01-Data Analysis with Python

课件：Data Analysis with Python\_PPT\_20220510\_LiuHong.pdf

试验环境链接：https://pan.baidu.com/s/1C43EqMM9oV66GLT-ikkFKg

提取码：n24t

培训录制链接：https://cisco.webex.com/recordingservice/sites/cisco/recording/playback/ae302d6ab22e103a97df005056810b1f

密码：9eBGFeGH

菜鸟网站关于python3相关的语法

<https://www.runoob.com/python3/python3-tutorial.html>

请回答下面题目

1. Python的基本数据类型有哪几种？各举2例。

|  |  |  |
| --- | --- | --- |
| 数值型 | 整型 int | 1, 345 (长整型) |
| 浮点 float | 1.2, 3e-2 |
| 复数 complex | 1+2j, 1.1+2.2j |
| 布尔 bool | True, False |
| 字符型 | 字符串 str | ‘123’, “abcdefg” |

1. Python中序列的数据类型有哪几种，分别列出他们的不同，并各举2例。

Table

Description automatically generated

|  |  |  |
| --- | --- | --- |
| list | 列表 | [1,2,3], [‘a’, ‘b’, ‘c’] |
| tuple | 元组 | (1,2,3), (‘a’, ‘b’, ‘c’) |
| dict | 字典 | {'a':'attribute', 'b': 'business'}, {'a':['abc', 'attri', 'arise']} |
| set | 集合 | {1,2,3}, {‘a’, ‘b’, ‘c’} |

元组和列表的区别在于元组中的数据不能修改

集合中的元素不能重复，如果重复，就自动忽略

字典中的键只能一个，且唯一，但是值可以相同，也可以多个，用[]括起来

1. 在Python中数据是有类型（具本类型见问题1，问题2），但是变量是没有类型的，只有当给变量赋值时（用=）变量才有类型，那么有什么方法确定变量的类型?

Python 中的变量不需要声明的，每个变量在使用前都必须赋值，在赋值之前变量是没有意义的，只有赋值以后该变量才会被创建。

在 Python中，变量就是变量，它没有类型，我们所说的"类型"指的是该变量所指的内存中数据或者对象的类型。

用函数type()可以确定变量的‘类型’

1. 与其它编程语言相比，Python的函数非常强大，包括参数和返回值，请列出它们的特点。

|  |  |  |
| --- | --- | --- |
| 参数 | 必填 | 只有参数名 |
| 缺省 | 参数名=缺省值 |
| 不定长 | \* tuple, 只有值，没有参数名 |
| 关键字 | \*\* dict，通过key表示参数名，通过value赋值 |
| 返回值 | 多个返回值 元组 |  |

上述的四种参数类型一定要搞清楚，特别是后二种：不定长和关键字

不定长参数，其实是以元组tuple的形式传入

关键字参数，其实是以字典dict的形式传入

1. Python中类的定义遵循“驼峰”定义，类的实例化对象采用小写。C++的类定义有构造函数和析构函数，Java中类的定义只有构造函数，没有析构函数，为什么？那么Python中的类有构造函数和析构函数吗

Python中类的定义如下：

class MyFirstClass:

类的相关实现

类对象的生成，则用类名+括号的形式

object = MyFirstClass()

|  |  |  |  |
| --- | --- | --- | --- |
|  | 类名 | 构造函数 | 析构函数 |
| C++ | CMyFirstClass | CMyFirstClass() | ~CMyFirstClass() |
| Java | MyFirstClass | MyFirstClass() | 无 （垃圾回收机制） |
| Python | MyFirstClass | **\_\_new\_\_ + \_\_init\_\_** | **\_\_del\_\_** |

