SAKARYA
ÜNİVERSİTESİ

BSM 461

INTRODUCTION TO BIG DATA

Lecture 2 – Intro to Python

Kevser Ovaz Akpınar, PhD

kovaz.sakarya.edu.tr
kovaz@sakarya.edu

Agenda

- Python for data analysis and visualization
- Python basics
- Popular libraries
- Data manipulation
- Plotting
- Pandas
- Exercises

Python

- Very popular general-purpose programming language
- Used from introductory programming courses to production systems

- Software programmer Guido van Rossum from Netherlands in 1990
- Name is given from a show called Flying Circus by English comedy group Monty Python
- Its not scripting language!!

Python supports:

- Structural programming
- Object oriented programming
- Functional programming

Python Programming

- Many IDEs available or
- Notepad + Python interpreter or
- Anaconda which has Spyder and Jupyter Notebook software for Python programming

- Two versions of Python in use - Python 2 and Python 3
- Python 3 not backward-compatible with Python 2
- A lot of packages are available for Python 2

- Check version using the following command

```
$ python -- version
```

Python Features

- Dynamically typed

(rather than statically typed like Java or C/C++)

- Interpreted

(rather than compiled like Java or C/C++)

Python programs are comparatively…

+ Quicker to write

+ Shorter

+ Ease of programming

+ Minimizes the time to develop and maintain code

+ Modular and object-oriented

+ Large community of users

+ A large standard and user-contributed library

− More error-prone

− Interpreted and therefore slower than compiled languages

− Decentralized with packages

Python for Data Analytics

- Fairly easy to read/write/process data using standard features
- Plus special packages for…
  - Numerical and statistical manipulations - numpy
  - Visualization (“plotting”) - matplotlib
  - Relational database like capabilities – pandas
  - Machine learning - scikit-learn
  - Network analysis - networkx
  - Unstructured data – re, nltk, PIL

More on Python

- Reference types and Object cloning
  - Most of the objects are Reference Type

- Functions are defined as ***“def”*** keyword

- Object oriented approach support
  - “scikit-learn” library is developed in object oriented manner. It contains many files like “naive_bayes.py”, which has classes.

Variable Types

- Numeric Types
- Strings
- Boolean Types
- Special Types

- Use the type function to determine variable type

```
>>type(log_file)
>>file
```

- Some keywords are reserved such as ‘and’, ‘assert’, ‘break’, ‘lambda’. A list of keywords are located at https://docs.python.org/2.5/ref/keywords.html

Data Structures

- List (starts from 0)
  - Negative indices allow access from tail to head
  - List slicing

    list[start_index:end_index:step]

    step 1 as default
  - *remove() append()*
- Dictionaries
  - Stores (key,value). Key is unique. Dictionaries support add, delete and search.
- Tuple

More on Python

- Lambda functions

  lambda *parameters* : *words*

```
#lambda function 1
fnc = lambda x : x + 1
print(fnc(1))
#Output: 2
print(fnc(fnc(1)))
#Output: 3

#lambda function 2
fnc2 = lambda x, y : x + y
print(fnc2(4,7))
#Output: 11

print(fnc2(4,fnc(1)))
#Output: 6
```

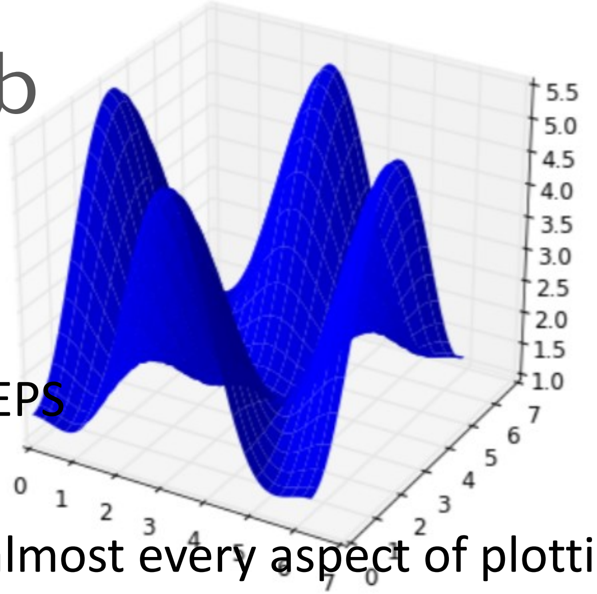
More on Python

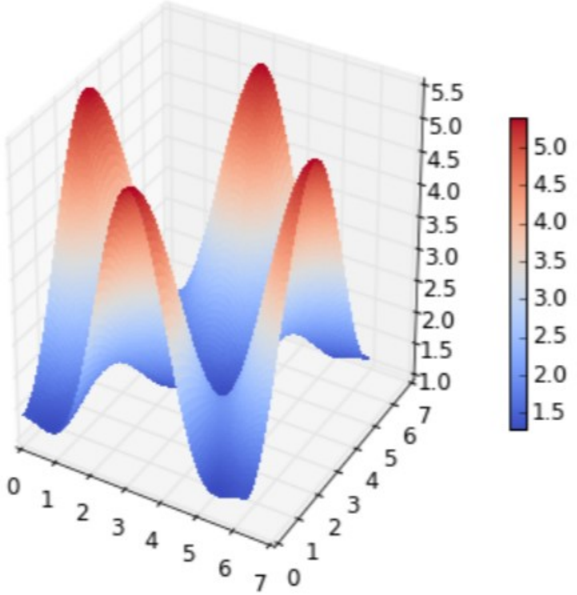
FAST!

