# Ashoka Summer School

Data Visualization

# Outline

Pandas library

Data structures

Visualization

Practical works

# Data Visualization - Brief

Pandas - **Pan**el **Da**ta **S**ystem

Used in production in many companies, especially in financial industries

Suited for many different kinds of data

Two primary data structures:

- Series (1 dimensional)
- DataFrame (2 dimensional). For R’s users, it’s like R’s data.frame on steroids.

# Adopt Python - Data Structures

- List → [1, 2, 3, 4, 5, “hello” ] : Ordered series of values
    - add data list.append(1)
- Dictionnary → { “key” : “value”, “hello” : 1 } : Key/Value data structure
    - add data dict[‘key’] = ‘value’
- Tuple → (1, 2, 3, 4, “hello”) : Like list but immutable

# Data Visualization - Series

1 dimensional

```
from pandas import Series
data = {'a' : 0., 'b' : 1., 'c' : 2.}
s = Series(data)
print(s)
a     0.0
b     1.0
c     2.0
```

Import pandas library

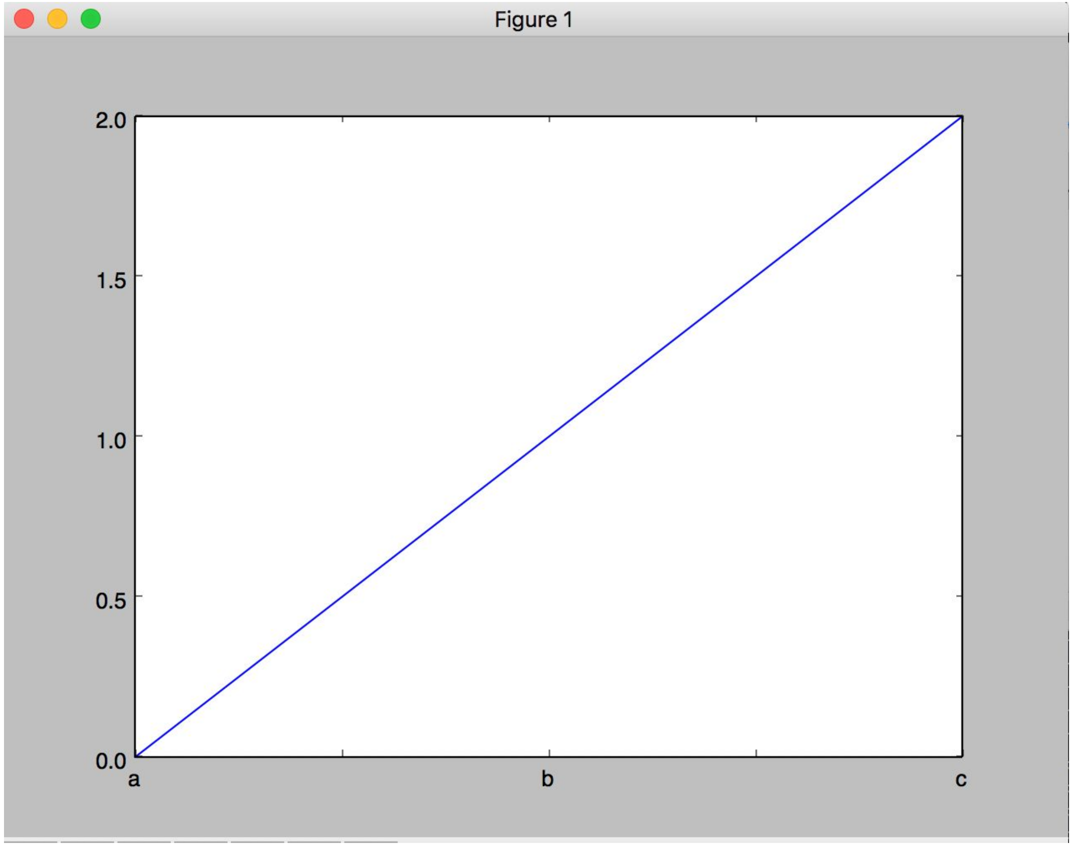
Create python ordered dictionary with data

Instantiate Series object

Show variable content

# Data Visualization - Series

```
import matplotlib.pyplot as plt
s.plot()
plt.show()
```

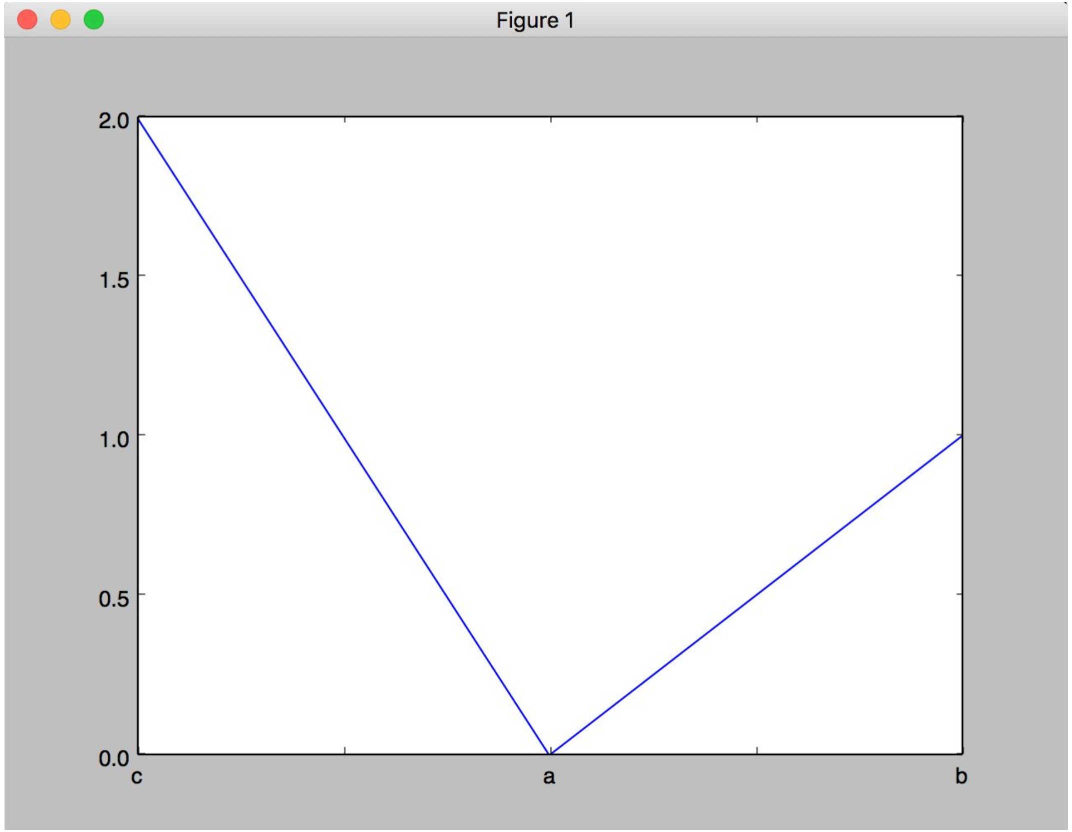


# Data Visualization - Series

```
s = s.reindex(['c','a','b'])

print(s)
c     2.0
a     0.0
b     1.0


s.plot()
plt.show()
```