1. 在利用Python进行数据分析时，常会用到二个工具库numpy和pandas，请问它们有什么作用？

Python进行数据分析时，最常用的二个科学库numpy和pandas。

numpy用于数值计算，可以想象成就是矩阵的运算

Graphical user interface, diagram

Description automatically generated

pandas用于数据处理，包括数据的加载，分片，合并，清洗，保存等

Diagram

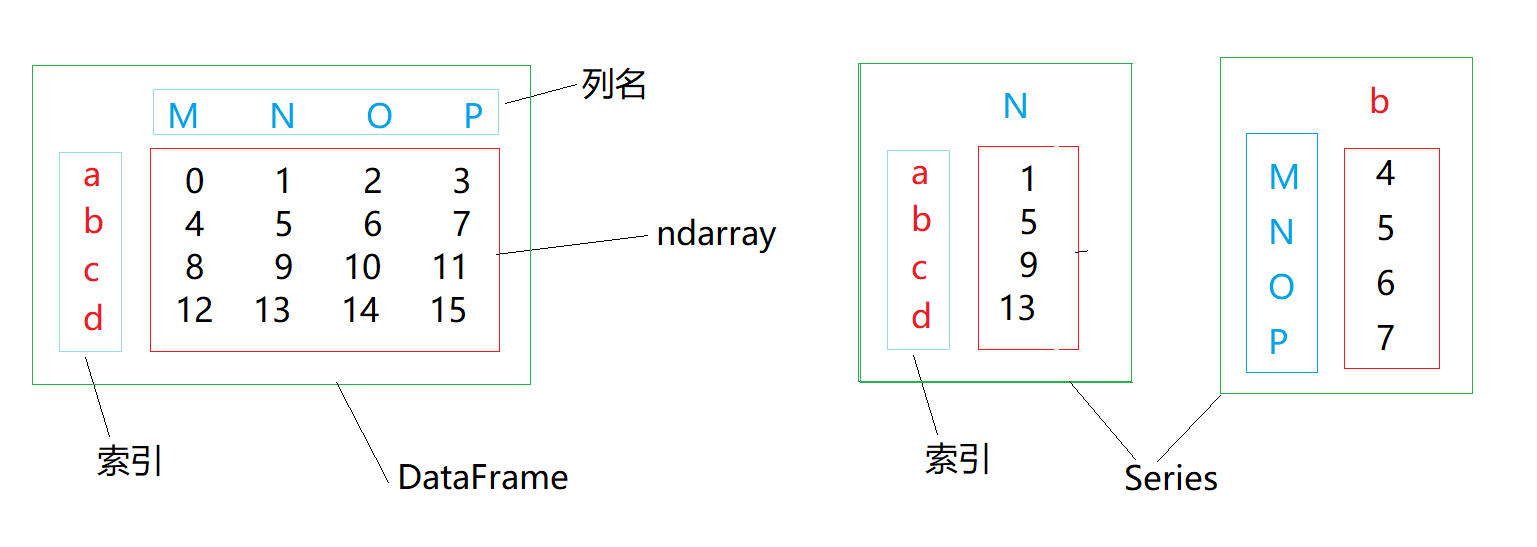
Description automatically generated

1. 除了Python的数据类型外，numpy和pandas又引入哪些数据类型？如何表示？

|  |  |
| --- | --- |
| numpy | ndarray |
| pandas | DataFrame |
| Series |

ndarray就是一个矩阵

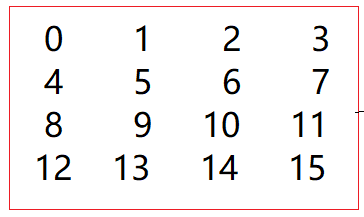
DataFrame就是在ndarray矩阵数据的基础上，增加了列名（用来描述列）和索引（用来描述行），也就是说ndarray是DataFrame的数据部分，DataFrame是在ndarray的基础上扩展了列名和索引，本质上是没有改变里面的数据部分，增加列名和索引只是为了方便存取数据。



在理解了ndarray和DataFrame后，再来对比DataFrame和Series，Series只是DataFrame中的一列或者一行，如果取了二列或者二行，则结果仍然是DataFrame，而不是Series。

1. 在numpy中，有几种方法获取数据元素？举例演示说明

从前面的学习中可以知道numpy中的数据类型是ndarray，相当于一个二维数组，或者更准确地说是一个二维矩阵，如下图所示



ndarray的元素可以行和列来表示: [行,列] (行和列从0开始)，比如t1[1, 3]

先了解一下数据切片操作，用:表示 [start:end] 这是一个半闭半开区间，包括start，不包括end

|  |  |  |
| --- | --- | --- |
| 取一个元素 |  | t1[1,3] |
| 取多个元素 | 1连续的行和连续的列 | t1[1:3, 1:3] |
| 2连续的行和不连续的列 | t1[1:3, [1,3]] |
| 3不连续的行和连续的列 | t1[[1,3], 1:3] |
| 4不连续的行和不连续的列 | 合并2与3 |

1. Dict, ndarray, dataframe, series，这四种数据类型有什么关系？最好用图示说明

关于ndarray, dataframe, series，这三者的关系，在问题7中已经回答了。

至于Dict，其实就是Key:Value的格式，其中Key只有一个元素，Value可以有多个元素，如果是多个元素，就需要用list来表示。

比如：

d = {’M’:[0,4,8,12],’N’:[1,5,9,13],’O’:[2,6,10,14],’P’:[3,7,11,15]}

df = pd.DataFrame(data=d,index=list(‘abcd’))