- **Easy**: You can write a Python program in one single line into the Python shell. So simple!
- **Numpy api:** Simple but not limited. Numpy: the main API used for what is called “scientific computing ecosystem.” Numpy handles linear algebra and matrix mathematics on a very large scale. Most machine learning algorithms and neural networks operate on these n-dimensional matrices.
  - ✓ Written in C and Fortran
  - ✓ Vectorized computations
- **Apache Spark has a Python shell**. You can open datasets, do transformations, and run algorithms in one easy command line. Without that you would have to package your program and then submit it to Spark using spark-submit. The disadvantage with **spark-submit**, as with any batch job, is you cannot inspect variables in real time. So can print values to a log. That’s OK for text, but when you use the Python shell that text is an object, which means you can further work with it. It’s not a static non-entity.

```
>>> def numpy_version () :
    t1 = time . time ()
     X = arange (10000000)
     Y = arange (10000000)
     Z = X + Y
     return time . time () - t1
>>> numpy_version ()
    0.059307098388671875
```

More on Python - Matplotlib

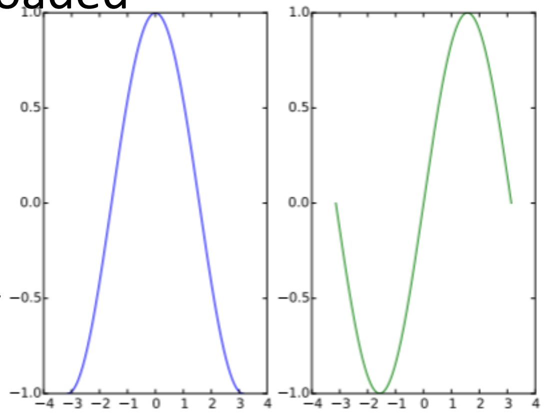




- Used for generating 2D and 3D scientific plots
- Support for LaTeX
- Fine-grained control over every aspect
- Many output file formats including PNG, PDF, SVG, EPS

- Configuration file ‘matplotlibrc’ used to customize almost every aspect of plotting
- On Linux, it looks in .config/matplotlib/matplotlibrc
- On other platforms, it looks in .matplotlib/matplotlibrc
- Use ‘matplotlib.matplotlib fname()’ to determine from where the current matplotlibrc is loaded
- Customization options can be found at http://matplotlib.org/users/customizing.html

- Matplotlib is the entire library
- **Pyplot** - a module within Matplotlib that provides access to the underlying plotting library
- **Pylab** - a convenience module that combines the functionality of Pyplot with Numpy

