

Introduction to Python for Data Science

数据科学入门 – 使用Python程序语言

Aug 2020 Microsoft Reactor | Ryan Chung

```
led by player
    ;.load_image("kg.png")
 idlize Dog object and create The Land Creater The Land Cr
self).__init__(image = r.v..
                                                                                                                                                                                           bottom = games, es
   re = games.Text(value = 0, size
                                          reen.add(self.score)
```



Ryan Chung

Instructor / DevelopIntelligence Founder / MobileDev.TW

@ryanchung403 on WeChat Ryan@MobileDev.TW





Reactor







developer.microsoft.com/reactor/
@MSFTReactor on Twitter

Data Science Workshop agenda 数据科学在线研讨会议程

Intro to Python

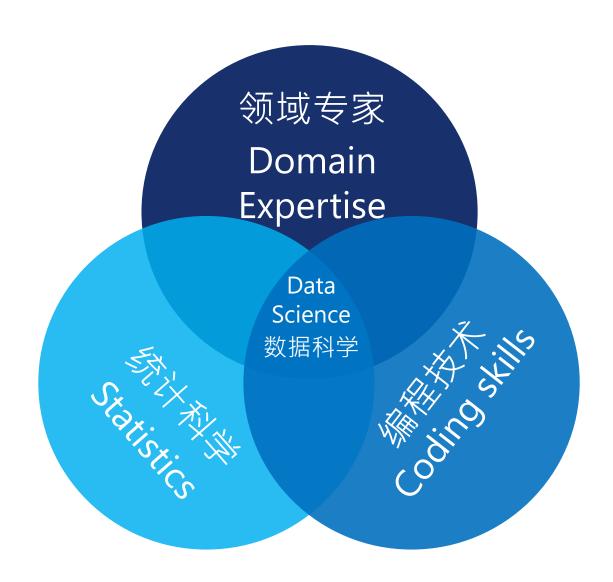
Python程序语言入门

19:30	Welcome 开场		
19:35	Introduction to Data Science 数据科学介绍		
20:10	Arithmetic and numeric types 数值运算与数据型态		
20:30	5-minute break 中场休息		
20:35	Strings 字符串操作		
20:50	Other data types 其他数据型态介绍		
21:00	Event end 研讨会结束		



What is Data Science?

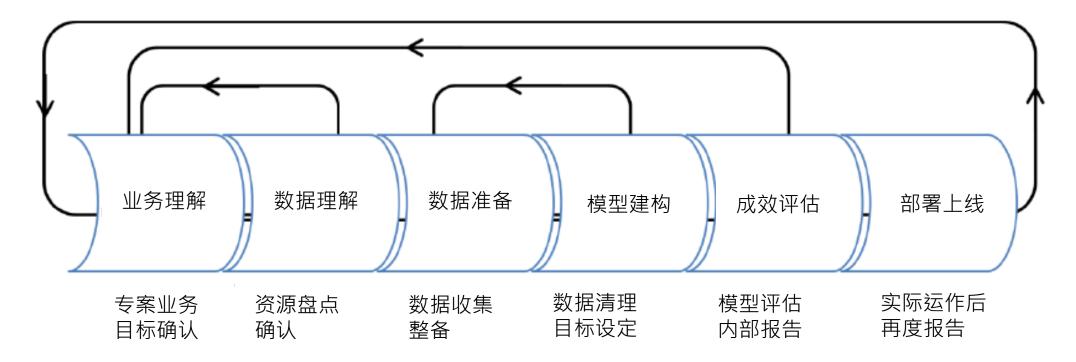
什么是数据科学?



Data Science Process 数据科学运作流程

首要挑战:从大数据中挖掘有意义的线索

Extracting value from large amounts of data and making human sense of it is the primary challenge of data science.



- from Introduction to Data Science on Microsoft Learn: https://docs.microsoft.com/learn/modules/intro-to-data-science-in-azure/2-data-science-process
- Helpful beginners' series: https://docs.microsoft.com/en-us/azure/machine-learning/studio/data-science-for-beginners-the-5-questions-data-science-answers



Specialized roles in Data Science 数据科学领域职业发展

技术导向

Data Scientist

Data Analyst

Data Architect

Data Engineer

Developer

数据科学家

数据分析师

数据架构师

数据工程师

开发者

领域导向

数据分析5个关键职务

商業決策建言

資料科學家

領域專家

預測分析

● 商業決策解讀

資料分析師

資料分析

- 資料探勘
- 資料視覺化

資料處理

資料工程師

軟體工程師

- 資料蒐集
- 資料整頓
- 資料儲存

(image source: managertoday)

關鍵職務	2015從業人數	2018需求人數	三年累計增幅	備註
領域專家	1,962人	3,201人	63%	_
資料科學家	1,092人	6, 296人	477%	增幅最大
資料分析師	8,159人	14,607人	79%	_
資料工程師	21,763人	45,843人	110%	缺口最多
軟體工程師	8,001人	30,576人	282%	_