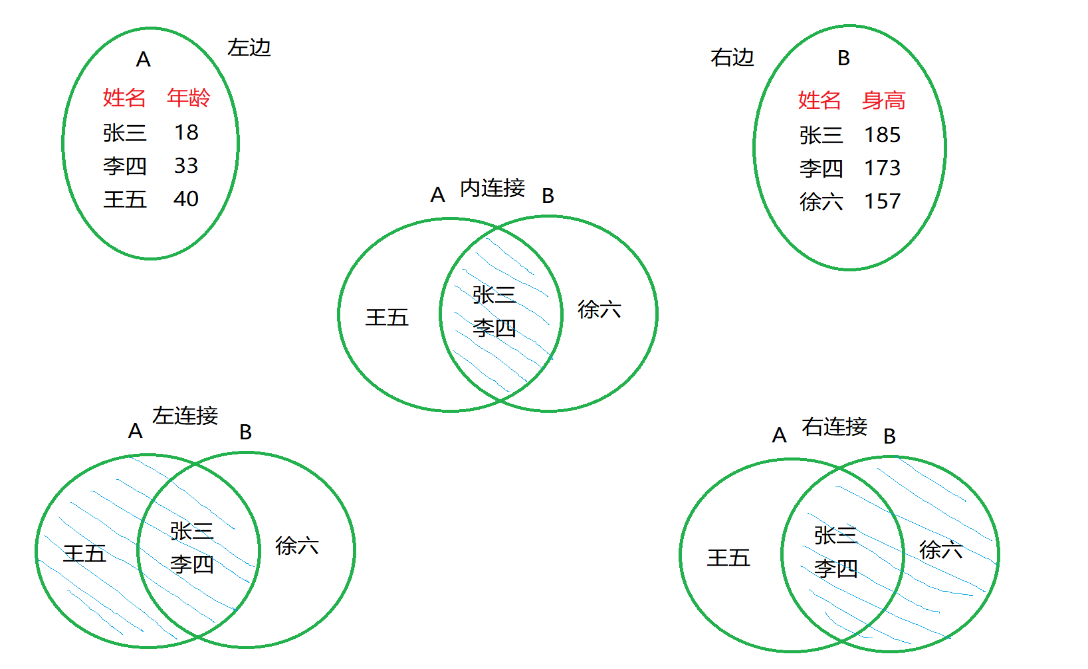
其实Dict和DataFrame是可以相互转换的

Dict 🡸🡺 DataFrame

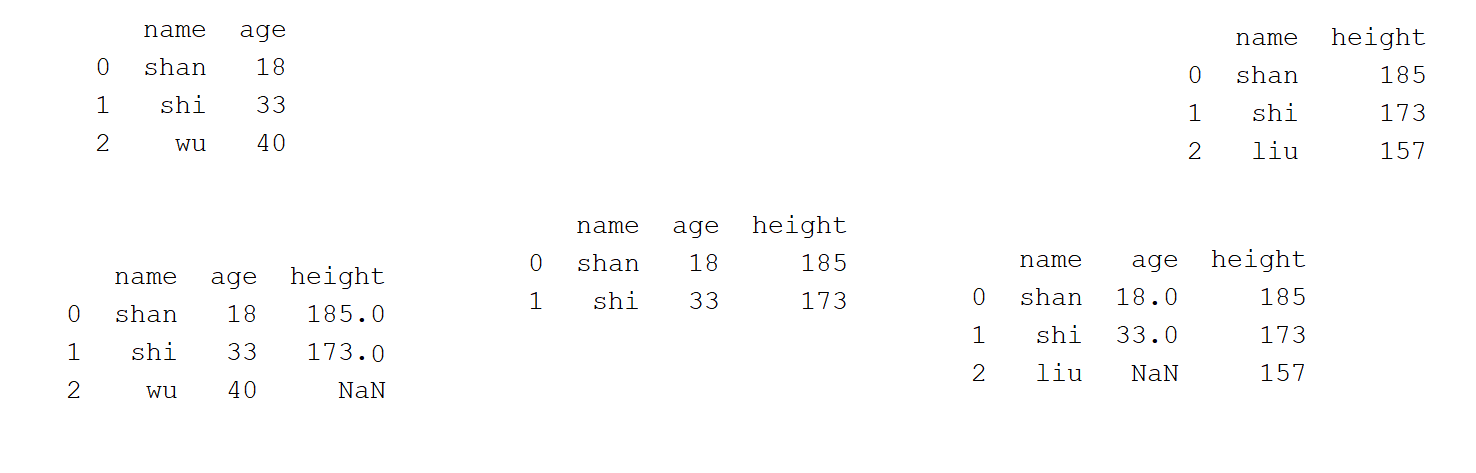
1. Dataframe中内连接，左连接，右连接是什么意思？

这个概念是来自于数据库中的SQL语句，有数据库基础的人很容易理解。

其实用集合来表示就很清楚直观：



对于Python的结果则是：



1. 什么是方差，标准差，协方差，皮尔逊系数？请用一个实际例子来解释

方差和标准差是用来描述一组数据波动的大小。

方差或者标准差越大，则说明这组数据波动越大；方差或标准差越小，则说明这组数据就越稳定。

标准差是方差的平方根，或者说方差是标准差的平方。

比如：

X： 50，100，100，60，50

Y： 73， 70， 75，72，70

平均值： E(X) = 72， E(Y) = 72

方差： S2(X) = 536.0，S2(Y) = 3.6

标准差：S(X) = 23.15167380558045，S(Y) = 1.8973665961010275

协方差用于衡量两组随机变量的相关性。

如果两组[变量](https://baike.baidu.com/item/%E5%8F%98%E9%87%8F/5271)的变化趋势一致，那么两组变量之间的协方差就是正值；如果两组变量的变化趋势相反，那么两组变量之间的协方差就是负值。

而[方差](https://baike.baidu.com/item/%E6%96%B9%E5%B7%AE/3108412" \t "_blank)是协方差的一种特殊情况，即当两个变量是相同的情况。

A picture containing text, screenshot

Description automatically generated 计算式

