网易出品 PRODUCE BY NETEASE

Pandas入门

讲师：萨缪尔 Samuel

萨缪尔老师

网易数据分析教研负责人，

前盛大游戏战略规划总监、前腾讯游戏商业智能中心Leader

上海交通大学高金硕士 、《哈佛管理导师》外部导师

- **知乎大V：「萨缪尔」**主要聚集与商业分析、行业与战略分析、Python数据分析等
- **BAT**互联网巨头商业洞察分析 & 咨询公司战略咨询背景
- 擅长行业趋势研究和战略管理咨询工具，**为20多家上市公司提供战略发展决策建议**

课程亮点

1. Pandas常见的数据类型
2. Pandas数据基本操作
3. Pandas数据分析操作
4. 案例

1

Pandas常见的数据类型

Pandas常见的数据类型

数据类型	性质	描述	使用方法
Series	一维数组	Series是Pandas中的基本对象，在NumPy的ndarray基础上进行扩展。Series支持下标存取元素和索引存取元素	index是索引对象，用于保存标签信息；values是保存元素值的数组
DataFrame	二维数组	DataFrame（数据框）类似于Excel电子表格，使用字典创建DataFrame实例时，将字典的键直接设置为列索引，并且指定一个列表作为字典的值	通过索引和行列取值

Pandas常见的数据基本操作

2

Dataframe数据的读取和操作

语句	示例	语句	示例
数据框的构建： df=pd.DataFrame({'x': ['a', 'b','c'], 'y':range(1,4), 'z':[2,5,3]})	x y z a 1 2 b 2 5 c 3 3	选取某一列： df['y'] df.y df.loc[:,['y']] df.iloc[:,[1]]	x y z a 1 2 b 2 5 c 3 3
选取多列： df[['x','y']] df.loc[:,['x','y']] df.iloc[:,[1,2]]	x y z a 1 2 b 2 5 c 3 3	选取某一行： df.loc[1,:] df.iloc[1,:]	x y z a 1 2 b 2 5 c 3 3
选取多行： df.loc[[0,1],:] df.iloc[[0,1],:]	x y z a 1 2 b 2 5 c 3 3	选取某个元素： df.loc[1,'y'] df.loc[[1],['y']] df.iloc[1,1]	x y z a 1 2 b 2 5 c 3 3
单条件过滤： df[df.z>=3]	x y z a 1 2 b 2 5 c 3 3	多条件过滤： df[(df.z>=3) & (df.z<=4)] df.query('z>=3 & z<=4')	x y z a 1 2 b 2 5 c 3 3

Pandas常见的数据分析操作

3

Pandas基础的数据描述

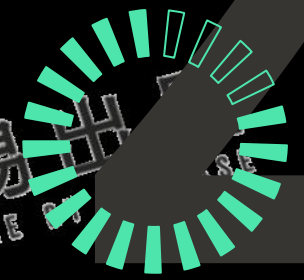
```
df.shape # (100, 6) 查看行数和列数

df.info() # 查看索引、数据类型和内存信息

df.describe() # 查看数值型列的汇总统计

df.dtypes # 查看各字段类型

df.axes # 显示数据行和列名

df.columns # 列名
```

类数据透视表操作

```
pd.pivot_table（data, values = None，index = None，columns = None，aggfunc ='mean'，fill_value = None，margin = False，dropna = True，margins_name ='All’ ）
```

data：DataFrame对象
values：要聚合的列或列的列表
index：数据透视表的index，从原数据的列中筛选
columns：数据透视表的columns，从原数据的列中筛选
aggfunc：用于聚合的函数，默认为numpy.mean，支持numpy计算方法
fill_value: 用于替换缺失值的值
margin: 添加所有行/列
dropna: 不包括条目为 NaN的列，默认为True
margin_name: 当margin为True时，将包含总计的行/列的名称

案例：世界500强数据



4

案例：世界500强数据基本操作

- 数据读取
- 描述型分析
- 数据选取
- 类数据透视表操作

课程总结

1. Pandas常见的数据类型，Series、DataFrame
2. DataFrame数据的基本操作：查询行、列；条件查询
3. Pandas的描述性数据分析，shape，info()，describe()
4. Pandas实现数据透视表pivot_table

谢谢观看