资料来源:104资讯科技



数据科学家

Harvard Business Review



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

2012年哈佛商业评论

数据科学家:21世纪最性感的职业

2015人力资源点评网Glassdoor调查 工作生活兼具薪水又高,数据科学家荣登最梦幻工作

「用数据解决真实问题的人」

工作生活平衡度排名前 25 职业薪资对照表 (Source: Glassdoor)

Job Title	Ranking	Salary	
Data Scientist	4.2	114,808	
SEO Manager	4.1	45,720	
Talent Acquisition Specialist	4	63,504	
Social Media Manager	4	40,000	
Substitute Teacher	3.9	24,380	
Recruiting Coordinator	3.9	44,700	
UX Designer	3.9	91,440	
Digital Marketing Manager	3.9	70,052	
Marketing Assistant	3.8	32,512	
Web Developer	3.8	<i>66</i> ,040	
Risk Analyst	3.8	<i>6</i> 9,088	
Civil Engineer	3.8	65,532	
Client Manager	3.8	71,120	
Instructional Designer	3.8	<i>66</i> ,040	
Marketing Analyst	3.8	60,000	
Software QA Engineer	3.8	91,440	
Web Designer	3.8	53.848	

生活工作平衡度前10名 - 2016

- ① Corporate Recruiter 人力招聘员
- ② UX Designer 使用者经验设计师
- ③ Data Scientist 数据科学家
- ④ Strategy Manager 策略经理
- ⑤ UI Designer 用户接口设计师
- ⑥ Recruiting Coordinator 招聘协调员
- ② Technical Account Manager 科技业务经理
- Mobile Developer 行动开发程序设计师
- ⑨ DevOps Engineer DevOps工程师
- ® Research Engineer 研究工程师



2020 工作与招聘趋势 - Glassdoor

- ① 人工智能将在管理政策上占有一席之地
- ② 打造公司文化,招聘相同气息的人
- ③ 因应经济紧缩,重新改写招聘规则
- ④ 更高的兼容,容许更多的不同
- ⑤ 高龄就职不是问题,只要能为公司产生价值
- ⑥ 越来越多人在行动装置上找寻工作

