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

第五讲

pandas

$$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$$

数据分析工具Pandas基础

--梁斌

目录

- Pandas的数据结构
- Pandas的数据操作
- Pandas统计计算和描述
- Pandas的绘图函数
- 实战案例：星际争霸II重放分析

目录

Pandas的数据结构

Series

- 类似一维数组的对象
- 通过list构建Series
  - ser_obj = pd.Series(range(10))
- 由数据和索引组成
  - 索引在左，数据在右
  - 索引是自动创建的
- 获取数据和索引
  - ser_obj.index, ser_obj.values
- 预览数据
  - ser_obj.head(n)

示例代码： 01_pandas_data_structures.ipynb

Pandas的数据结构

Series (续)

- 通过索引获取数据
  - ser_obj[idx]
- 索引与数据的对应关系仍保持在数组运算的结果中
- 通过dict构建Series
- name属性
  - ser_obj.name, ser_obj.index.name

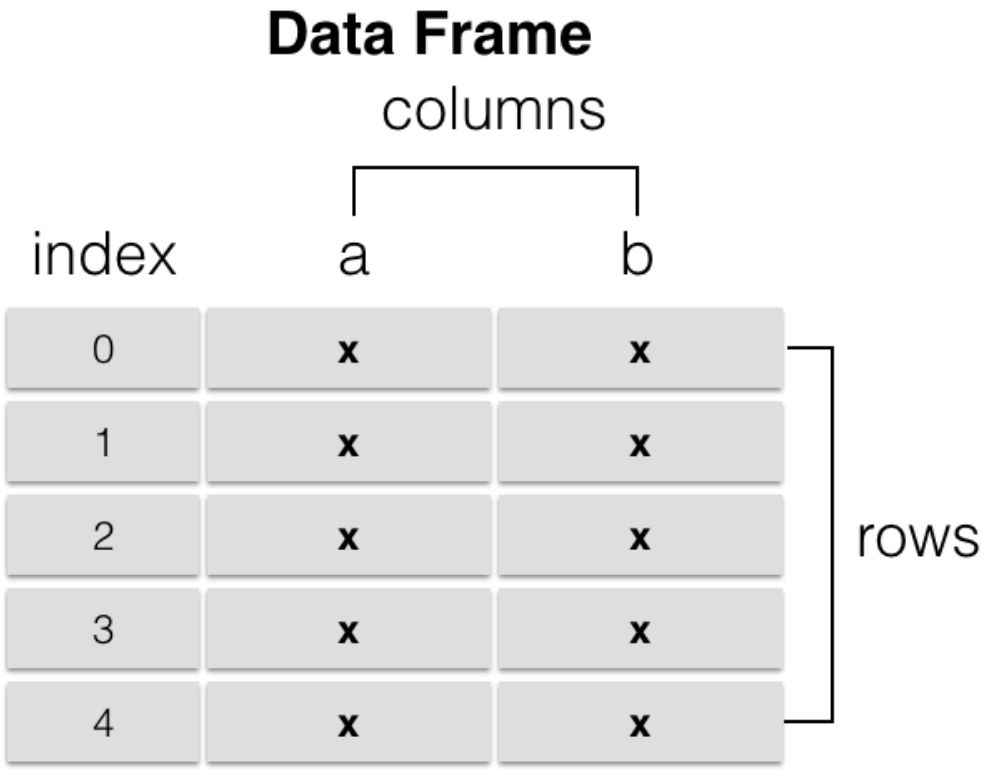
示例代码： 01_pandas_data_structures.ipynb

Pandas的数据结构

DataFrame

示例代码： 01_pandas_data_structures.ipynb

- 类似多维数组/表格数据 (如，excel, R中的data.frame)
- 每列数据可以是不同的类型，what about ndarray?
- 索引包括列索引和行索引



Pandas的数据结构

示例代码： 01_pandas_data_structures.ipynb

DataFrame

- 通过ndarray构建DataFrame
- 通过dict构建DataFrame
- 通过列索引获取列数据（Series类型 ）
  - df_obj[col_idx] 或 df_obj.col_idx
- 增加列数据，类似dict添加key-value
  - df_obj[new_col_idx] = data
- 删除列
  - del df_obj[col_idx]

Pandas的数据结构

索引对象Index

- Series和DataFrame中的索引都是Index对象
- 不可变(immutable)
  - 保证了数据的安全
- 常见的Index种类
  - Index
  - Int64Index
  - MultiIndex，“层级”索引
  - DatetimeIndex，时间戳类型

示例代码： 01_pandas_data_structures.ipynb

目录

Pandas的数据操作

索引操作

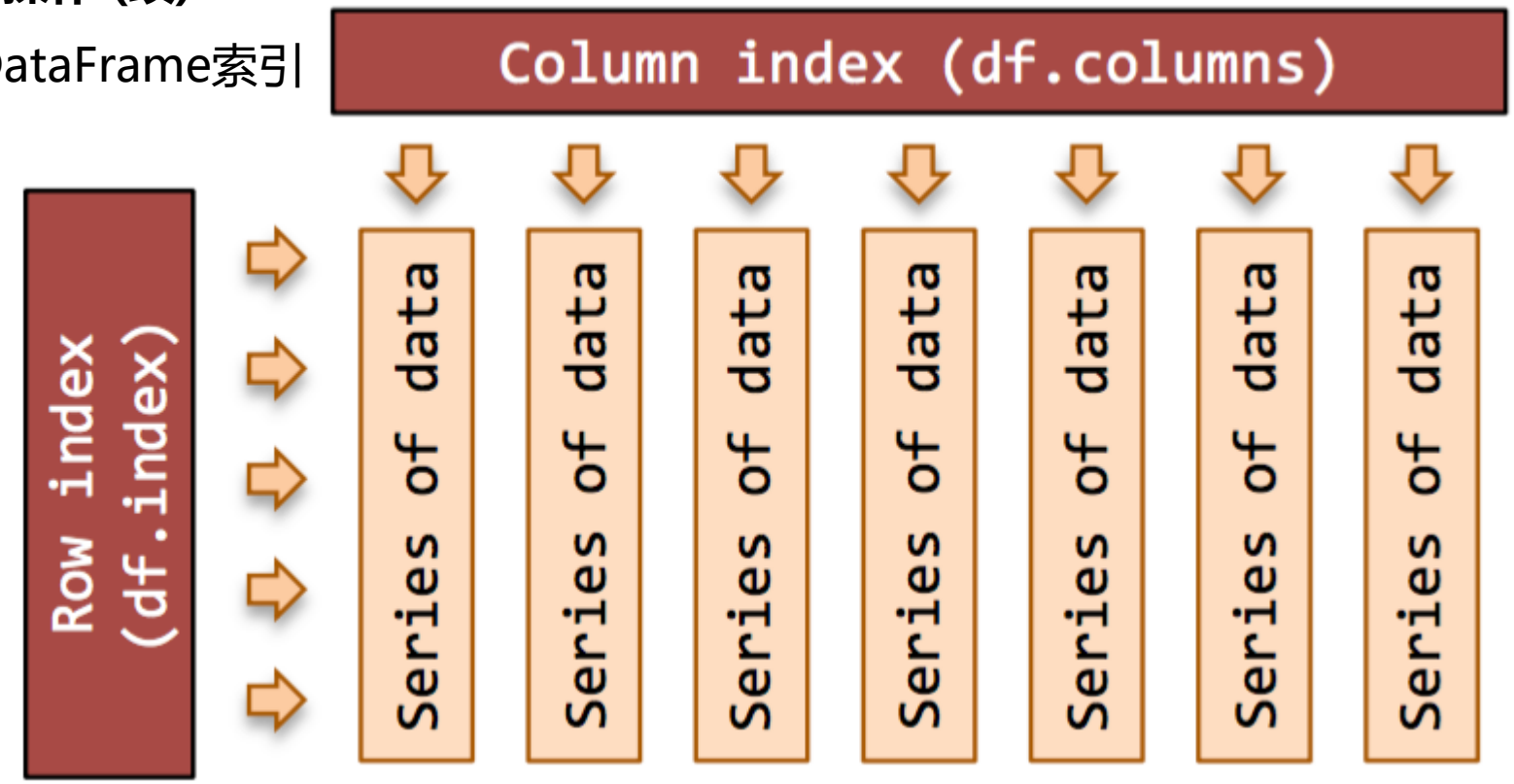
- Series索引
  - 行索引，ser_obj[‘label’], ser_obj[pos]
  - 切片索引，ser_obj[2:4], ser_obj[‘label1’:’label3’]
    - 注意，按索引名切片操作时，是包含终止索引的。
  - 不连续索引，ser_obj[[‘label1’,’label2’, ‘label3’]]
    ser_obj[[pos1, pos2, pos3]]
  - 布尔索引

示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

索引操作 (续)

- DataFrame索引



示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

索引操作 (续)

- DataFrame索引
  - 列索引
    - df_obj[‘label’]
  - 不连续索引
    - df_obj[[‘label1’ , ‘label2’]]

示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

索引操作总结

- Pandas的索引可归纳为3种
- .loc，标签索引
- .iloc，位置索引
- .ix，标签与位置混合索引
  - 先按标签索引尝试操作，然后再按位置索引尝试操作
- 注意
  - DataFrame索引时可将其看作ndarray操作
  - 标签的切片索引是包含末尾位置的

示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

运算与对齐

- 按索引对齐运算，没对齐的位置补NaN
  - Series 按行索引对齐
  - DataFrame按行、列索引对齐
- 填充未对齐的数据进行运算
  - 使用add, sub, div, mul
  - 同时通过fill_value指定填充值
- 填充NaN
  - fillna

示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

函数应用

- 可直接使用NumPy的ufunc函数，如abs等
- 通过apply将函数应用到行或列上
  - 注意指定轴的方向，默认axis=0
- 通过applymap将函数应用到每个数据上

示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

排序

- sort_index，索引排序
  - 对DataFrame操作时注意轴方向
- 按值排序
  - sort_values(by= ‘label’)

示例代码： 02_pandas_data_process.ipynb

Pandas的数据操作

处理缺失数据

- 判断是否存在缺失值
  - ser_obj.isnull(), df_obj.isnull()
- dropna
  - 丢弃缺失数据
- fillna
  - 填充缺失数据

示例代码： 02_pandas_data_process.ipynb

目录

Pandas统计计算和描述

常用的统计计算

- sum, mean, max, min...
- axis=0 按列统计，axis=1按行统计
- skipna 排除缺失值， 默认为True
- idmax, idmin, cumsum

统计描述

- describe 产生多个统计数据

示例代码： 03_pandas_stats.ipynb

Pandas统计计算和描述

方法	说明
count	非NA值的数量
describe	针对Series或各DataFrame列计算汇总统计
min、max	计算最小值和最大值
argmin、argmax	计算能够获取到最小值和最大值的索引位置（整数）
idxmin、idxmax	计算能够获取到最小值和最大值的索引值
quantile	计算样本的分位数（0到1）
sum	值的总和
mean	值的平均数
median	值的算术中位数（50%分位数）
mad	根据平均值计算平均绝对离差
var	样本值的方差
std	样本值的标准差

Pandas统计计算和描述

方法	说明
skew	样本值的偏度（三阶矩）
kurt	样本值的峰度（四阶矩）
cumsum	样本值的累计和
cummin、cummax	样本值的累计最大值和累计最小值
cumprod	样本值的累计积
diff	计算一阶差分（对时间序列很有用）
pct_change	计算百分数变化

