Coding seminar

Lesson 4: UNIX commands and Statistics

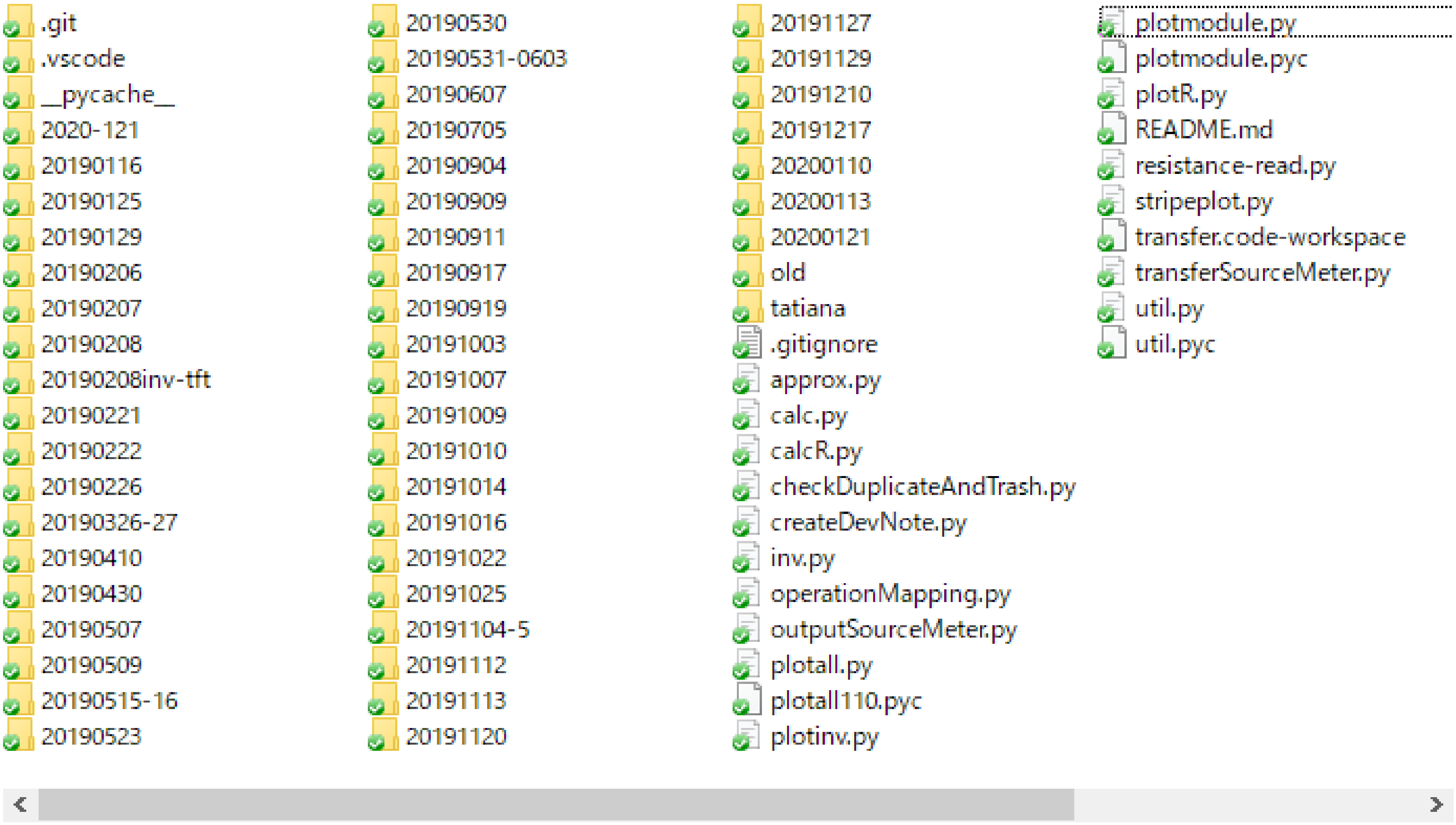
Ikue Hirata, PhD

Contents

Folder structure and UNIX commands

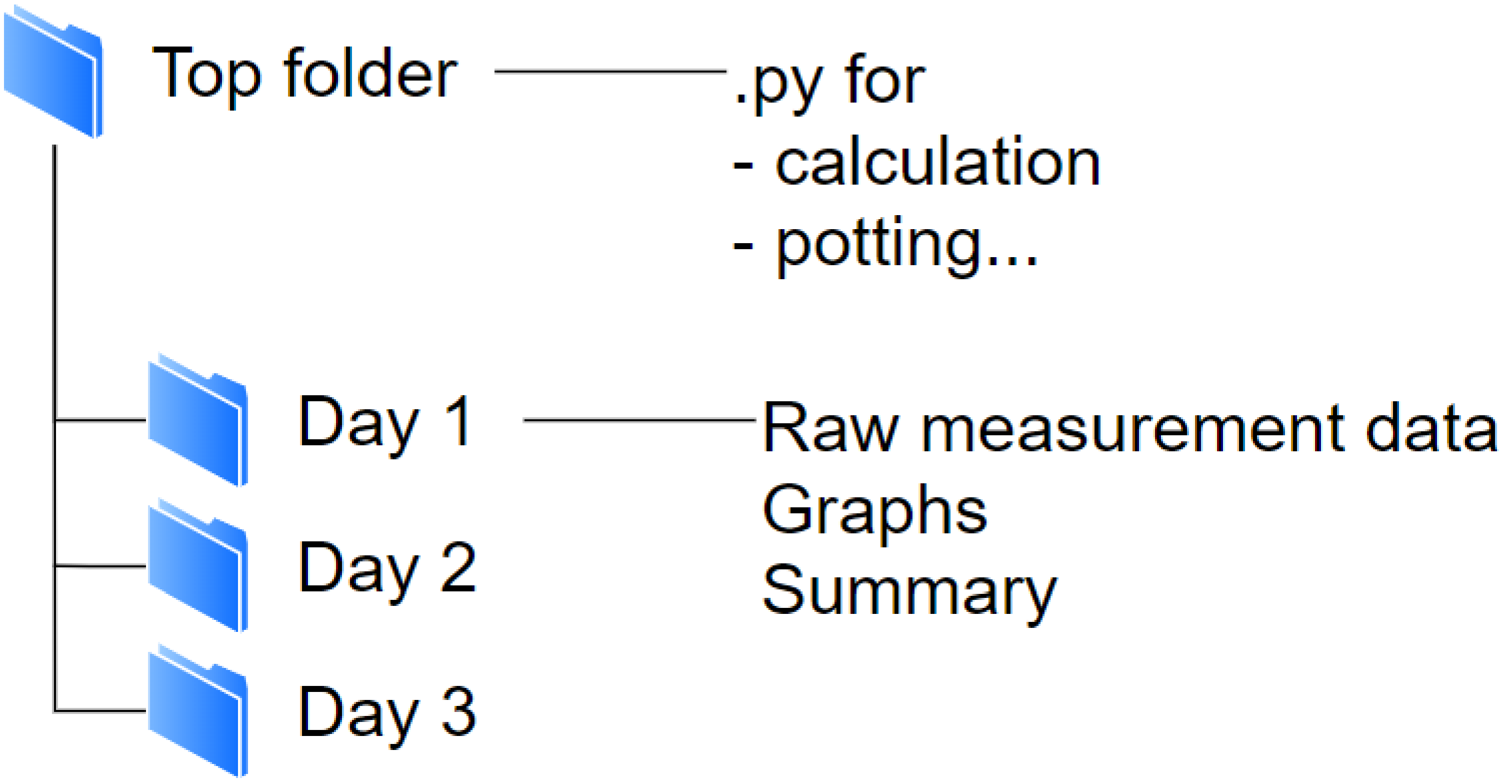
Pandas, seaborn

Getting more practical/systematic



```
.git
.vscode
__pycache__
2020-121
20190116
20190125
20190129
20190206
20190207
20190208
20190208inv-tft
20190221
20190222
20190226
20190326-27
20190410
20190430
20190507
20190509
20190515-16
20190523
20190530
20190531-0603
20190607
20190705
20190904
20190909
20190911
20190917
20190919
20191003
20191007
20191009
20191010
20191014
20191016
20191022
20191025
20191104-5
20191112
20191113
20191120
20191127
20191129
20191210
20191217
20200110
20200113
20200121
old
tatiana
.gitignore
approx.py
calc.py
calcR.py
checkDuplicateAndTrash.py
createDevNote.py
inv.py
operationMapping.py
outputSourceMeter.py
plotall.py
plotall110.pyc
plotinv.py
plotmodule.py
plotmodule.pyc
plotR.py
README.md
resistance-read.py
stripeplot.py
transfer.code-workspace
transferSourceMeter.py
util.py
util.pyc
```

Directory structure



Top folder

.py for
- calculation
- potting...

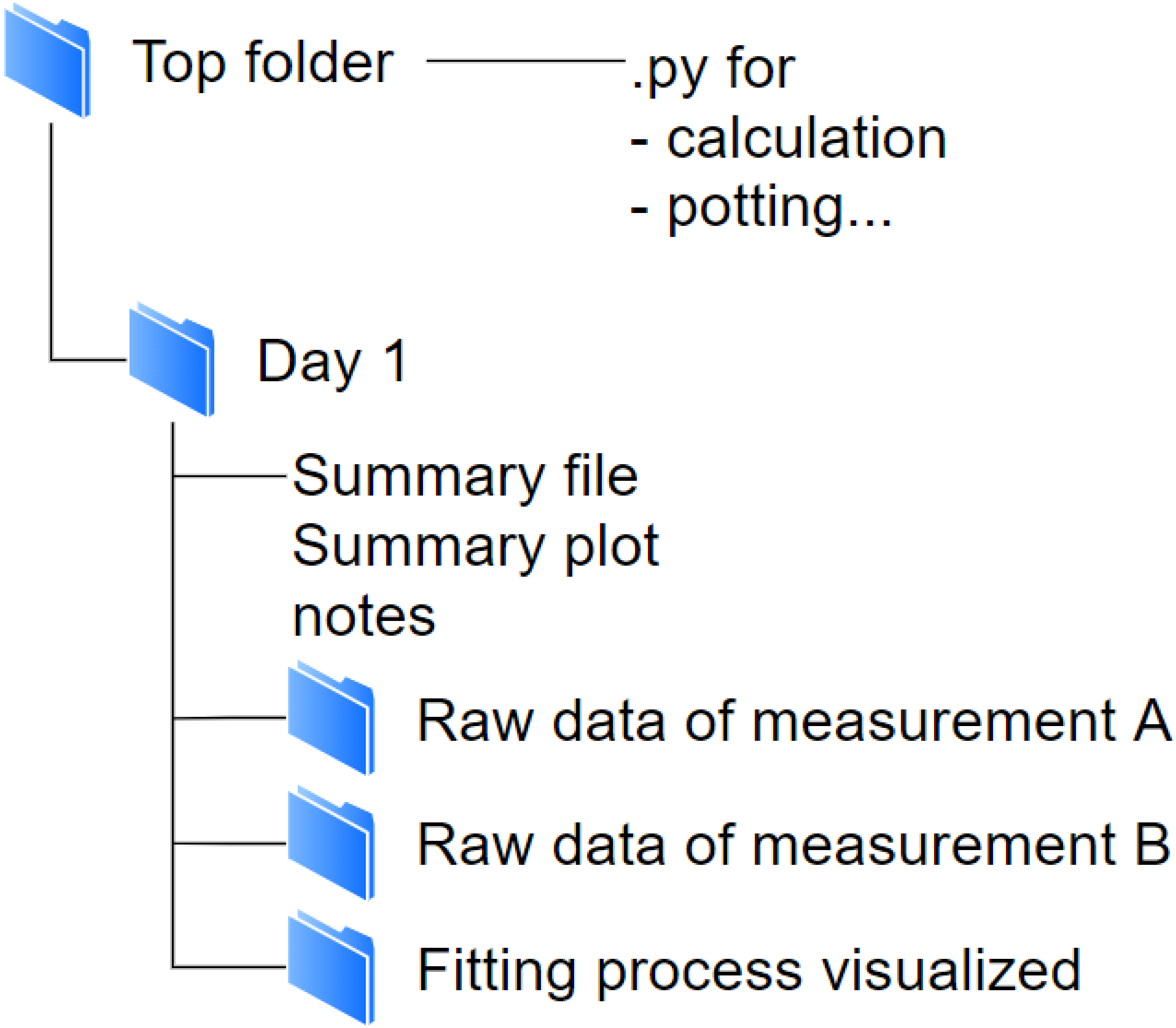
Day 1

Day 2

Day 3

Raw measurement data
Graphs
Summary

Directory structure



Top folder

.py for
- calculation
- potting...

Day 1

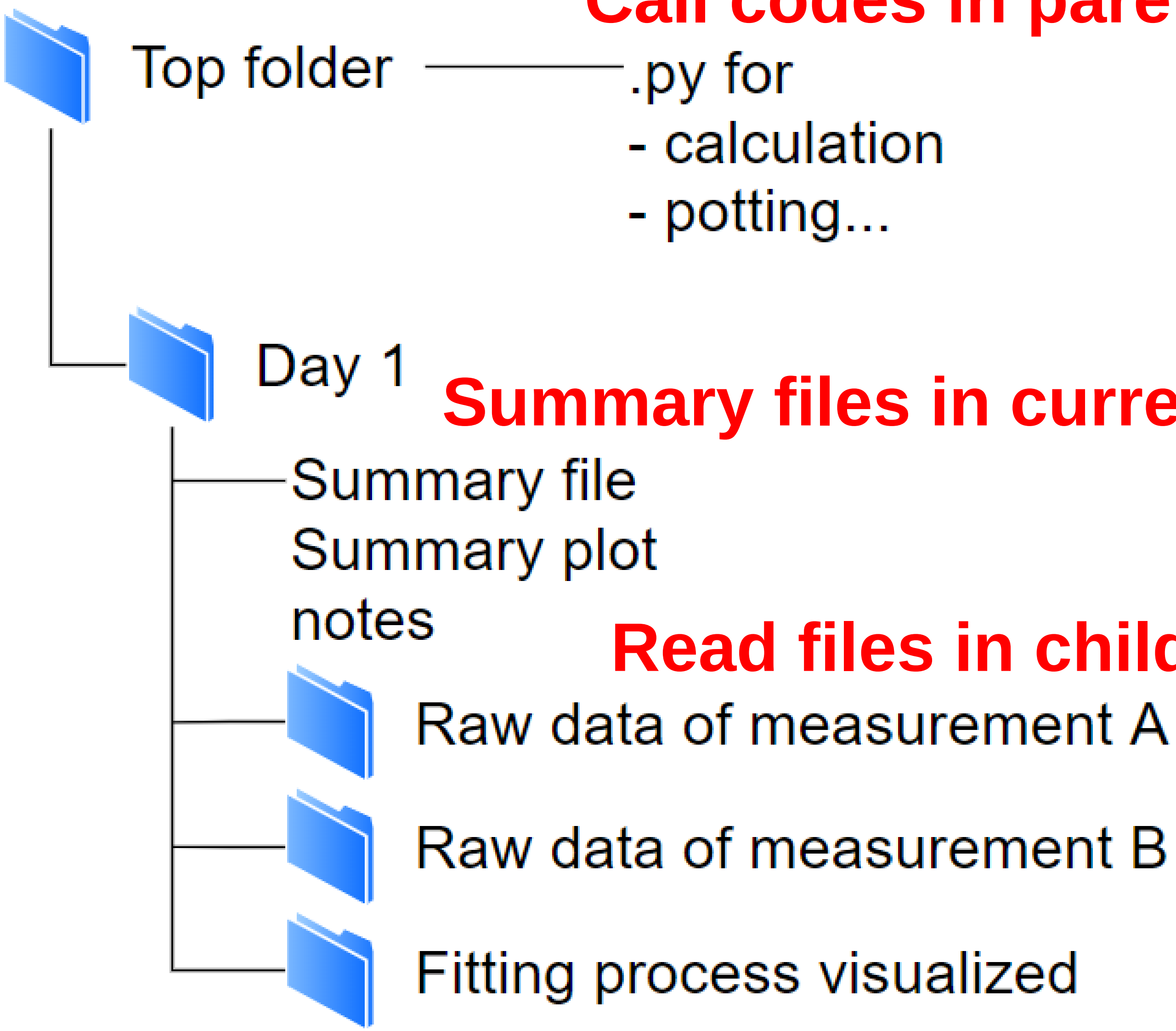
Summary file
Summary plot
notes

Raw data of measurement A

Raw data of measurement B

Fitting process visualized

Directory structure



Call codes in parent

Top folder

.py for
- calculation
- potting...

