Python for Data Science

Prof. Dr. Wolfgang Ecker, Lorenzo Servadei,
Elena Zennaro
19th October 2018

Content

Copyright © Infineon Technologies AG 2018. All rights reserved.

Python Libraries for Data Science

Data Analysis

- NumPy
- SciPy
- Pandas
- SciKit-Learn

Data Visualization

- Matplotlib
- Seaborn

These are the libraries we will use during our Course!

Copyright © Infineon Technologies AG 2018. All rights reserved.

NumPy

- Fundamental package for scientific computing with Python
- Objects for multidimensional arrays and matrices, as well as functions that allow to easily perform advanced mathematical and statistical operations on those objects
- Provides vectorization of mathematical operations on arrays and matrices which significantly improves the performance

Link: http://www.numpy.org/

Copyright © Infineon Technologies AG 2018. All rights reserved.

SciPy

- Collection of algorithms for linear algebra, differential equations, numerical integration, optimization, statistics and more

- It adds significant power to the interactive Python session by providing the user with high-level commands and classes for manipulating and visualizing data

- Routines for computing integrals numerically, solving differential equations, optimization and sparse matrices

Link: https://www.scipy.org/scipylib/

Copyright © Infineon Technologies AG 2018. All rights reserved.

Pandas

- Adds data structures and tools designed to work with table-like data (similar to Series and Data Frames in R)
- Provides tools for data manipulation: reshaping, merging, sorting, slicing, aggregation etc.
- Allows handling missing data

	rank	discipline	phd	service	sex	salary
0	Prof	B	56	49	Male	186960
1	Prof	A	12	6	Male	93000
2	Prof	A	23	20	Male	110515
3	Prof	A	40	31	Male	131205
4	Prof	B	20	18	Male	104800

Link: http://pandas.pydata.org/

Copyright © Infineon Technologies AG 2018. All rights reserved.

SciKit-Learn

- Provides machine learning algorithms: classification, regression, clustering, model validation etc.

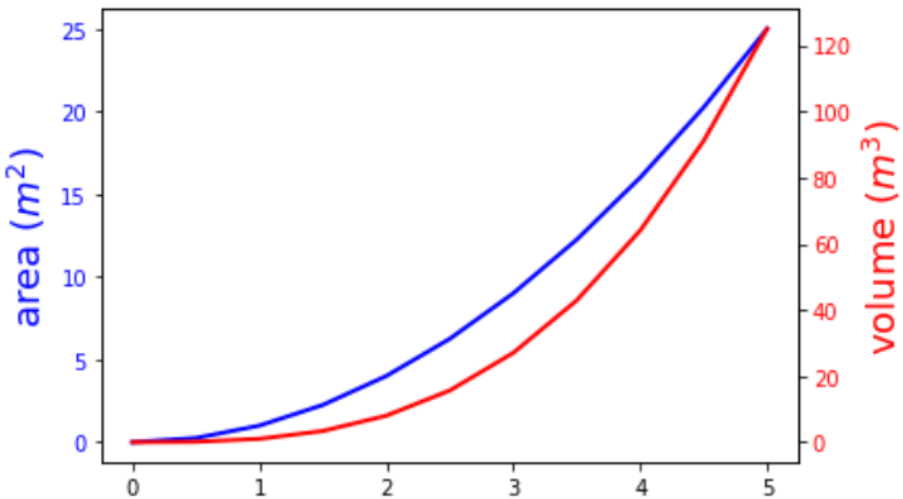
- Built on NumPy, SciPy and Matplotlib

Link: http://scikit-learn.org/

Copyright © Infineon Technologies AG 2018. All rights reserved.

Matplotlib

- Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats
- A set of functionalities similar to those of MATLAB
- Line plots, scatter plots, barcharts, histograms, pie charts etc.
- Relatively low-level; some effort needed to create advanced visualization

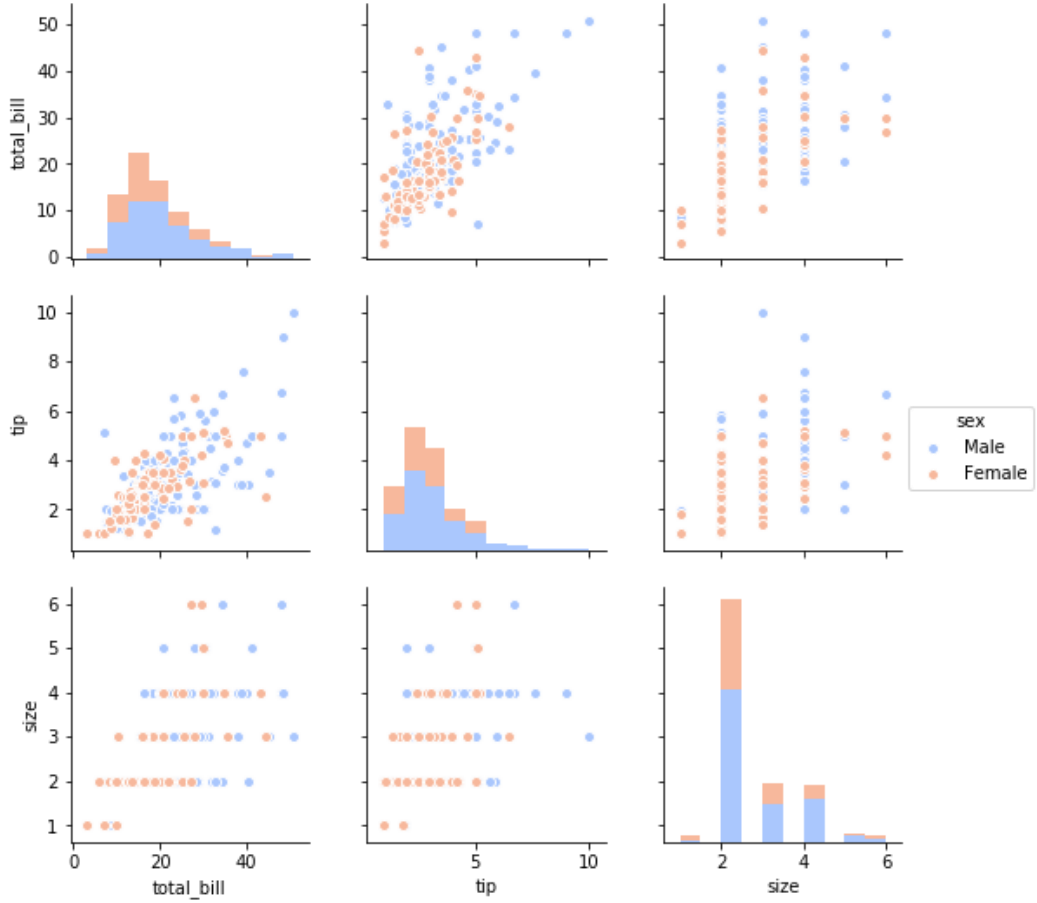


Link: https://matplotlib.org/

Copyright © Infineon Technologies AG 2018. All rights reserved.

Seaborn

- Based on Matplotlib
- Provides high level interface for drawing attractive statistical graphics