目录

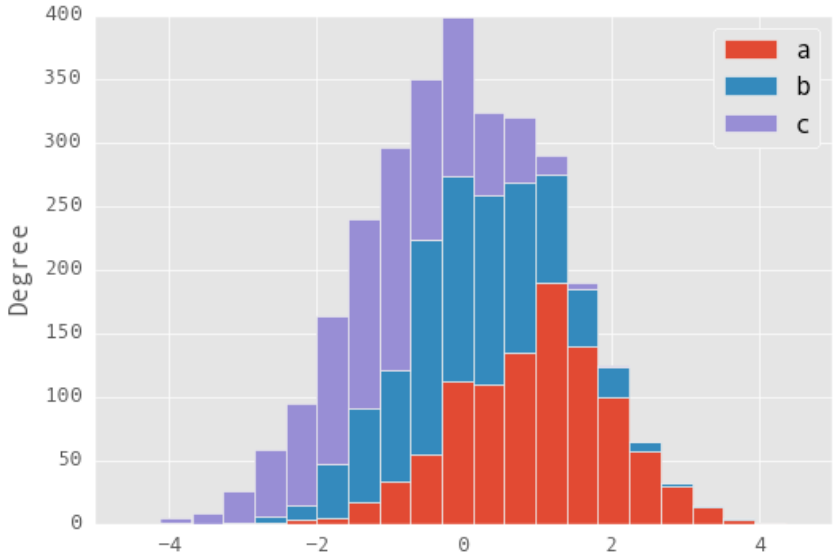
Pandas的绘图函数

Matplotlib

- 相对“低级” 的绘图工具
- 需要自己完成基础组件的组装，如图例、标题、标签等。

Pandas绘图函数

- 高效、简单
- 根据数据的索引、标签进行绘图



Pandas的绘图函数

Pandas常用的绘图函数

- 线形图， ser_obj.plot(), df_obj.plot()
- 柱状图，ser_obj.plot(kind= ‘bar’), df_obj.plot(kind= ‘bar’)
  - barh，水平柱状图
- 散布矩阵
  - pd.scatter_matrix(df_obj)
- 更多绘图函数请参考最后的链接

示例代码： 04_pandas_plot.ipynb

目录

实战案例

项目介绍

- https://www.kaggle.com/sfu-summit/starcraft-ii-replay-analysis
- 战队的各属性分析

项目任务

- 分析各战队的统计信息
- 可视化分析结果

涉及知识点

- Pandas数据操作
- Matplotlib绘图

示例代码：lecture05_proj.zip

实战案例

分析步骤

1．查看数据

2．明确分析目标

- 分析各战队的属性
- 可视化属性统计信息

3．处理缺失数据（可选）

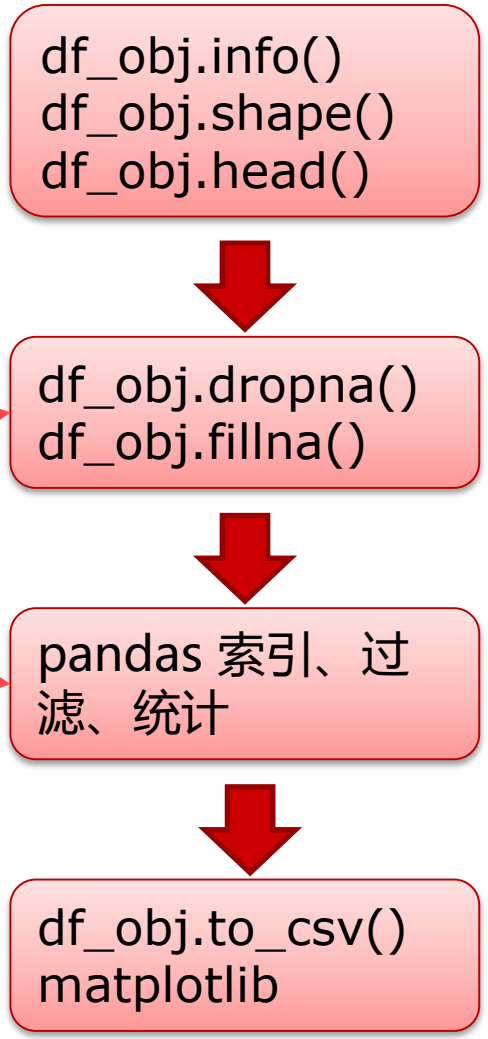
4．数据统计分析

- 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

- 10分钟了解Pandas

  http://pandas.pydata.org/pandas-docs/stable/10min.html
- Pandas的索引操作

  http://pandas.pydata.org/pandas-docs/stable/indexing.html
- Pandas处理缺失数据

  http://pandas.pydata.org/pandas-docs/stable/missing_data.html
- Pandas绘图

  http://pandas.pydata.org/pandas-docs/version/0.18.1/visualization.html
- 项目参考

  https://www.kaggle.com/jonlee317/d/sfu-summit/starcraft-ii-replay-analysis/starcraft-simple-data-exploration

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