Day 1

Summary files in current

Summary file
Summary plot
notes

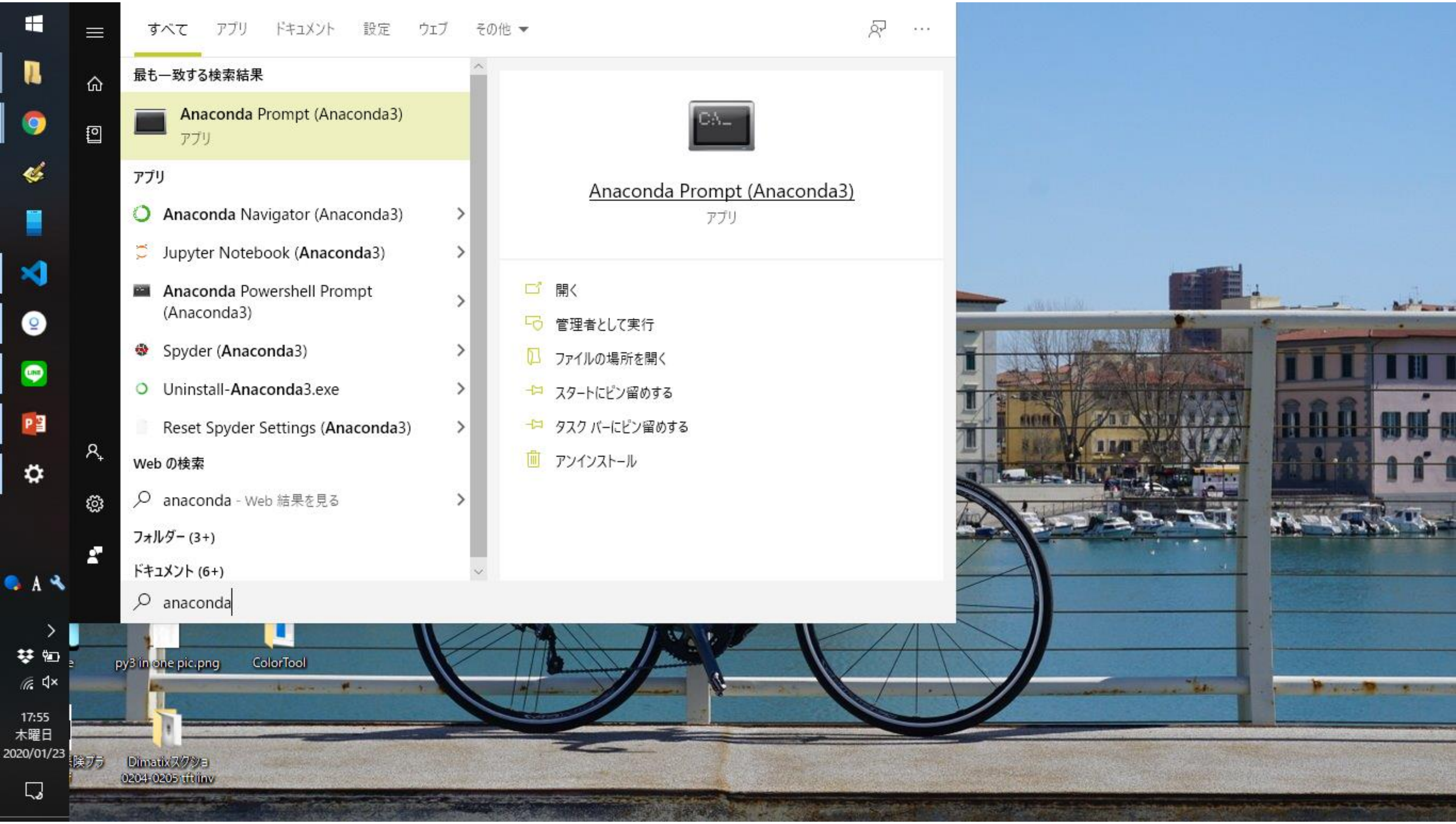
Read files in children

Raw data of measurement A

Raw data of measurement B

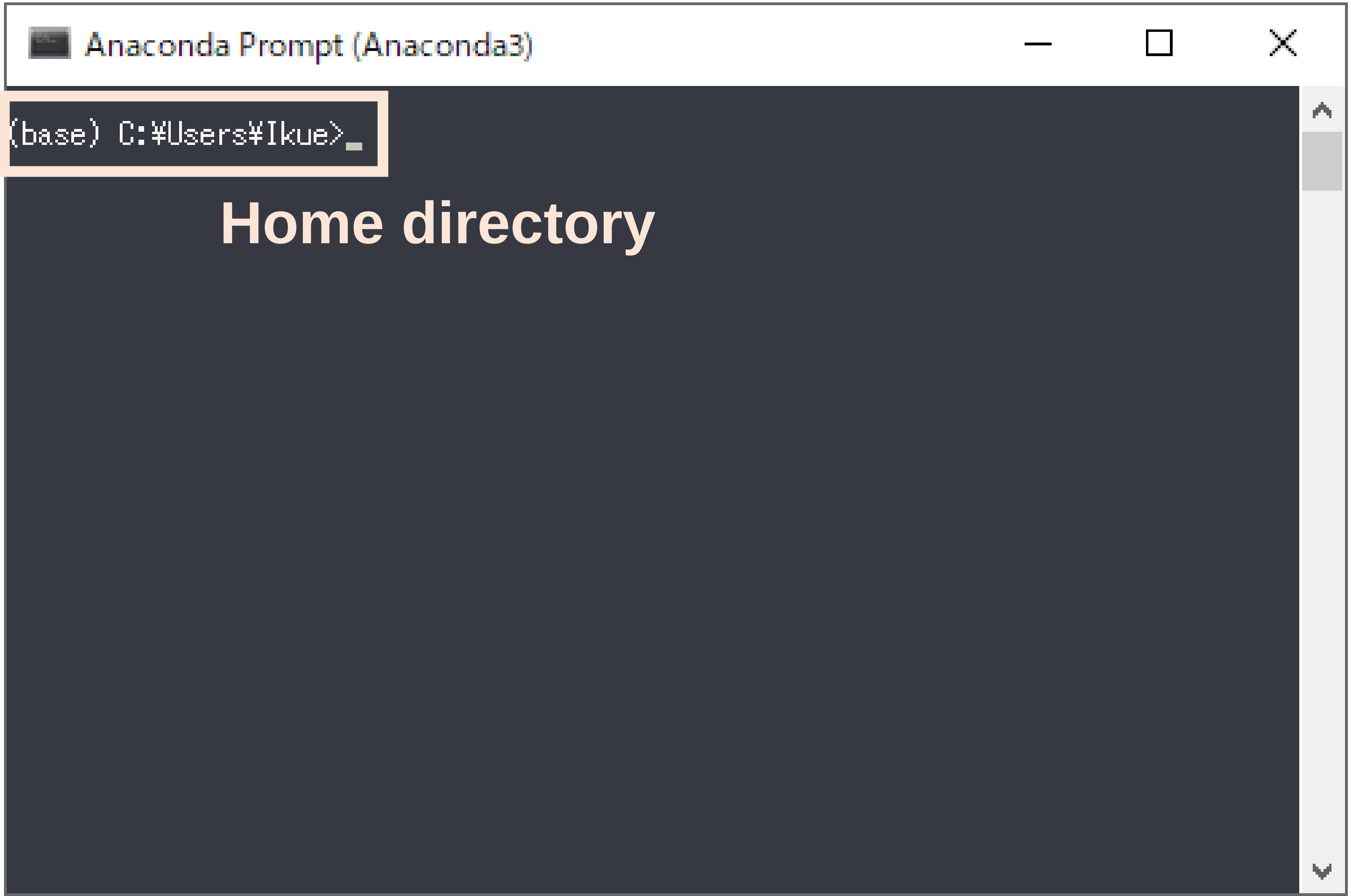
Fitting process visualized

Launch Anaconda Prompt



anaconda

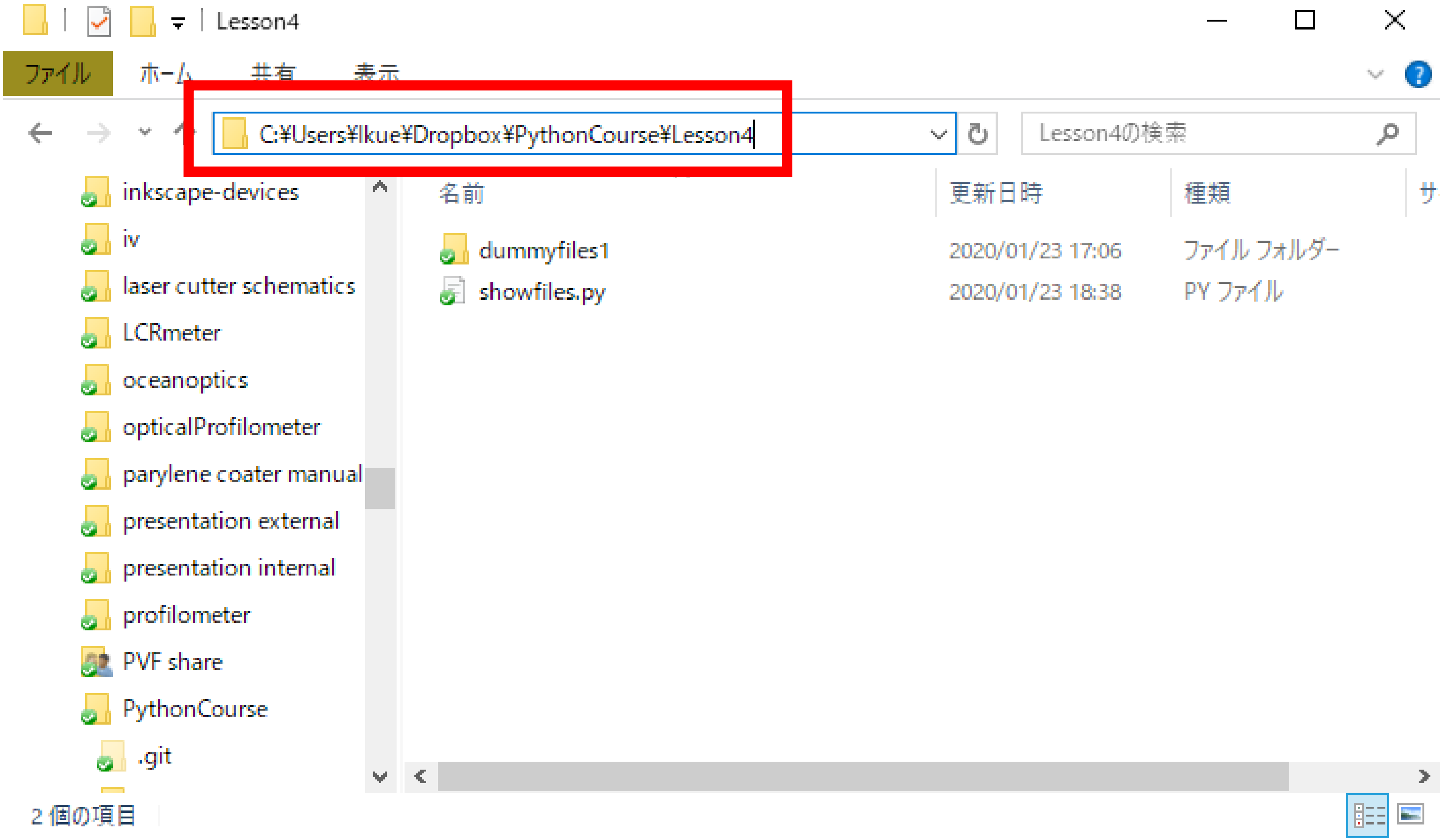
Anaconda Prompt



```
(base) C:¥Users¥Ikue>_
```

Home directory

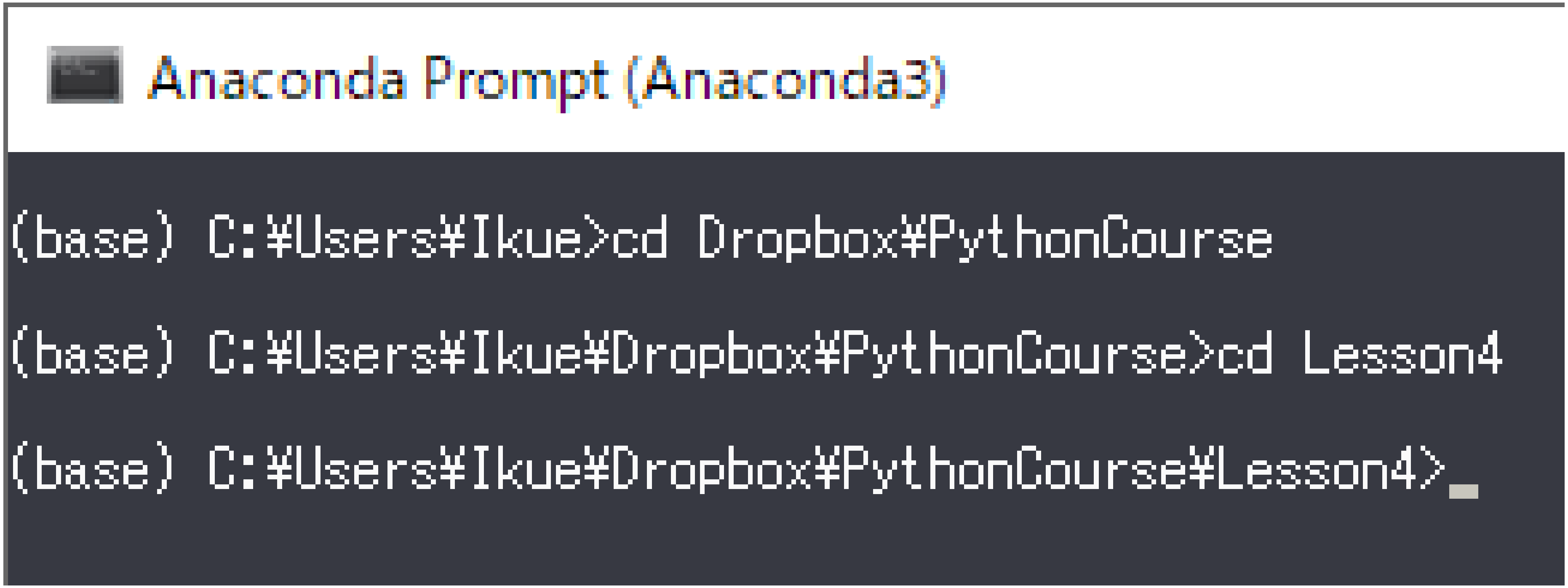
Move to other directory



C:¥Users¥Ikue¥Dropbox¥PythonCourse¥Lesson4

名前	更新日時	種類	サ
dummyfiles1	2020/01/23 17:06	ファイル フォルダー	
showfiles.py	2020/01/23 18:38	PY ファイル	

