日照职业技术学院
RIZHAO POLYTECHNIC

数据采集与治理

项目三　景区游客量统计

主讲：赵娜

项目三 景区游客量统计

使用 Python 进行数据分析，既可以使用 python 语言，也可以使用第三方扩展包，例如 numpy 和 pandas 等。

本项目使用 python 和第三方扩展包进行数据分析，通过案例展示，用户可以掌握使用 python、numpy 和 pandas 进行数据分析的技术，了解不同工具的特点。

知识目标和技能目标

【知识目标】

- 了解数据分析的含义
- 了解使用 Python 和第三方包进行数据分析的优势

【技能目标】

- 能够使用 Python 进行数据分析
- 能够使用 numpy、pandas 包进行数据分析
- 能够使用 numpy 的函数读取 csv 文件
- 能够使用 pandas 的函数读取 csv 文件

情境描述

哼哼唧唧旅游公司承接游客的国内旅游业务。公司成立半年来，成功组织了多次国内旅游团。近期，为了优化公司资源配置，决定对半年来公司的业务数据进行梳理。找到游客青睐的旅游点，加大投入，同时对游客较少的旅游点进行改进升级。

为此，该公司找到了欢喜科技。欢喜科技公司将这项任务交给了小刘。经分析，小刘决定对这些数据进行分析，并使用 Python 和 numpy 作为数据分析的工具。

大纲

1 任务分析

2 使用python进行分析

3 使用numpy进行分析

4 使用pandas进行分析

5 三种实现方法比较

任务分析

数据表的前15行

日期	九寨沟	张家界	香港	东部华侨城	上海迪士尼
2017/1/1	30	17	17	3	28
2017/1/2	12	21	8	3	9
2017/1/3	14	22	15	1	32
2017/1/4	6	29	13	7	15
2017/1/5	31	15	5	3	10
2017/1/6	26	23	17	6	25
2017/1/7	9	18	12	7	28
2017/1/8	9	19	12	4	15
2017/1/9	22	24	11	10	14
2017/1/10	32	9	6	3	20
2017/1/11	10	19	16	3	26
2017/1/12	17	17	17	7	31
2017/1/13	21	25	12	10	6
2017/1/14	17	27	15	1	32

数据表的最后15行

2017/9/7	21	8	16	6	32
2017/9/8	13	10	11	9	6
2017/9/9	16	32	16	3	31
2017/9/10	29	13	15	10	32
2017/9/11	6	13	14	8	11
2017/9/12	13	18	12	5	6
2017/9/13	29	19	10	10	24
2017/9/14	25	27	18	9	18
2017/9/15	31	17	7	10	31
2017/9/16	23	7	15	6	13
2017/9/17	19	17	22	4	28
2017/9/18	29	20	16	2	3
2017/9/19	26	19	11	4	15
2017/9/20	12	10	17	4	32
2017/9/21	18	31	9	1	14

数据表的特点

- 需要处理的核心数据，即每日的游客数量，是整数类型。

- 数据表中有汉字字符和特殊格式的日期需要处理。

开发工具选择

- 使用纯Python语言，需要使用文件操作、函数设计、列表操作、数据格式转换等相关技术。

- 使用numpy或者pandas技术开发，需要使用文件读写、数组操作、DataFrame操作等相关技术。

使用python进行分析

创建求和函数

定义函数getTotalTourist（），用以求解数组元素的累加和，并将计算结果total作为返回值返回。该函数具有一个参数place，是需要进行处理的数组对象。

```
def getTotalTourist( place ):
        total = 0
    for dayTourist in place:
        total += dayTourist
    return total
```

从文件中读取数据

使用open( )函数，以读方式打开文件，然后使用read( )函数将数据读取到变量all_data中。

```
data_file =
open('tourist_data.csv','r')
all_data = csv.reader(data_file)
for day_data in all_data:
    print( day_data )
```