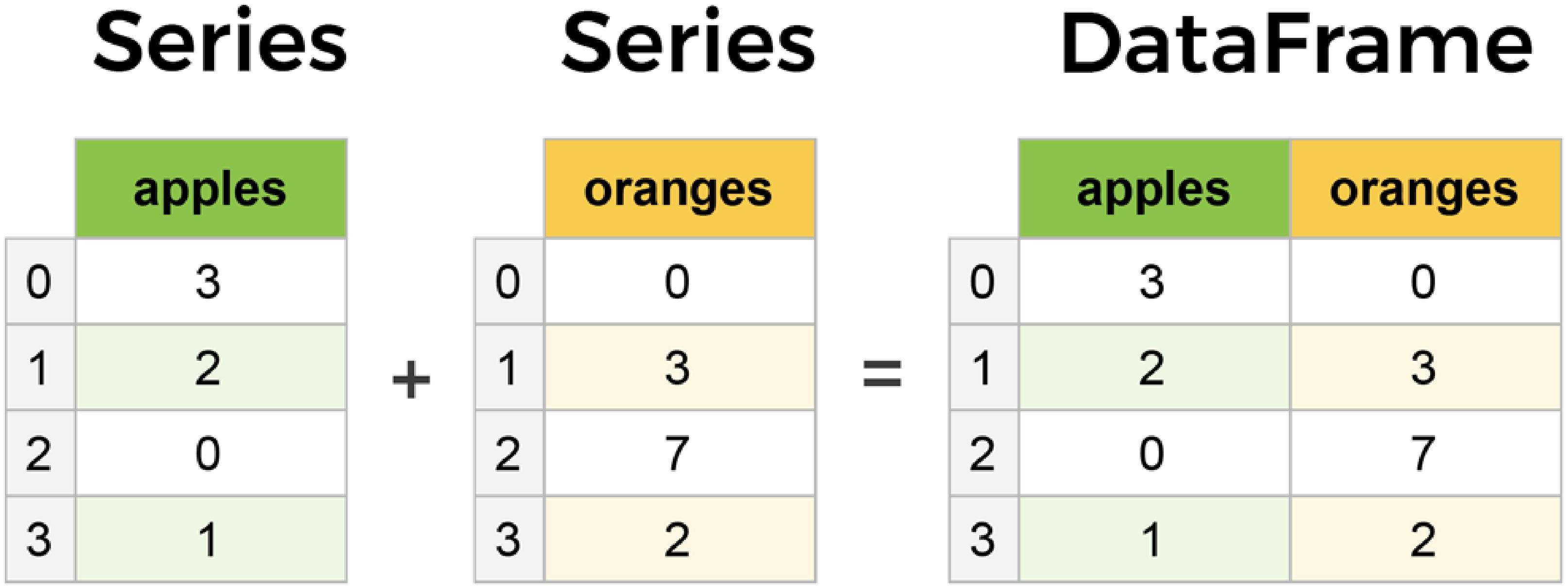


More on Python

- **The Python Pip Toolkit:** Programmers contribute to its open source repository, the Python Package Index (PIP). Sample pip packages read and write to JSON and **requests** to work with web services.
- **Pandas:** Open-source library! Transform data from one format to another and run these algorithms at scale, meaning across a cluster. For example, older algorithms that existed before distributed computing (i.e., big data) like scikit-learn would not work with distributed data frames and other objects run across a cluster. They are designed to work with one file on one computer. So that is an issue to keep in mind as you figure out which framework to use. With Pandas, for very large data sets you might have a hybrid of tools

No support of parallel processing!!

More on Python: Pandas



Series

	apples
0	3
1	2
2	0
3	1

+

Series

	oranges
0	0
1	3
2	7
3	2

=

DataFrame

	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2

More on Python: Pandas Comparison with SQL

```
SELECT total_bill, tip, smoker, time
FROM tips
LIMIT 5;
```

```
tips[['total_bill', 'tip', 'smoker', 'time']].head(5)
```

```
SELECT *
FROM tips
WHERE time = 'Dinner'
LIMIT 5;
```

```
tips[tips['time'] == 'Dinner'].head(5)
```

```
SELECT city, rank
FROM df1
UNION ALL
SELECT city, rank
FROM df2;
```

```
pd.concat([df1, df2])
```

More on Python

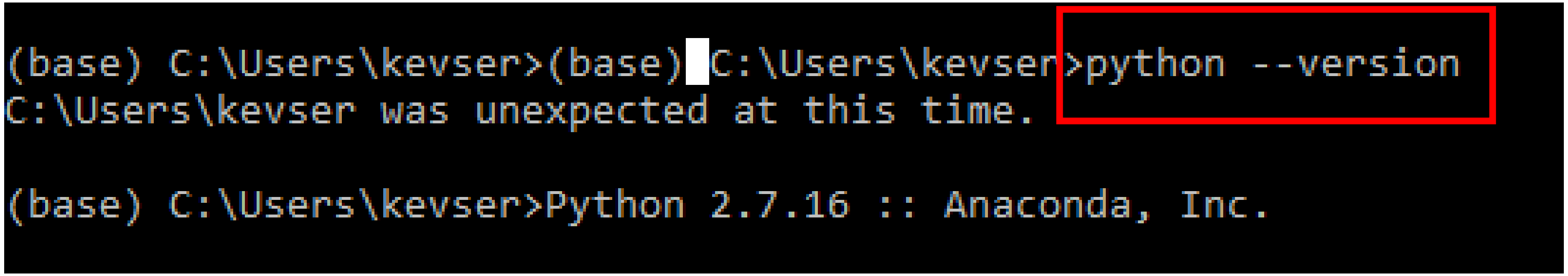
- **Python Notebooks (IPYTHON):** Jupyter is used for notebooks. It is an interactive computational environment, in which you can combine code execution, rich text, mathematics, plots and rich media

Python on OS

- MacOS X, High Sierra has a preloaded version of Python 2.7 out-of-the-box. If you have macOS X, you will not have to install or configure anything else in order to use Python 2. If you want to use Python3, then installation is required