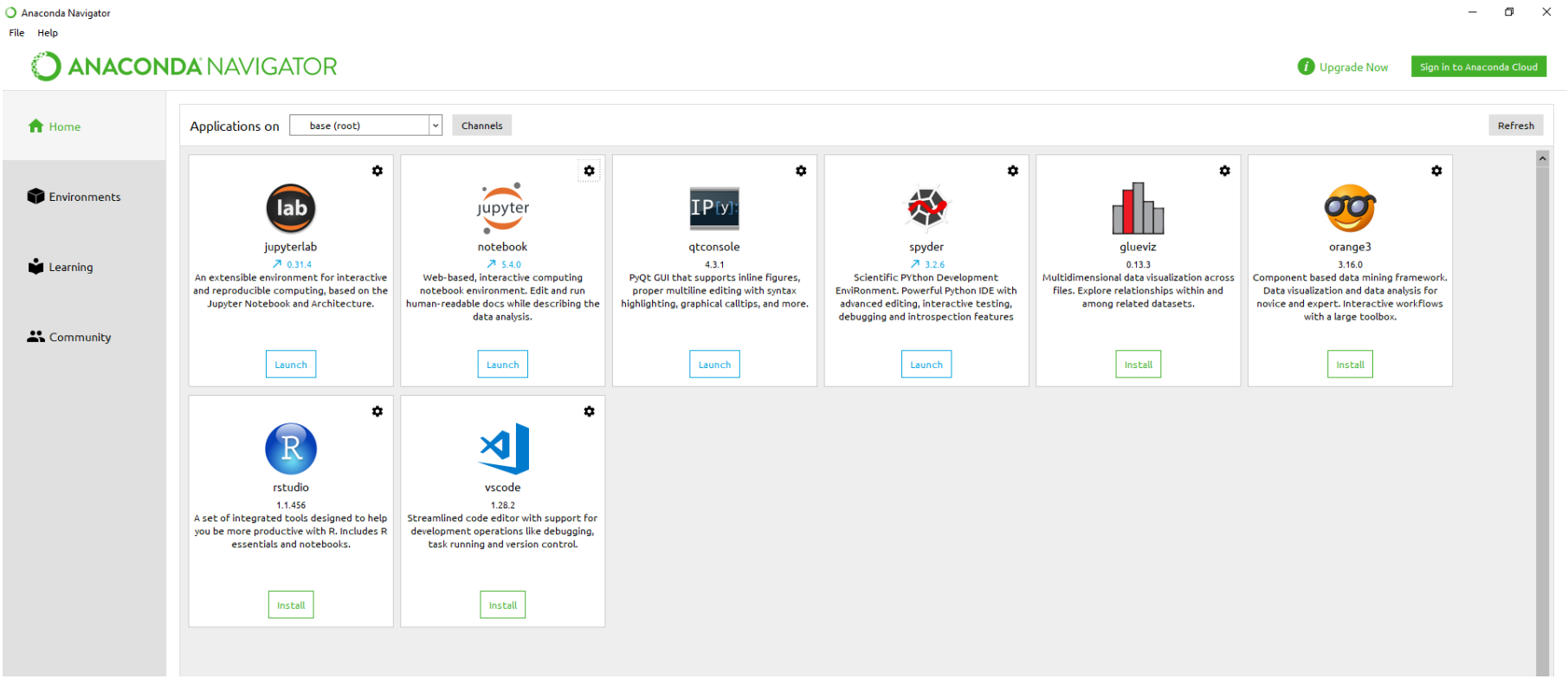


Link: https://seaborn.pydata.org/

Copyright © Infineon Technologies AG 2018. All rights reserved.

Anaconda

› Python Data Science Platform



To install Anaconda in Ubuntu: https://www.digitalocean.com/community/tutorials/how-to-install-anaconda-on-ubuntu-18-04-quickstart

Copyright © Infineon Technologies AG 2018. All rights reserved.

Jupyter Notebook

Jupyter Notebook is an open-source web application to create documents that contain:

- Data cleaning and transformation
- Numerical simulations
- Statistical modeling
- Data visualization
- Machine Learning

Link: https://jupyter.org/

Copyright © Infineon Technologies AG 2018. All rights reserved.

How it looks like

Copyright © Infineon Technologies AG 2018. All rights reserved.

How it looks like

jupyter NumPy Last Checkpoint: a few seconds ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

Using NumPy

Once you've installed NumPy you can import it as a library:

In [1]:
```
import numpy as np
```

Numpy has many built-in functions and capabilities. We won't cover them all but instead we will focus on some of the most important aspects of Numpy: vectors,arrays,matrices, and number generation. Let's start by discussing arrays.

Numpy Arrays

NumPy arrays are the main way we will use Numpy throughout the course. Numpy arrays essentially come in two flavors: vectors and matrices. Vectors are strictly 1-d arrays and matrices are 2-d (but you should note a matrix can still have only one row or one column).

Let's begin our introduction by exploring how to create NumPy arrays.

Creating NumPy Arrays

From a Python List

We can create an array by directly converting a list or list of lists:

In [2]:
```
my_list = [1,2,3]
my_list
```

Out[2]: [1, 2, 3]

Copyright © Infineon Technologies AG 2018. All rights reserved.

Part of your life. Part of tomorrow.