

[Triển khai Python cho Data Mining]

Cài đặt Python trên windows
cùng một số thư viện thường dùng và thử nghiệm
với một Project đơn giản

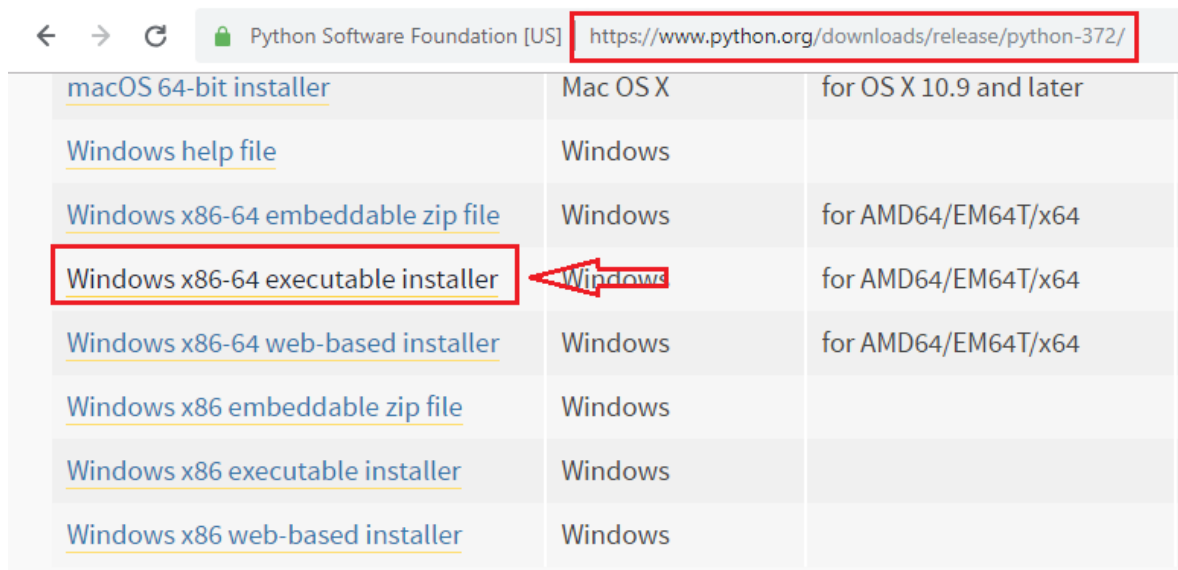
Mục lục

Phần I: Hướng dẫn cài đặt Python.....	3
1.Download Python 3.7.2	3
2.Tiến hành cài đặt Python 3.7.2 trên Windows.....	3
Phần II: Hướng dẫn cài đặt một số thư viện thường dùng	6
1. Hướng dẫn cài đặt một số thư viện cho python	6
2. Hướng dẫn kiểm tra thư viện trong python.....	7
PHẦN III: Cài đặt PyCharm IDE cho Python trên windows	9
PHẦN IV: Thử nghiệm với một Project đơn giản trong Python.....	13

Phần I: Hướng dẫn cài đặt Python

1. Download Python 3.7.2

Chọn phiên bản phù hợp với máy tính của bạn và download Python tại trang chủ python.org hoặc một link được chia sẻ nào đó.


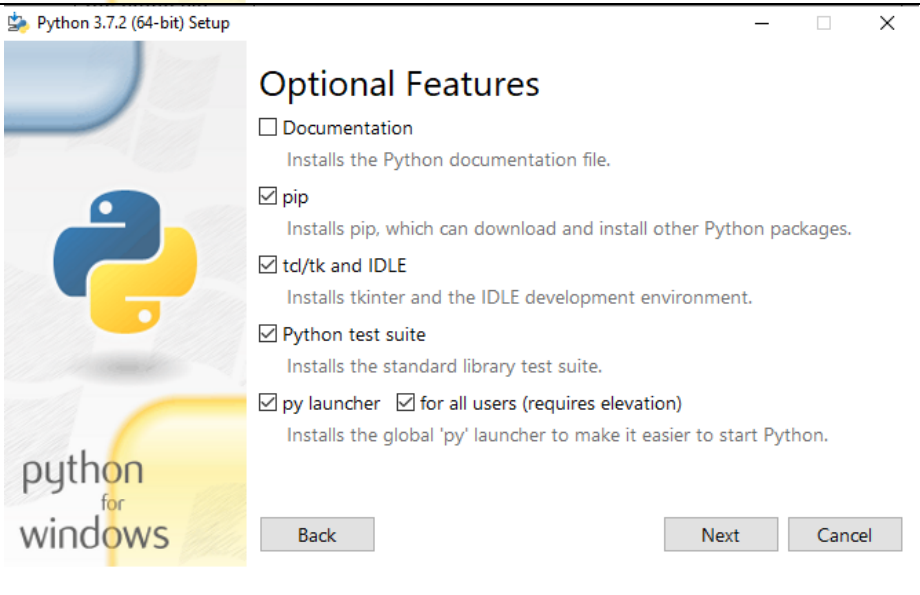
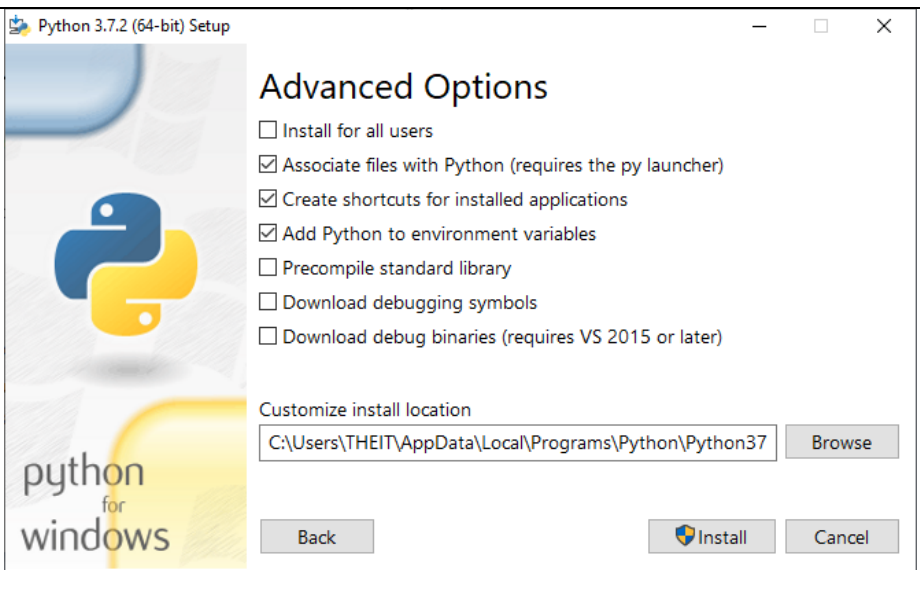


