法律声明

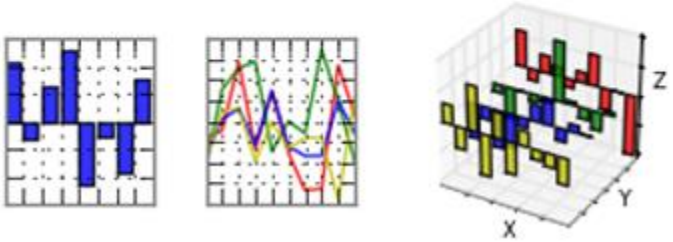
- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：大数据分析挖掘
  - 新浪微博：ChinaHadoop

第六讲

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



数据分析工具Pandas进阶

--梁斌

目录

- Pandas层级索引
- 数据的分组与聚合
- 数据的分组运算
- Pandas透视表与交叉表
- 实战案例：互联网电影资料库分析

目录

Pandas层级索引

层级索引（hierarchical indexing）

- MultiIndex对象
- 选取子集
  - 外层选取 ser_obj[‘outer_label’]
  - 内层选取 ser_obj[:, ‘inner_label’]
- 常用于分组操作、透视表的生成等
- 交换分层顺序
  - swaplevel()
- 排序分层
  - sortlevel()

示例代码： 01_pandas_multi_index.ipynb

Pandas层级索引

层级索引（续）

		0	1	2	3
bar	one	-1.133800	0.548640	1.109034	0.643708
	two	-0.792654	0.518681	-0.611958	0.913413
baz	one	0.775624	-2.520829	-0.472691	-0.557803
	two	0.190005	0.435193	1.635680	1.584821
foo	one	-0.592235	-0.361735	1.336444	-1.280014
	two	-1.016622	1.409086	0.114743	0.408211
qux	one	0.662941	-1.258482	-0.373214	-0.974658
	two	-0.931004	0.596507	0.148323	0.475039

示例代码： 01_pandas_multi_index.ipynb

目录

Pandas分组与聚合

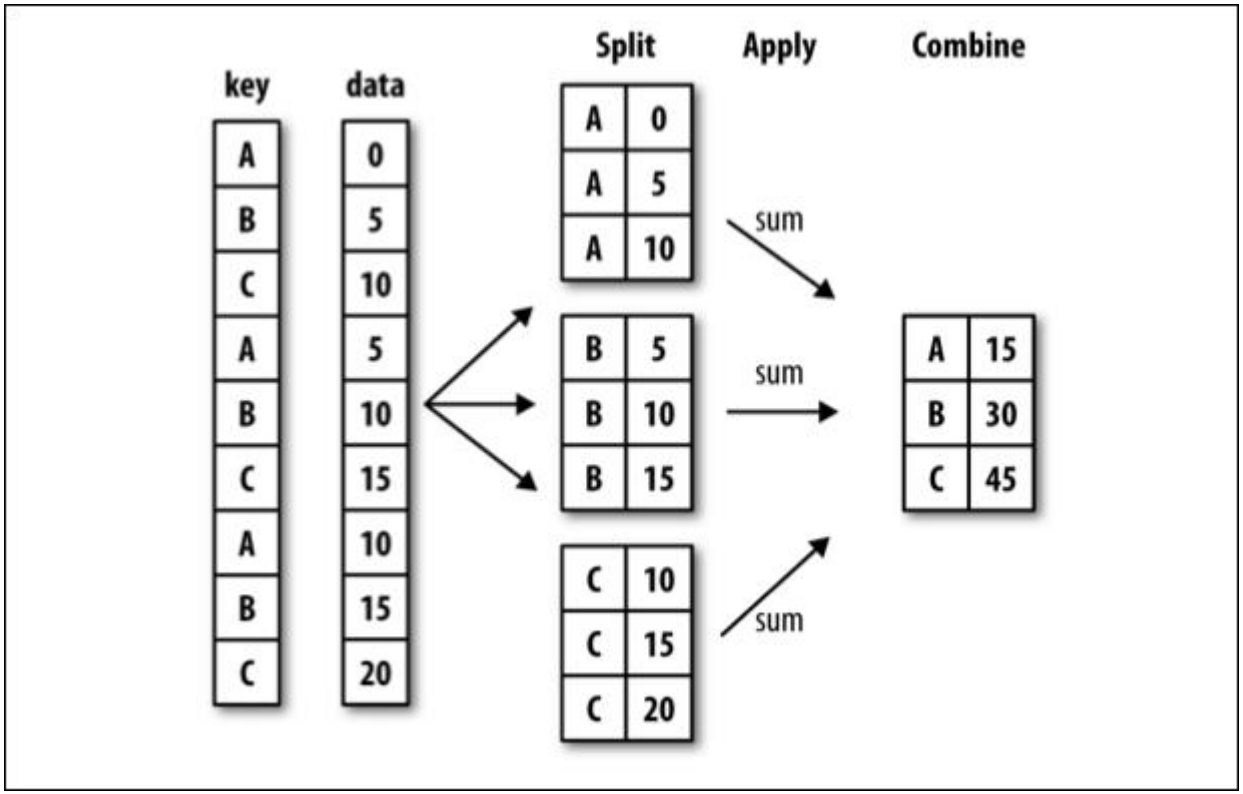
分组 (groupby)

