

Πρώτο Εργαστήριο - MEM205 Περιγραφική Στατιστική

1. Jupyter Notebook

Το Jupyter Notebook είναι μια δικτυακή εφαρμογή ανοικτού κώδικα που μας επιτρέπει να δημιουργούμε και να μοιράζομαστε έγγραφα που περιέχουν κώδικα, μαθηματικές εξισώσεις, γραφήματα και κείμενα.

1.1. Εκκίνηση του jupyter notebook

Στο εργαστήριο Γ109

1. Κάνουμε login ως Guest.
2. Εκκινούμε ένα terminal.
3. Πληκτρολογούμε `/opt/pyenv-3.7.5/bin/jupyter notebook` και πατάμε enter.
4. Θα εμφανισθεί στο terminal ένα url.
 - για παράδειγμα <http://localhost:8888/?token=7f98aa75c0cdd2a229d5d71484f6f8a0132e87309440f54a>
(<http://localhost:8888/?token=7f98aa75c0cdd2a229d5d71484f6f8a0132e87309440f54a>)
5. Κράταμε πατημένο το **Ctrl** και κάνουμε κλικ στο url, τότε θα ξεκινήσει ο **firefox** με ένα tab με το jupyter notebook.
6. Το terminal πρέπει να παραμείνει σε λειτουργία ωστέ να μπορούμε να δουλέψουμε με το jupyter notebook.

1.2. Γνωριμία με το jupyter notebook

Jupyter Modes

- Command Mode
- Edit Mode
 - Markdown
 - Code

Hello Jupyter Notebook - Markdown

$$\sum_{i=1}^N x_i$$

Hello Jupyter Notebook - Python

1. Γράφουμε κωδικά **Python** (μία ή περισσότερες γραμμές).
2. Εκτελούμε τον κώδικα με **Shift + Enter**.

```
In [1]: print("Hello, Jupyter Notebook!")
```

```
Hello, Jupyter Notebook!
```

```
In [2]: "Hello, Jupyter Notebook!"
```

```
Out[2]: 'Hello, Jupyter Notebook!'
```

2. Εισαγωγή στο Pandas

```
In [3]: import pandas as pd
```

2.1. DataFrame

```
In [4]: data_1 = \
{
  'name': ['Nick', 'Helen', 'Steve', 'Mary', 'George'],
  'age': [20, 19, 22, 21, 20],
  'field_of_study': ['Mathematics', 'Physics', 'Literature', 'Mathematics', 'Chemistry'],
  'os': ['Windows', 'Windows', 'Windows', 'Linux', 'MacOS']
}
```

```
In [5]: dataset_1 = pd.DataFrame(data_1)
```

```
In [6]: dataset_1
```

```
Out[6]:
```

	name	age	field_of_study	os
0	Nick	20	Mathematics	Windows
1	Helen	19	Physics	Windows
2	Steve	22	Literature	Windows
3	Mary	21	Mathematics	Linux
4	George	20	Chemistry	MacOS

```
In [7]: dataset_1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 4 columns):
name                5 non-null object
age                 5 non-null int64
field_of_study      5 non-null object
os                  5 non-null object
dtypes: int64(1), object(3)
memory usage: 288.0+ bytes
```

```
In [8]: data_2 = \
[['Nick', 20, 'Mathematics', 'Windows'],
 ['Helen', 19, 'Physics', 'Windows'],
 ['Steve', 22, 'Literature', 'Windows'],
 ['Mary', 21, 'Mathematics', 'Linux'],
 ['George', 20, 'Chemistry', 'MacOS']]
```

```
In [9]: dataset_2a = pd.DataFrame(data_2)
```

```
In [10]: dataset_2a
```

```
Out[10]:
```

	0	1	2	3
0	Nick	20	Mathematics	Windows
1	Helen	19	Physics	Windows
2	Steve	22	Literature	Windows
3	Mary	21	Mathematics	Linux
4	George	20	Chemistry	MacOS

```
In [11]: var_names = ["name", 'age', 'field_of_study', 'os']
```

```
In [12]: dataset_2b = pd.DataFrame(data_2, columns=var_names)
```

```
In [13]: dataset_2b
```

```
Out[13]:
```

	name	age	field_of_study	os
0	Nick	20	Mathematics	Windows
1	Helen	19	Physics	Windows
2	Steve	22	Literature	Windows
3	Mary	21	Mathematics	Linux
4	George	20	Chemistry	MacOS

DataFrame και αρχείο .csv

1. Δημιουργούμε το αρχείο students.csv με έναν editor.
2. Γράφουμε το έξηξ περιεχόμενο και κάνουμε save.

```
name,age,field_of_stydy,os  
Nick,20,Mathematics,Windows  
Helen,19,Physics,Windows  
Steve,22,Literature,Windows  
Mary,21,Mathematics,Linux  
George,20,Chemistry,MacOS
```

```
In [14]: dataset_3a = pd.read_csv('./students.csv')
```

```
In [15]: dataset_3a
```

```
Out[15]:
```

	name	age	field_of_stydy	os
0	Nick	20	Mathematics	Windows
1	Helen	19	Physics	Windows
2	Steve	22	Literature	Windows
3	Mary	21	Mathematics	Linux
4	George	20	Chemistry	MacOS

```
In [16]: dataset_3a.index.values
```

```
Out[16]: array([0, 1, 2, 3, 4])
```

```
In [17]: dataset_3a.loc[0]
```

```
Out[17]: name                Nick
age                20
field_of_stydy    Mathematics
os                 Windows
Name: 0, dtype: object
```

```
In [18]: dataset_3a.iloc[0]
```

```
Out[18]: name                Nick
age                20
field_of_stydy    Mathematics
os                 Windows
Name: 0, dtype: object
```

```
In [19]: dataset_3b = dataset_3a.set_index('name')
```

```
In [23]: dataset_3b
```

```
Out[23]:
```

	age	field_of_stydy	os
name			
Nick	20	Mathematics	Windows
Helen	19	Physics	Windows
Steve	22	Literature	Windows
Mary	21	Mathematics	Linux
George	20	Chemistry	MacOS

```
In [22]: dataset_3b.index.values
```

```
Out[22]: array(['Nick', 'Helen', 'Steve', 'Mary', 'George'], dtype=object)
```

```
In [24]: dataset_3b.loc['Nick']
```

```
Out[24]: age                20  
field_of_stydy    Mathematics  
os                Windows  
Name: Nick, dtype: object
```

```
In [ ]: dataset_3b.iloc[0]
```

```
In [25]: dataset_3c = pd.read_csv('./students.csv', index_col='name')
```

```
In [26]: dataset_3c
```

```
Out[26]:
```

	age	field_of_stydy	os
name			
Nick	20	Mathematics	Windows
Helen	19	Physics	Windows
Steve	22	Literature	Windows
Mary	21	Mathematics	Linux
George	20	Chemistry	MacOS

```
In [28]: dataset_3d = dataset_3c.drop('os', axis=1)
```

```
In [30]: dataset_3d
```

```
Out[30]:
```

	age	field_of_stydy
name		
Nick	20	Mathematics
Helen	19	Physics
Steve	22	Literature
Mary	21	Mathematics
George	20	Chemistry

```
In [31]: dataset_3e = dataset_3c.drop('Nick')
```

```
In [33]: dataset_3e
```

```
Out[33]:
```

	age	field_of_stydy	os
name			
Helen	19	Physics	Windows
Steve	22	Literature	Windows
Mary	21	Mathematics	Linux
George	20	Chemistry	MacOS