cd Move to other directory

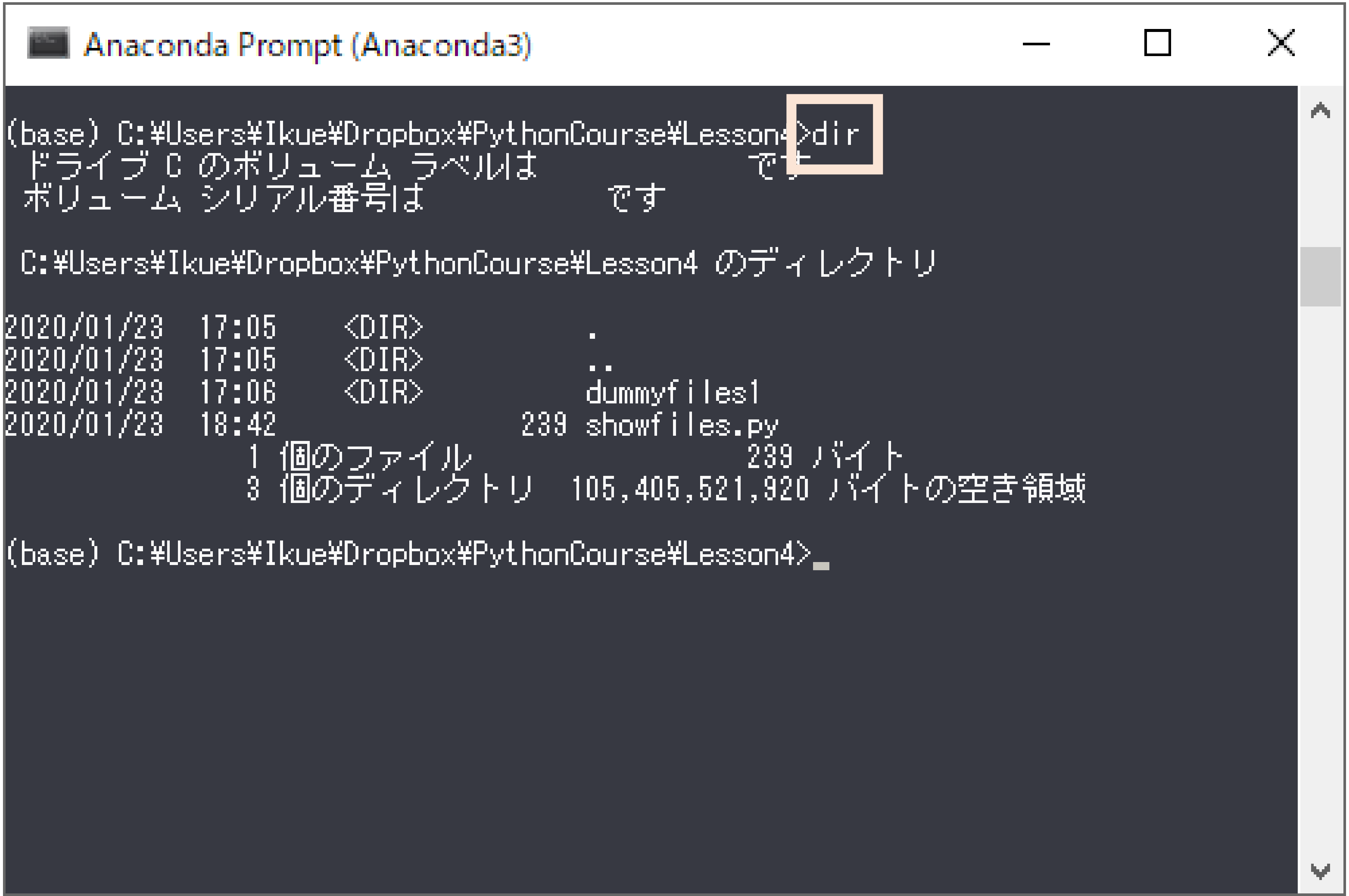


```
(base) C:¥Users¥Ikue>cd Dropbox¥PythonCourse

(base) C:¥Users¥Ikue¥Dropbox¥PythonCourse>cd Lesson4

(base) C:¥Users¥Ikue¥Dropbox¥PythonCourse¥Lesson4>
```

dir Show info on the directory



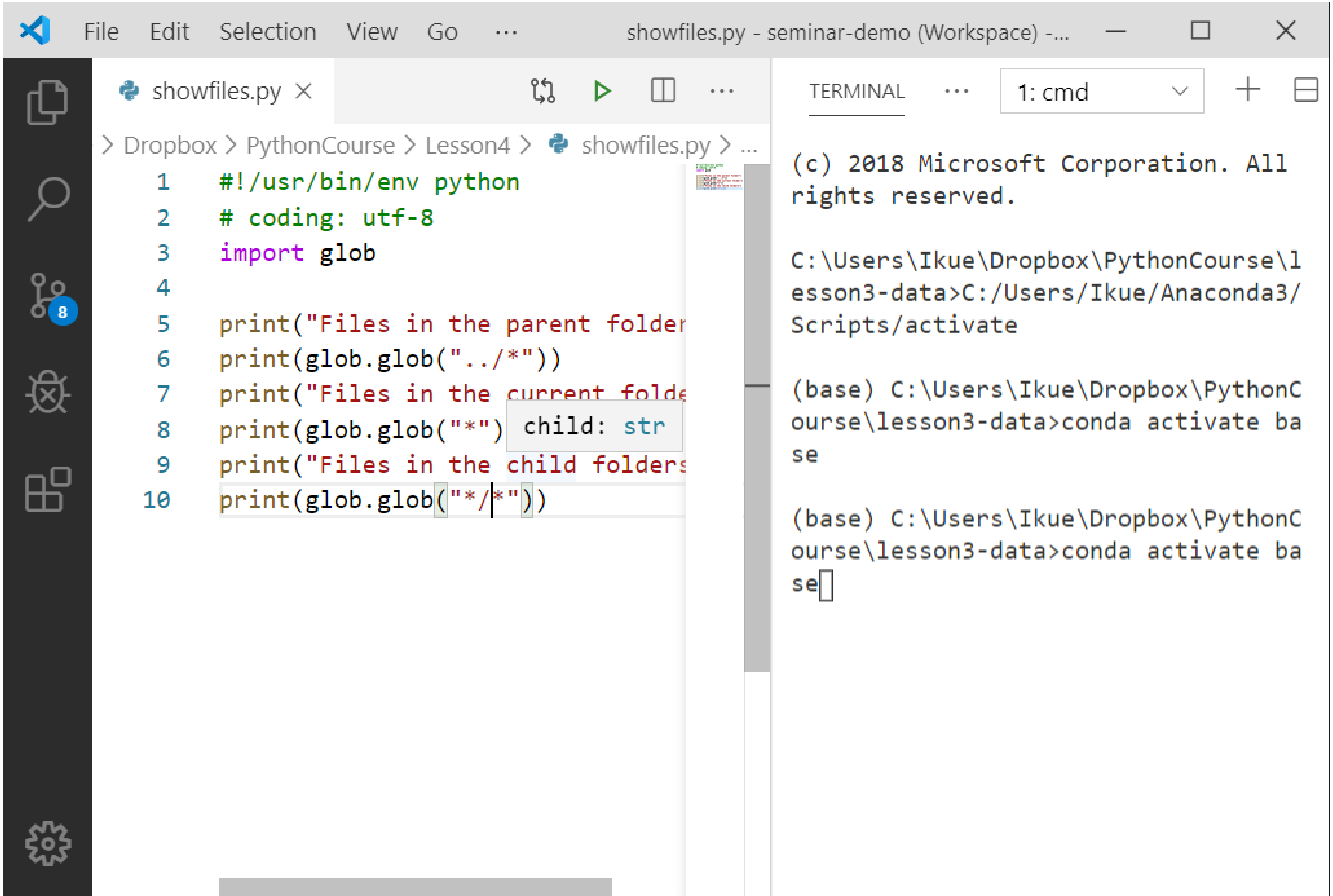
```
(base) C:¥Users¥Ikue¥Dropbox¥PythonCourse¥Lesson4>dir
 ドライブ C のボリューム ラベルは          です
 ボリューム シリアル番号は          です

 C:¥Users¥Ikue¥Dropbox¥PythonCourse¥Lesson4 のディレクトリ

2020/01/23  17:05    <DIR>          .
2020/01/23  17:05    <DIR>          ..
2020/01/23  17:06    <DIR>          dummyfiles1
2020/01/23  18:42               239 showfiles.py
               1 個のファイル                 239 バイト
               3 個のディレクトリ  105,405,521,920 バイトの空き領域

(base) C:¥Users¥Ikue¥Dropbox¥PythonCourse¥Lesson4>_
```

Terminal in VS Code is similar



File
Edit
Selection
View
Go
···
showfiles.py - seminar-demo (Workspace) -...
showfiles.py
Dropbox > PythonCourse > Lesson4 > showfiles.py > ...
TERMINAL
1: cmd
8

```python
#!/usr/bin/env python
# coding: utf-8
import glob

print("Files in the parent folder
print(glob.glob("../*"))
print("Files in the current folde
print(glob.glob("*")
print("Files in the child folders
print(glob.glob("*/*"))
```

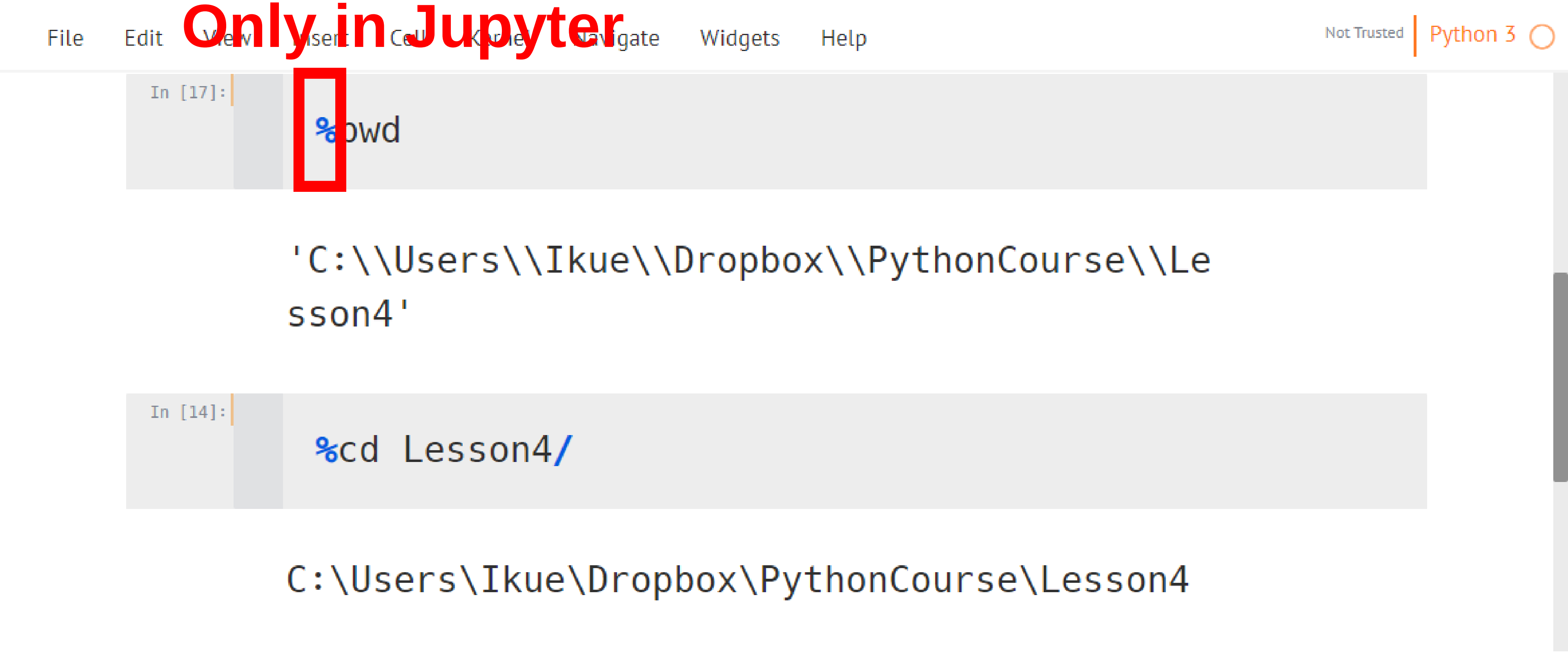
```
(c) 2018 Microsoft Corporation. All
rights reserved.

C:\Users\Ikue\Dropbox\PythonCourse\l
esson3-data>C:/Users/Ikue/Anaconda3/
Scripts/activate

(base) C:\Users\Ikue\Dropbox\PythonC
ourse\lesson3-data>conda activate ba
se

(base) C:\Users\Ikue\Dropbox\PythonC
ourse\lesson3-data>conda activate ba
se
```

(optional) from Jupyter Notebook



Only in Jupyter

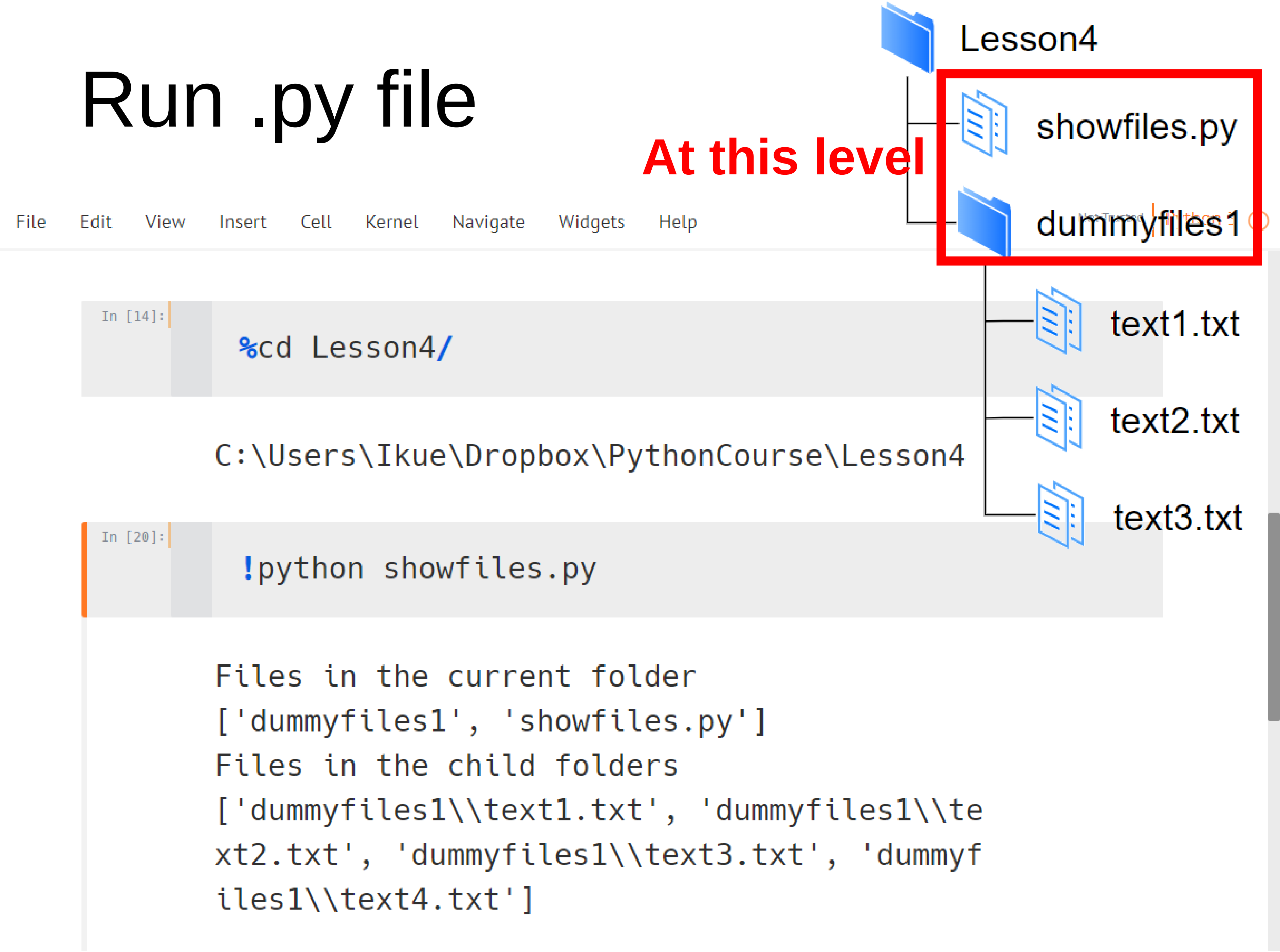
```
In [17]: %pwd
```

```
'C:\\Users\\Ikue\\Dropbox\\PythonCourse\\Le
sson4'
```

```
In [14]: %cd Lesson4/
```

```
C:\Users\Ikue\Dropbox\PythonCourse\Lesson4
```

Run .py file



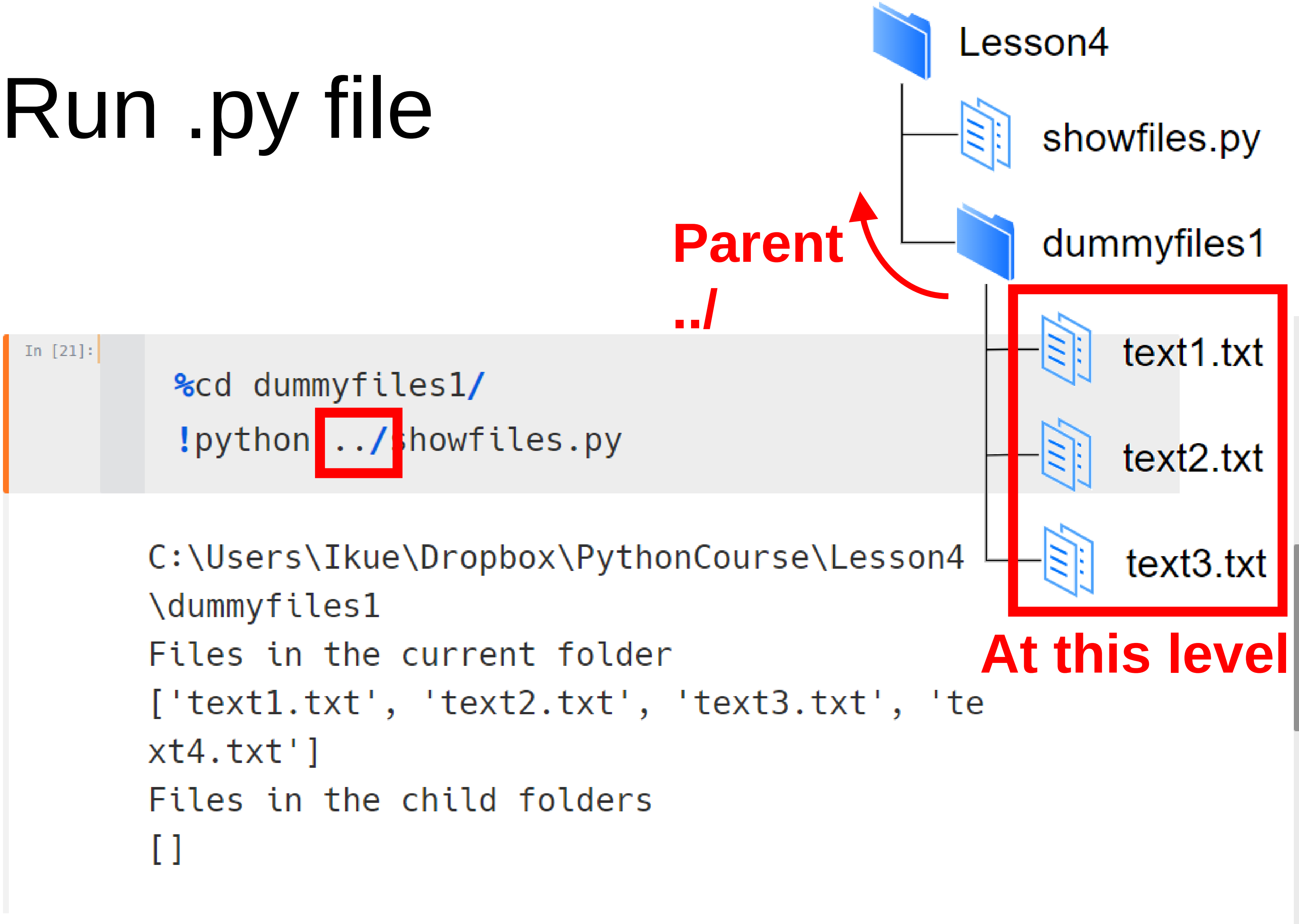
```
%cd Lesson4/
```

```
C:\Users\Ikue\Dropbox\PythonCourse\Lesson4
```

```
!python showfiles.py
```

```
Files in the current folder
['dummyfiles1', 'showfiles.py']
Files in the child folders
['dummyfiles1\\text1.txt', 'dummyfiles1\\te
xt2.txt', 'dummyfiles1\\text3.txt', 'dummyf
iles1\\text4.txt']
```

Run .py file

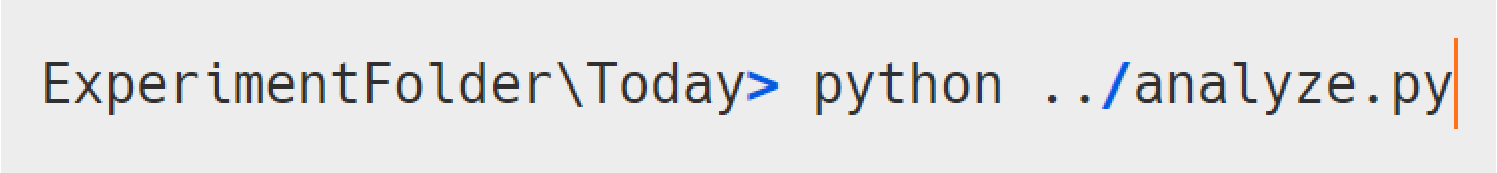


In [21]:

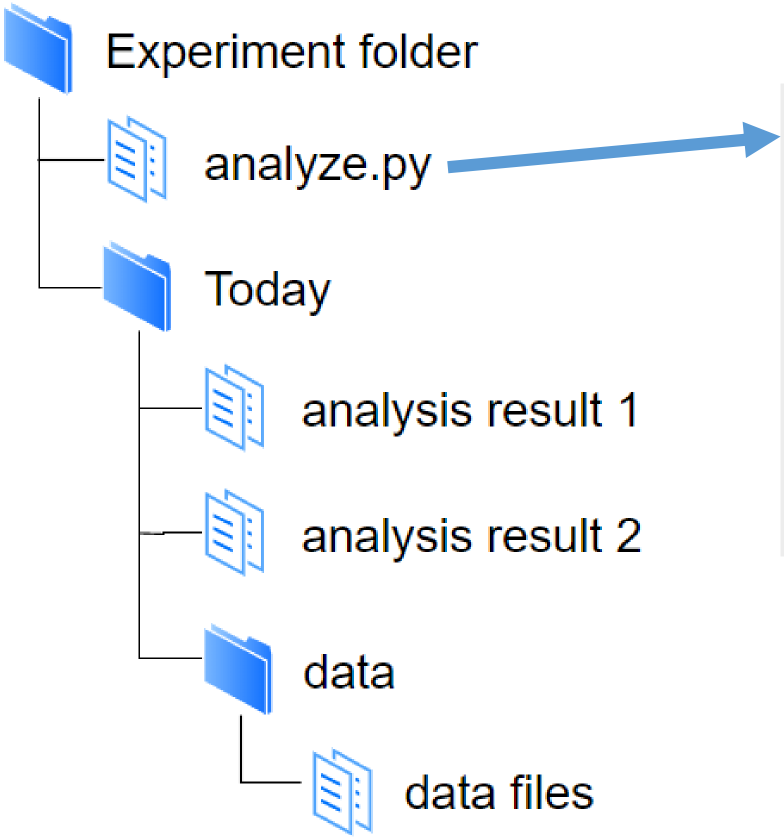
```
%cd dummyfiles1/
!python ../showfiles.py
```

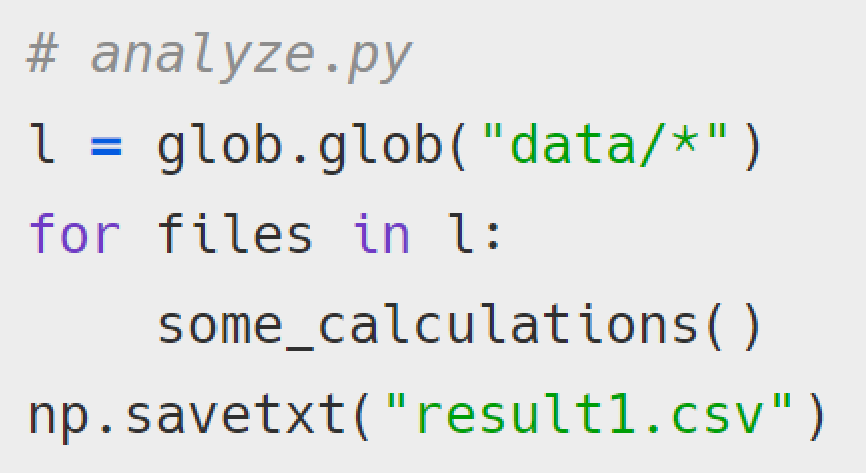
```
C:\Users\Ikue\Dropbox\PythonCourse\Lesson4
\dummyfiles1
Files in the current folder
['text1.txt', 'text2.txt', 'text3.txt', 'te
xt4.txt']
Files in the child folders
[]
```

In practical case



```
ExperimentFolder\Today> python ../analyze.py
```





```
# analyze.py
l = glob.glob("data/*")
for files in l:
    some_calculations()
np.savetxt("result1.csv")
```

UNIX commands

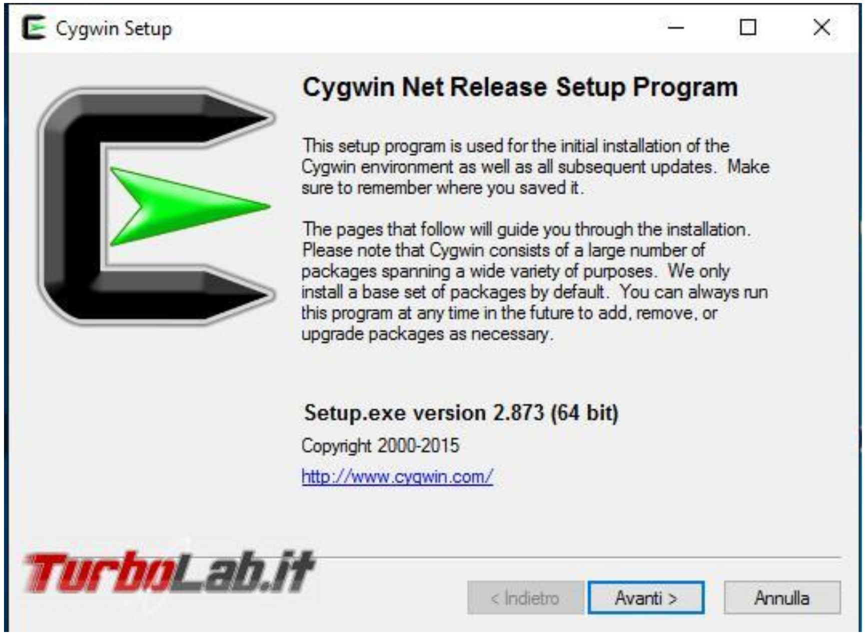
Function	UNIX	Python
show current path	pwd	os.getcwd()
change directory	cd	os.chdir()
copy file(s)	cp	shutil.copyfile()

mv file(s)	mv	os.rename()
		shutil.move()

make directory(ies)	mkdir	os.mkdir()
search words	grep	glob.glob() + find()
replace words	sed awk	re.sub() …?

Run UNIX commands on Windows

Option 1: Cygwin (and other shells)



Cygwin Net Release Setup Program

This setup program is used for the initial installation of the Cygwin environment as well as all subsequent updates. Make sure to remember where you saved it.

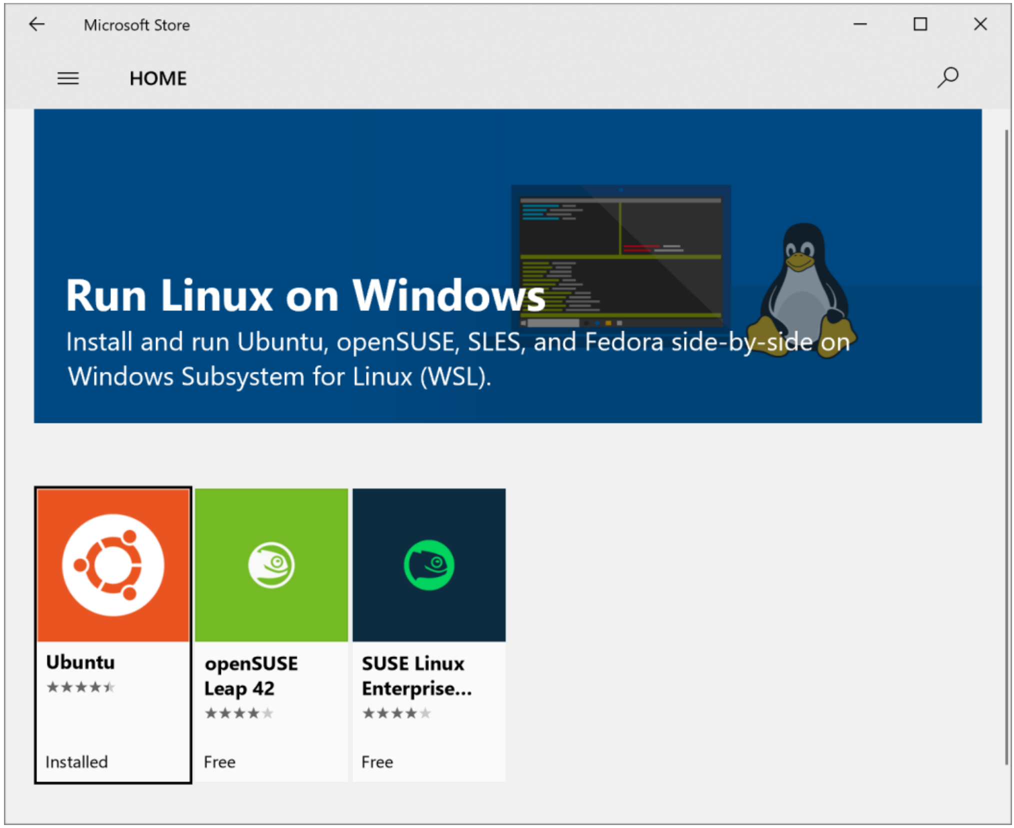
The pages that follow will guide you through the installation. Please note that Cygwin consists of a large number of packages spanning a wide variety of purposes. We only install a base set of packages by default. You can always run this program at any time in the future to add, remove, or upgrade packages as necessary.

Setup.exe version 2.873 (64 bit)

Copyright 2000-2015

http://www.cygwin.com/

Option 2: WSL

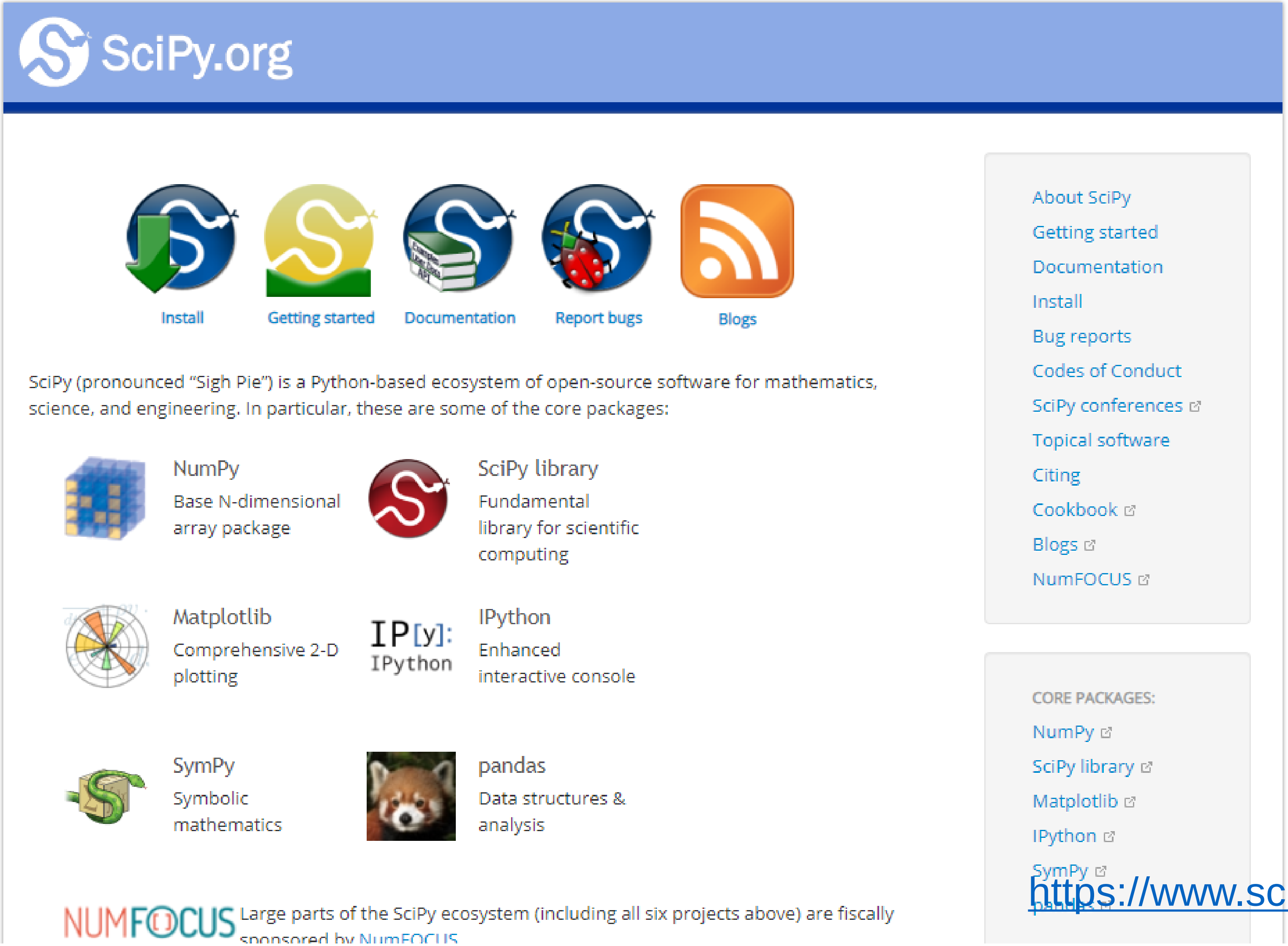


Run Linux on Windows

Install and run Ubuntu, openSUSE, SLES, and Fedora side-by-side on Windows Subsystem for Linux (WSL).

More in Appendix

SciPy package family



SciPy (pronounced "Sigh Pie") is a Python-based ecosystem of open-source software for mathematics, science, and engineering. In particular, these are some of the core packages:

NumPy
Base N-dimensional array package

SciPy library
Fundamental library for scientific computing

Matplotlib
Comprehensive 2-D plotting

IPython
Enhanced interactive console

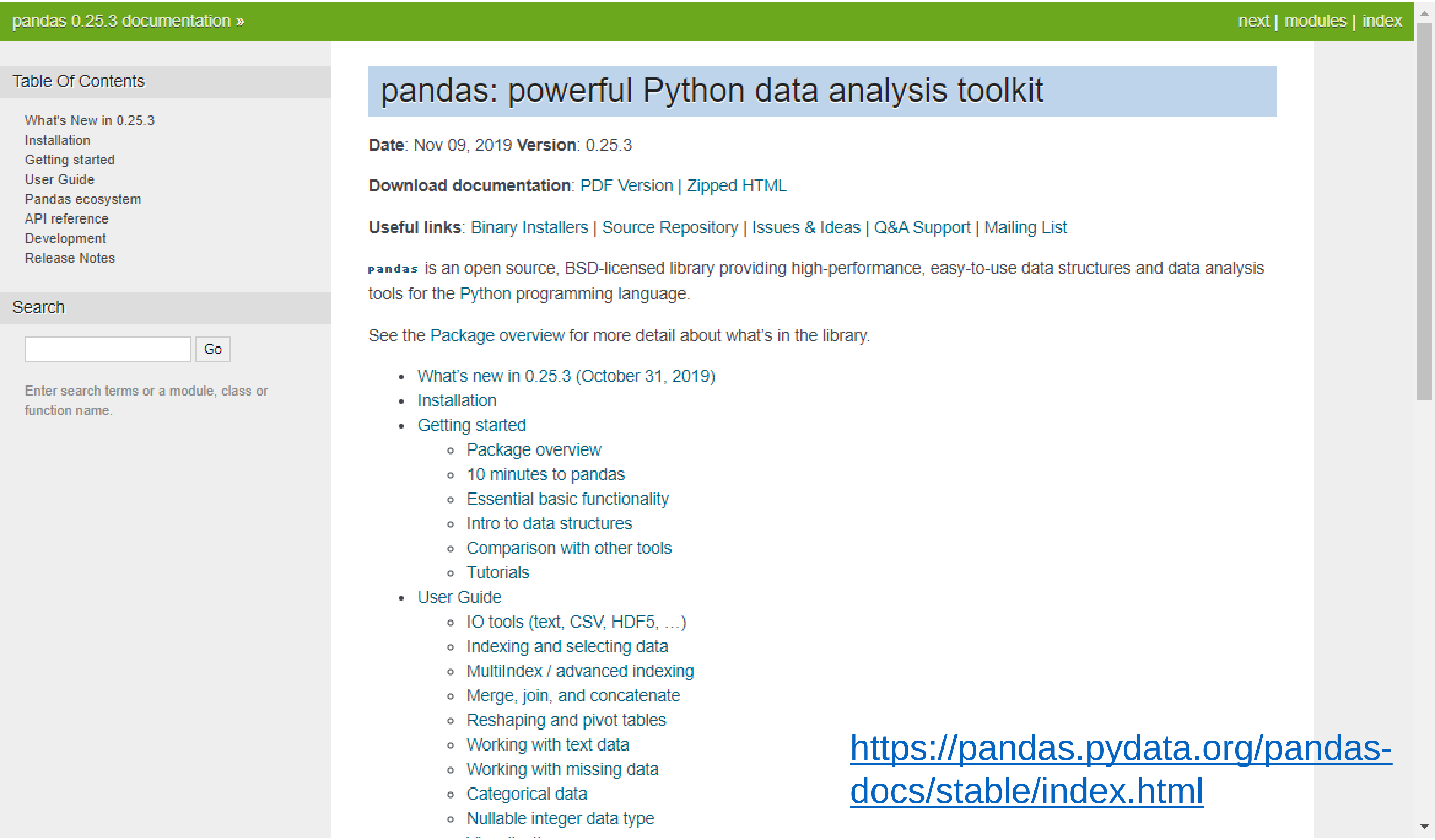
SymPy
Symbolic mathematics

pandas
Data structures & analysis

Large parts of the SciPy ecosystem (including all six projects above) are fiscally sponsored by NumFOCUS

https://www.scipy.org/

pandas: database handling



pandas 0.25.3 documentation »
next | modules | index

Table Of Contents

Search
Go
Enter search terms or a module, class or function name.

pandas: powerful Python data analysis toolkit

Date: Nov 09, 2019 **Version**: 0.25.3

Download documentation: PDF Version | Zipped HTML

Useful links: Binary Installers | Source Repository | Issues & Ideas | Q&A Support | Mailing List

`pandas` is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

See the Package overview for more detail about what's in the library.

- What's new in 0.25.3 (October 31, 2019)
- Installation
- Getting started
  - Package overview
  - 10 minutes to pandas
  - Essential basic functionality
  - Intro to data structures
  - Comparison with other tools
  - Tutorials
- User Guide
  - IO tools (text, CSV, HDF5, …)
  - Indexing and selecting data
  - MultiIndex / advanced indexing
  - Merge, join, and concatenate
  - Reshaping and pivot tables
  - Working with text data
  - Working with missing data
  - Categorical data
  - Nullable integer data type

https://pandas.pydata.org/pandas-docs/stable/index.html

Benefit of pandas

numpy: for numerical array, like x-y

loadtxt() - no missing data

genfromtext() - with missing data

column name -> tuple (cannot access by index)

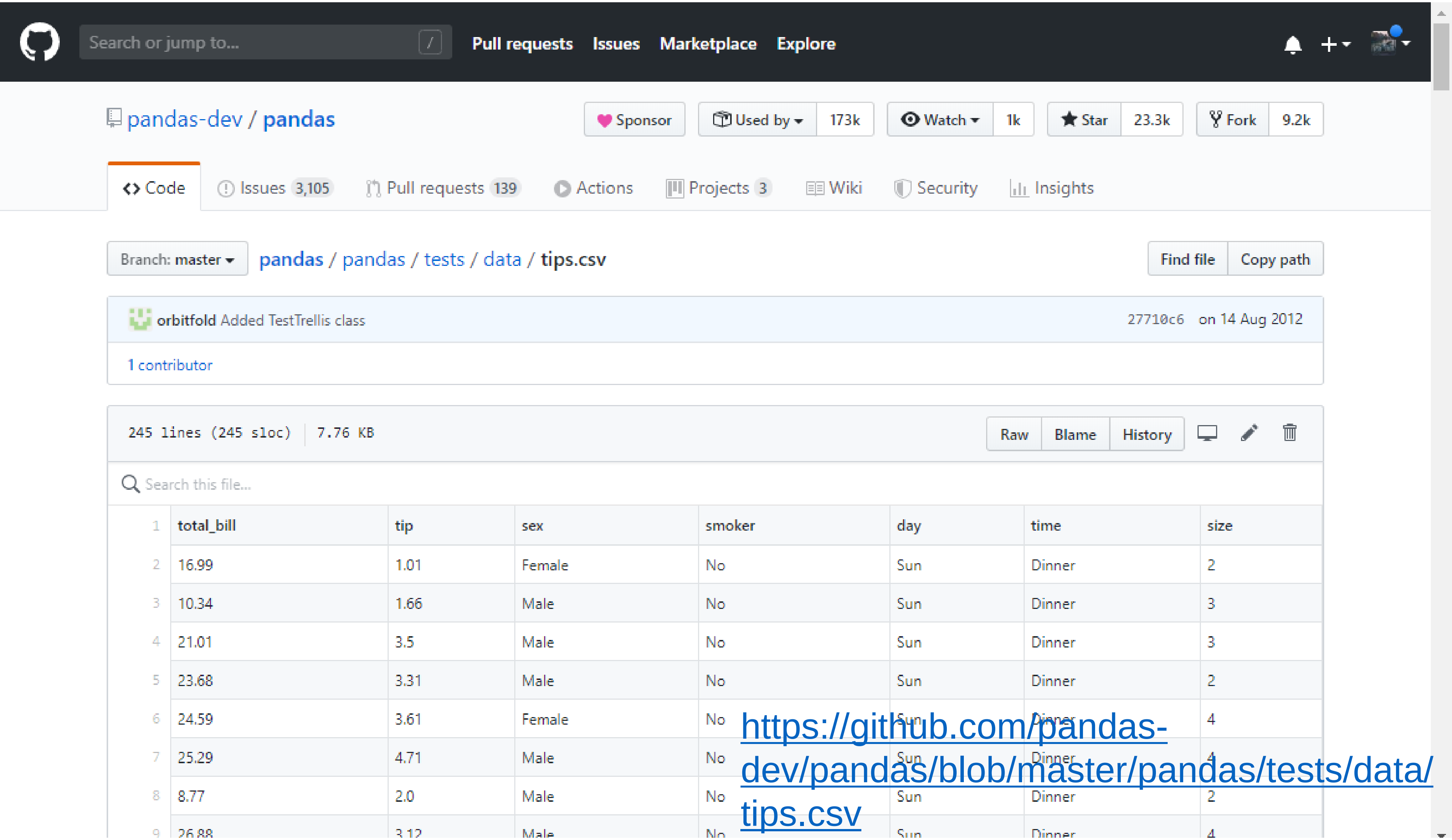
pandas: for data set, including `str`

load_csv() accessible by index and/or label

generic column name

stronger for statistics

Sample data



pandas-dev / pandas

pandas / pandas / tests / data / tips.csv

orbitfold Added TestTrellis class 27710c6 on 14 Aug 2012

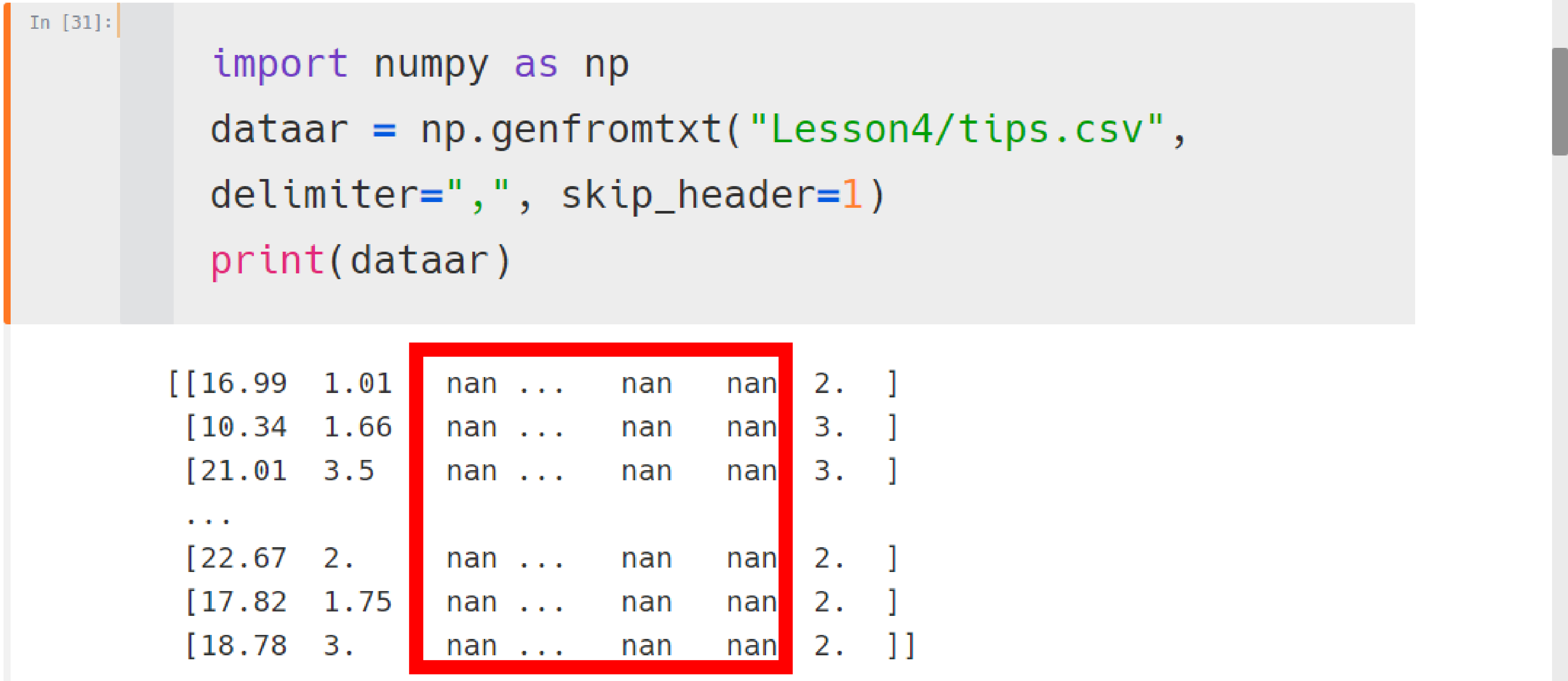
1 contributor

245 lines (245 sloc) 7.76 KB

	total_bill	tip	sex	smoker	day	time	size
2	16.99	1.01	Female	No	Sun	Dinner	2
3	10.34	1.66	Male	No	Sun	Dinner	3
4	21.01	3.5	Male	No	Sun	Dinner	3
5	23.68	3.31	Male	No	Sun	Dinner	2
6	24.59	3.61	Female	No	Sun	Dinner	4
7	25.29	4.71	Male	No	Sun	Dinner	4
8	8.77	2.0	Male	No	Sun	Dinner	2
9	26.88	3.12	Male	No	Sun	Dinner	4

https://github.com/pandas-dev/pandas/blob/master/pandas/tests/data/tips.csv

By Numpy

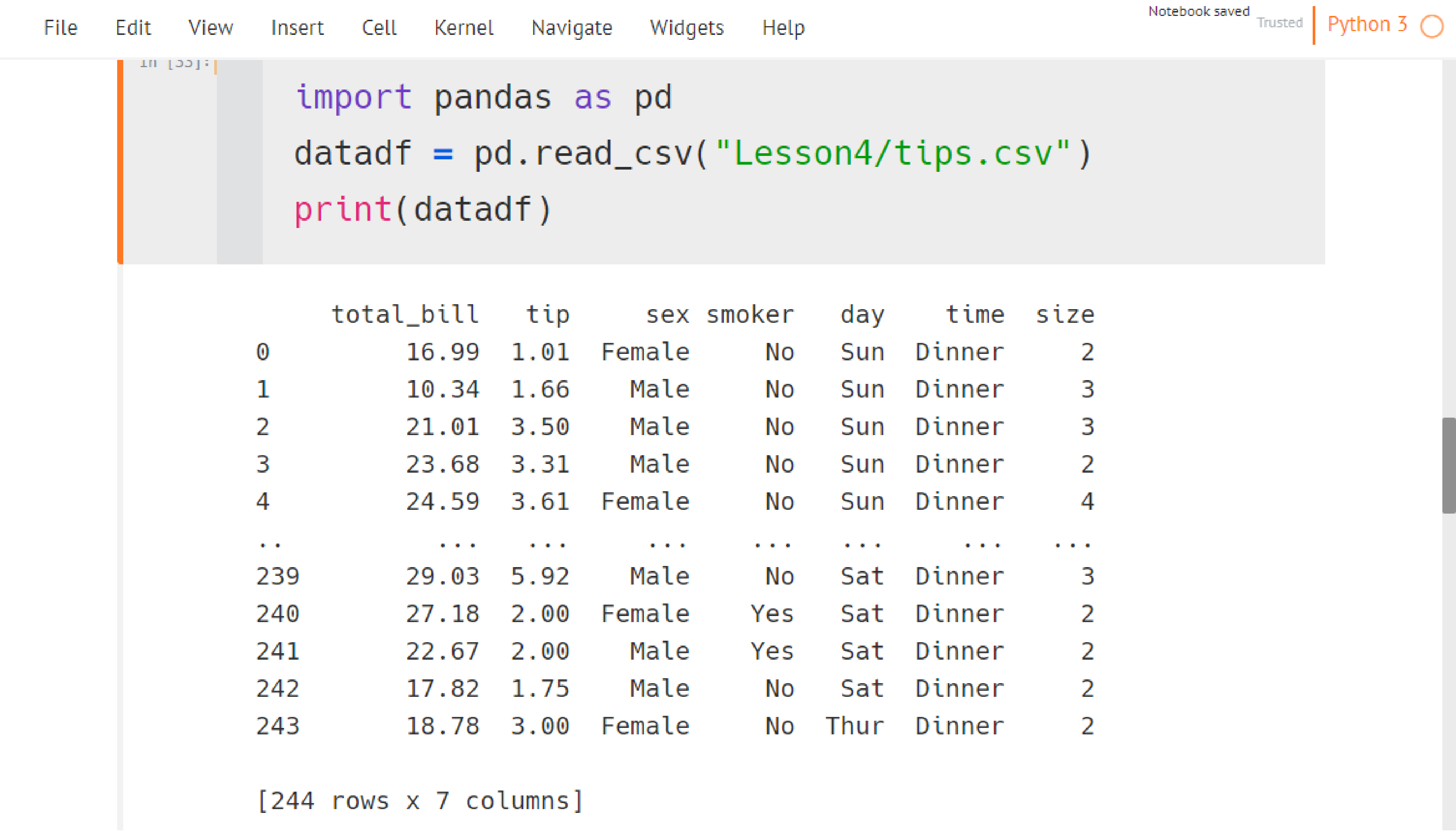


```python
import numpy as np
dataar = np.genfromtxt("Lesson4/tips.csv",
delimiter=",", skip_header=1)
print(dataar)
```

```
[[16.99  1.01   nan ...   nan   nan  2.  ]
 [10.34  1.66   nan ...   nan   nan  3.  ]
 [21.01  3.5    nan ...   nan   nan  3.  ]
 ...
 [22.67  2.     nan ...   nan   nan  2.  ]
 [17.82  1.75   nan ...   nan   nan  2.  ]
 [18.78  3.     nan ...   nan   nan  2.  ]]
```

Text data lost
(needs special treatment)

By pandas

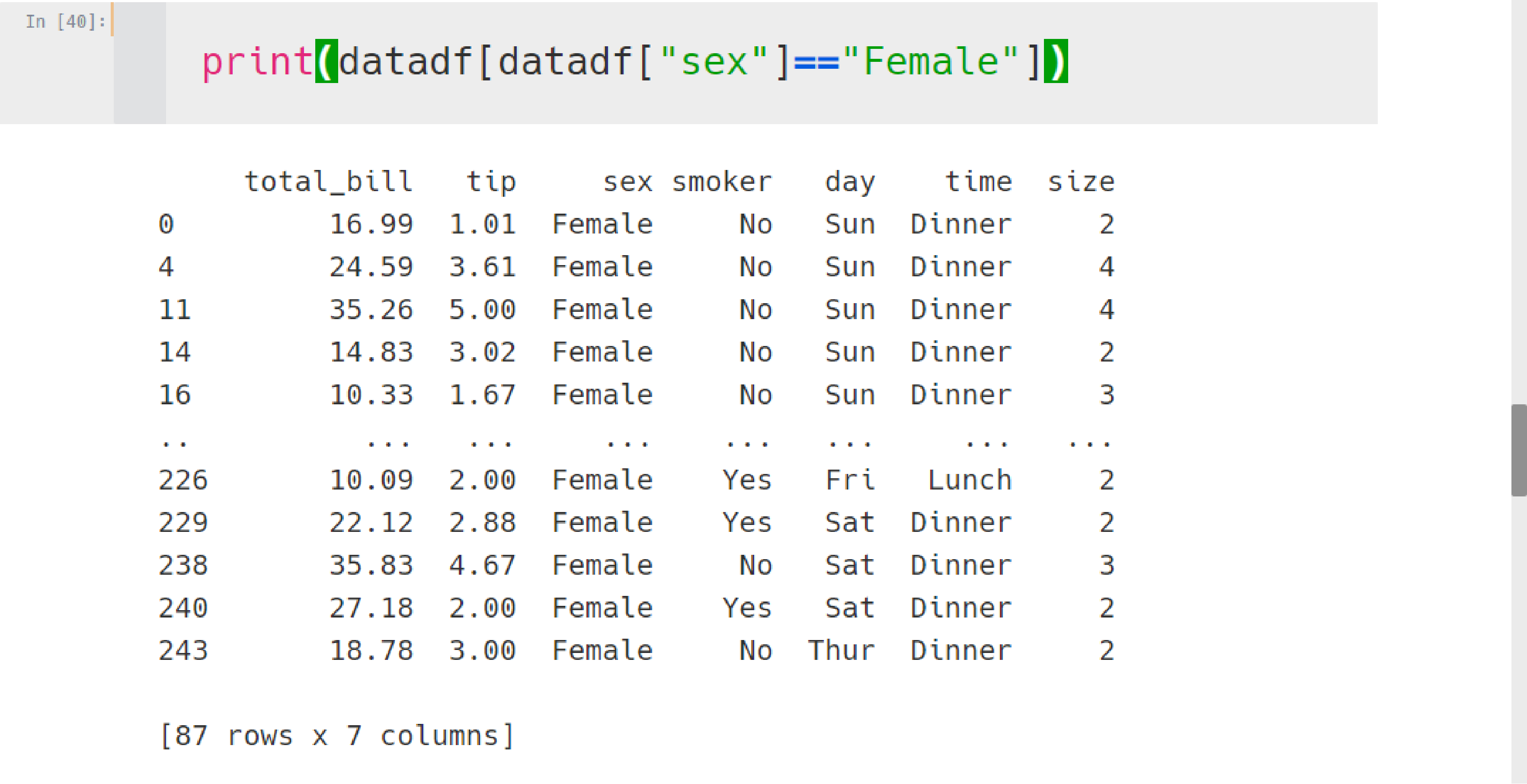


```python
import pandas as pd
datadf = pd.read_csv("Lesson4/tips.csv")
print(datadf)
```

```
     total_bill   tip     sex smoker   day    time  size
0         16.99  1.01  Female     No   Sun  Dinner     2
1         10.34  1.66    Male     No   Sun  Dinner     3
2         21.01  3.50    Male     No   Sun  Dinner     3
3         23.68  3.31    Male     No   Sun  Dinner     2
4         24.59  3.61  Female     No   Sun  Dinner     4
..          ...   ...     ...    ...   ...     ...   ...
239       29.03  5.92    Male     No   Sat  Dinner     3
240       27.18  2.00  Female    Yes   Sat  Dinner     2
241       22.67  2.00    Male    Yes   Sat  Dinner     2
242       17.82  1.75    Male     No   Sat  Dinner     2
243       18.78  3.00  Female     No  Thur  Dinner     2

[244 rows x 7 columns]
```

Partial data by conditions



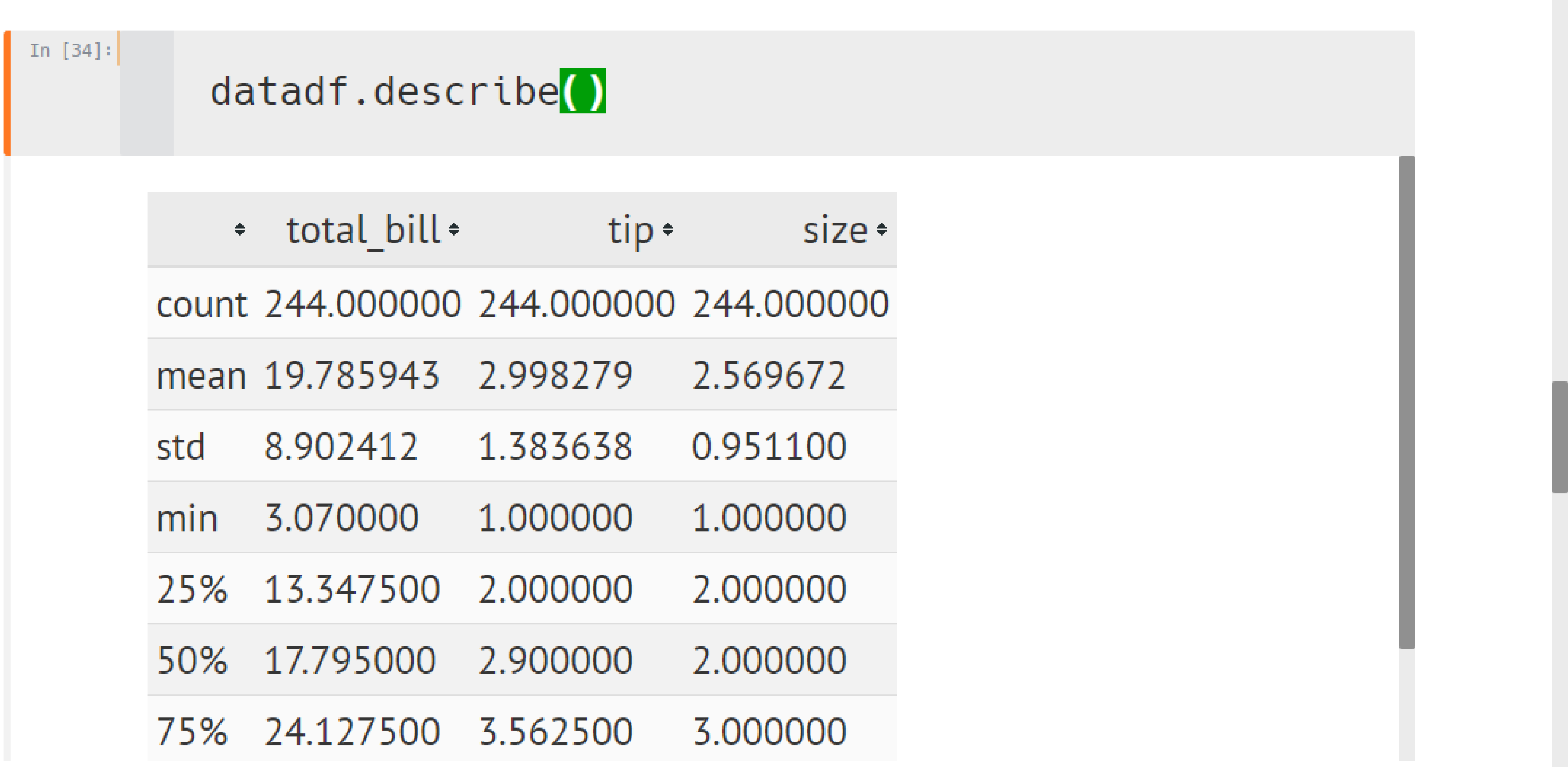
In [40]:

```
print(datadf[datadf["sex"]=="Female"])
```

```
     total_bill   tip     sex smoker   day    time  size
0         16.99  1.01  Female     No   Sun  Dinner     2
4         24.59  3.61  Female     No   Sun  Dinner     4
11        35.26  5.00  Female     No   Sun  Dinner     4
14        14.83  3.02  Female     No   Sun  Dinner     2
16        10.33  1.67  Female     No   Sun  Dinner     3
..          ...   ...     ...    ...   ...     ...   ...
226       10.09  2.00  Female    Yes   Fri   Lunch     2
229       22.12  2.88  Female    Yes   Sat  Dinner     2
238       35.83  4.67  Female     No   Sat  Dinner     3
240       27.18  2.00  Female    Yes   Sat  Dinner     2
243       18.78  3.00  Female     No  Thur  Dinner     2

[87 rows x 7 columns]
```

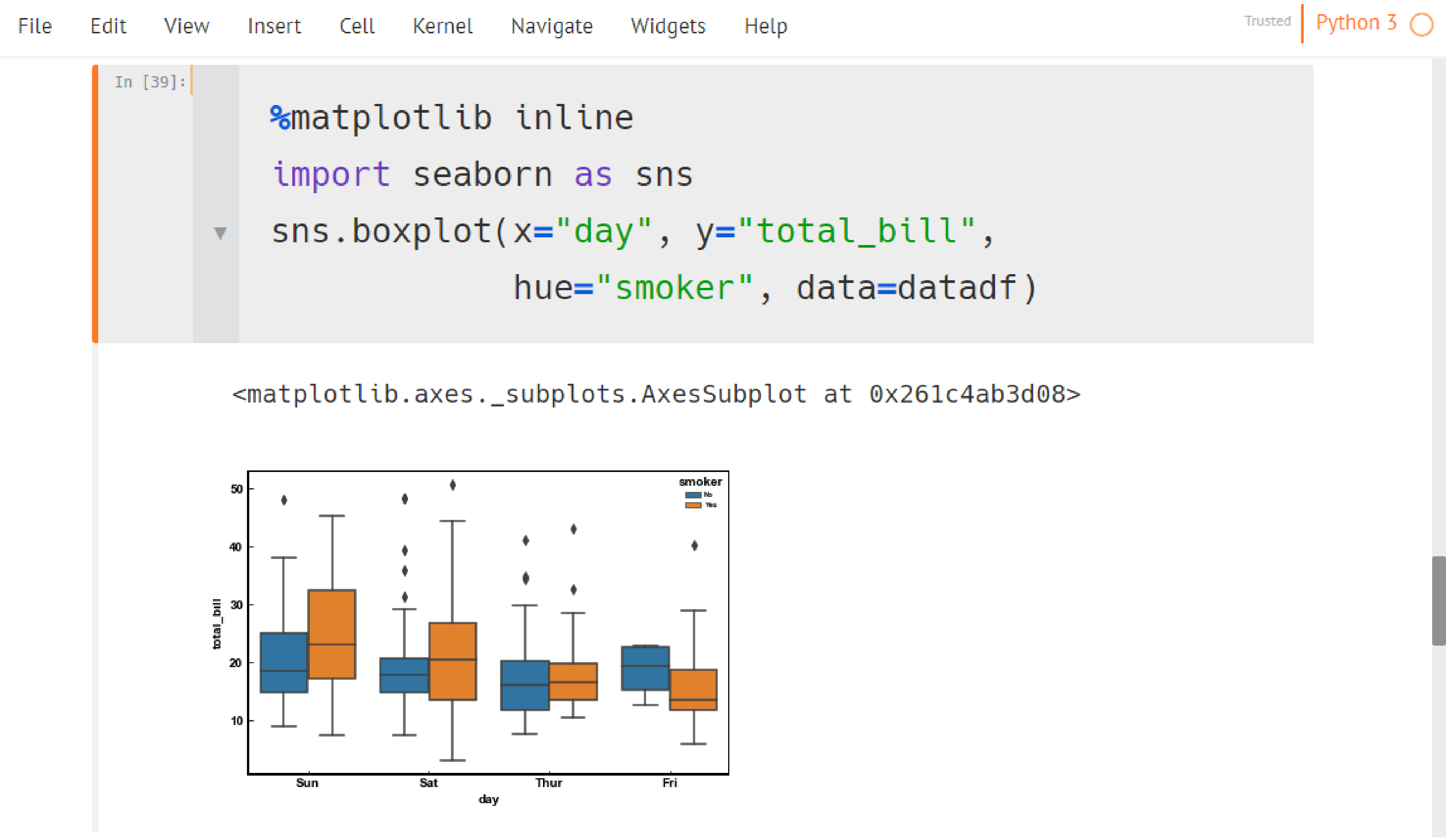
Some statistics



```
In [34]: datadf.describe()
```

	total_bill	tip	size
count	244.000000	244.000000	244.000000
mean	19.785943	2.998279	2.569672
std	8.902412	1.383638	0.951100
min	3.070000	1.000000	1.000000
25%	13.347500	2.000000	2.000000
50%	17.795000	2.900000	2.000000
75%	24.127500	3.562500	3.000000