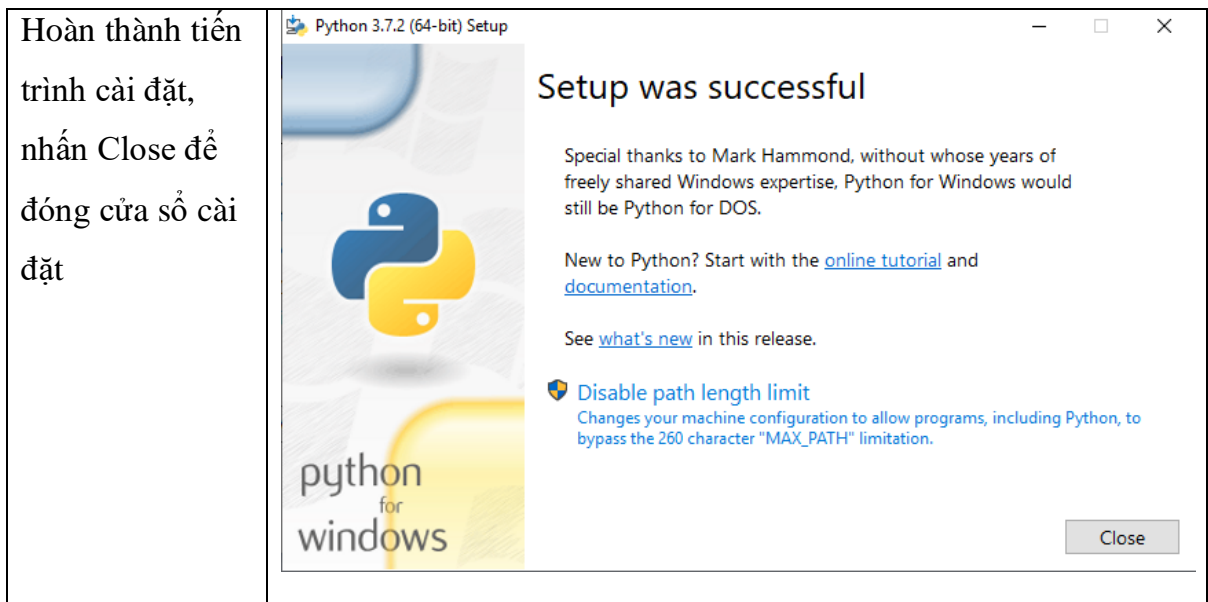
macOS 64-bit installer	Mac OS X	for OS X 10.9 and later
Windows help file	Windows	
Windows x86-64 embeddable zip file	Windows	for AMD64/EM64T/x64
Windows x86-64 executable installer	Windows	for AMD64/EM64T/x64
Windows x86-64 web-based installer	Windows	for AMD64/EM64T/x64
Windows x86 embeddable zip file	Windows	
Windows x86 executable installer	Windows	
Windows x86 web-based installer	Windows	

2. Tiến hành cài đặt Python 3.7.2 trên Windows

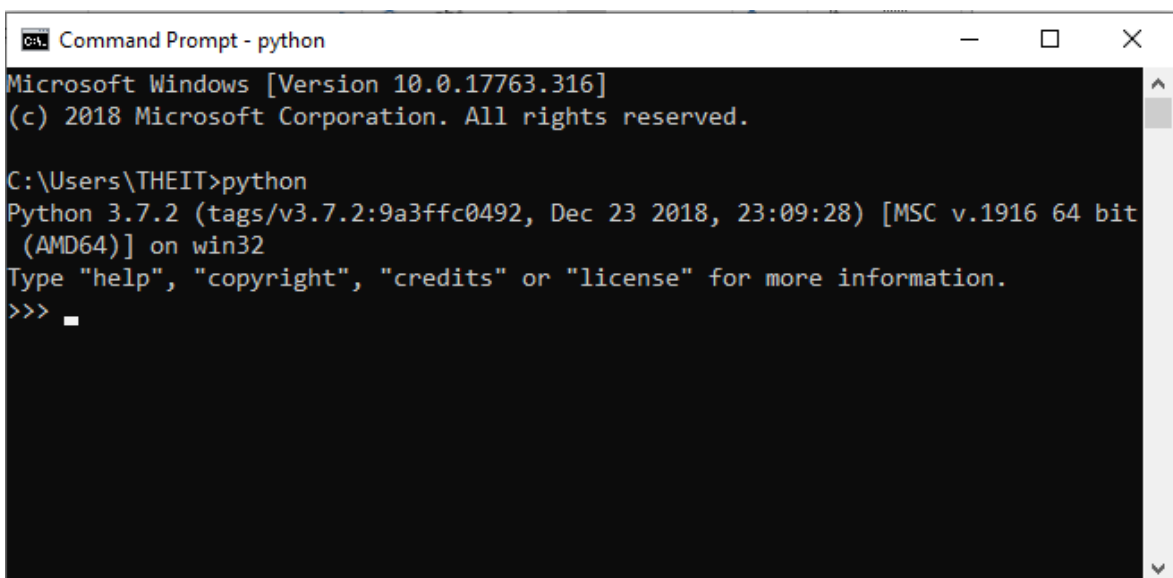
Chạy file setup sau khi bạn đã download về. Tích vào Add Python 3.7 to PATH để thêm Python vào biến môi trường



<p>Click chọn Install Now để thực hiện cài đặt ngay, Click chọn Customize installation để tùy chỉnh cài đặt.</p>	
<p>Sau khi chọn Customize installation, cửa sổ Optional Features sẽ hiện ra, tích chọn 1 số tính năng cần thiết và nhấn Next</p>	
<p>Để tùy chọn như mặc định, chọn Browse để thay đổi đường Path cài đặt nếu bạn muốn, tiếp tục Click vào Install để tiến hành cài đặt Python 3.7.2</p>	



- Tại cửa sổ command line, gõ “python” để kiểm tra Python đã được cài đặt thành công chưa.



