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。

int: 1, 3

str: ‘Chengwen’, ‘Li’

float: 3.33, 5.89

bool: true, false

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。

list: [1,2,3], [3,6,8] 可读可操作，有序且支持重复

tuple: (1,2,3,4), (2, 4, 6, 2) 只能读，不能操作，有序且支持重复

set: {1,2,3,4}, {5,6,7,8} 没有下标,直接操作数据，且不支持重复

dict: {‘j’ :’java’, ‘p’:’python’} 有无序下标，通过key:value的方式使用key去操作数据

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?

使用type(变量)

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。

参数可以给定一个缺省值，也可以通过名称赋值。使用\*修饰词代表不定长参数，\*\*表示关键字参数

返回值可以返回多个值，用tuple进行保存。若没有返回值，默认为none

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，没有构造函数，为什么？那么Python中的类有构造函数和析构函数吗？

析构函数往往是用来善后的，而Java有自动回收机制所以不需要。

python中有构造函数和析构函数 init\_(self), del\_(self)

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？

numpy 是Python创建高层数据的基础，更有效地使用其他工具类似Panda。

pandas 是进行数据清洗/整理的最好工具

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？

ndarray， dataframe, series

1. 在numpy中，有几种方法获取数据元素？举例演示说明

四种

获取其中一个值t1[行，列]

获取连续的行与列t1[行:行，列:列]

获取连续的行不连续的列t1[:,[1,3]]

获取不连续的行与列 temp = t1[:,[1,3]] temp[0,2] 或者 t1[:,[1,3]][[0,2]]

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

dict 是一个索引值一个value来呈现

ndarray 是一个数组

dataframe 是一个二维数组且有行列索引

series类似于dataframe但是是一维的

1. Dataframe中内连接，左连接，右连接是什么意思？

内连接是只保留关键字段相同的那些行

左连接行顺序上原原本本遵从数据集1的关键字段的出现顺序

右链接行顺序上原原本本遵从数据集2的关键字段的出现顺序

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

方差就是数据的波动大小，标准差就是方差开根号。协方差是衡量两组数据的相关性。皮尔逊系数是描述两个变量间线性相关强弱程度的统计量

比如3此考试同学A考了80，90，100.同学B考了85，85，100.

同学A的方差=（90-100）2+（90-90）2+（90-80）2

标准差=(（90-100）2+（90-90）2+（90-80）2)

协方差=（（90-100）（90-100）+（90-90）（90-85）+（90-80）（90-85））/3

皮尔逊系数=协方差/标准差A\*标准差B

1. 什么是透视图？

可以自行更改选择你想要的行列名进行排列

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 勇敢者的游戏（1995）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