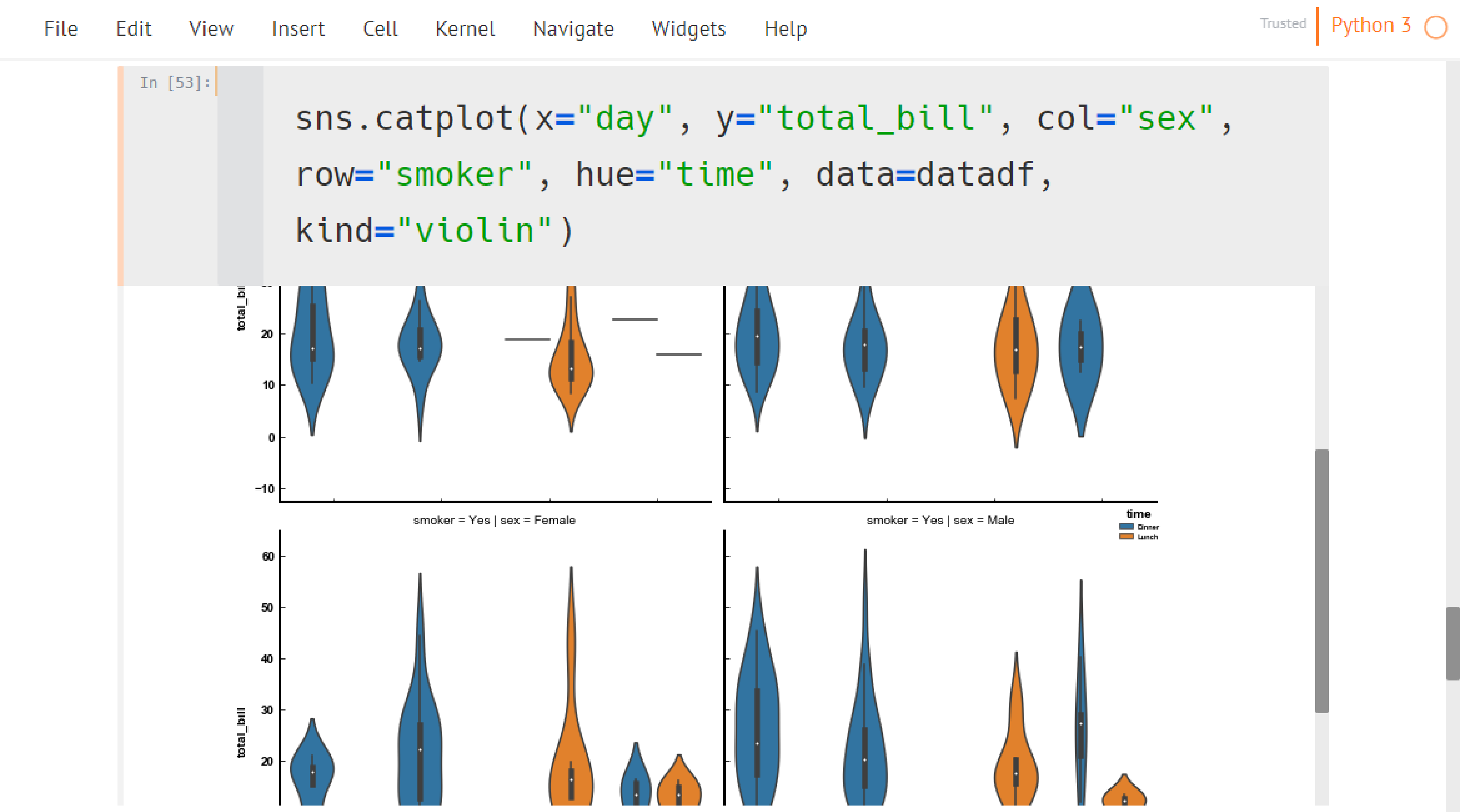
Plotting by seaborn



```
%matplotlib inline
import seaborn as sns
sns.boxplot(x="day", y="total_bill",
            hue="smoker", data=datadf)
```

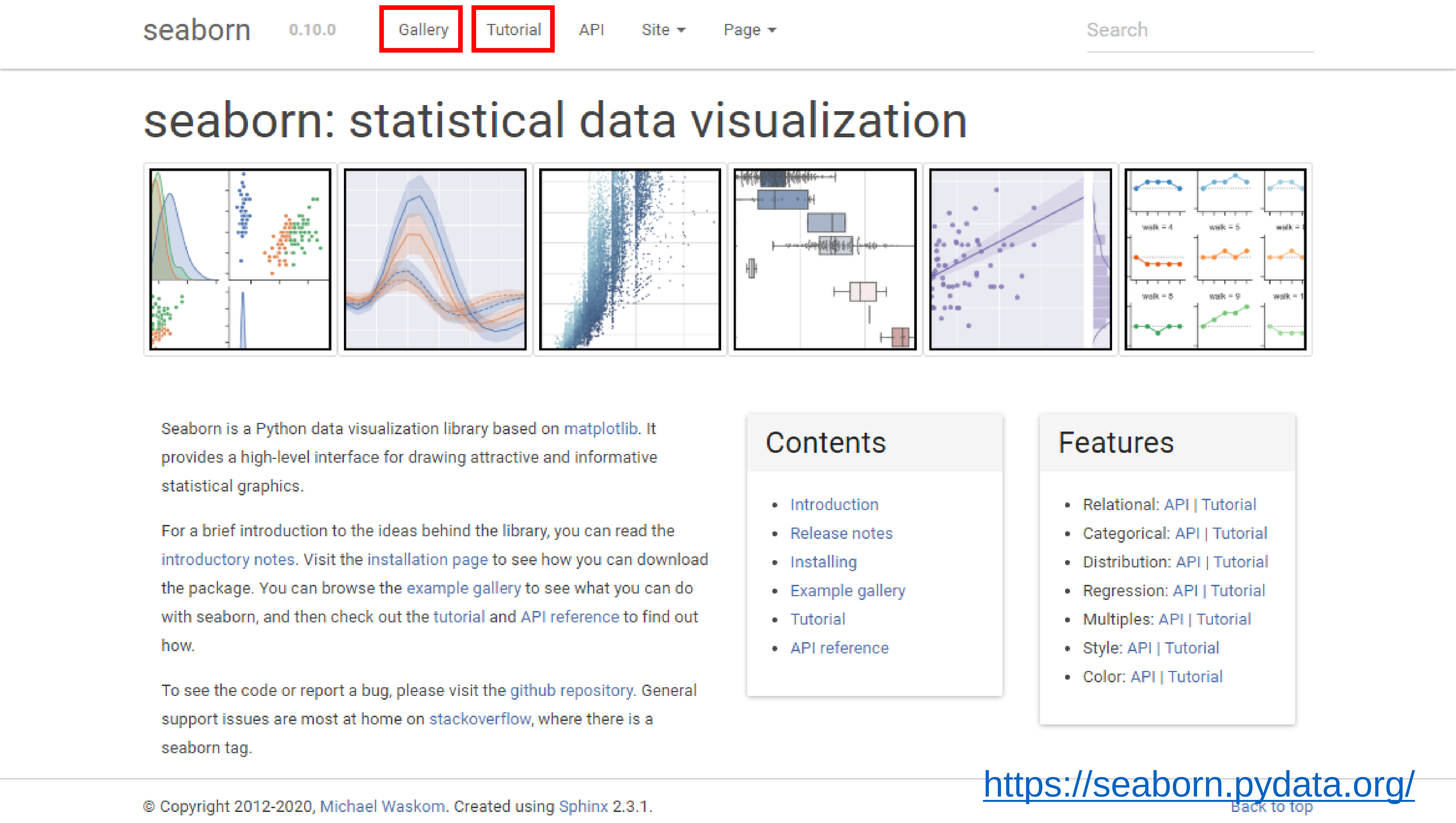
```
<matplotlib.axes._subplots.AxesSubplot at 0x261c4ab3d08>
```

More complex plot



```
sns.catplot(x="day", y="total_bill", col="sex",
row="smoker", hue="time", data=datadf,
kind="violin")
```

More about seaborn



seaborn: statistical data visualization

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

For a brief introduction to the ideas behind the library, you can read the introductory notes. Visit the installation page to see how you can download the package. You can browse the example gallery to see what you can do with seaborn, and then check out the tutorial and API reference to find out how.

To see the code or report a bug, please visit the github repository. General support issues are most at home on stackoverflow, where there is a seaborn tag.

Contents

- Introduction
- Release notes
- Installing
- Example gallery
- Tutorial
- API reference

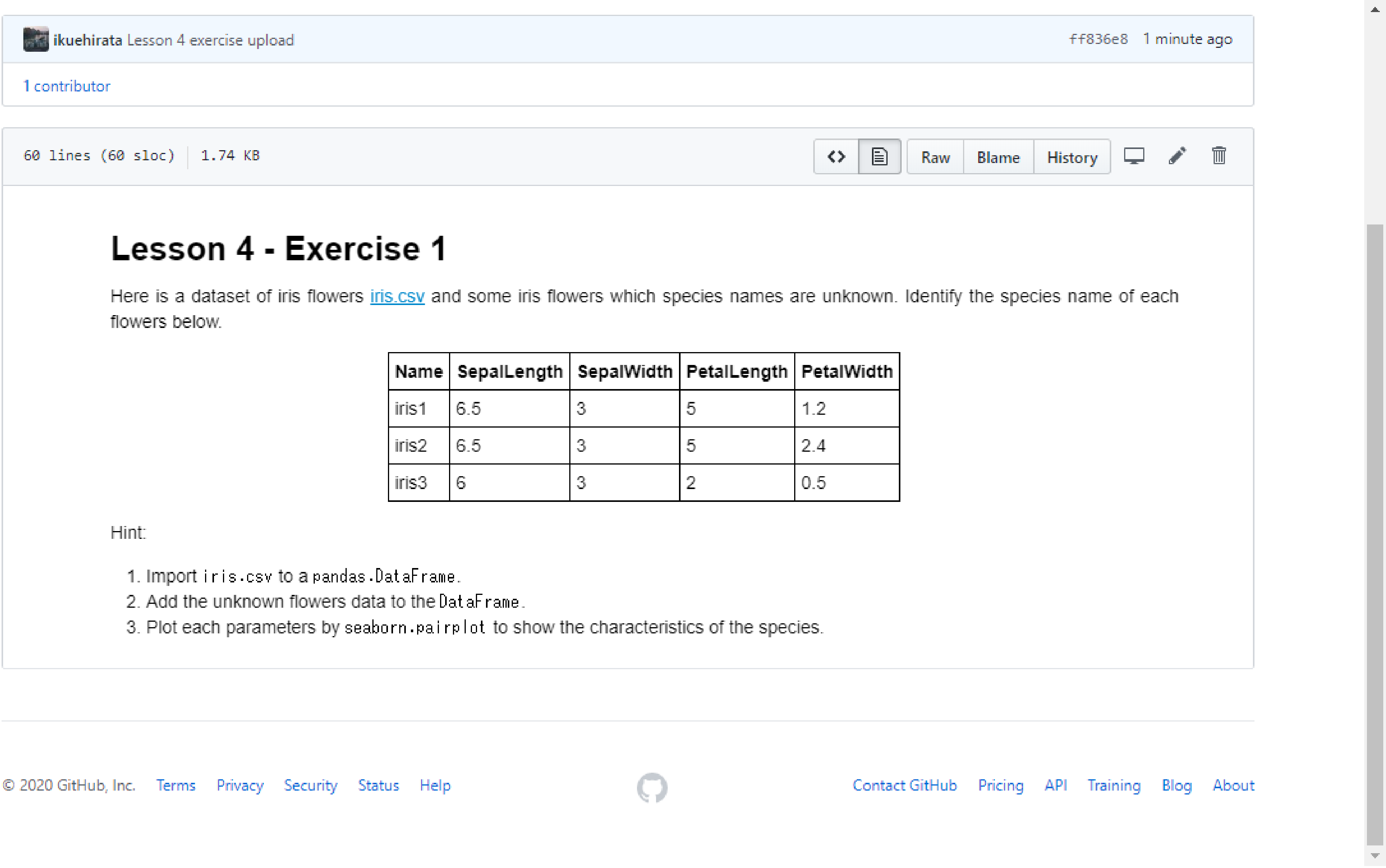
Features

- Relational: API | Tutorial
- Categorical: API | Tutorial
- Distribution: API | Tutorial
- Regression: API | Tutorial
- Multiples: API | Tutorial
- Style: API | Tutorial
- Color: API | Tutorial

Back to top

https://seaborn.pydata.org/

Exercise 1


ikuehirata Lesson 4 exercise upload
ff836e8 1 minute ago
1 contributor
60 lines (60 sloc) | 1.74 KB
Raw
Blame
History

Lesson 4 - Exercise 1

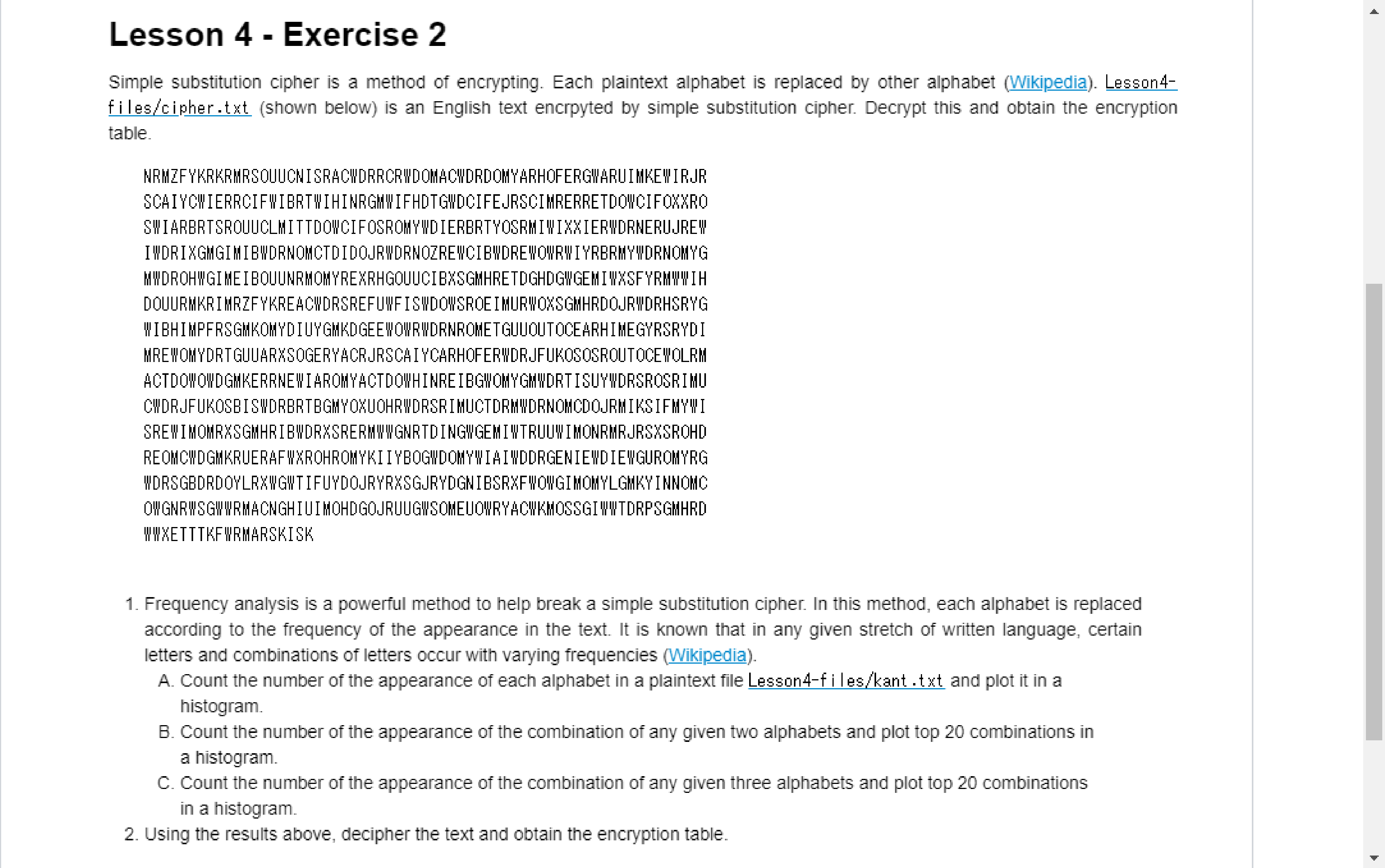
Here is a dataset of iris flowers iris.csv and some iris flowers which species names are unknown. Identify the species name of each flowers below.