- 对数据集进行分组，然后对每组进行统计分析
- SQL能够对数据进行过滤，分组聚合
- pandas能利用groupby进行更加复杂的分组运算
- 分组运算过程
  - split->apply->combine
  - 拆分：进行分组的根据
  - 应用：每个分组运行的计算规则
  - 合并：把每个分组的计算结果合并起来

Pandas分组与聚合

分组 (续)

- 分组运算过程
  - split->apply->combine



Pandas分组与聚合

分组 (续)

- GroupBy对象：DataFrameGroupBy，SeriesGroupBy
- GroupBy对象没有进行实际运算，只是包含分组的中间数据
- 对GroupBy对象进行分组运算/多重分组运算，如mean()
  - 非数值数据不进行分组运算
- size() 返回每个分组的元素个数

示例代码： 02_pandas_groupby.ipynb

Pandas分组与聚合

分组 (续)

- 按列名分组
  - obj.groupby(‘label’)
- 按列名多层分组
  - obj.groupby([‘label1’ , ‘label2’])->多层dataframe
- 按自定义的key分组
  - obj.groupby(self_def_key)
  - 自定义的key可为列表或多层列表
- unstack可以将多层索引的结果转换成单层的dataframe

示例代码： 02_pandas_groupby.ipynb

Pandas分组与聚合

分组 (续)

- GroupBy对象支持迭代操作
  - 每次迭代返回一个元组 (group_name, group_data)
  - 可用于分组数据的具体运算
- GroupBy对象可以转换成列表或字典
- Pandas也支持按列分组
- 其他分组方法
  - 通过字典分组
  - 通过函数分组，函数传入的参数为行索引或列索引
  - 通过索引级别分组

示例代码： 02_pandas_groupby.ipynb

Pandas分组与聚合

聚合 (aggregation)

- 数组产生标量的过程，如mean()、count()等
- 常用于对分组后的数据进行计算
- 内置的聚合函数
  - sum(), mean(), max(), min(), count(), size(), describe()
- 可自定义函数，传入agg方法中
  - grouped.agg(func)
  - func的参数为groupby索引对应的记录

示例代码： 02_pandas_groupby.ipynb

Pandas分组与聚合

聚合 (续)

- 应用多个聚合函数
  - 同时应用多个函数进行聚合操作，使用函数列表
  - 对不同的列分别作用不同的聚合函数，使用dict

示例代码： 02_pandas_groupby.ipynb

Pandas分组与聚合

聚合 (续)

- 常用的内置聚合函数

函数名	说明
count	分组中非NA值的数量
sum	非NA值的和
mean	非NA值的平均值
median	非NA值的算术中位数
std、var	无偏（分母为n - 1）标准差和方差
min、max	非NA值的最小值和最大值
prod	非NA值的积
first、last	第一个和最后一个非NA值

目录

数据的分组运算

分组运算

- 原因:
  - 聚合运算改变了原始数据的shape
  - 如何保持原始数据的shape?
    - 使用merge的外连接，比较复杂
    - transform
- transform的计算结果和原始数据的shape保持一致
  - 如：grouped.transform(np.mean)
  - 也可传入自定义函数

示例代码： 03_pandas_grouped_apply_transform.ipynb

数据的分组运算

分组运算 (续)