从文件中读取数据

打印输出all_data。

```
['日期', '九寨沟', '张家界', '香港', '东部华侨城', '上海迪士尼']
['2017/1/1', '30', '17', '17', '3', '28']
['2017/1/2', '12', '21', '8', '3', '9']
['2017/1/3', '14', '22', '15', '1', '32']
['2017/1/4', '6', '29', '13', '7', '15']
['2017/1/5', '31', '15', '5', '3', '10']
['2017/1/6', '26', '23', '17', '6', '25']
['2017/1/7', '9', '18', '12', '7', '28']
['2017/1/8', '9', '19', '12', '4', '15']
['2017/1/9', '22', '24', '11', '10', '14']
['2017/1/10', '32', '9', '6', '3', '20']
['2017/1/11', '10', '19', '16', '3', '26']
['2017/1/12', '17', '17', '17', '7', '31']
['2017/1/13', '21', '25', '12', '10', '6']
['2017/1/14', '17', '27', '15', '1', '32']
```

读取九寨沟的旅游数据

九寨沟的每日游客数量位于数据文件中的第二列，由all_data中每行的第2个元素组成。在这里，读取每行数据组成的列表的第二个元素，并存储在列表jzg_data中。

```
jzg_data = []
for row in all_data:
    jzg_data.append(row[1])
```

或者

```
jzg_data = [row[1] for row in all_data]
```

读取九寨沟的旅游数据

打印输出jzg_data。

['九寨沟', '30', '12', '14', '6', '31', '26', '9', '9', '22', '32', '10', '17',
'21', '17', '13', '30', '22', '9', '4', '28', '16', '8', '5', '24', '13', '26',
'6', '31', '22', '18', '28', '19', '6', '12', '10', '25', '14', '4', '17', '18',
'18', '5', '15', '30', '25', '20', '13', '18', '27', '14', '11', '5', '14', '25'
, '14', '17', '18', '11', '25', '25', '28', '25', '31', '27', '7', '11', '5', '1
7', '14', '10', '20', '21', '31', '22', '12', '26', '20', '3', '20', '18', '28',
'30', '27', '7', '32', '6', '7', '24', '3', '31', '6', '25', '12', '23', '15', '
3', '12', '28', '22', '30', '32', '21', '32', '9', '20', '3', '9', '16', '15', '
27', '25', '27', '8', '31', '15', '23', '17', '27', '29', '24', '30', '3', '8',
'31', '10', '30', '31', '22', '27', '17', '15', '14', '15', '27', '25', '18', '4
', '26', '3', '32', '10', '32', '31', '24', '3', '29', '6', '25', '11', '32', '1
8', '20', '25', '30', '8', '8', '27', '9', '26', '18', '16', '5', '10', '28', '3
2', '4', '15', '31', '32', '16', '24', '12', '27', '30', '22', '8', '6', '13', '
31', '6', '3', '15', '13', '19', '30', '21', '20', '10', '13', '19', '21', '18',
'12', '23', '24', '27', '28', '3', '21', '6', '6', '23', '26', '23', '20', '26',
'25', '31', '19', '6', '9', '18', '10', '23', '28', '31', '17', '6', '27', '15',
'26', '12', '13', '30', '5', '25', '32', '24', '28', '31', '3', '24', '25', '11'
, '28', '26', '4', '22', '15', '23', '32', '22', '13', '11', '26', '20', '28', '
9', '13', '21', '13', '16', '29', '6', '13', '29', '25', '31', '23', '19', '29',
'26', '12', '18']

游客数据类型转换

- 该列表的第一个元素是旅游点名称，该数值对于计算总人数而言是冗余值。
- 为了计算累加和，需要把字符串型数据转换为整数类型数据。

```
jzg_data_str = jzg_data[1:]
jzg_data = list( map(int, jzg_data_str) )
```