Name	SepalLength	SepalWidth	PetalLength	PetalWidth
iris1	6.5	3	5	1.2
iris2	6.5	3	5	2.4
iris3	6	3	2	0.5

Hint:

1. Import `iris.csv` to a `pandas.DataFrame`.
2. Add the unknown flowers data to the `DataFrame`.
3. Plot each parameters by `seaborn.pairplot` to show the characteristics of the species.

Exercise 2 (difficult!)



Lesson 4 - Exercise 2

Simple substitution cipher is a method of encrypting. Each plaintext alphabet is replaced by other alphabet (Wikipedia). `Lesson4-files/cipher.txt` (shown below) is an English text encrpyted by simple substitution cipher. Decrypt this and obtain the encryption table.

```
NRMZFYKRKRMRSOUUCNISRACWDRRCRWDOMACWDRDOMYARHOFERGWARUIMKEWIRJR
SCAIYCWIERRCIFWIBRTWIHINRGMWIFHDTGWDCIFEJRSCIMRERRETDOWCIFOXXRO
SWIARBRTSROUUCLMITTDOWCIFOSROMYWDIERBRTYOSRMIWIXXIERWDRNERUJREW
IWDRIXGMGIMIBWDRNOMCTDIDOJRWDRNOZREWCIBWDREWOWRWIYRBRMYWDRNOMYG
MWDROHWGIMEIBOUUNRMOMYREXRHGOUUCIBXSGMHRETDGHDGWGEMIWXSFYRMWWIH
DOUURMKRIMRZFYKREACWDRSREFUWFISWDOWSROEIMURWOXSGMHRDOJRWDRHSRYG
WIBHIMPFRSGMKOMYDIUYGMKDGEEWOWRWDRNROMETGUUOUTOCEARHIMEGYRSRYDI
MREWOMYDRTGUUARXSOGERYACRJRSCAIYCARHOFERWDRJFUKOSOSROUTOCEWOLRM
ACTDOWOWDGMKERRNEWIAROMYACTDOWHINREIBGWOMYGMWDRTISUYWDRSROSRIMU
CWDRJFUKOSBISWDRBRTBGMYOXUOHRWDRSRIMUCTDRMWDRNOMCDOJRMIKSIFMYWI
SREWIMOMRXSGMHRIBWDRXSRERMWWGNRTDINGWGEMIWTRUUWIMONRMRJRSXSROHD
REOMCWDGMKRUERAFWXROHROMYKIIYBOGWDOMYWIAIWDDRGENIEWDIEWGUROMYRG
WDRSGBDRDOYLRXWGWTIFUYDOJRYRXSGJRYDGNIBSRXFWOWGIMOMYLGMKYINNOMC
OWGNRWSGWWRMACNGHIUIMOHDGOJRUUGWSOMEUOWRYACWKMOSSGIWWTDRPSGMHRD
WWXETTTKFWRMARSKISK
```

1. Frequency analysis is a powerful method to help break a simple substitution cipher. In this method, each alphabet is replaced according to the frequency of the appearance in the text. It is known that in any given stretch of written language, certain letters and combinations of letters occur with varying frequencies (Wikipedia).
    - A. Count the number of the appearance of each alphabet in a plaintext file `Lesson4-files/kant.txt` and plot it in a histogram.
    - B. Count the number of the appearance of the combination of any given two alphabets and plot top 20 combinations in a histogram.
    - C. Count the number of the appearance of the combination of any given three alphabets and plot top 20 combinations in a histogram.
2. Using the results above, decipher the text and obtain the encryption table.

Study by yourself

Pandas official document

https://pandas.pydata.org/pandas-docs/stable/index.html

Pandas unofficial tutorial

https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/

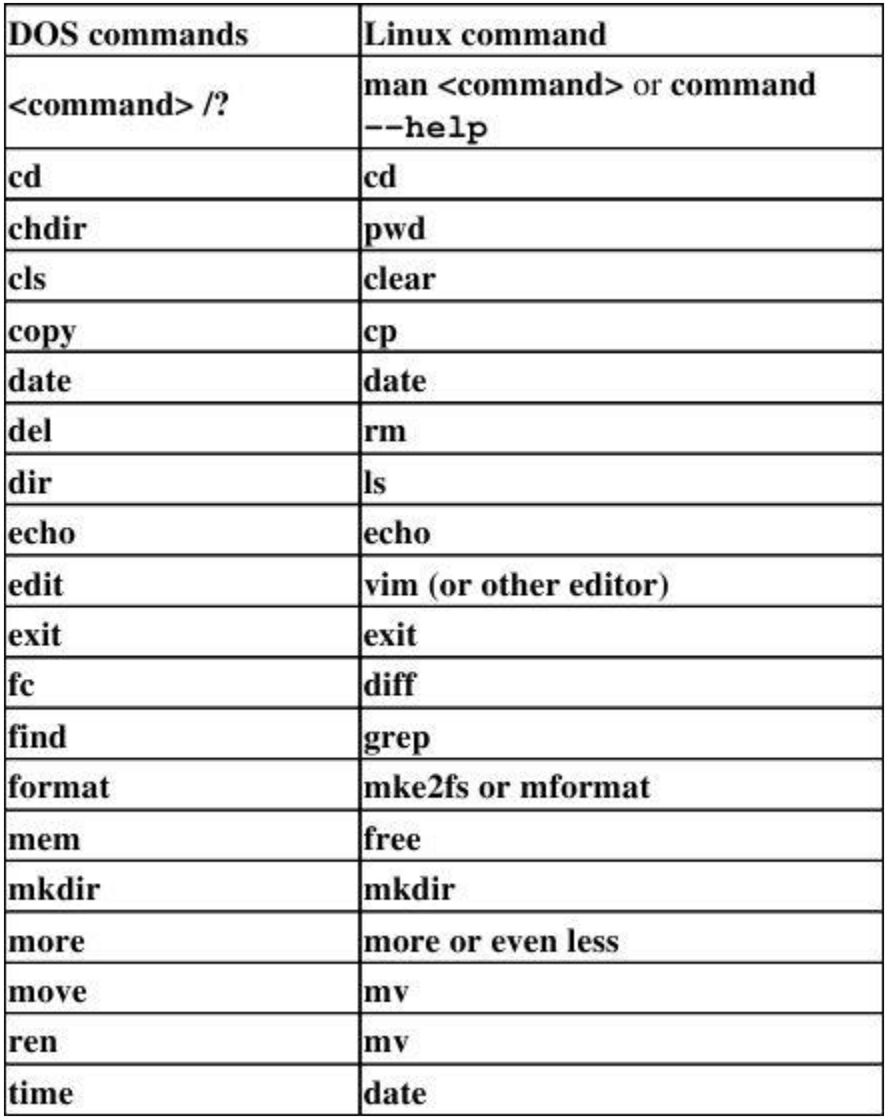
seaborn official document

https://seaborn.pydata.org/

Appendix for Windows users

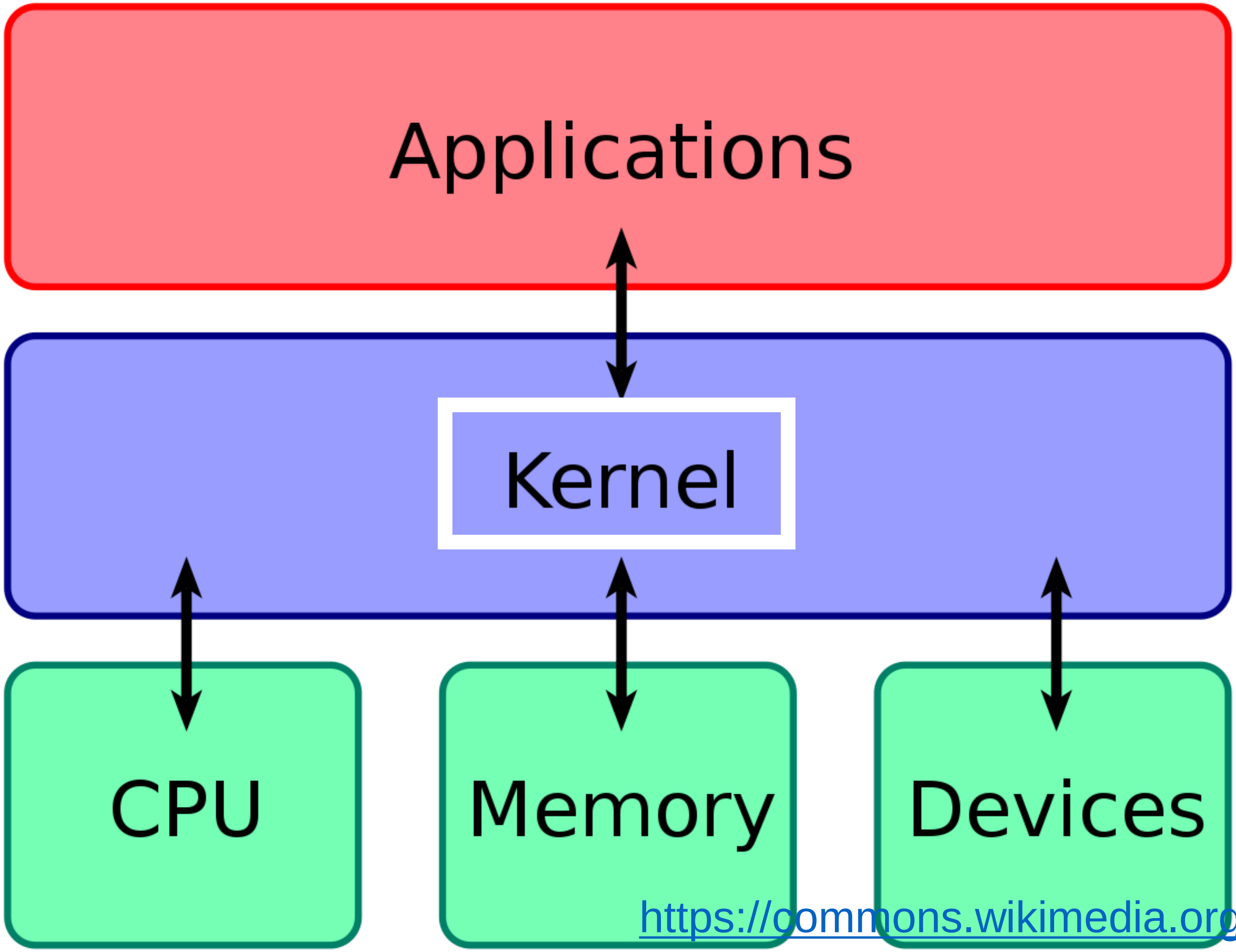
Terminals and shells

Why commands are different?



DOS commands	Linux command
<command> /?	man <command> or command `--help`
cd	cd
chdir	pwd
cls	clear
copy	cp
date	date
del	rm
dir	ls
echo	echo
edit	vim (or other editor)
exit	exit
fc	diff
find	grep
format	mke2fs or mformat
mem	free
mkdir	mkdir
more	more or even less
move	mv
ren	mv
time	date

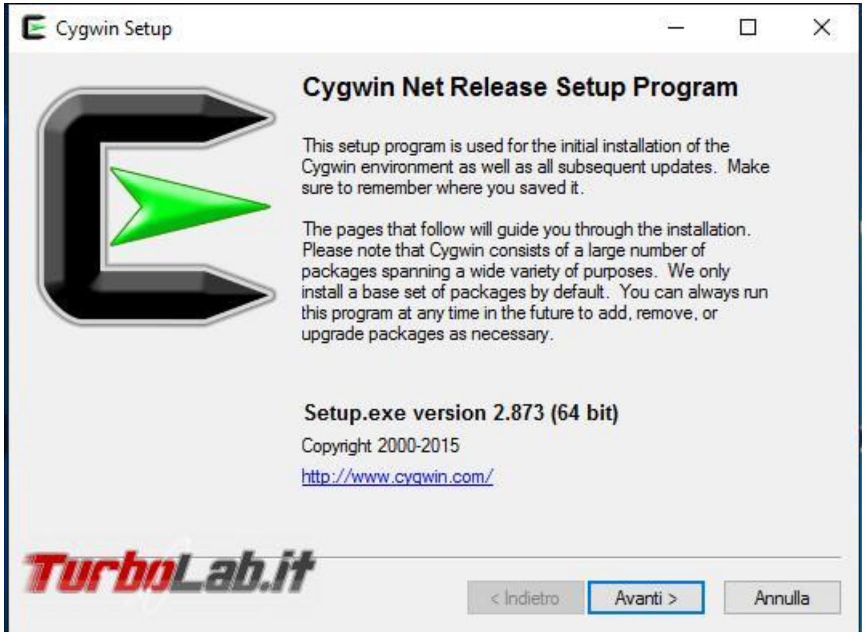
Because of kernel



https://commons.wikimedia.org/wiki/File:Kernel_Layout.svg

Install UNIX-like environment

Option 1: Cygwin (and other shells)



Cygwin Setup

Cygwin Net Release Setup Program

This setup program is used for the initial installation of the Cygwin environment as well as all subsequent updates. Make sure to remember where you saved it.

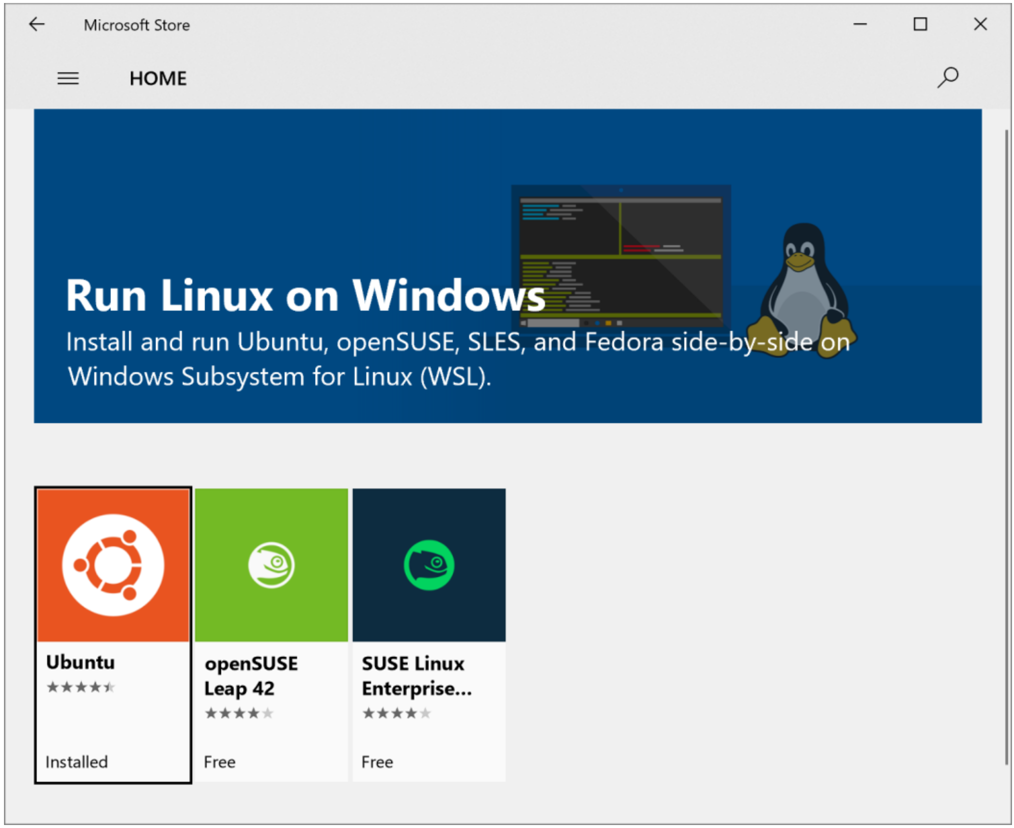
The pages that follow will guide you through the installation. Please note that Cygwin consists of a large number of packages spanning a wide variety of purposes. We only install a base set of packages by default. You can always run this program at any time in the future to add, remove, or upgrade packages as necessary.

Setup.exe version 2.873 (64 bit)

Copyright 2000-2015

http://www.cygwin.com/

Option 2: WSL



Run Linux on Windows

Install and run Ubuntu, openSUSE, SLES, and Fedora side-by-side on Windows Subsystem for Linux (WSL).

Ubuntu	openSUSE Leap 42	SUSE Linux Enterprise...
★★★★½	★★★★☆	★★★★☆
Installed	Free	Free

Keywords to learn further

About CUI:

- stdin/stdout
- redirect
- pipe
- alias
- PATH

About OS:

- kernel
- terminal
- shell
- Linux/UNIX

Have fun! :)

/cygdrive/c/Users/Ikue/Dropbox/PythonCourse/exercises_and_model_answers/Lesson3-data