- grouped.apply(func)
  - func函数在各分组上调用，然后结果通过pd.concat组装到一起
  - 产生层级索引
    - 外层索引是分组名
    - 内层索引是df_obj的行索引
  - 禁止层级索引, group_keys=False
- apply可以用来处理不同分组内的缺失数据填充
  - 如：填充该分组的均值

示例代码： 03_pandas_grouped_apply_transform.ipynb

目录

Pandas透视表与交叉表

透视表 (pivot table)

- 根据一个或多个键对数据进行聚合
- 根据行和列的分组键将数据划分到各个区域中
- pd.pivot_table(df_data)
- index参数：透视表中的索引值
- columns参数：分组的列
- aggfunc：应用在每个区域的聚合函数，默认为np.mean
- fill_value：替换结果中的缺失值

示例代码： 04_pandas_pivottab_crosstab.ipynb

Pandas透视表与交叉表

交叉表 (cross table)

- 用于计算分组频率的特殊透视表
- pd.crosstab(index, columns)
  - index: 分组数据，交叉表的行索引
  - columns: 交叉表的列索引

示例代码： 04_pandas_pivottab_crosstab.ipynb

目录

实战案例

项目介绍

- https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

实战案例

项目介绍

- https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset
- 背景：
  1. 是否有统一的方法判断一部电影的好坏？
  2. 能否通过以往评论预测下一部电影的评分？
  3. 海报人脸的个数与电影评分是否有关系？

示例代码：lecture06_proj.zip

实战案例

项目任务

- 使用分组统计数据集的基本信息
  - 查看票房统计信息
  - imdb评分统计
  - 电影产量趋势
- 基于电影类型的分析
- 可视化分析结果

涉及知识点

- Pandas分组操作与统计
- Pandas绘图

示例代码：lecture06_proj.zip

实战案例

分析步骤

1．查看数据

2．明确分析目标

- 分组统计基本信息
- 统计电影类型信息

3．处理缺失数据（可选）

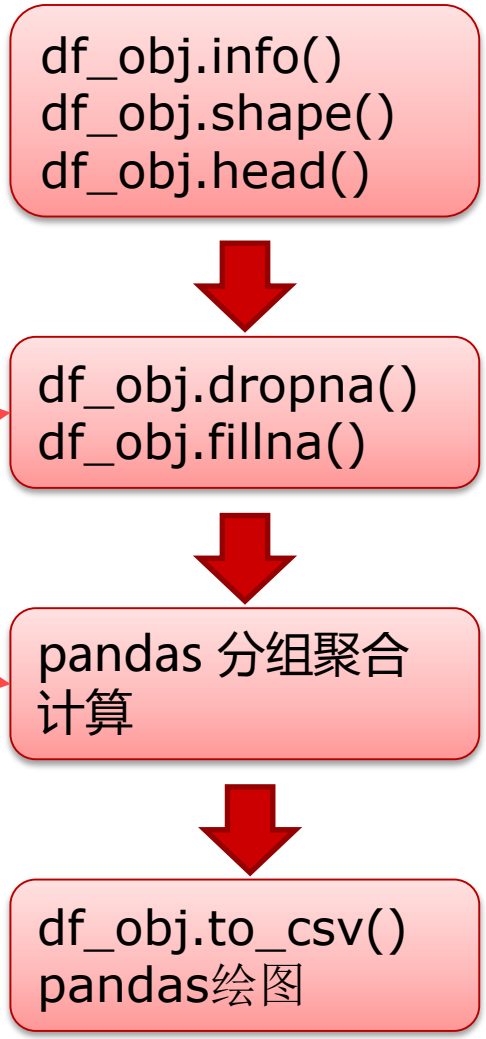
4．数据统计分析

- 模块化常用功能

5．保存分析结果

1．分析结果数据

2．可视化结果



参考

- Pandas高级索引/层级索引

  http://pandas.pydata.org/pandas-docs/stable/advanced.html
- Pandas中的GroupBy

  http://pandas.pydata.org/pandas-docs/stable/groupby.html
- Pandas透视表

  http://pandas.pydata.org/pandas-docs/stable/reshaping.html
- 《Python for Data Analysis》

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

小象问答 @Robin_TY

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop