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

第七讲

数据的规整与可视化

--梁斌

目录

- 数据清洗、连接、合并、重构和转换
- 常用的Python数据可视化工具
  - Matplotlib回顾及扩充
  - Seaborn绘图
  - 交互式数据可视化—Bokeh绘图
- 实战案例：空难历史数据分析

目录

- 数据清洗、连接、合并、重构和转换
- 常用的Python数据可视化工具
  - Matplotlib回顾及扩充
  - Seaborn绘图
  - 交互式数据可视化—Bokeh绘图
- 实战案例：空难历史数据分析

数据清洗

- 数据清洗是数据分析关键的一步，直接影响之后的处理工作
- 数据需要修改吗？有什么需要修改的吗？数据应该怎么调整才能适用于接下来的分析和挖掘？
- 是一个迭代的过程，实际项目中可能需要不止一次地执行这些清洗操作
- 处理缺失数据
  - pd.fillna()，pd.dropna()

数据连接

pd.merge

- 根据单个或多个键将不同DataFrame的行连接起来
- 类比数据库的连接操作 (第三课)
- 默认将重叠列的列名作为“外键”进行连接
  - on显示指定“外键”
  - left_on，左侧数据的“外键”
  - right_on，右侧数据的“外键”
- 默认是“内连接”(inner)，即结果中的键是交集

示例代码： 01_data_merge.ipynb

数据连接

pd.merge (续)

- how指定连接方式
- “外连接”(outer)，结果中的键是并集
- “左连接”(left)
- “右连接”(right)
- 处理重复列名
  - suffixes，默认为_x, _y
- 按索引连接
  - left_index=True或right_index=True

示例代码： 01_data_merge.ipynb

数据合并

pd.concat

- 沿轴方向将多个对象合并到一起
- NumPy的concat
  - np.concatenate
- pd.concat
  - 注意指定轴方向，默认axis=0
  - join指定合并方式，默认为outer
  - Series合并时查看行索引
  - DataFrame合并时同时查看行索引和列索引

示例代码： 02_data_concat.ipynb

数据重构

重构

- stack
  - 将列索引旋转为行索引，完成层级索引
  - DataFrame->Series
- unstack
  - 将层级索引展开
  - Series->DataFrame
  - 默认操作内层索引，即level=-1

示例代码： 03_data_reshape.ipynb

数据重构

重构

- stack
  - 将列索引旋转为行索引，完成层级索引
  - DataFrame->Series
- unstack
  - 将层级索引展开
  - Series->DataFrame
  - 默认操作内层索引，即level=-1

示例代码： 03_data_reshape.ipynb

数据转换

处理重复数据

- duplicated() 返回布尔型Series表示每行是否为重复行
- drop_duplicates() 过滤重复行
  - 默认判断全部列
  - 可指定按某些列判断

map

- Series根据map传入的函数对每行或每列进行转换

数据替换

- replace

示例代码： 04_data_transform.ipynb

目录

- 数据清洗、连接、合并、重构和转换
- 常用的Python数据可视化工具
  - Matplotlib回顾及扩充
  - Seaborn绘图
  - 交互式数据可视化—Bokeh绘图
- 实战案例：空难历史数据分析

目录

- 数据清洗、连接、合并、重构和转换
- 常用的Python数据可视化工具
  - Matplotlib回顾及扩充
  - Seaborn绘图
  - 交互式数据可视化—Bokeh绘图
- 实战案例：空难历史数据分析

Matplotlib

颜色、标记、线型

- ax.plot(x, y, ‘r--’)
  - 等价于ax.plot(x, y, linestyle= ‘--’, color= ‘r’)

颜色

- b: blue
- g: green
- r: red
- c: cyan
- m: magenta
- y: yellow
- k: black
- w: white

标记

marker	description
"."	point
","	pixel
"o"	circle
"v"	triangle_down
"^"	triangle_up
"<"	triangle_left

线型

linestyle	description
'-' or 'solid'	solid line
'--' or 'dashed'	dashed line
'-.' or 'dashdot'	dash-dotted line
':' or 'dotted'	dotted line
'None'	draw nothing
' '	draw nothing
''	draw nothing

示例代码： 05_matplotlib.ipynb

Matplotlib

刻度、标签、图例

- 设置刻度范围
  - plt.xlim(), plt.ylim()
  - ax.set_xlim(), ax.set_ylim()
- 设置显示的刻度
  - plt.xticks(), plt.yticks()
  - ax.set_xticks(), ax.set_yticks()
- 设置刻度标签
  - ax.set_xticklabels(), ax.set_yticklabels()
- 设置坐标轴标签
  - ax.set_xlabel(), ax.set_ylabel()

示例代码： 05_matplotlib.ipynb

Matplotlib

刻度、标签、图例 (续)

- 设置标题
  - ax.set_title()
- 图例
  - ax.plot(label= ‘legend’)
  - ax.legend(), plt.legend()
    - loc= ‘best’ 自动选择放置图例最佳位置

matplotlib设置

- plt.rc()
- http://matplotlib.org/users/customizing.html

示例代码： 05_matplotlib.ipynb

目录

- 数据清洗、连接、合并、重构和转换
- 常用的Python数据可视化工具
  - Matplotlib回顾及扩充
  - Seaborn绘图
  - 交互式数据可视化—Bokeh绘图
- 实战案例：空难历史数据分析

Seaborn

什么是Seaborn

- Python中的一个制图工具库，可以制作出吸引人的、信息量大的统计图
- 在Matplotlib上构建，支持numpy和pandas的数据结构可视化，甚至是scipy和statsmodels的统计模型可视化

特点

- 多个内置主题及颜色主题
- 可视化单一变量、二维变量用于比较数据集中各变量的分布情况
- 可视化线性回归模型中的独立变量及不独立变量

Seaborn

特点 (续)

- 可视化矩阵数据，通过聚类算法探究矩阵间的结构
- 可视化时间序列数据及不确定性的展示
- 可在分割区域制图，用于复杂的可视化

安装

- conda install seaborn
- pip install seaborn

Seaborn

数据集分布可视化

- 单变量分布 sns.distplot()
  - 直方图 sns.distplot(kde=False)
  - 核密度估计 sns.distplot(hist=False) 或 sns.kdeplot()
  - 拟合参数分布 sns.distplot(kde=False, fit=)
- 双变量分布
  - 散布图 sns.jointplot()
  - 二维直方图 Hexbin sns.jointplot(kind= ‘hex’)
  - 核密度估计 sns.jointplot(kind= ‘kde’)
- 数据集中变量间关系可视化 sns.pairplot()

示例代码： 06_seaborn.ipynb

Seaborn

类别数据可视化

- 类别散布图
  - sns.stripplot() 数据点会重叠
  - sns.swarmplot() 数据点避免重叠
  - hue指定子类别
- 类别内数据分布
  - 盒子图 sns.boxplot(), hue指定子类别
  - 小提琴图 sns.violinplot(), hue指定子类别
- 类别内统计图
  - 柱状图 sns.barplot()
  - 点图 sns.pointplot()

示例代码： 06_seaborn.ipynb

目录

- 数据清洗、连接、合并、重构和转换
- 常用的Python数据可视化工具
  - Matplotlib回顾及扩充
  - Seaborn绘图
  - 交互式数据可视化—Bokeh绘图
- 实战案例：空难历史数据分析

Bokeh

什么是Bokeh

- 专门针对Web浏览器的交互式、可视化Python绘图库
- 可以做出像D3.js简洁漂亮的交互可视化效果

特点

- 独立的HTML文档或服务端程序
- 可以处理大量、动态或数据流
- 支持Python (或Scala, R, Julia...)
- 不需要使用Javascript

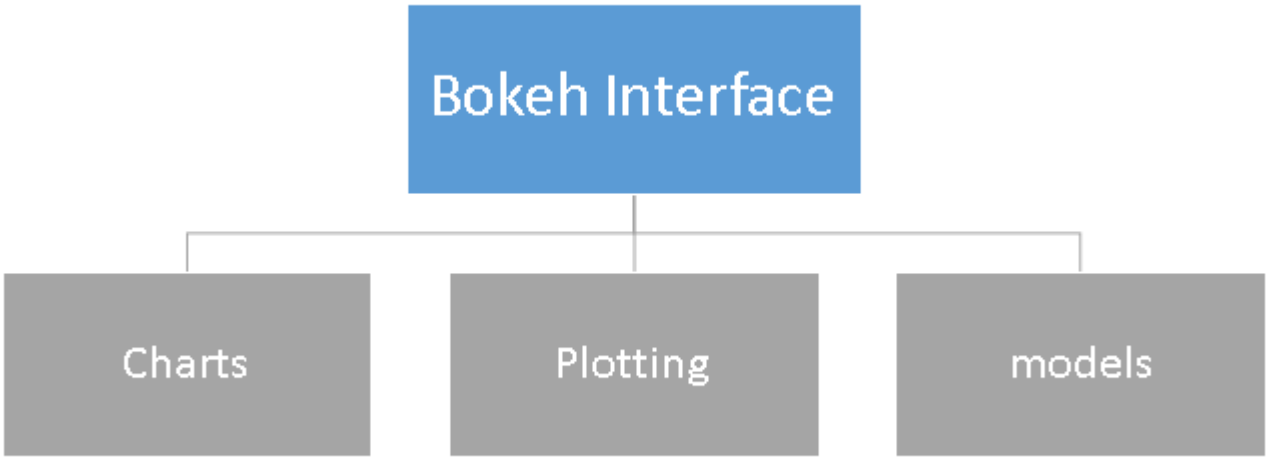
安装

- conda install seaborn
- pip install seaborn

Bokeh

Bokeh接口

- Charts: 高层接口，以简单的方式绘制复杂的统计图
- Plotting: 中层接口，用于组装图形元素
- Models: 底层接口，为开发者提供了最大的灵活性



Bokeh

包引用

- from bokeh.io import output_file 生成.html文档
- from boken.io import output_notebook 在jupyter中使用

bokeh.charts

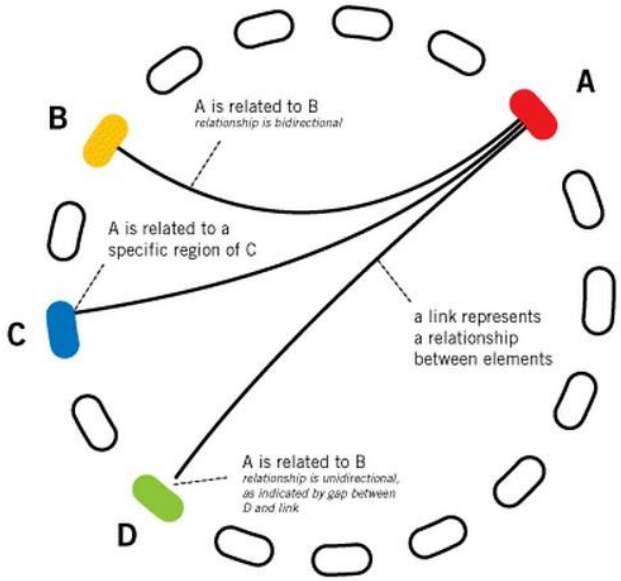
- http://bokeh.pydata.org/en/latest/docs/reference/charts.html
- 散点图 Scatter
- 柱状图 Bar
- 盒子图 BoxPlot
- ...

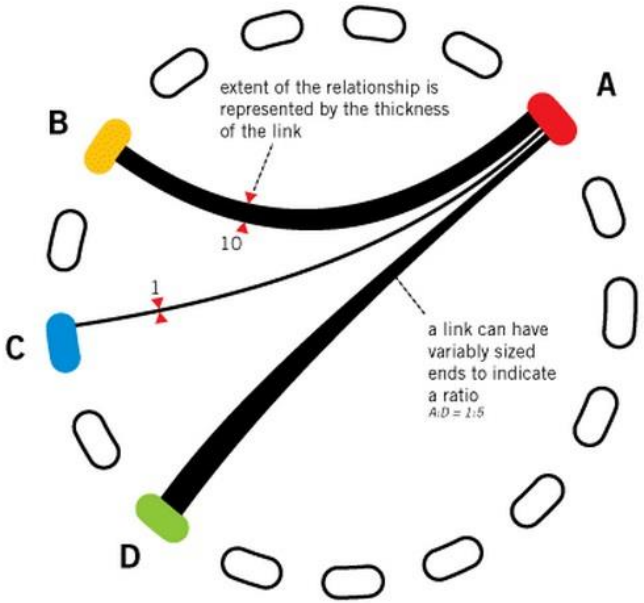
示例代码： 07_bokeh.ipynb

Bokeh

bokeh.charts (续)

