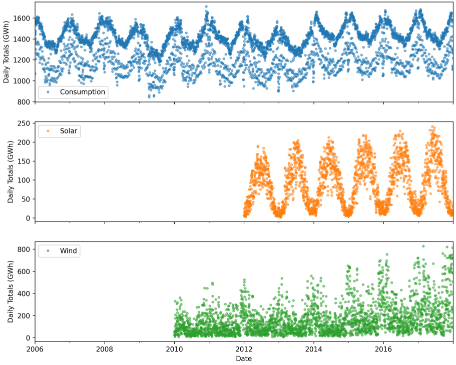
Motivation II



Motivation III

Date	Consumption	Wind	Solar	Wind+Solar
2012-01-01	948.128	227.465	6.587	234.052
2012-01-02	1269.581	207.327	6.574	213.901
2012-01-03	1334.745	473.468	24.679	498.147
2012-01-04	1347.136	499.804	14.681	514.485
2012-01-05	1376.658	523.851	5.071	528.922
2012-01-06	1291.215	286.265	13.16	299.425
2012-01-07	1175.688	368.288	4.115	372.403
2012-01-08	1103.383	220.851	8.44	229.291
2012-01-09	1443.371	151.837	5.264	157.101
2012-01-10	1434.631	175.995	17.827	193.822
2012-01-11	1449.768	197.434	10.849	208.283
2012-01-12	1442.448	446.327	18.023	464.35
2012-01-13	1403.402	415.106	18.778	433.884
2012-01-14	1203.165	174.69	26.772	201.462
2012-01-15	1150.92	34.468	36.609	71.077
2012-01-16	1487.782	52.345	39.682	92.027
2012-01-17	1518.074	76.43	31.036	107.466
2012-01-18	1498.809	225.266	40.924	266.19
2012-01-19	1470.066	282.584	3.885	286.469

vs.



Data Analysis

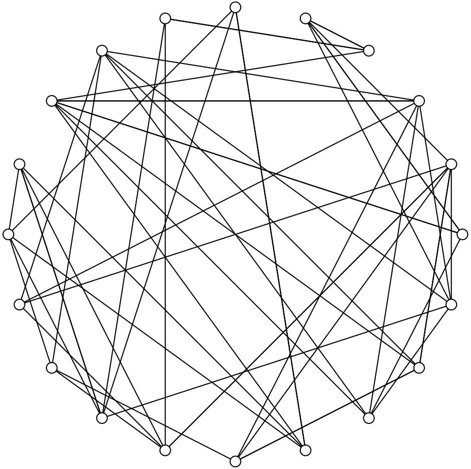
Exploratory data analysis field of statistics that[3]

- “. . . has been an influential back-to-basics movement, eschewing probability models and focusing on graphical visualization of data.”
- “. . . along with a general view of data science as going beyond statistical theory . . . ”
- “. . . focused on discovery . . . ”

[3] Andrew Gelman and Aki Vehtari. “What are the most important statistical ideas of the past 50 years?” In: *arXiv:2012.00174 [stat]* (June 2021). (Visited on 07/06/2021).

Data Sources

- Databases
- APIs
- Web Scraping
- Data Streams
- (Flat Files)



Data Analysis with Python?

Scientific Computing Languages: *R*, *SAS*, *Stat*, *MATLAB*, *SPSS*, . . .

Which Language to use?

- `https://www.tiobe.com/tiobe-index/`
- `https://bootcamp.berkeley.edu/blog/most-in-demand-programming-languages/`
- `https://statisticstimes.com/tech/top-computer-languages.php`
- `...`

Language Introduction

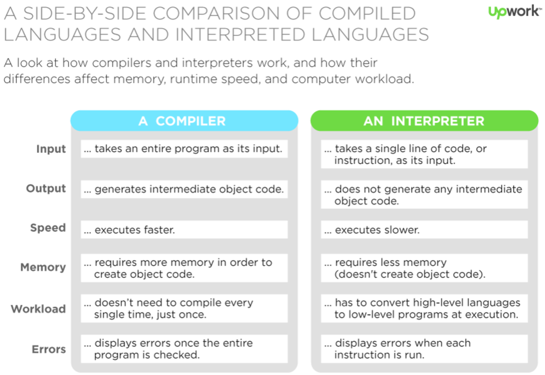
Python

Python is an easy-to-use language that makes it simple to get your program working. This makes Python ideal for prototype development and other ad-hoc programming tasks. However, Python as well supports object-oriented programming with classes and multiple inheritance. Code can be grouped into modules and packages.- *python.org*[4]

[4] Guido van Rossum and Fred L. Drake. *The Python Language Reference*. https://docs.python.org/3/reference/index.html. 2011. (Visited on 09/11/2023).

Language Introduction

Python is an **interpreted language**. Thus, it is similar to Matlab, but opposed to C, for instance, which is a compiled language.



A SIDE-BY-SIDE COMPARISON OF COMPILED LANGUAGES AND INTERPRETED LANGUAGES

upwork

A look at how compilers and interpreters work, and how their differences affect memory, runtime speed, and computer workload.

	A COMPILER	AN INTERPRETER
Input	... takes an entire program as its input.	... takes a single line of code, or instruction, as its input.
Output	... generates intermediate object code.	... does not generate any intermediate object code.
Speed	... executes faster.	... executes slower.
Memory	... requires more memory in order to create object code.	... requires less memory (doesn't create object code).
Workload	... doesn't need to compile every single time, just once.	... has to convert high-level languages to low-level programs at execution.
Errors	... displays errors once the entire program is checked.	... displays errors when each instruction is run.

Language Introduction

The two major Python versions, Python 2 and **Python 3**, are quite different from each other.

Python3.x

This courses uses Python 3, because it more semantically correct and supports newer features. Be aware of the two versions when searching for code snippets online.

For an in-depth overview about the differences between the two major versions, be referred to Sebastian Rashka’s Blog.

Alternate Language Implementations:

- CPython
- Jython
- Python for .NET
- IronPython
- PyPy

Each of these implementations varies in some way from the language as documented in this manual, or introduces specific information beyond what's covered in the standard Python documentation. Please refer to the implementation-specific documentation to determine what else you need to know about the specific implementation you're using.

Environments

- Environments manage the packages for a certain project or application
- when activated only using packages from environment
- possible to have different versions for same package on different projects
- e.g. *conda* or *virtualenv*

Package Manager

- Have you ever struggled with your packages installed?
- Package manager take care about installing and managing packages (or libraries).
- e.g. *conda* or *pip*

IDE vs. Notebooks I

Jupyter Notebooks

- All in One:
  - Code
  - Visualizations
  - Text (Markdown and LaTeX support)
- Report look-and-feel

Attention

Integrated Development Environment

- similar to what known already (Eclipse, Spyder, etc.)
- Development of applications
- integration of external dependencies
- can take care about environments
- Notebook Support (CAVE)

IDE vs. Notebooks II

Notebooks are cool for quick tests (and data science), but shall not be applied for development, see this video.

Python

Community and Conferences/Meetups

Communities	Meetups
pydata (google group)	PyCon and EuroPython
pystatsmodels	regional PyCon conferences
scikit-learn mailinglist	SciPy and EuroSciPy
etc.	PyData

To Dos for you

Minimal Requirements:

- □ Install Conda and create environment
- □ Install Jupyter Notebook
- □ Download and run first Notebook
- □ Complete Wrap-up Exercises
- □ Do the quiz (Questions from Wrap-up) in the elearning course

Setup for Lab

A missing or non-working environment leads to a not accountant of attendance.