Al in Management

智能指导系统

运用大数据产生实时回馈与 建议给予员工

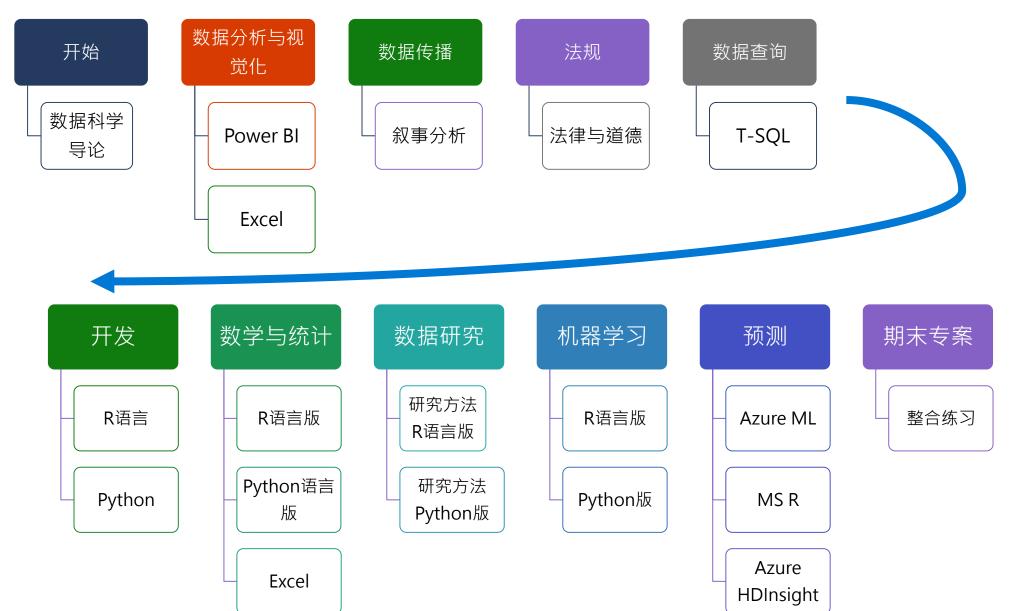
业务员工作流程优化

工作排程建议/活动监视

客户服务介入

实时提供更周到的响应建议情绪指数、处理速度监视

数据科学学习路径

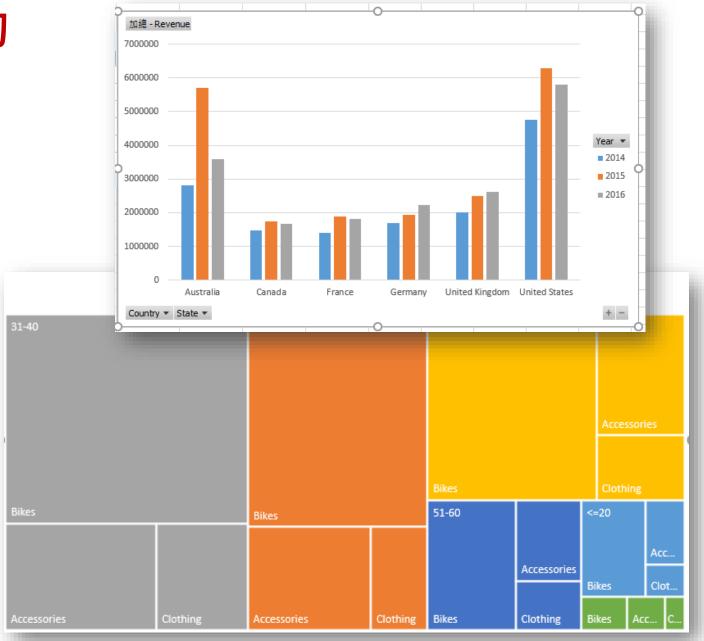


15

Source: 微软MPP资料科学导论

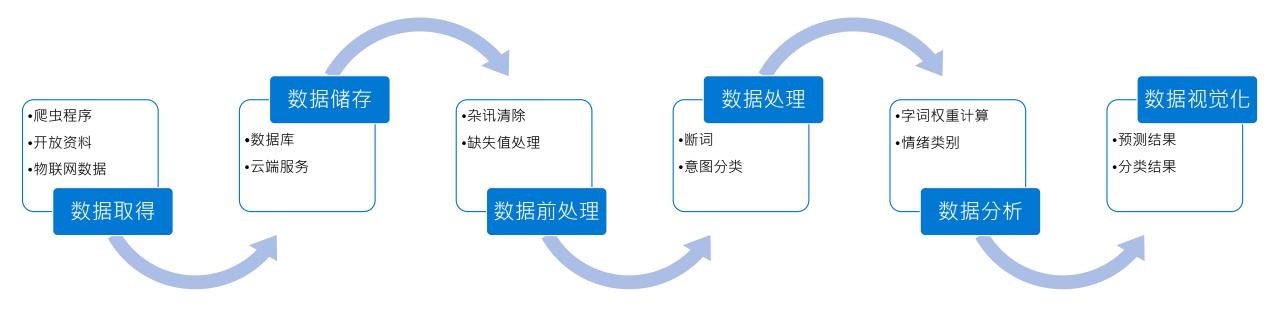
数据科学家应具备的能力

- ·统计学、数学
- ・程序语言
 - · R · Python
- · 视觉化工具呈现与讲解能力
 - · Power BI, Tableau, Qlik, Excel
- ·建模技术、汇整工具
 - · Azure machine learning, Spark