游客数据类型转换

输出打印转换后的数据列表jzg_data。

```
[30, 12, 14, 6, 31, 26, 9, 9, 22, 32, 10, 17, 21, 17, 13, 30, 22, 9, 4, 28, 16,
8, 5, 24, 13, 26, 6, 31, 22, 18, 28, 19, 6, 12, 10, 25, 14, 4, 17, 18, 18, 5, 15
, 30, 25, 20, 13, 18, 27, 14, 11, 5, 14, 25, 14, 17, 18, 11, 25, 25, 28, 25, 31,
27, 7, 11, 5, 17, 14, 10, 20, 21, 31, 22, 12, 26, 20, 3, 20, 18, 28, 30, 27, 7,
32, 6, 7, 24, 3, 31, 6, 25, 12, 23, 15, 3, 12, 28, 22, 30, 32, 21, 32, 9, 20, 3,
9, 16, 15, 27, 25, 27, 8, 31, 15, 23, 17, 27, 29, 24, 30, 3, 8, 31, 10, 30, 31,
22, 27, 17, 15, 14, 15, 27, 25, 18, 4, 26, 3, 32, 10, 32, 31, 24, 3, 29, 6, 25,
11, 32, 18, 20, 25, 30, 8, 8, 27, 9, 26, 18, 16, 5, 10, 28, 32, 4, 15, 31, 32, 1
6, 24, 12, 27, 30, 22, 8, 6, 13, 31, 6, 3, 15, 13, 19, 30, 21, 20, 10, 13, 19, 2
1, 18, 12, 23, 24, 27, 28, 3, 21, 6, 6, 23, 26, 23, 20, 26, 25, 31, 19, 6, 9, 18
, 10, 23, 28, 31, 17, 6, 27, 15, 26, 12, 13, 30, 5, 25, 32, 24, 28, 31, 3, 24, 2
5, 11, 28, 26, 4, 22, 15, 23, 32, 22, 13, 11, 26, 20, 28, 9, 13, 21, 13, 16, 29,
6, 13, 29, 25, 31, 23, 19, 29, 26, 12, 18]
```

求解总游客数

调用函数getTotalTourist（），计算jzg_data列表元素的累加和，累加所得的结果即为这段时期通过该旅行社赴九寨沟游客的总量。

```
jzg_total = getTotalTourist(jzg_data)
print("这段时期到九寨沟旅游的总人数是：",jzg_total)
```

代码运行结果，如下：

```
这段时期到九寨沟旅游的总人数是：  4957
```

完整的代码清单

- 读取csv文件需要使用csv包中的reader()函数，因而引入csv包。

- 在进行了文件读取操作之后，需要关闭文件。使用close()函数实现。

```
import csv
def getTotalTourist( place ):
    total = 0
    for dayTourist in place:
        total += dayTourist
    return total
data_file = open('tourist_data.csv','r')
all_data = csv.reader(data_file)
#for day_data in all_data:
#    print( day_data )
#jzg_data = []
#for row in all_data:
#    jzg_data.append(row[1])
jzg_data = [row[1] for row in all_data]
#print(jzg_data)
jzg_data_str = jzg_data[1:]
jzg_data = list( map(int, jzg_data_str) )
#print(jzg_data)
jzg_total = getTotalTourist(jzg_data)
print("这段时期到九寨沟旅游的总人数是：",jzg_total)
data_file.close()
```

计算张家界的游客总量

张家界的每日游客量数据位于数据文件的第三列，因此需要对数据表中的第三列数据进行处理。

```
data_file = open('tourist_data.csv','r')
all_data = csv.reader(data_file)
zjj_data = [row[2] for row in all_data]
zjj_data_str = zjj_data[1:]
zjj_data = list( map(int, zjj_data_str) )
zjj_total = getTotalTourist(zjj_data)
print("这段时期到张家界旅游的总人数是:",
zjj_total )
data_file.close()
```

计算香港的游客总量

香港的每日游客量数据位于数据文件的第四列，因此需要对数据表中的第四列数据进行处理。

```
data_file = open('tourist_data.csv','r')
all_data = csv.reader(data_file)
hk_data = [row[3] for row in all_data]
hk_data_str = hk_data[1:]
hk_data = list( map(int, hk_data_str) )
hk_total = getTotalTourist(hk_data)
print("这段时期到香港旅游的总人数是:", hk_total )
data_file.close()
```