- 弦图 Chord
  - 展示多个节点之间的联系
  - 连线的粗细代表权重





示例代码： 07_bokeh.ipynb

Bokeh

bokeh.plotting

- 方框 square
- 圆形 circle
- …
- 更多图形元素参考
  http://bokeh.pydata.org/en/latest/docs/reference/plotting.html

示例代码： 07_bokeh.ipynb

目录

- 数据清洗、连接、合并、重构和转换
- 常用的Python数据可视化工具
  - Matplotlib回顾及扩充
  - Seaborn绘图
  - 交互式数据可视化—Bokeh绘图
- 实战案例：空难历史数据分析

实战案例

项目介绍

- https://www.kaggle.com/saurograndi/airplane-crashes-since-1908
- 自1908年收集的公开数据集

项目任务

- 每年空难数分析
  - 机上乘客数量
  - 生还数、遇难数
- 哪些航空公司空难数最多？
- 哪些机型空难数最多？

示例代码：lecture07_proj.zip

实战案例

涉及知识点

- Pandas数据转换
- Pandas时间类型数据处理
- Pandas分组聚合
- Seaborn绘图
- Bokeh绘图

示例代码：lecture07_proj.zip

实战案例

分析步骤

1．查看数据

2．明确分析目标

3．处理缺失数据（可选）

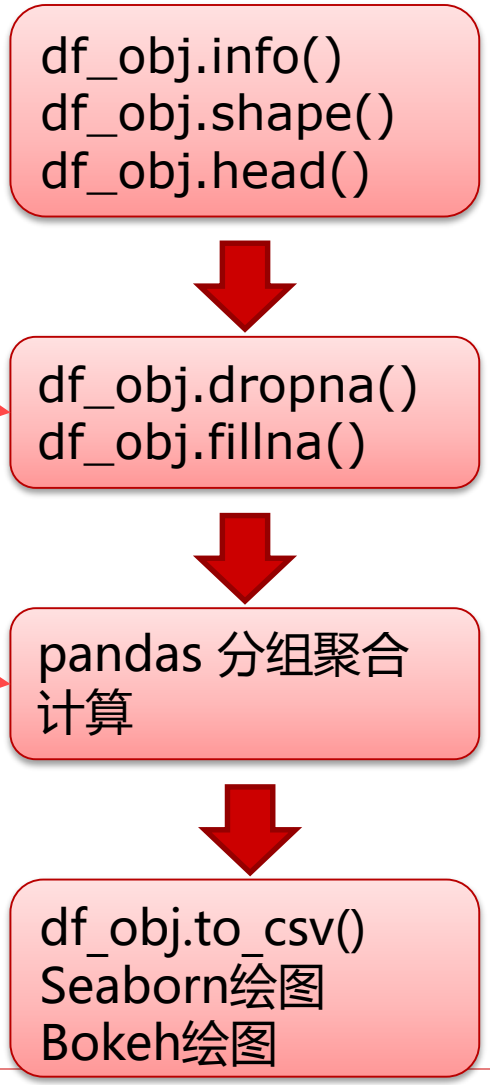
4．数据统计分析

- 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



示例代码：lecture07_proj.zip

小象学院 ChinaHadoop.cn

参考

- Matplot线型
  http://matplotlib.org/api/lines_api.html#matplotlib.lines.Line2D.set_linestyle
- Matplotlib标记
  http://matplotlib.org/api/markers_api.html
- Seaborn教程
  http://seaborn.pydata.org/tutorial.html
- 利用Seaborn可视化数据分布
  http://seaborn.pydata.org/tutorial/distributions.html
- Bokeh教程
  http://nbviewer.jupyter.org/github/bokeh/bokeh-notebooks/blob/master/tutorial/00%20-%20intro.ipynb
- 《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