Source: 微软MPP资料科学导论

数据处理 - 以文字探勘流程为例



Data 数据

 Answer 答案

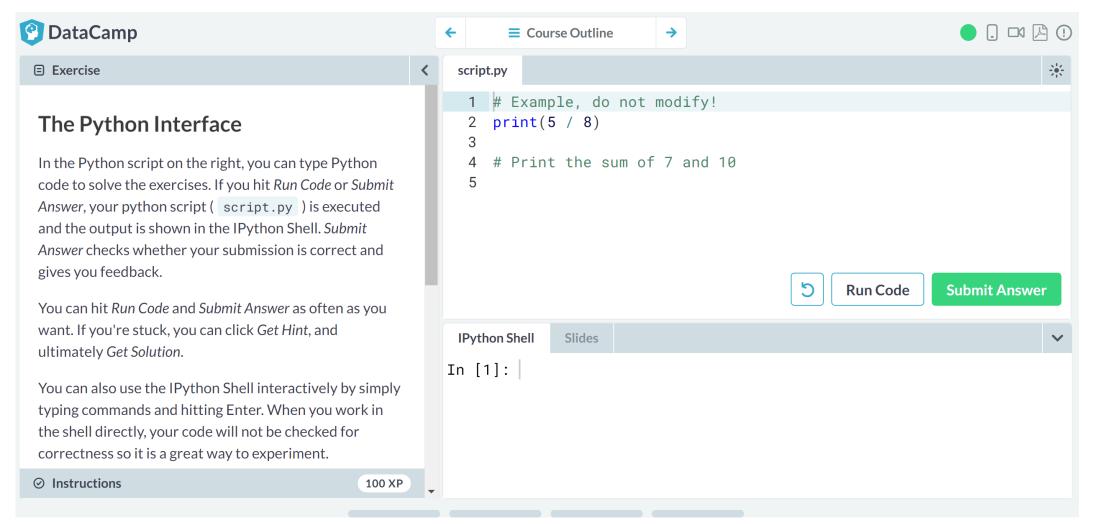


Python 程序设计基础

Section 1 第一节

云端练习环境

DataCamp - Microsoft On-line Training Partner



19

本机开发环境设置

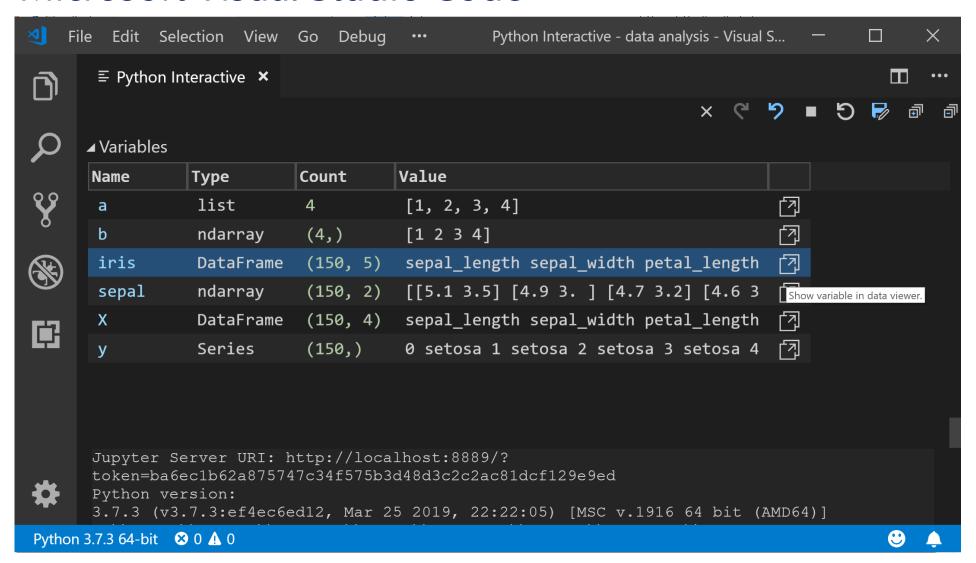
- ·安装Python
 - · 下载对应操作系统的版本
- ・确认安装
 - · 开启命令提示字符
 - 輸入python
 - · 如有出现执行环境代表确认安装完成
 - ・按下Ctrl+Z退出



```
Python 3.7.3 (v3.7.3:ef4ec6ed12, Mar 25 2019, 22:22:05) [MSC v.1916 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

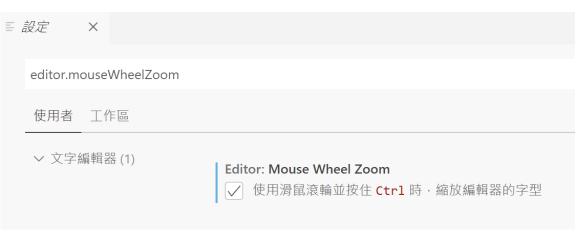
本机开发环境选择

Microsoft Visual Studio Code



安装扩充套件与设定

- ·按下左边 Extensions图示 或 Ctrl + Shift + X
 - · Chinese (Simplified) Language Pack for Visual Studio Code
 - Python
- ·设定Ctrl+鼠标滚轴控制编辑器字号
 - editor.mouseWheelZoom
- · 设定编辑时自动储存
 - ・档案 -> 自动储存





Section 1 overview 第一节 综览

- ·Python的数值运算
- ・字符串操作
- ·其他资料型态
 - · 串行 List、序对 Tuple、集合 Set、字典 Dictionary
- ・是否在群里
- ・串行建构
- ・模组汇入

Section 1 overview 第一节 综览

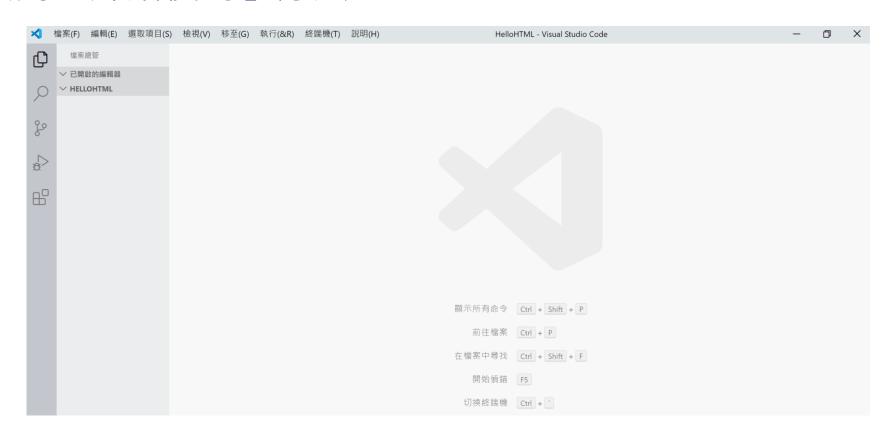
为什么使用Python在数据科学?

- · Easy to learn 容易学习
- · Flexible 弹性大
- · Powerful libraries 强大的函式库

知名社群网路、影片平台以及搜寻引擎公司都大量使用Python在他们的核心技术,数据科学也是其中一个应用项目。

第一个 HelloWorld.py

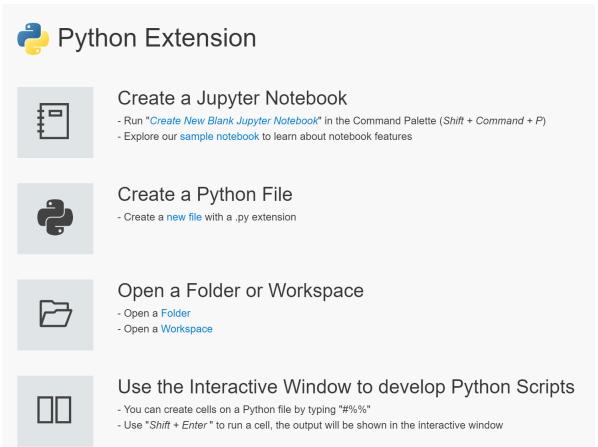
- · 在计算机中新增一个文件夹: HelloPython
- · 打开VS Code, 档案(F) -> 开启文件夹...
- · 选择刚才建立的文件夹
- · 关闭「开始使用」分页



第一个 HelloWorld.py

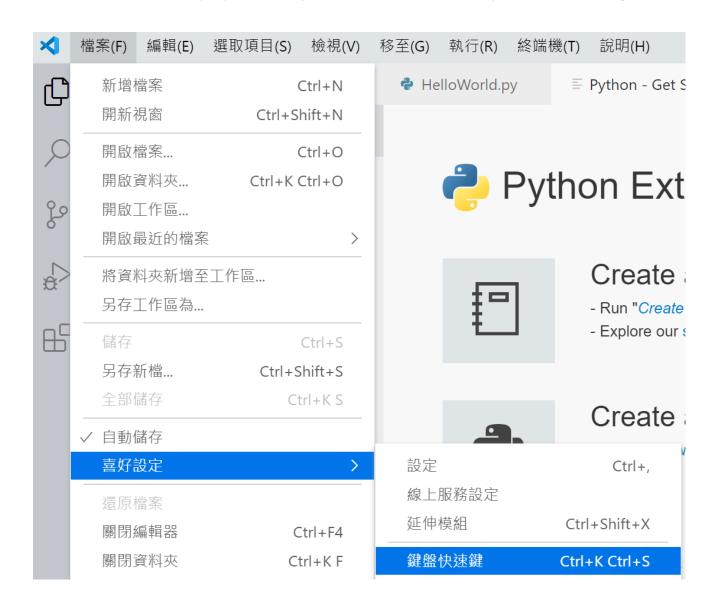
- ·左边档案总管区,按下鼠标右键 -> 新增档案
- ·输入HelloWorld.py
- ·档案前会出现Python图案





设定执行快捷键

・档案 -> 喜好设定 -> 键盘快捷方式



设定执行快捷键

·输入Python进行搜寻



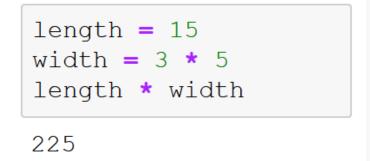
28

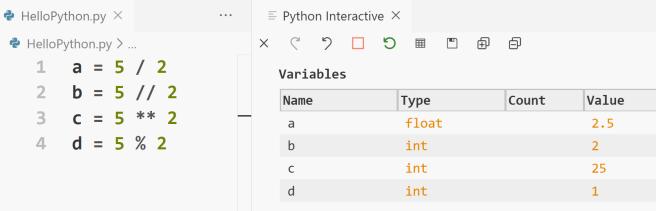
数值运算

· Python numeric operators: + - * / // ** % 数值运算子 商数 指数 余数

· Variables:

变数





• Expressions:

运算式/表达式

True

29

批注方式

·单行 #

· 多行
Python会忽略没有指定给变量的字符串
""

这边放批注

文字处理

```
• String literals: '"Isn\'t," she said.'
                  '"Isn\'t," she said.'
```

```
♣ HelloPython.py × 单引号、双引号、三引号
♣ HelloPython.py > ...
                                            单引号、双引号皆可
      userName = "王小明"
      userNamewithLocation = '上海來的"王小明"' 单引号可以包着双引号
      userNamewithLocationAndCompany = """
                                            三引号可以换行
      上海來的王小明,
      他很會寫Python."""

    ■ Python Interactive ×
     Variables
 Name
                          Type
                                     Count
                                                 Value
  userName
                          str
                                                 王小明
  userNamewithLocation
                                                 上海來的"王小明"
                                      9
                          str
  userNamewithLocationAndCompany
                                                 上海來的王小明, 他很會寫Python.
                          str
                                      21
```

Reactor

· Concatenating strings: | 3 * 'un' + 'ium'

'unununium'



Variables

Name	Туре	Count	Value
badPersonName	str	3	李大同
hateWord	str	4	我討厭你
wantToTell	str	15	李大同我討厭你我討厭你我討厭你

String indices:

取得单一字符,用[位置],从0开始

```
word = 'Python'
word[0] # Character in position 0.
'P'
```

Slicing strings:

取得部分字符串,[开始位置,结束位置-不包含]