The screenshot shows a 'Command Prompt - python' window. The title bar reads 'Command Prompt - python'. The window content shows the following text: 'Microsoft Windows [Version 10.0.17763.316] (c) 2018 Microsoft Corporation. All rights reserved. C:\Users\THEIT>python Python 3.7.2 (tags/v3.7.2:9a3ffc0492, Dec 23 2018, 23:09:28) [MSC v.1916 64 bit (AMD64)] on win32 Type "help", "copyright", "credits" or "license" for more information. >>>'. The prompt '>>>' is followed by a cursor.

Phần II: Hướng dẫn cài đặt một số thư viện thường dùng

1. Hướng dẫn cài đặt một số thư viện cho python

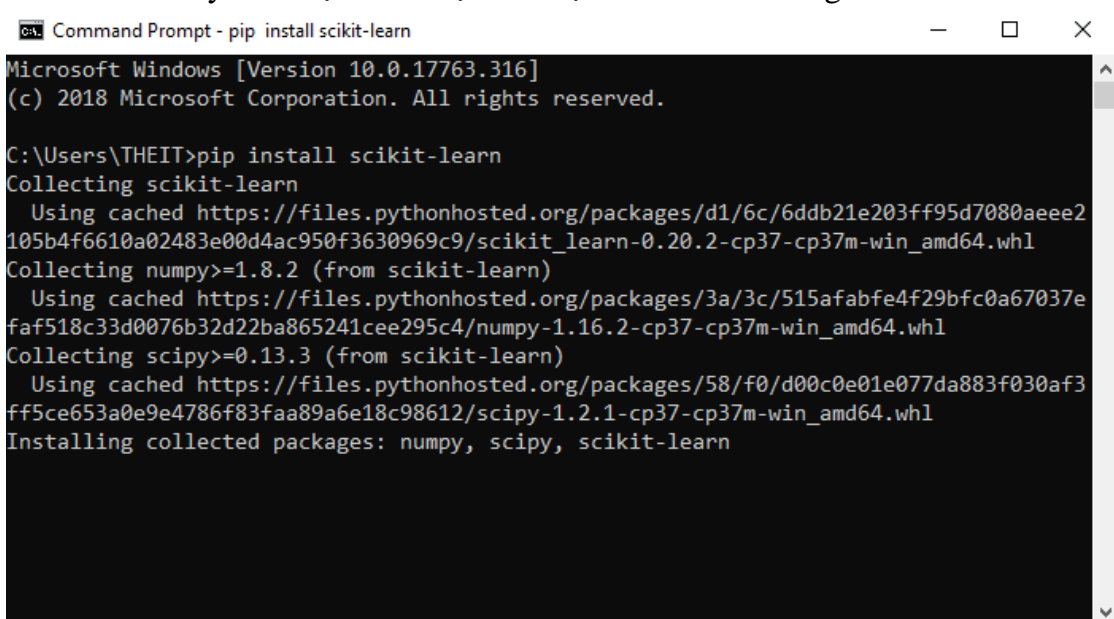
MỘT SỐ THƯ VIỆN PYTHON 3 VÀ CÁCH CÀI ĐẶT

Cú pháp cài đặt chung: `pip install <Tên thư viện>`

Gỡ bỏ cài đặt: `pip uninstall <Tên thư viện>`

Libraries	Description	Installation
scikit-learn	Working with classical ML algorithms	<code>pip install scikit-learn</code>
tensorflow	Deep Learning	<code>pip install tensorflow</code>
theano	Deep Learning	<code>pip install theano</code>
pandas	Data structures & analysis	<code>pip install pandas</code>
matplotlib	Data visualization	<code>pip install matplotlib</code>
seaborn	Data visualization	<code>pip install seaborn</code>
numpy	Scientific computing	<code>pip install numpy</code>
NLTK	Natural Langue Processing	<code>pip install nltk</code>
scipy	scientific computing	<code>pip install scipy</code>

- Dưới đây là ví dụ về cài đặt thư viện scikit-learn bằng cmd



```
Command Prompt - pip install scikit-learn
Microsoft Windows [Version 10.0.17763.316]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\THEIT>pip install scikit-learn
Collecting scikit-learn
  Using cached https://files.pythonhosted.org/packages/d1/6c/6ddb21e203ff95d7080ae2105b4f6610a02483e00d4ac950f3630969c9/scikit_learn-0.20.2-cp37-cp37m-win_amd64.whl
Collecting numpy>=1.8.2 (from scikit-learn)
  Using cached https://files.pythonhosted.org/packages/3a/3c/515afabfe4f29bfc0a67037efaf518c33d0076b32d22ba865241cee295c4/numpy-1.16.2-cp37-cp37m-win_amd64.whl
Collecting scipy>=0.13.3 (from scikit-learn)
  Using cached https://files.pythonhosted.org/packages/58/f0/d00c0e01e077da883f030af3ff5ce653a0e9e4786f83faa89a6e18c98612/scipy-1.2.1-cp37-cp37m-win_amd64.whl
Installing collected packages: numpy, scipy, scikit-learn
```

- Scikit-learn yêu cầu:

- Python (≥ 2.7 or ≥ 3.4),
- NumPy ($\geq 1.8.2$),
- SciPy ($\geq 0.13.3$).

Hệ thống sẽ kiểm tra và tự động download và cài đặt các thư viện yêu cầu song song nếu các thư viện yêu cầu chưa được cài đặt

- Cài đặt thành công

```
Select Command Prompt
Microsoft Windows [Version 10.0.17763.316]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\THEIT>pip install scikit-learn
Collecting scikit-learn
  Using cached https://files.pythonhosted.org/packages/d1/6c/6ddb21e203ff95d7080aeeee2105b4f6610a02483e00d4ac950f3630969c9/scikit_learn-0.20.2-cp37-cp37m-win_amd64.whl
Collecting numpy>=1.8.2 (from scikit-learn)
  Using cached https://files.pythonhosted.org/packages/3a/3c/515afabfe4f29bfc0a67037efaf518c33d0076b32d22ba865241cee295c4/numpy-1.16.2-cp37-cp37m-win_amd64.whl
Collecting scipy>=0.13.3 (from scikit-learn)
  Using cached https://files.pythonhosted.org/packages/58/f0/d00c0e01e077da883f030af3ff5ce653a0e9e4786f83faa89a6e18c98612/scipy-1.2.1-cp37-cp37m-win_amd64.whl
Installing collected packages: numpy, scipy, scikit-learn
Successfully installed numpy-1.16.2 scikit-learn-0.20.2 scipy-1.2.1

C:\Users\THEIT>
```