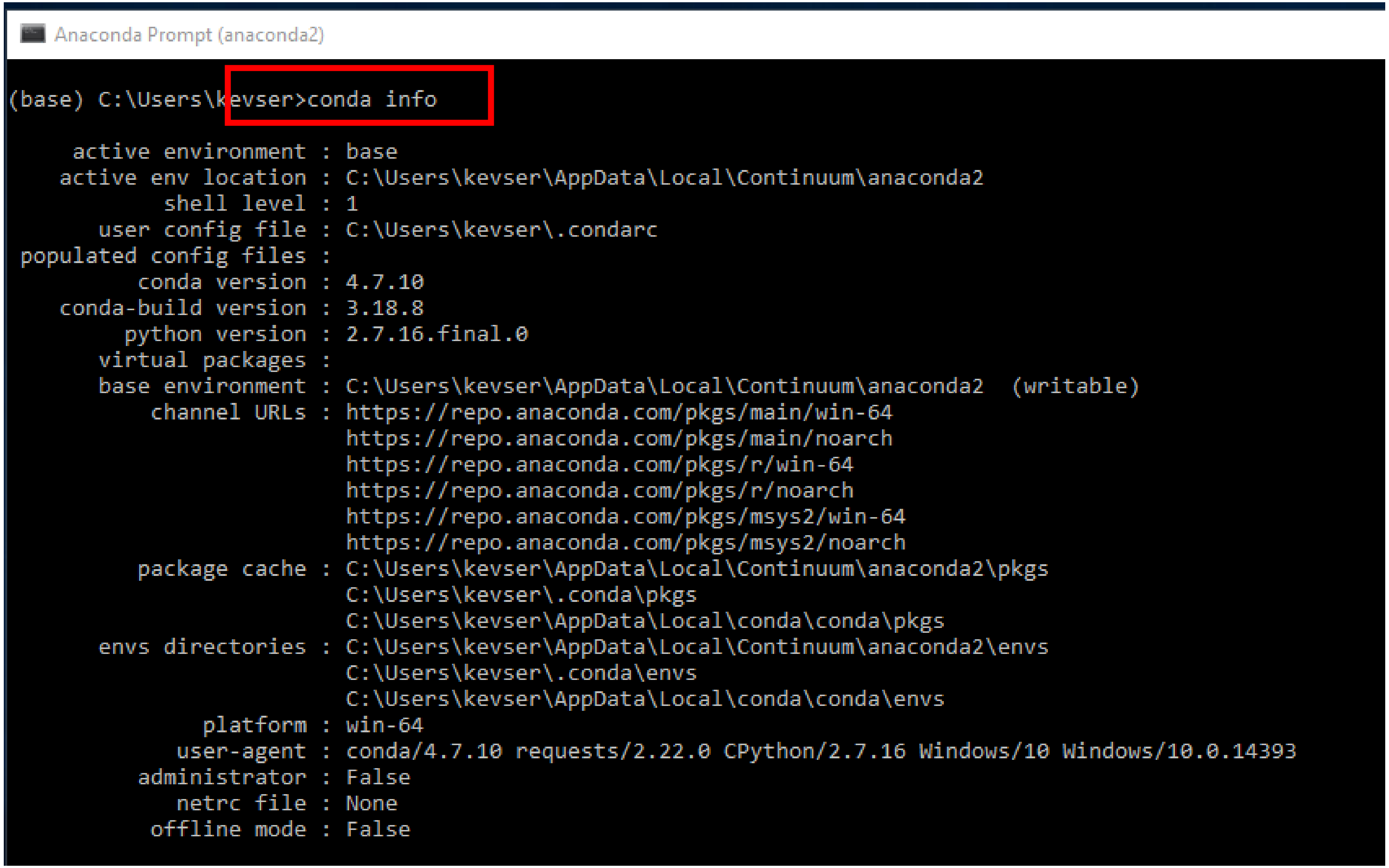
- Python doesn’t come prepackaged with Windows. Download the installer and follow the wizard.