```
word[0:2] # Characters from position 0 (included) to 2 (excluded).
'Py'
```

```
♦ StringArray.py > ...
      hostSay = "下雨天留客天留我不留"
      mean1 = hostSay[0:3]+", "+hostSay[3:6]+", "+hostSay[6:9]+"? "+hostSay[9]
      mean2 = hostSay[0:2]+", "+hostSay[2:5]+", "+hostSay[5:7]+", "+hostSay[7:]

    ≡ Python Interactive ×
 白
                                                                             Jupyter Server:
 Variables
 Name
                          Count
                                      Value
             Type
 hostSay
                                      下雨天留客天留我不留
              str
                          10
 mean1
              str
                          16
                                      下雨天,留客天,留我不?留
                          16
                                      下雨,天留客,天留,我不留
 mean2
              str
```

Other data types 其他资料型态

资料型态	中文	符号	是否有顺序性	内容是否可改变	概念
List	串行	[]	0	0	类似其他语言的阵列(数组),但内容的资料型态可以不同
Tuple	序对	()	Ο	X	用在一组固定顺序的常数集合上
Set	集合	{}	X	0	没有顺序概念,一堆资料放在一起
Dictionary	字典	{}	X	Ο	Key-Value成对,中间是冒号

```
dataType.py > ...
      roomGuest = ["王明", "柳宇", "陳尚"]
      roomKey = (1111, 2222, 3333)
      breakfastChoice = {"中式", "西式", "法式"}
      guestBreakfast = {
          "王明":"中式",
     "柳宇":"法式",
    "陳尚":"中式"

    ≡ Python Interactive ×
     Variables
                                Value
 Name
               Type
                        Count
  breakfastChoice
                                {'西式', '法式', '中式'}
                set
                                {'王明': '中式', '柳宇': '法式', '陳尚': '中式'}
  guestBreakfast
                dict
  roomGuest
               list
                                ['王明', '柳宇', '陳尚']
                                (1111, 2222, 3333)
  roomKey
               tuple
```

Membership testing 是否在群里

```
dataType.py > ...
      roomGuest = ["王明", "柳宇", "陳尚"]
      roomKey = (1111, 2222, 3333)
      breakfastChoice = {"中式", "西式", "法式"}
      guestBreakfast = {
         "王明":"中式",
   6 "柳宇":"法式",
     "陳尚":"中式"
      print("王明有住房嗎?" + str("王明" in roomGuest))
      print("陳尚應該沒來住房吧?" + str("陳尚" not in roomGuest))

≡ Python Interactive ×
[11]▶ roomGuest = ["王明", "柳宇", "陳尚"]...
   王明有住房嗎? True
   陳尚應該沒來住房吧? False
```

Reactor

List comprehension 串行建构

Programmatically create lists: 程序化自动建立串行内容

```
ListComprehension.py > ...
        numbers = [x for x in range(1,11)] #注意不包含range結束
        odd_numbers = [x for x in range(1,11) if x %2 != 0]
        even_numbers = [x \text{ for } x \text{ in range}(1,11) \text{ if } x \%2 == 0]
        odd_squares = [x*x for x in range(1,11) if x % 2 != 0]