计算东部华侨城的游客总量

东部华侨城的每日游客量数据位于数据文件的第五列，因此需要对数据表中的第五列数据进行处理。

```
data_file = open('tourist_data.csv','r')
all_data = csv.reader(data_file)
dbhqc_data = [row[4] for row in all_data]
dbhqc_data_str = dbhqc_data[1:]
dbhqc_data = list( map(int, dbhqc_data_str) )
dbhqc_total = getTotalTourist(dbhqc_data)
print("这段时期到东部华侨城旅游的总人数是:", dbhqc_total )
data_file.close()
```

计算上海迪士尼的游客总量

上海迪士尼的每日游客量数据位于数据文件的第六列，因此需要对数据表中的第六列数据进行处理。

```
data_file = open('tourist_data.csv','r')
all_data = csv.reader(data_file)
shdisney_data = [row[5] for row in all_data]
shdisney_data_str = shdisney_data[1:]
shdisney_data = list( map(int, shdisney_data_str) )
shdisney_total = getTotalTourist(shdisney_data)
print('通过该旅行社赴上海迪士尼的游客总数为:', shdisney_total)
data_file.close()
```

运行程序输出结果


```
这段时期到九寨沟旅游的总人数是：  4957
这段时期到九寨沟旅游的总人数是：4911
这段时期到香港旅游的总人数是：3425
这段时期到东部华侨城旅游的总人数是：1459
这段时期到上海迪士尼旅游的总人数是：4799
```

使用numpy进行分析

引入numpy包

需要使用numpy中的函数，如loadtxt()，因此需要导入numpy包，并取别名为“np”。

```
import numpy as np
```

从文件读取数据

使用loadtxt( )函数读取数据，其返回值是一个数组对象，默认数据类型是浮点数。一次性读取其第2、3、4、5、6列，存储在数组jzg_data、zjj_data、hk_data、dbhqc_data、shdisney_data中。

```
(jzg_data,
 zjj_data,
 hk_data,
 dbhqc_data,
 shdisney_data) = np.loadtxt('tourist_data.csv', \
                             skiprows = 1, \
                             dtype = 'int', \
                             delimiter = ',', \
                             usecols = (1,2,3,4,5), \
                             unpack = True)
print( shdisney_data)
```

从文件读取数据

输出打印shdisney_data。

```
[28  9 32 15 10 25 28 15 14 20 26 31  6 32  3 32  4 30 31 25  4 20  6 21
 32 16  8  5 19 31 25 12  4 17 10 15 28 22 26 21 30 25 13  6 15 18 17 16
 19 27 28 10  8 17 29 25 10 21 32 25  5 15 16 19 26 16 14 28 21 28  3 31
 22 19 29 22 21  4 11 32  9 14  6 23 14 25 10 15 26  9 23 29 15  6 29  9
 16  6 30 10 19 27  7  9 10 20  9 27 27 30  6  4 27  5 16 19 30 32 11  7
  6 11 27 21 32 22  9 20 24  8 17 29  8 16 32 18 14 30 17 20 21 21 13  8
 23 30 14 22 18 14  5 22 32  6 27 10 13 15  4 32 30  7 20 11 13 26 19 18
 30 22 29  8 23  3 31 16 32 12 29  9 32  6  5  4 31 32 10 27  5 15 24 22
 25 32 10 32 15 20 18 25 31 26  5 21  8  7 30 19  7 18 19 31  5  5 17 27
 15 19 21 11 12 26 23 20 31  5 11  9 25  8  5  3 19 19  5 10  3 25 11  7
  3 24 21  8 32 21 12 26 31 32  6 31 32 11  6 24 18 31 13 28  3 15 32 14]
```