Python



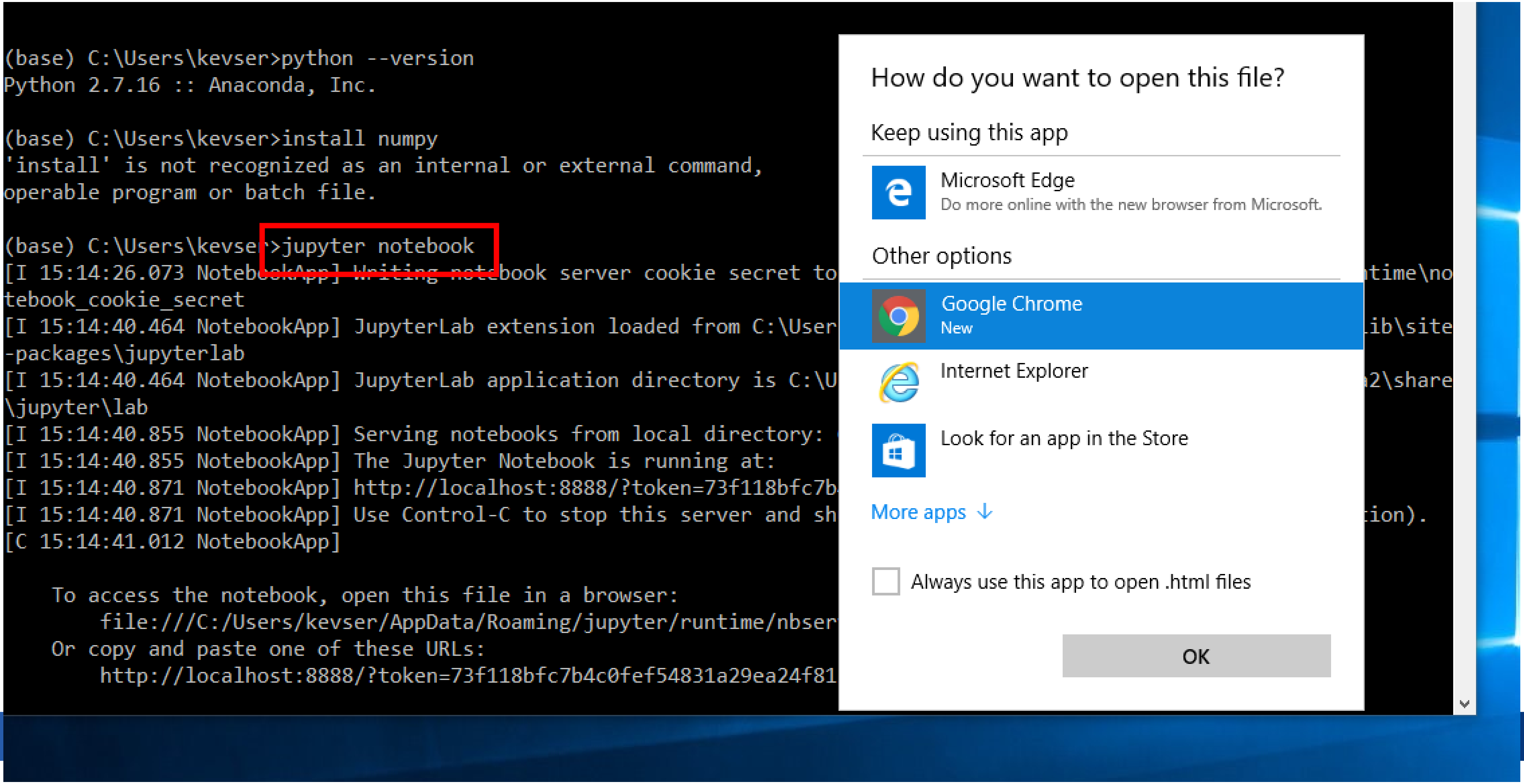
```
(base) C:\Users\kevser>(base) C:\Users\kevser>python --version
C:\Users\kevser was unexpected at this time.

(base) C:\Users\kevser>Python 2.7.16 :: Anaconda, Inc.
```

Python



```
(base) C:\Users\kevser>conda info

     active environment : base
    active env location : C:\Users\kevser\AppData\Local\Continuum\anaconda2
            shell level : 1
       user config file : C:\Users\kevser\.condarc
 populated config files :
          conda version : 4.7.10
    conda-build version : 3.18.8
         python version : 2.7.16.final.0
       virtual packages :
       base environment : C:\Users\kevser\AppData\Local\Continuum\anaconda2  (writable)
           channel URLs : https://repo.anaconda.com/pkgs/main/win-64
                          https://repo.anaconda.com/pkgs/main/noarch
                          https://repo.anaconda.com/pkgs/r/win-64
                          https://repo.anaconda.com/pkgs/r/noarch
                          https://repo.anaconda.com/pkgs/msys2/win-64
                          https://repo.anaconda.com/pkgs/msys2/noarch
          package cache : C:\Users\kevser\AppData\Local\Continuum\anaconda2\pkgs
                          C:\Users\kevser\.conda\pkgs
                          C:\Users\kevser\AppData\Local\conda\conda\pkgs
       envs directories : C:\Users\kevser\AppData\Local\Continuum\anaconda2\envs
                          C:\Users\kevser\.conda\envs
                          C:\Users\kevser\AppData\Local\conda\conda\envs
               platform : win-64
             user-agent : conda/4.7.10 requests/2.22.0 CPython/2.7.16 Windows/10 Windows/10.0.14393
          administrator : False
             netrc file : None
           offline mode : False
```

Python



```
(base) C:\Users\kevser>python --version
Python 2.7.16 :: Anaconda, Inc.

(base) C:\Users\kevser>install numpy
'install' is not recognized as an internal or external command,
operable program or batch file.

(base) C:\Users\kevser>jupyter notebook
[I 15:14:26.073 NotebookApp] Writing notebook server cookie secret to                              time\no
tebook_cookie_secret
[I 15:14:40.464 NotebookApp] JupyterLab extension loaded from C:\User                              ib\site
-packages\jupyterlab
[I 15:14:40.464 NotebookApp] JupyterLab application directory is C:\U                              a2\share
\jupyter\lab
[I 15:14:40.855 NotebookApp] Serving notebooks from local directory:
[I 15:14:40.855 NotebookApp] The Jupyter Notebook is running at:
[I 15:14:40.871 NotebookApp] http://localhost:8888/?token=73f118bfc7b
[I 15:14:40.871 NotebookApp] Use Control-C to stop this server and sh                              ion).
[C 15:14:41.012 NotebookApp]

    To access the notebook, open this file in a browser:
        file:///C:/Users/kevser/AppData/Roaming/jupyter/runtime/nbser
    Or copy and paste one of these URLs:
        http://localhost:8888/?token=73f118bfc7b4c0fef54831a29ea24f81
```

How do you want to open this file?