≡ Python Interactive ×
Variables
                                       Value
  Name
               Type
                           Count
   even numbers
               list
                                        [2, 4, 6, 8, 10]
                           5
  numbers
              list
                                        [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
                           10
   odd numbers
              list
                                        [1, 3, 5, 7, 9]
   odd squares
               list
                           5
                                        [1, 9, 25, 49, 81]
```

Importing modules 模组汇入

汇入整个模组进行使用 import ...



import 惯例写法

import numpy as np import pandas as pd import matplotlib.pyplot as plt import seaborn as sns import statsmpodels as sm

汇入部分模组进行使用 from ... import ...

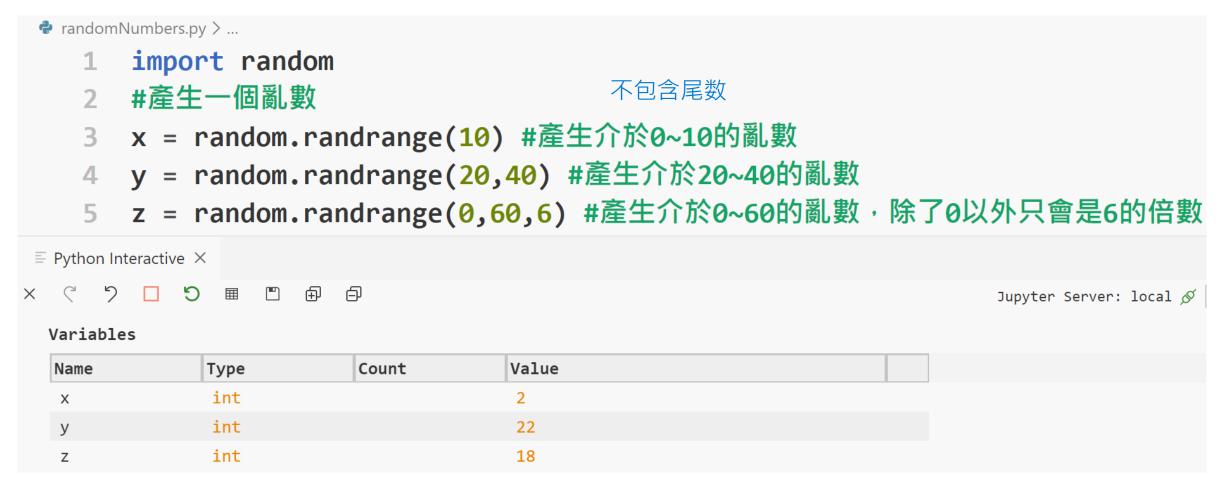


07.import_modules.py

自订汇入模组名称 import ... as ...



产生一个乱数



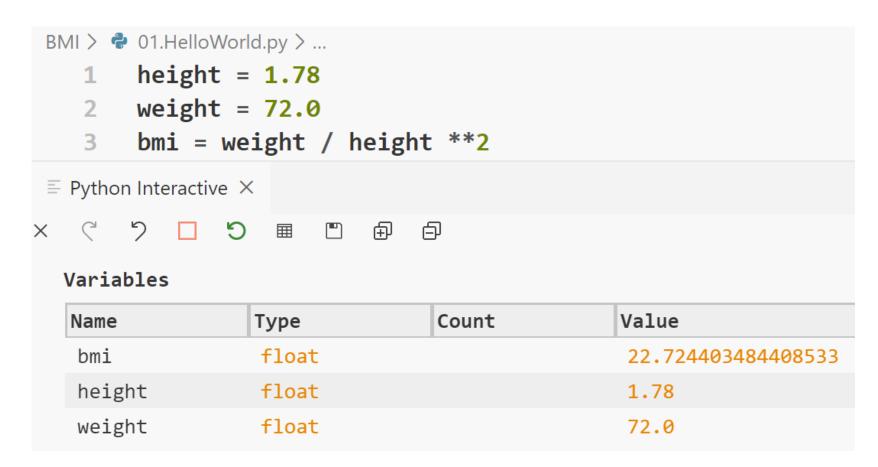
综合练习:家庭BMI值计算

案例:BMI值计算

- 1. 直接计算BMI值
- 2. 显示BMI值对应的结果
- 3. 用List建立全家人的BMI资料
- 4. 帮全家人检验,判断是否落于正常范围
- 5. 用BMI资料画图,观察一年的体重变化
- 6. 引用儿童的BMI指标来进行比对

使用Python计算BMI值

- ·变数直接命名、直接给值
- ·**n:乘以自己n次



BMI值标准

成人的体重分级与标准			
分 级	身体质量指数		
体重过轻	BMI < 18.5		
正常范围	18.5 ≤ BMI < 24		
过 重	24 ≤ BMI < 27		
轻度肥胖	27 ≤ BMI < 30		
中度肥胖	30 ≤ BMI < 35		
重度肥胖	BMI ≥ 35		
资料来源:食品资讯网/肥胖及体重控制			

显示判断结果

- ·if
- · elif
- · else

```
BMI > • 01.HelloWorld.py > ...
       height = 1.78
       weight = 72.0
       bmi = weight / height **2
       print('你的BMI值為'+"%.2f" % bmi)
                                       小数点后留两位数的浮点数
       if bmi <18.5:
           print('體重過輕')
       elif 18.5 <= bmi < 24.0:
                                       可以直接用一个范围来写
           print('正常範圍')
       elif 24.0 <= bmi < 27.0:
           print('體重過重')
  10
                                       所有条件式不用加()
       elif 27.0 <= bmi < 30.0:
                                       最后面打上冒号
           print('輕度肥胖')
  12
       elif 30.0 <= bmi < 35.0:
           print('中度肥胖')
  14
  15
       else:
           print('重度肥胖')
  16

≡ Python Interactive ×

      り <a>り</a>
               ⊕
                         [8] height = 1.78...
                                                          44
   你的BMI值為22.72
                                          BMI/01.HelloWorld.py
   正常範圍
```

用List建立全家的BMI资料

·名称、身高、体重、BMI值

```
BMI > • 02.FamilyBMI.py > ...
       family_data = [
            ['Dad',178,72],
       ['Mom',155,44],
           ['Kid',117,19]
   5
   6
       for each_one in family_data:
            this_bmi = each_one[2] / ((each_one[1]/100) **2)
            each_one.append(this_bmi)
  10
       family data

≡ Python Interactive ×
  [['Dad', 178, 72, 22.724403484408533],
  ['Mom', 155, 44, 18.314255983350673],
   ['Kid', 117, 19, 13.879757469501062]]
```

计算完之后,加回原本的List

用List建立全家的BMI资料

·名称、身高、体重、BMI值、BMI指标

```
10
    for each_one in family_data:
         if each one[3] <18.5:</pre>
12
13
             bmi_index = '體重過輕'
14
        elif 18.5 <= each one[3] < 24.0:</pre>
             bmi index = '正常範圍'
15
        elif 24.0 <= each_one[3] < 27.0:</pre>
16
             bmi_index = '體重過重'
17
18
        elif 27.0 <= each_one[3] < 30.0:</pre>
             bmi_index = '輕度肥胖'
19
        elif 30.0 <= each_one[3] < 35.0:</pre>
20
21
             bmi index = '中度肥胖'
22
        else:
             bmi_index = '重度肥胖'
23
24
        each one.append(bmi index)
25
    family data
26
```

把判断完的结果也放回List

```
[['Dad', 178, 72, 22.724403484408533, '正常範圍'], ['Mom', 155, 44, 18.314255983350673, '體重過輕'], ['Kid', 117, 19, 13.879757469501062, '體重過輕']]
```

用List建立全家的BMI资料

·名称、身高、体重、BMI值、BMI指标、是否落于正常范围

```
25
     for each_one in family_data:
26
          if each_one[4] == '正常範圍':
27
28
               is_normal = True
                                                 增加一个布尔值,来记录是否落在正常范围
29
          else:
30
               is_normal = False
          each_one.append(is_normal)
31
32
                                                [['Dad', 178, 72, 22.724403484408533, '正常範圍', True],
                                                 ['Mom', 155, 44, 18.314255983350673, '體重過輕', False],
     family_data
                                                 ['Kid', 117, 19, 13.879757469501062, '體重過輕', False]]
```

前面用到的语法

- ・资料型态
 - ·字符串、整数、浮点数、布尔值、list
- ・运算符
 - · 等于、大于、大于等于、小于等于、小于
- ・流程控制

 - · if..else if..else
- · List 运算
 - ・新増元素

检查资料型态

```
32

33 for each_one in family_data[0]:

print(type(each_one)) 用type() 检查资料型态
```

0	1	2	3	4	5
Dad	178	72	22.7244034844	正常範圍	true

```
<class 'str'>
<class 'int'>
<class 'int'>
<class 'int'>
<class 'float'>
<class 'str'>
<class 'bool'>
```

观察list中的部分元素

- ・起始:结束
- · 取出元素不包含结束
- ·不写开头,代表从最前面开始
- ·不写结尾,代表从那一项开始走到全部结束

```
family_data[0:2] 第0笔、第1笔

[['Dad', 178, 72, 22.724403484408533, '正常範圍', True], ['Mom', 155, 44, 18.314255983350673, '體重過輕', False]]

family_data[:2] 第0笔、第1笔

[['Dad', 178, 72, 22.724403484408533, '正常範圍', True], ['Mom', 155, 44, 18.314255983350673, '體重過輕', False]]
```

```
[['Dad', 178, 72, 22.724403484408533, '正常範圍', True],
['Mom', 155, 44, 18.314255983350673, '體重過輕', False],
['Kid', 117, 19, 13.879757469501062, '體重過輕', False]]
     family_data[1:] 第1笔、第2笔
     [['Mom', 155, 44, 18.314255983350673, '體重過輕', False],
      ['Kid', 117, 19, 13.879757469501062, '體重過輕', False]]
```

全部的资料

family_data

函数/函式 Function

- ·关键词def
- ・注意缩排、冒号

```
def functionName([parameters]):
    statements
    [return | return value]
    [statements]
```

练习:温度单位转换运算

· 输入:温度、单位

· 输出: 摄氏温度、华式温度

练习:温度单位转换运算

```
def DegreeTransfer(degree, unit):
   if unit == "C" or unit == "c":
       degreeF = degree * (9/5) + 32
       degreeC = float(degree)
   elif unit == "F" or unit == "f":
       degreeC = (5/9) * (degree - 32)
       degreeF = float(degree)
   else:
       print("输入值有误,请重新输入")
       return
   print("华氏"+str(degreeF)+"°F、摄氏"+str(degreeC)+"°C")
inputDegree, inputUnit = input("请输入温度与单位,例如摄氏25度c可输入'25 c'").split()
DegreeTransfer(eval(inputDegree), inputUnit)
```



Reactor







developer.microsoft.com/reactor/
@MSFTReactor on Twitter



议程结束感谢的



请记得填写课程回馈问卷 https://aka.ms/ReactorFeedback

© 2019 Microsoft Corporation. All rights reserved. The text in this document is available under the Creative Commons Attribution 3.0 License, additional terms may apply. All other content contained in this document (including, without limitation, trademarks, logos, images, etc.) are not included within the Creative Commons license grant. This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal, reference purposes.

This document is provided "as-is." Information and views expressed in this document, including URL and other Internet Web site references, may change without notice. You bear the risk of using it. Some examples are for illustration only and are fictitious. No real association is intended or inferred. Microsoft makes no warranties, express or implied, with respect to the information provided here.