求解游客总数

使用方法sum()，求解数组元素的累加和。求解数组jzg_data元素累加和。

```
jzg_total = jzg_data.sum()
```

求解游客总数

使用相同方法求解其他数组对象元素的累加和。

```
jzg_total = jzg_data.sum()
zjj_total = zjj_data.sum()
hk_total = hk_data.sum()
dbhqc_total = dbhqc_data.sum()
shdisney_total = shdisney_data.sum()

print("(numpy)这段时期到九寨沟旅游的总人数是:",\
      jzg_total)
print("(numpy)这段时期到张家界旅游的总人数是:",\
      zjj_total)
print("(numpy)这段时期到香港旅游的总人数是:",\
      hk_total)
print("(numpy)这段时期到东部华侨城旅游的总人数是:",\
      dbhqc_total)
print("(numpy)这段时期到上海迪士尼旅游的总人数是:",\
      shdisney_total)
```

求解游客总数

输出打印这5个城市的旅客总量。

```
(Numpy)这段时期到九寨沟旅游的总人数是:  4957
(Numpy)这段时期到张家界旅游的总人数是:  4911
(Numpy)这段时期到香港旅游的总人数是:  3425
(Numpy)这段时期到东部华侨城旅游的总人数是:  1459
(Numpy)这段时期到上海迪士尼旅游的总人数是:  4799
```

使用pandas进行分析

引入pandas包

需要使用pandas中的函数，因此需要导入pandas包，并取别名为“pd”。

```
import pandas as pd
```

从文件读取数据

使用pandas的read_csv( ) 函数读取数据并存储在data中。

```
data = pd.read_csv('tourist_data.csv',\
                   index_col = u'日期',\
                   header = 0,\
                   encoding='gb2312' )
print("data的数据类型是: ", type(data))
print(data)
```

从文件读取数据

输出打印data。

```
data的数据类型是: <class 'pandas.core.frame.DataFrame'>
            九寨沟  张家界  香港  东部华侨城  上海迪士尼
日期
2017/1/1     30    17   17      3      28
2017/1/2     12    21    8      3       9
2017/1/3     14    22   15      1      32
2017/1/4      6    29   13      7      15
2017/1/5     31    15    5      3      10
2017/1/6     26    23   17      6      25
2017/1/7      9    18   12      7      28
2017/1/8      9    19   12      4      15
2017/1/9     22    24   11     10      14
```

求解各城市游客总数

可以使用sum()函数计算DataFrame对象中某一列所有元素的累加和。获取DataFrame这种二维表的某一列，可以通过使用列名字或列索引实现。

```
jzg_total = data['九寨沟'].sum()
zjj_total = data['张家界'].sum()
hk_total = data['香港'].sum()
dbhqc_total = data['东部华侨城'].sum()
shdisney_total = data['上海迪士尼'].sum()
print("(pandas)这段时期到九寨沟旅游的总人数是:",\
      jzg_total)
print("(pandas)这段时期到张家界旅游的总人数是:",\
      zjj_total)
print("(pandas)这段时期到香港旅游的总人数是:",\
      hk_total)
print("(pandas)这段时期到东部华侨城旅游的总人数是:",\
      dbhqc_total)
print("(pandas)这段时期到上海迪士尼旅游的总人数是:",\
      shdisney_total)
```

求解各城市游客总数

输出打印计算结果。

```
(Pandas)这段时期到九寨沟旅游的总人数是: 4957
(Pandas)这段时期到张家界旅游的总人数是: 4911
(Pandas)这段时期到香港旅游的总人数是: 3425
(Pandas)这段时期到东部华侨城旅游的总人数是: 1459
(Pandas)这段时期到上海迪士尼旅游的总人数是: 4799
```

三种实现方法比较

代码量不同

使用纯Python的实现，具有最多行的代码， numpy次之，pandas最少。普遍的共识是，使用第三方扩展包可以更优雅的完成数据分析的工作。

处理的数据对象不同

纯Python的实现中，处理的数据存储在序列对象中。对序列对象进行处理，通常需要进行多次循环实现、创建多个函数。这也是导致代码行数更多的原因。

numpy处理的对象是数组对象，具有多个预定义的数据分析函数，高效的实现常用数据分析开发工作。

pandas处理的对象是DateFrame或Series对象，相比较numpy的数组而言，这些对象用于处理矩阵或者二维数组更加方便。

日照职业技术学院
RIZHAO POLYTECHNIC

感谢观看

主讲：赵娜