Keep using this app

Microsoft Edge
Do more online with the new browser from Microsoft.

Other options

Google Chrome
New

Internet Explorer

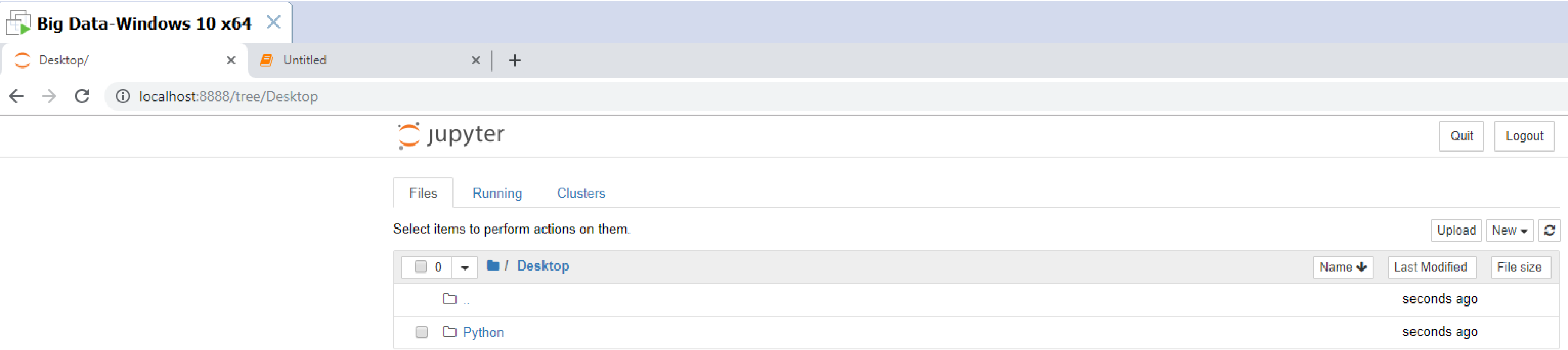
Look for an app in the Store

More apps ↓

☐ Always use this app to open .html files

OK

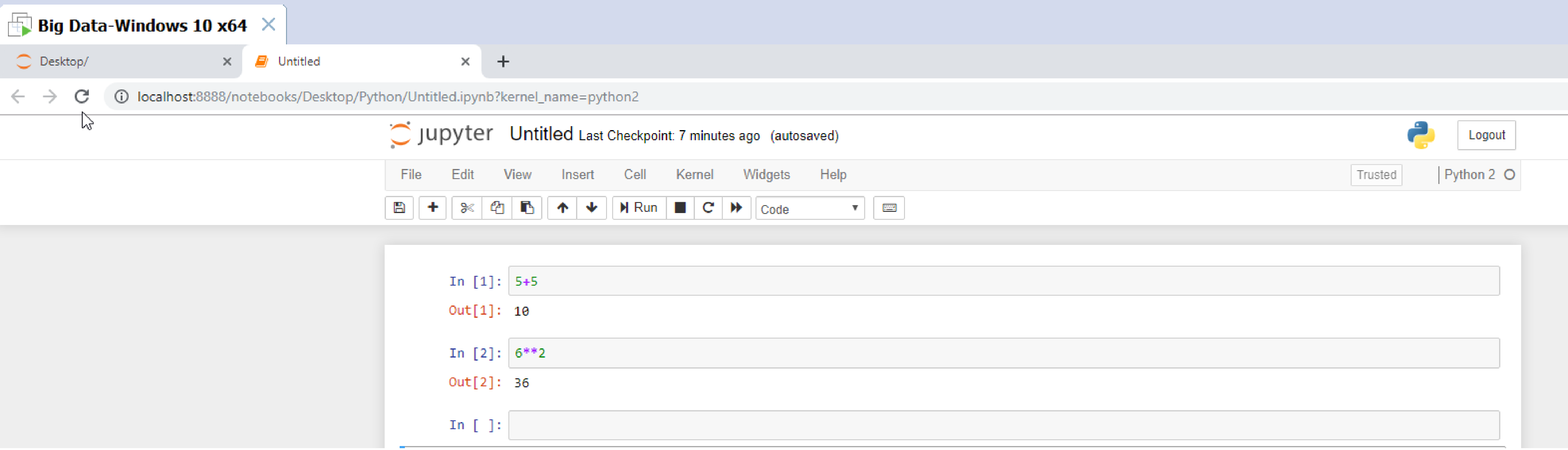
Python



Select items to perform actions on them.

