Python for Analysts

Laura Richter, Aruba
PyConZA 2019

Why Python

Plan for today: Topics

- Jupyter notebook
- Python quick start
- Data modules:
  - Numpy
  - Matplotlib
  - Pandas
  - Sklearn

Plan for today: Sessions

- Session 0: Intro
- Session 1: Jupyter notebook & Python quick start (~1.5 hour)
- Session 2: Numpy (~2 hour)
- Session 3: Matplotlib (~1 hour)
- Session 4: Pandas (~2 hour)
- Session 5: Sklearn (~1 hour)
- Session 6: Round up (~0.5 hour)

Session 1: Python

- General purpose
- High level
- Object orientated
- Dynamically typed
- Large package ecosystem

Session 1: Jupyter notebooks

- What are Jupyter notebooks? [demo]
- What is Jupyter lab? [demo]
- Local or cloud? [demo]
- The kernel

Session 1: Python Built-in types

- Basic [demo]
  - Integer
  - Float
  - Complex
  - Boolean
  - String

- Collections [demo]
  - List
  - Dictionary

Session 1: Python Syntax

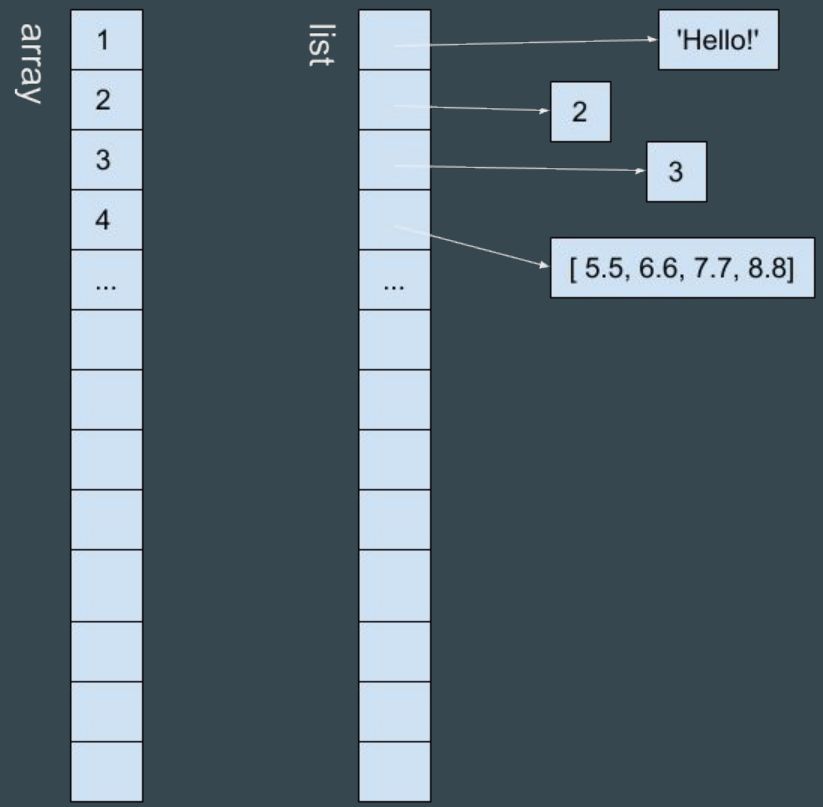
- Flow control [demo]
  - for
  - if
- Functions [demo]
- Comments and Docstrings [demo]
- Accessing data from files [demo]

Session 1: Python packages & PyPI

- What is a Python package?
- Python Package Index
- Install packages using pip [demo]
  - Versions
  - pip vs pip3
  - pip in a notebook

Session 2: NumPy

- Numpy Arrays



Session 2: NumPy

- Vectorisation [demo]
- Multi-dimensional arrays
- Boolean indexing
- Broadcasting
- Slicing & Views

Session 3: Matplotlib

- Python plotting packages [demo]
  - Bokeh
  - Plot.ly
  - Seaborn
  - Matplotlib

Session 3: Matplotlib

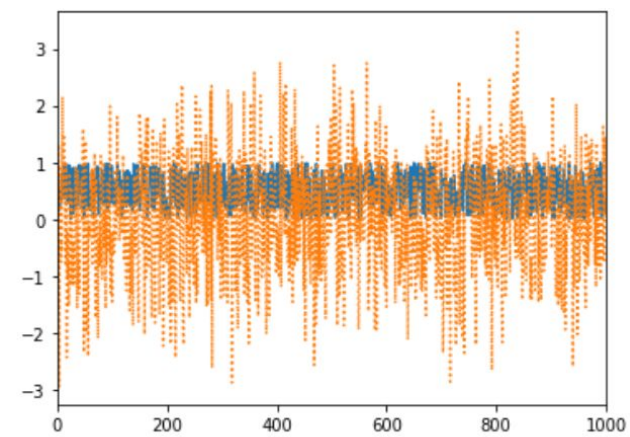
- Imperative versus Object orientated

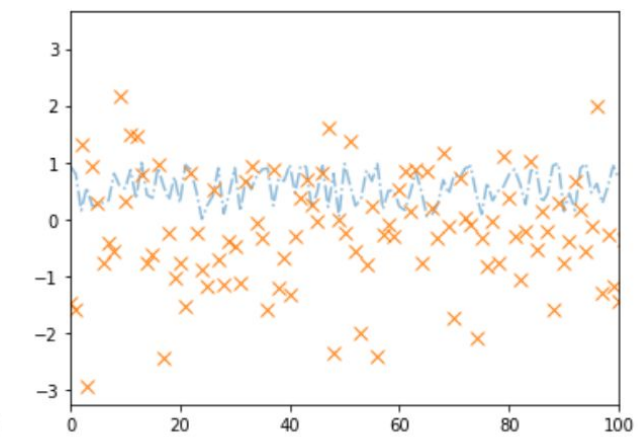
```
from matplotlib import *

plot(x, y)
xlabel('X')
ylabel('Y')
```