# Data Visualization - DataFrame

2 dimensional table data structure

Like R’s data.frame

Data manipulation with integrated indexing

Support heterogeneous type of columns

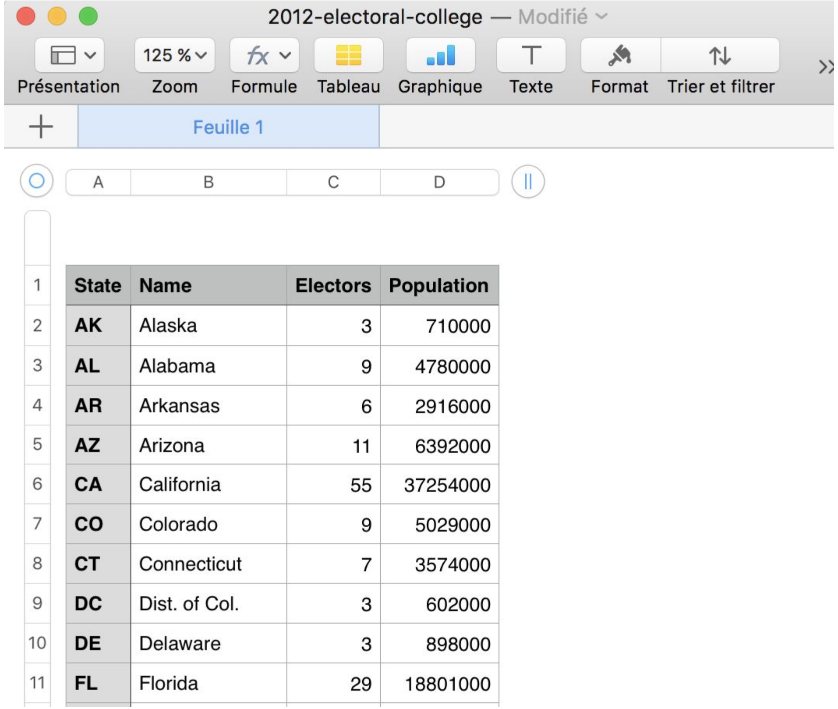
# Data Visualization - DataFrame

## File input/output

```
import pandas as pd

data = pd.read_csv('2012-electoral-college.csv',
sep=';', index_col='State')

data.head()
            Name  Electors    Population
State
AK          Alaska      3           710000
AL    Alabama     9           4780000
AR          Arkansas    6           2916000
AZ    Arizona     11          6392000
CA          California  55          37254000
```



| State | Name | Electors | Population |
|---|---|---|---|
| AK | Alaska | 3 | 710000 |
| AL | Alabama | 9 | 4780000 |
| AR | Arkansas | 6 | 2916000 |
| AZ | Arizona | 11 | 6392000 |
| CA | California | 55 | 37254000 |
| CO | Colorado | 9 | 5029000 |
| CT | Connecticut | 7 | 3574000 |
| DC | Dist. of Col. | 3 | 602000 |
| DE | Delaware | 3 | 898000 |
| FL | Florida | 29 | 18801000 |

# Data Visualization - DataFrame

## Calculation and statistics

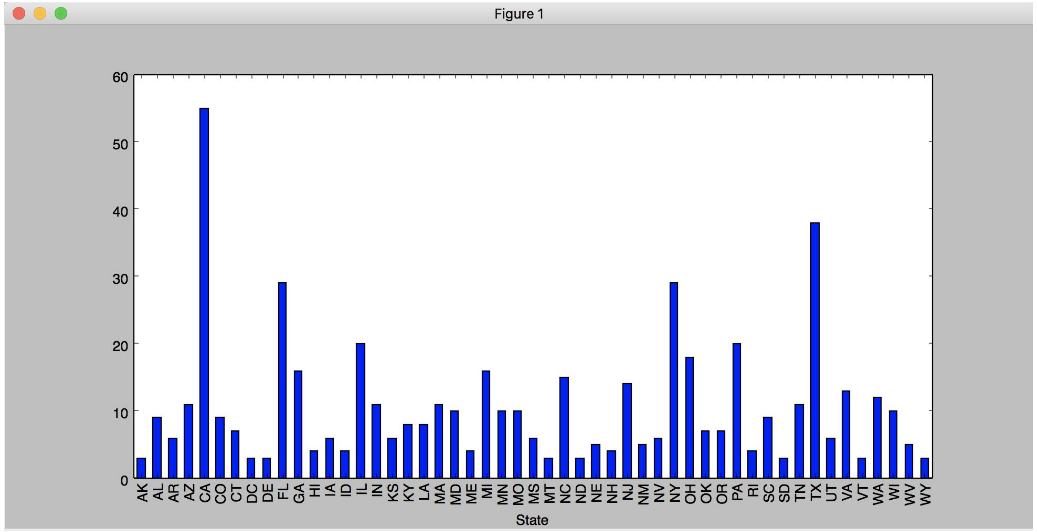
```
>>> data.Electors.mean()
10.549019607843137
>>> data.Electors.max()
55
>>> data.loc[data.Electors.argmax(), 'Name']
'California'
>>> data.Population.sum()
308746000
>>> data['ratio'] = data['Electors']/data['Population']
>>> data
                  Name  Electors    Population        ratio
State
AK                Alaska      3           710000            0.000004
AL                Alabama     9           4780000           0.000002
AR                Arkansas    6           2916000           0.000002
AZ                Arizona     11          6392000           0.000002
[...]
```

# Data Visualization - DataFrame

## Visualization with matplotlib

```
import matplotlib.pyplot() as plt

data.Electors.plot.bar()
plt.show()
```



# Data Visualization - Go further

And much more…

- Group By
- Merge, join, aggregation
- Reshaping and Pivot Tables
- Time based series, date functions
- Multi-index
- ...

# Let’s play !

https://ashoka.cdsp.sciences-po.fr