	/ Desktop	
	..	seconds ago
□	Python	seconds ago

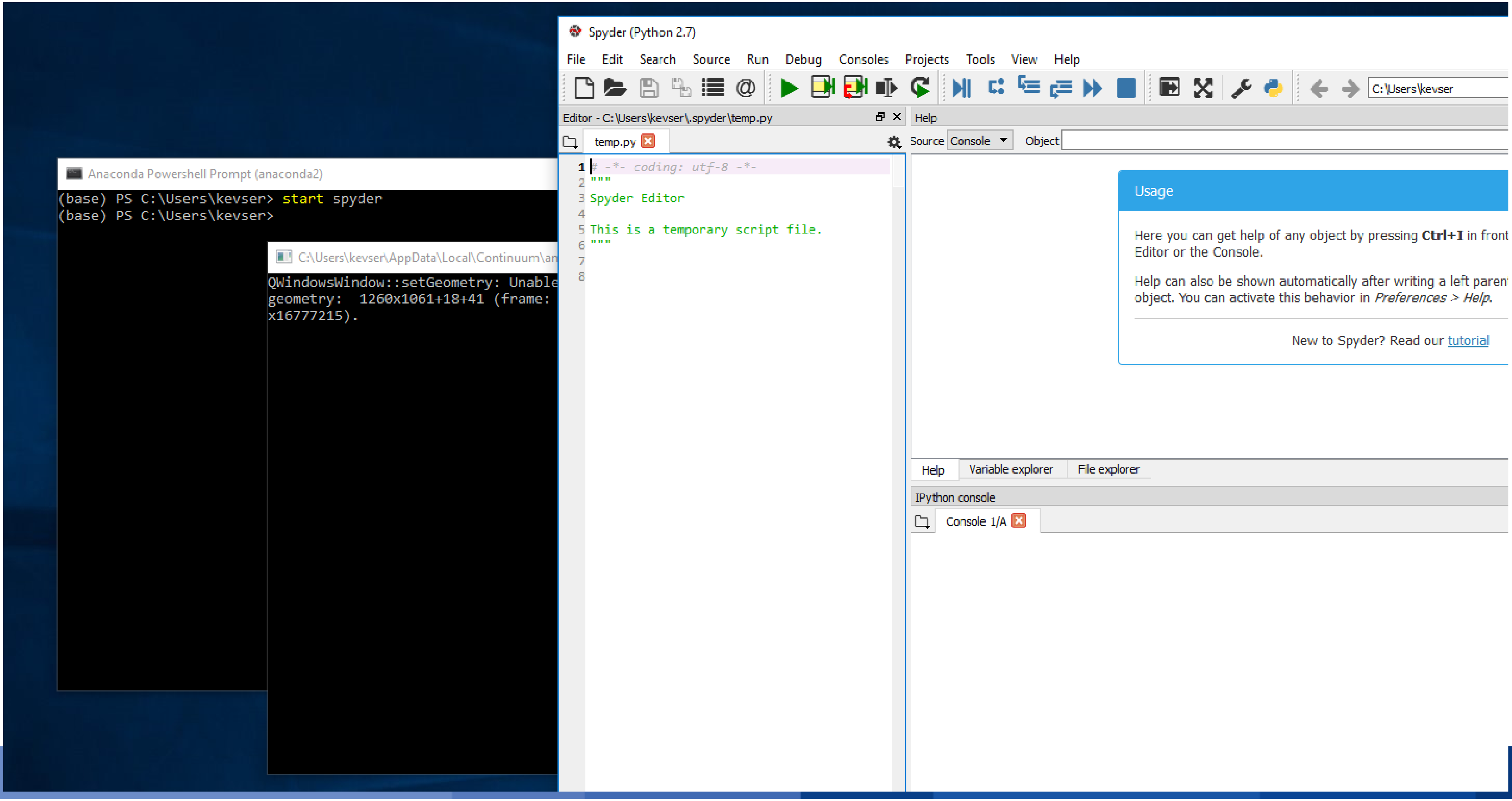
Python



Big Data-Windows 10 x64
Desktop/
Untitled
localhost:8888/notebooks/Desktop/Python/Untitled.ipynb?kernel_name=python2
jupyter
Untitled Last Checkpoint: 7 minutes ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 2
Run
Code

```
In [1]: 5+5
Out[1]: 10

In [2]: 6**2
Out[2]: 36

In [ ]:
```

Python -Spyder



```
(base) PS C:\Users\kevser> start spyder
(base) PS C:\Users\kevser>
```

```
QWindowsWindow::setGeometry: Unable
geometry:  1260x1061+18+41 (frame:
x16777215).
```