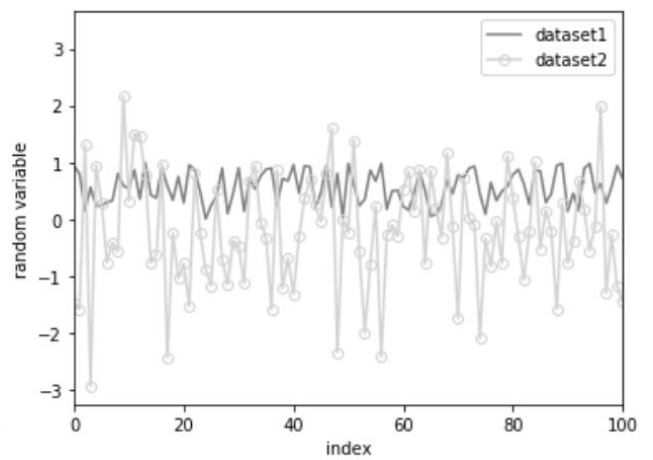
```
import matplotlib.pyplot as plt

fig, ax = plt.subplots(1, 1)
ax.plot(x, y)
ax.set_xlabel('X')
ax.set_ylabel('Y')
```

Session 3: Matplotlib

- Object orientated aproach







Session 3: Matplotlib

Plot Types

- Lines
- Scatter plot
- Histograms
- Bar charts (vertical, horizontal)
- Filled areas
- Lines
- Box plots
- ...

Plot styling

- Colour
- Line styles and width
- Marker styles and sizes
- Alpha (transparency)
- Axis labels
- Titles
- Grids
- ...

Session 4: Pandas

- Built on NumPy arrays

Session 4: Pandas

- DataFrame [demo]

	Date	Project id	Project	Project lead	Location	Venue	Marketing	Photography	Catering	Decor	Printing	Other
0	25/03/2019	2019-01	Cryptocurrency trade show	Sarah Ferris	Cape Town	63466	29367	6944	1487	35660	624	2600
1	28/03/2019	2019-02	Mills wedding	Blake van Rensburg	Somerset West	44114	0	7264	19558	7532	2993	1106
2	14/04/2019	2019-03	Bradley graduation	Lomile Moreki	Stellenbosch	65354	13050	1207	11327	7980	1879	0
3	21/04/2019	2019-04	Kitchen tea at Kirstenbosch	Craig Russell	Cape Town	33362	0	2327	7224	25823	5641	0
4	01/05/2019	2019-05	Atmosphere conference	Blake van Rensburg	Stellenbosch	64979	30310	8509	1329	31663	773	30550
5	01/05/2019	2019-06	Green construction indaba	Khanyisa Matyolo	Cape Town	14831	29097	4888	17985	10094	11791	34010
6	08/05/2019	2019-07	Greyville horse races	Blake van Rensburg	Cape Town	68757	35682	9177	17305	7091	0	13568
7	30/05/2019	2019-08	Greg and Hayley wedding	Sarah Ferris	Somerset West	12050	0	1082	18983	2101	3727	0
8	03/06/2019	2019-09	Adams family braai	Sarah Ferris	Stellenbosch	40070	0	7392	0	16492	10274	0
9	04/06/2019	2019-10	Soil Indaba	Hannah Grey	Cape Town	43039	35882	6789	4775	27432	2875	31944
10	20/06/2019	2019-11	Matyolo wedding	Lomile Moreki	Robertson	33033	0	3500	17372	26651	9241	38934
11	28/06/2019	2019-12	SACOF expo	Lomile Moreki	Stellenbosch	76129	0	8270	19318	6752	4237	2788

Session 4: Pandas

- Indexing [demo]
  - Indexing columns
  - Indexing rows and columns: .loc
  - Indexing into underlying data directly: .iloc
- Working with columns
  - Vectorised operations
- DataFrame methods and attributes

Session 4: Pandas

- Data imports and Exports
  - csv
  - Excel
  - JSON & JSONL
  - Google bigquery [demo]
  - ... many others

Session 4: Pandas

- Element wise operations: .apply
- Boolean indexing and filtering
- NaN / empty values
- Columns of different types:
  - Numeric
  - String

Session 4: Pandas

- Aggregations and pivots
- Merging
- Timeseries

Session 5: Scikit-Learn

- Scipy toolkit for Machine Learning
- Functionality:
  - Preprocessing
  - Regression
  - Classification
  - Clustering
  - Pipelining
  - Evaluation
  - Dimensionality reduction
  - Model selection
  - .....

Session 5: Scikit-Learn

- Sklearn patterns
- Regression
- Clustering

Session 6: Round up

Thank you