- Các thư viện khác cũng tương tự, có thể cài đặt đồng thời các thư viện 1 lúc bằng dấu cách:

```
Select Command Prompt - pip install scikit-learn tensorflow theano pandas matplotlib seaborn numpy nltk
Microsoft Windows [Version 10.0.17763.316]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\THEIT>pip install scikit-learn tensorflow theano pandas matplotlib seaborn numpy nltk
Requirement already satisfied: scikit-learn in c:\users\theit\appdata\local\programs\python\python37\lib\site-packages (0.20.2)
Collecting tensorflow
  Using cached https://files.pythonhosted.org/packages/7b/14/e4538c2bc3ae9f4ce6f6ce7ef1180da05abc4a617afb798268232b01d0d/tensorflow-1.13.1-cp37-cp37m-win_amd64.whl
```

2. Hướng dẫn kiểm tra thư viện trong python

- Trong cửa sổ command line, gõ python -c "help('modules');" để kiểm tra tất cả các thư viện đã được cài đặt trong máy của bạn. Danh sách thư viện/modules sẽ hiện ra:

```
Command Prompt
Microsoft Windows [Version 10.0.17763.316]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\THEIT>python -c "help('modules');"

Please wait a moment while I gather a list of all available modules...

__future__      array      fileinput      random
__abc__          ast        fnmatch        re
__ast__          astroid    formatter      reprlib
__asyncio        asynchat   fractions      rlcompleter
__bisect         asyncio   ftplib         runpy
__blake2         asyncore   functools      sched
__bootlocale     atexit    gc             secrets
__bz2            audioop    genericpath    select
__codecs         base64     getopt         selectors
__codecs_cn      bdb        getpass        setuptools
__codecs_hk      binascii   gettext        shelve
__codecs_iso2022 binhex     glob           shlex
__codecs_jp      bisect     gzip           shutil
__codecs_kr      brain_argparse hashlib        signal
__codecs_tw      brain_attrs heapq          site
__collections    brain_builtin_inference hmac           six
__collections_abc brain_collections html          smtpd
```

- Gõ `python -c "help('<Tên thư viện>');"` để kiểm tra một thư viện nào đó đã được cài đặt hay chưa

```
Command Prompt - python -c "help('numpy');"
Microsoft Windows [Version 10.0.17763.316]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\THEIT>python -c "help('numpy');"
Help on package numpy:

NAME
    numpy

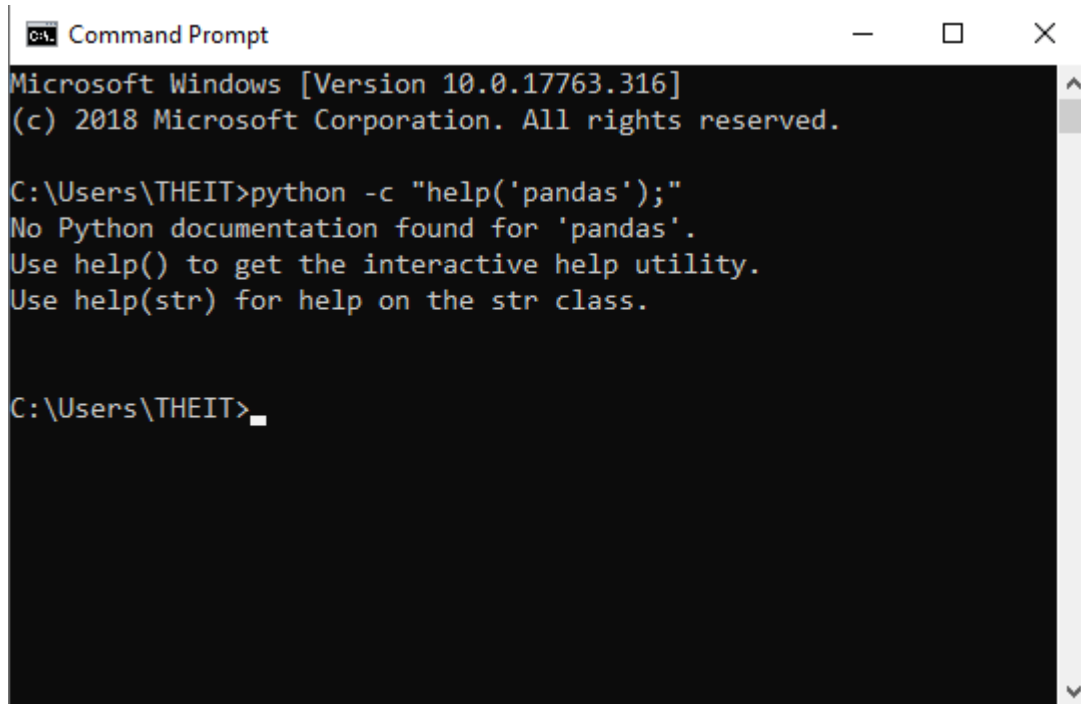
DESCRIPTION
    NumPy
    =====

    Provides
    1. An array object of arbitrary homogeneous items
    2. Fast mathematical operations over arrays
    3. Linear Algebra, Fourier Transforms, Random Number Generation

    How to use the documentation
    -----
    Documentation is available in two forms: docstrings provided
    with the code, and a loose standing reference guide, available from
    `the NumPy homepage <https://www.scipy.org>`_.

    We recommend exploring the docstrings using
```

- Thông báo “No Python documentation found” sẽ hiện ra nếu thư viện cần kiểm tra chưa được cài đặt



```
C:\> Command Prompt
Microsoft Windows [Version 10.0.17763.316]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\THEIT>python -c "help('pandas');"
```

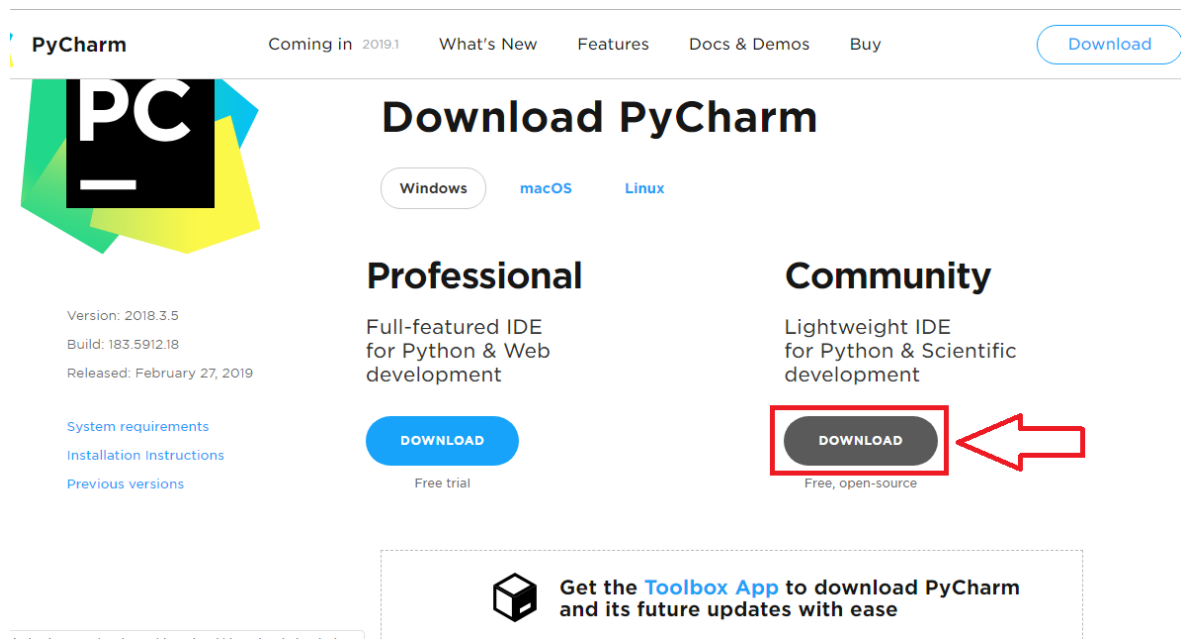
No Python documentation found for 'pandas'.
Use help() to get the interactive help utility.
Use help(str) for help on the str class.

```
C:\Users\THEIT>
```

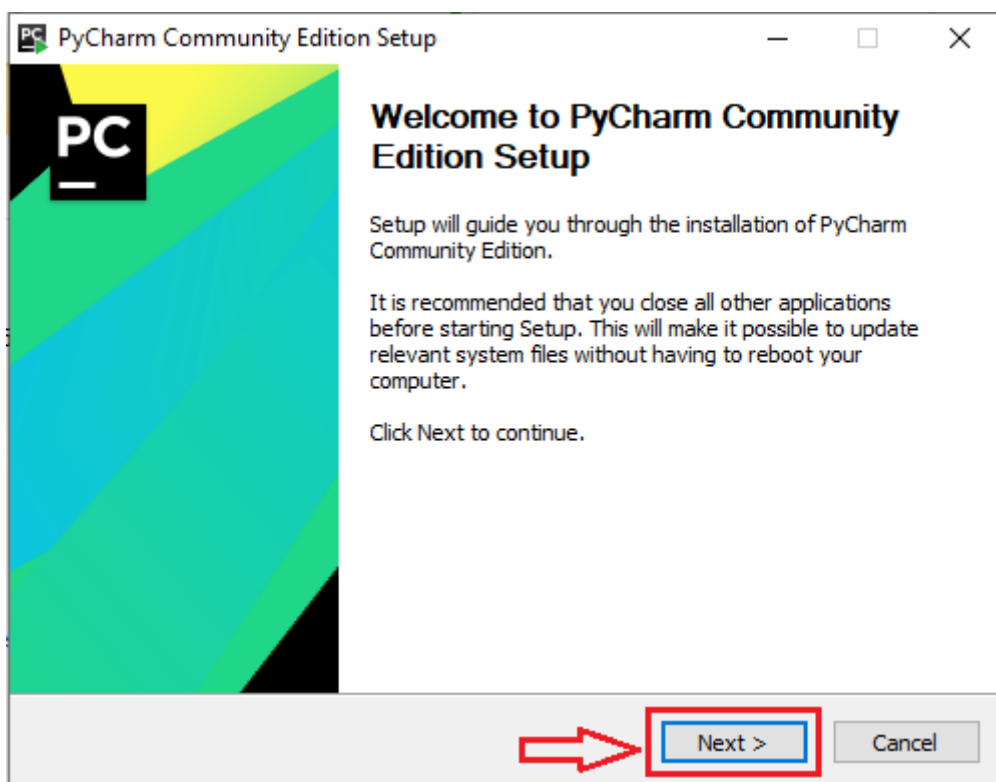
PHẦN III: Cài đặt PyCharm IDE cho Python trên windows

- Tải xuống PyCharm từ trang web. Chọn Phiên bản Community hoặc Phiên bản Professional theo lựa chọn.

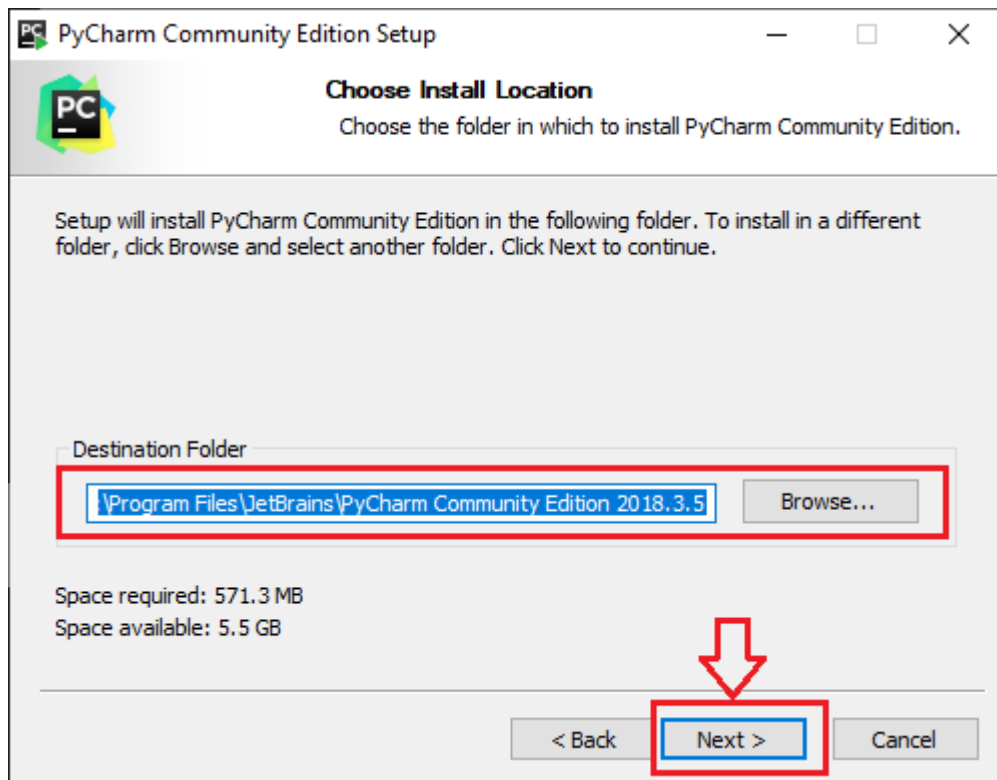
<https://www.jetbrains.com/pycharm/download/#section=windows>



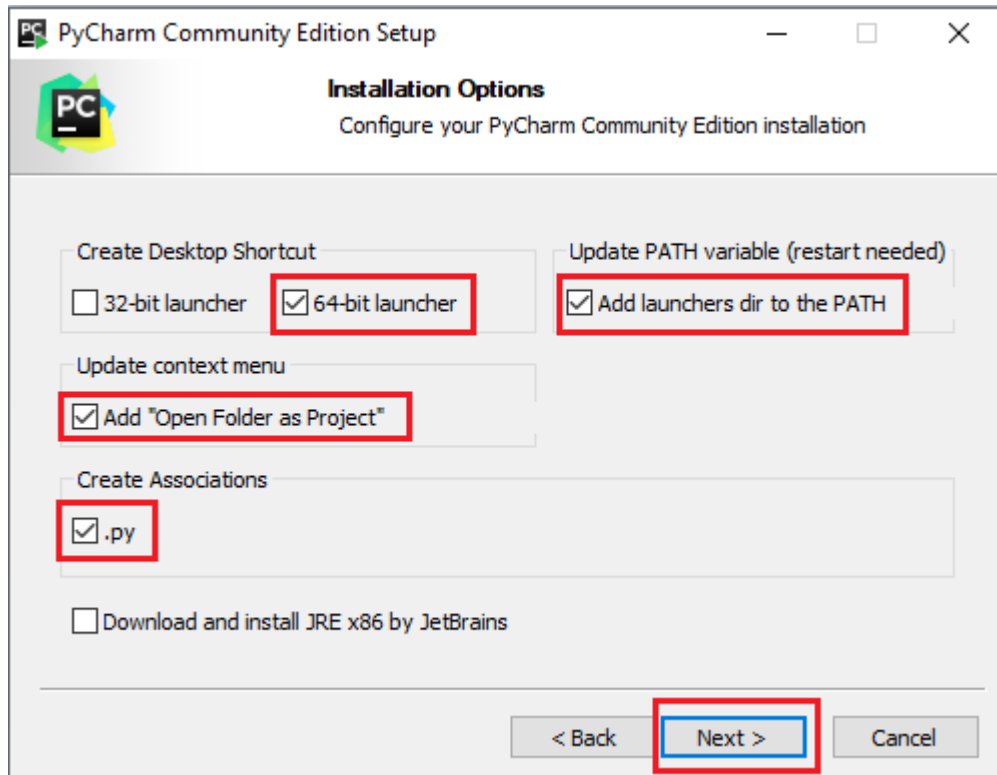
- Sau khi hoàn tất tải xuống, hãy chạy tệp .exe. Cửa sổ cài đặt sẽ bắt đầu.



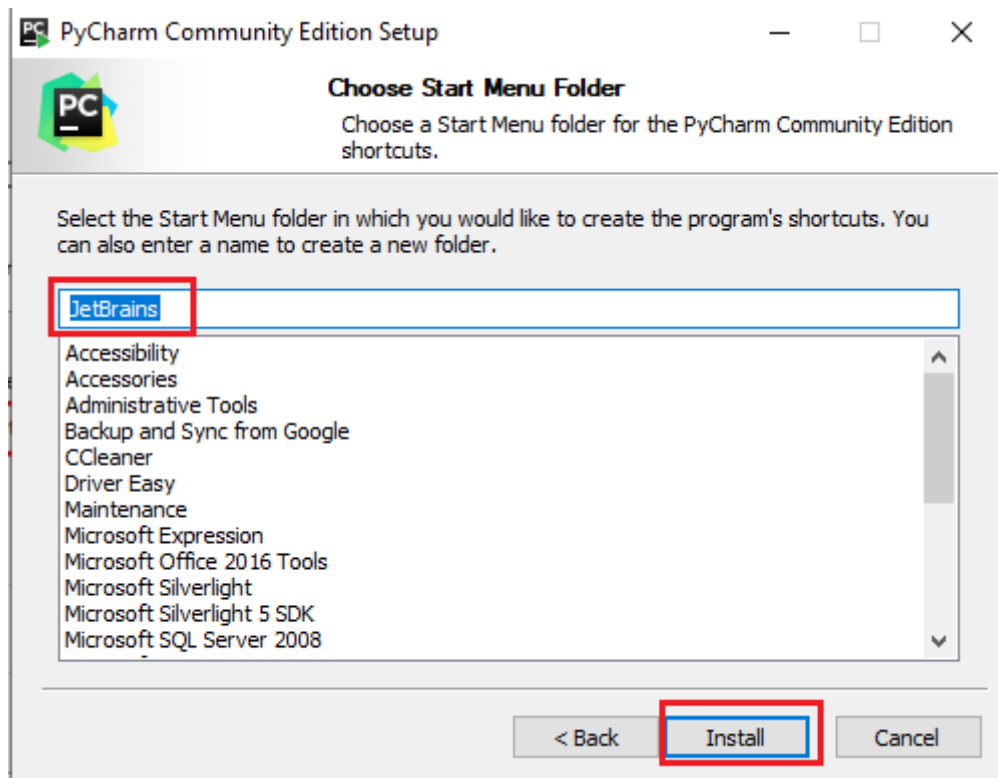
- Nhấn vào nút Next để đến bước tiếp theo.
- Bây giờ cửa sổ cài đặt sẽ hỏi về đường dẫn mà người dùng muốn cài đặt nó.



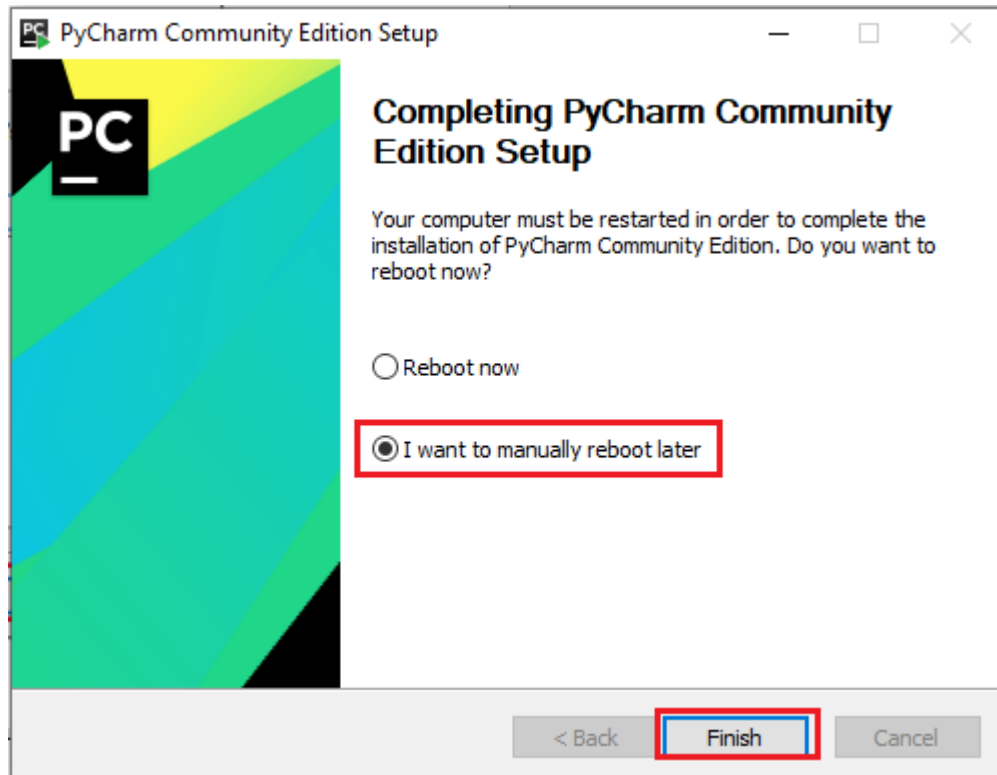
- Lựa chọn các checkbox và nhấn Next



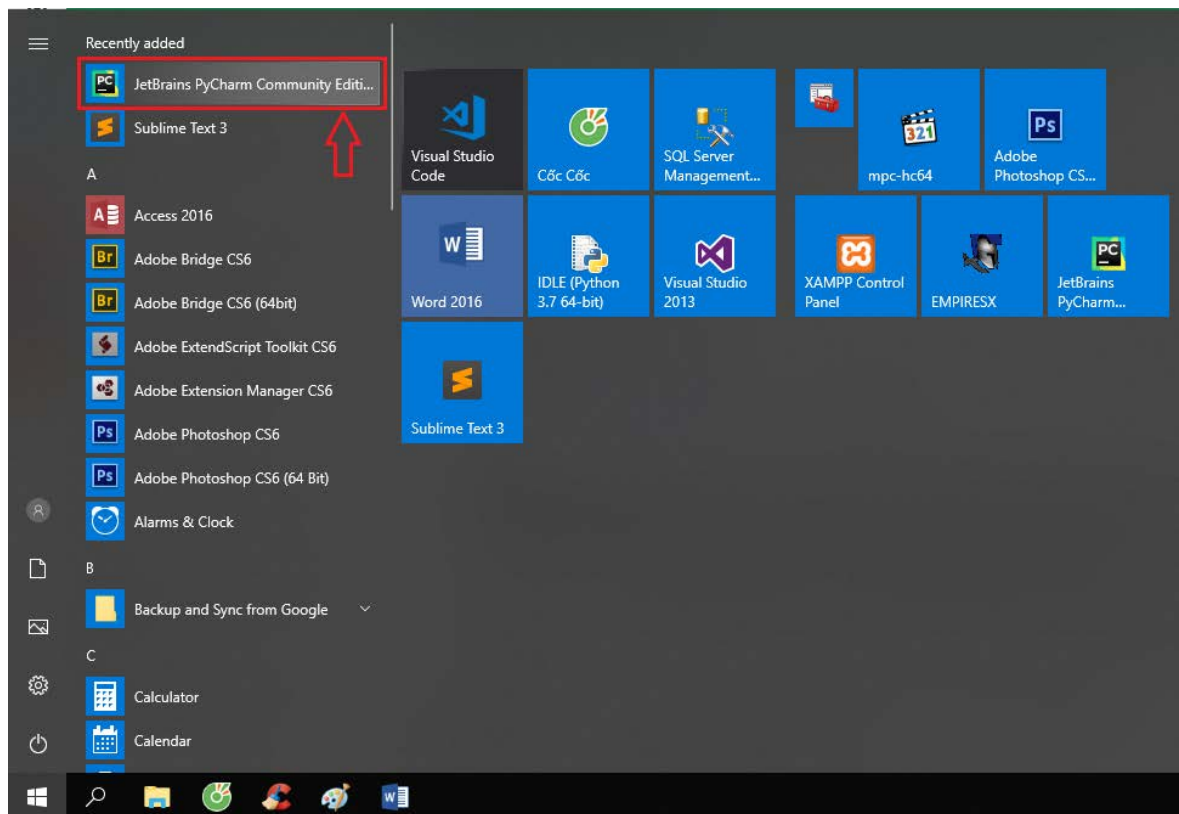
- Chọn tên thư mục sẽ xuất hiện trong Start Menu. Sau đó bấm vào nút install để bắt đầu cài đặt.



- Sau khi cài đặt hoàn tất, cửa sổ trên sẽ xuất hiện. Nhấp vào Finish để hoàn tất cài đặt và mở PyCharm IDE.



- Khởi động PyCharm trên Start Menu



PHẦN IV: Thử nghiệm với một Project đơn giản trong Python

Tạo lập, chạy thử 1 project đơn giản trong python, đưa ra màn hình những giá trị thống kê: mean, median, mode, variance, standard deviation của một tập dữ liệu (data set) đơn giản chứa 10 giá trị ngẫu nhiên trong đoạn [0, 20]

Với python, bài toán này là đơn giản, chỉ mất vài dòng lệnh để giải quyết bài toán bằng cách sử dụng các libraries/modules như numpy, scipy, random

Trước hết, bạn phải cài đặt các package numpy, scipy cho máy tính của bạn bằng cách vào *chế độ dòng lệnh command line* và thực hiện cú pháp sau (bỏ qua nếu trước đó bạn đã cài rồi):

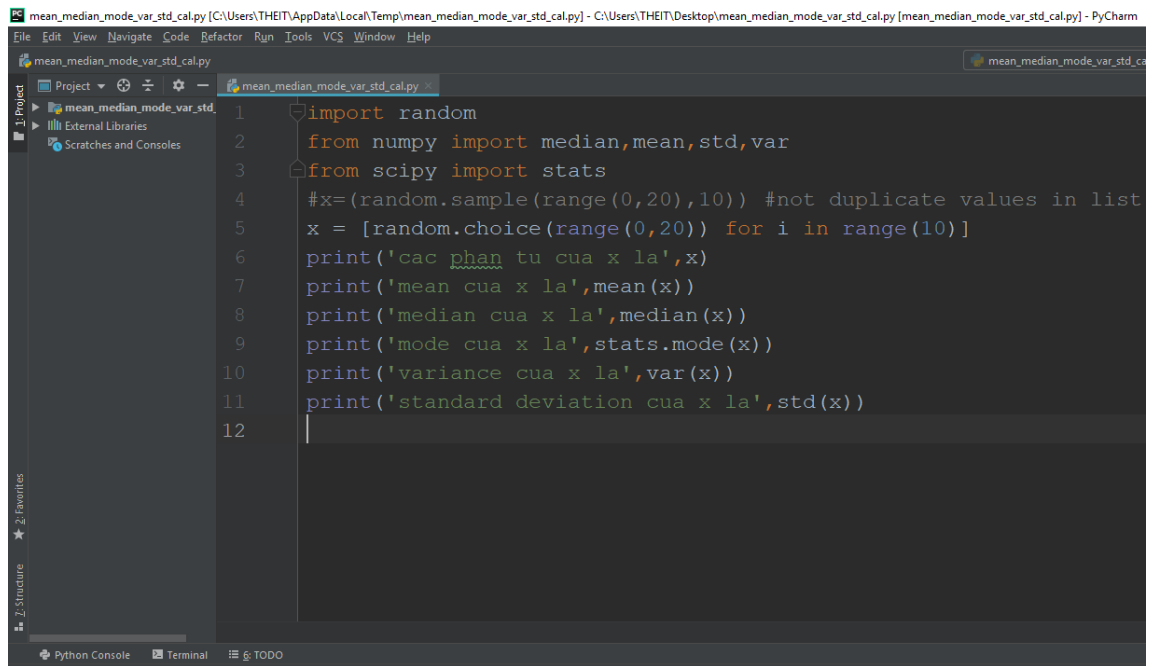
Cú pháp: pip install numpy scipy

#program:

```
import random #using random module
from numpy import median,mean,std,var #using numpy package
from scipy import stats #using scipy package
#x=(random.sample(range(0,20),10)) #not duplicate values
x = [random.choice(range(0,20)) for i in range(10)]
print('cac phan tu cua x la:', x)
print('mean cua x la:', mean(x))
print('median cua x la:', median(x))
print('mode cua x la:', stats.mode(x))
print('variance cua x la:', var(x))
print('standard deviation cua x la:', std(x))
```

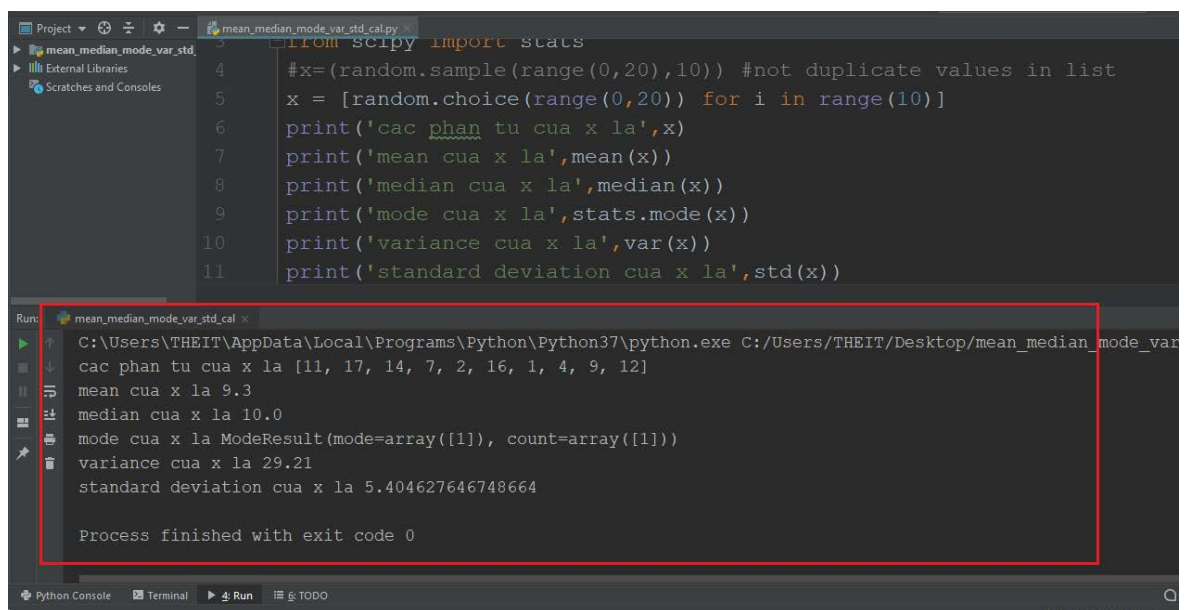
- Mở PyCharm IDE và tạo mới 1 file python bằng cách *File* → *New* và chọn *Python File* và tiến hành đặt tên cho file, file python được tạo sẽ có đuôi định

dạng là .py



```
1 import random
2 from numpy import median, mean, std, var
3 from scipy import stats
4 #x=(random.sample(range(0,20),10)) #not duplicate values in list
5 x = [random.choice(range(0,20)) for i in range(10)]
6 print('cac phan tu cua x la',x)
7 print('mean cua x la',mean(x))
8 print('median cua x la',median(x))
9 print('mode cua x la',stats.mode(x))
10 print('variance cua x la',var(x))
11 print('standard deviation cua x la',std(x))
12
```

- Nhấn tổ hợp phím alt + shift + F10 để thực thi chương trình, kết quả:



```
Run: mean_median_mode_var_std_cal x
C:\Users\THEIT\AppData\Local\Programs\Python\Python37\python.exe C:/Users/THEIT/Desktop/mean_median_mode_var_std_cal.py
cac phan tu cua x la [11, 17, 14, 7, 2, 16, 1, 4, 9, 12]
mean cua x la 9.3
median cua x la 10.0
mode cua x la ModeResult(mode=array([1]), count=array([1]))
variance cua x la 29.21
standard deviation cua x la 5.404627646748664

Process finished with exit code 0
```