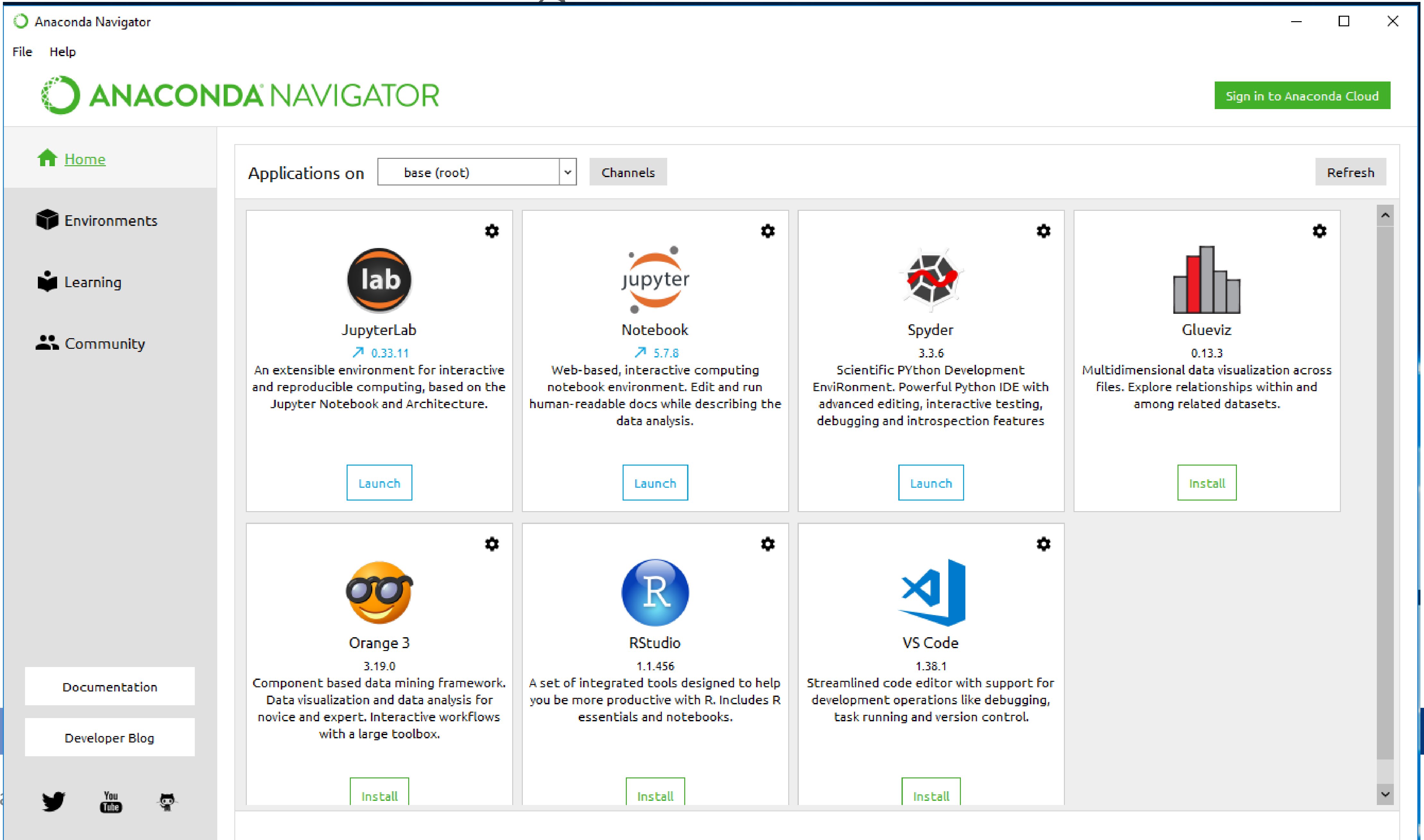
```
# -*- coding: utf-8 -*-
"""
Spyder Editor

This is a temporary script file.
"""
```

Here you can get help of any object by pressing **Ctrl+I** in front
Editor or the Console.

Help can also be shown automatically after writing a left paren
object. You can activate this behavior in *Preferences > Help*.

New to Spyder? Read our tutorial

Python – Anaconda Navigator other tools



JupyterLab
↗ 0.33.11
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Notebook
↗ 5.7.8
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Spyder
3.3.6
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Glueviz
0.13.3
Multidimensional data visualization across files. Explore relationships within and among related datasets.

Orange 3
3.19.0
Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

RStudio
1.1.456
A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

VS Code
1.38.1
Streamlined code editor with support for development operations like debugging, task running and version control.

Python – Samples

ADDITIONAL REFERENCES

Python and SQL Comparison,
https://pandas.pydata.org/pandas-docs/stable/getting_started/comparison/comparison_with_sql.html

Python ile Veri Biliminie Giriş,
https://medium.com/deep-learning-turkiye/python-ile-veri-bilimine-dal%C4%B1%C5%9F-3f069260ebda

Matplotlib Tutorials,
https://matplotlib.org/tutorials/introductory/pyplot.html

BYU, Big Data Science & Capstone Lecture Notes - Python

Stanford University Lecture Notes,
http://web.stanford.edu/class/cs102/lecturenotes/PythonData2.txt

Big Data Analytics in Python Programming, https://www.youtube.com/watch?v=G8VvTp0zgC0

Python for Big Data Analytics – 1, https://www.youtube.com/watch?reload=9&v=BiRXCLKLxrc

www.kaggle.com, “sf_salaries” Dataset