Text, letter

Description automatically generated 推导式

上述数组的协方差cov(X, Y) = 10.0

皮尔逊相关系数r是用来描述两个变量间线性相关强弱程度的统计量，r的绝对值越大，表明相关性越强；r取值范围为【-1,1】，为正代表两个变量存在正相关，为负代表两个变量存在负相关。

Text

Description automatically generated

上述数组的皮尔逊系数r = 0.2276493186283971

1. 什么是数据透视？

在数据分析过程中，图表是最直观的一种数据分析方式，数据透视表/图具有很强的动态交互性。

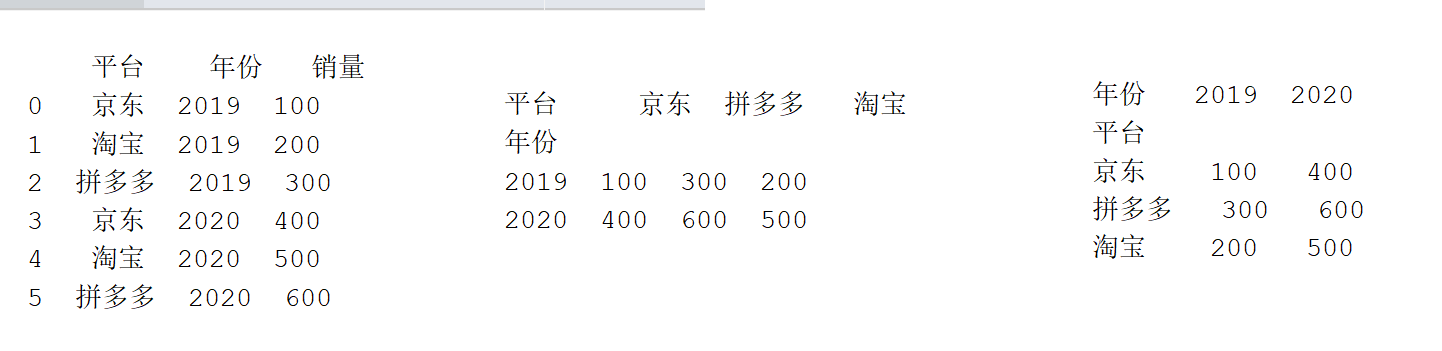
为了讲清楚数据透视表/图，我们先用Excel来示例一个数据例子。

1. 先产生原始数据
2. 选择任一原始数据，插入透视图 PivotChart
3. 系统会自动生成透视表 PivotTable以及和透视表对应的透视图 PivotChart
4. 此时的透视图是空的，需要选择数据

Graphical user interface, application

Description automatically generated

Pandas中有个很有用的方法 pivot()，是指定相应的列分别作为行标签index和列标签columns，并指定相应的列作为值values，然后重新生成一个新的DataFrame对象，这样的好处是使得数据更加的直观和容易分析，俗称数据透视。



1. 用培训时提供的数据，做一个影视推荐系统

要求：

1，基于皮尔逊系数r

2，数据基于 电影.xlsx 和 评分.xlsx

3，为了防止刷分，我们需要设置评分次数阈值为100

4，推荐 《阿甘正传（1994）》给相关的用户

5，完成代码，并上传到git上

6，代码中要有注释，说清楚每一次的作用

