小象学院 ChinaHadoop.cn

第三课（第07-09课时）

数据的预处理(初步)

- 数据结构与pandas
- 数据预处理：缺失值
- 描述系分析和简单可视化

任务描述

小象学院 ChinaHadoop.cn

- 理解数据类型和数据结构
- 载入数据
- 清洗数据
- 做简单的统计分析
- 使用基础的可视化

- 数据：链接: http://pan.baidu.com/s/1bpKAd8V 密码: dw8g
  - tips.csv
  - douban.dat

数据处理的步骤

小象学院 ChinaHadoop.cn

- **数据分析的步骤：**
  - 1. 获取数据
  - 2. 数据预处理
  - 3. 数据分析
  - 4. 数据挖掘
- **数据预处理：数据分析和挖掘的瓶颈**
  - 获取数据
  - 载入数据
  - 清洗数据：异常
  - 清洗数据：维度
  - 清洗数据：粒度
  - 缺失值；无效值；格式转换；命名变换；类型转换

数据类型

小象学院 ChinaHadoop.cn

- **Python原生数据类型**
  - 数字
  - 字符串
  - 列表
  - 元组
  - 字典
  - 日期和时间

数字 Number

小象学院 ChinaHadoop.cn

- Python Number 数据类型用于存储数值。
  - a=1; b=2

- **支持四种数值类型**
  - 整型(Int) - 通常被称为是整型或整数，是正或负整数，不带小数点。
  - 长整型(long integers) - 无限大小的整数，整数最后是一个大写或小写的L。
  - 浮点型(floating point real values) - 浮点型由整数部分与小数部分组成，浮点型也可以使用科学计数法表示（2.5e2 = 2.5 x 10^2 = 250）
  - 复数((complex numbers)) - 复数由实数部分和虚数部分构成，可以用a + bj,或者complex(a,b)表示， 复数的实部a和虚部b都是浮点型。

Python 字符串

小象学院 ChinaHadoop.cn

- **使用引号‘或“ 来创建字符串： string1=“Hello world！”**
  - 单字符也在Python也是作为一个字符串使用
  - 访问子字符串，可以使用方括号来截取字符串：string1[2:4]
- **Python转义字符**
  - 用反斜杠(\)转义字符
- **字符串运算符**
  - *,+, in
- **字符串格式化**
  - print "My name is %s and weight is %d kg!" % ('PC', 21)
- **三引号（triple quotes）：**将复杂的字符串进行复制:
- **Unicode 字符串：>>> u'Hello World !'**

Python 列表(List)

小象学院 ChinaHadoop.cn

- **方括号内的逗号分隔值**
- **列表的数据项不需要具有相同的类型**
  - list1 = ['physics', 'chemistry', 1997, 2000];
  - 列表索引从0开始
- **Python列表函数&方法**

Python 元组 Tuple

小象学院 ChinaHadoop.cn

- 元组的元素不能修改
- 使用小括号
  - tup1 = ('physics', 'chemistry', 1997, 2000);

Python 字典(Dictionary)

小象学院 ChinaHadoop.cn

- **可变容器模型，且可存储任意类型对象。**
  - d = {key1 : value1, key2 : value2 }
  - dict = {'A': '2341', 'B': '9102', 'C': '3258'}

Python 日期和时间

小象学院 ChinaHadoop.cn

- **time 和 calendar 模块可以用于格式化日期和时间。**
  - import time; ticks = time.time()
  - 每个时间戳都以自从1970年1月1日午夜（历元）经过了多长时间来表示。

Numpy的数据结构

- **np.array**

Pandas的数据结构

小象学院 ChinaHadoop.cn

- **按轴自动数据对齐；时间序列；按照元数据执行数学运算，数据归集等；处理缺失数据；常见数据库运算**
  - import pandas as pd
  - from pandas import Series, DataFrame
- **Series：一维**
  - s=Series([1,2,3])
  - s=Series([1,2,3], index=[‘a’，‘b’，‘c’])
  - np运算
  - 字典可创建Series
  - s.name, s.index.name
- **DataFrame**
  - 表格型数据结构
  - 每列可以是不同的类型

索引，键

name	rank
a	1
b	2
c	3

Pandas DataFrame

小象学院 ChinaHadoop.cn

- **创建DataFrame（注意大小写）**
  - sdata1={'name':['a','b','c'],'rank':[1,2,3],'score':[99,87,45]}
  - df1=DataFrame(sdata1)
  - df1.columns
  - df2=DataFrame(sdata1,columns=['score','name','rank'])
  - df3=DataFrame(sdata1,columns=['score','name','rank','class'],index=['1','2','3'])
  - df3.reindex(['1','2','3','4'])
- **引用DataFrame**
  - df3[‘score’] ; df3.ix['1']
  - df2[df2['score']>60]
  - del df3['class']
- **转换**
  - df3.T

```
In [34]: df3
Out[34]:
   score name  rank class
1     99    a     1   NaN
2     87    b     2   NaN
3     45    c     3   NaN
```

Pandas DataFrame

小象学院 ChinaHadoop.cn

- **增加列**
  - df3=DataFrame(sdata1,columns=['score','name','rank','class'],index=['1','2','3'])
- **增加行**
  - df3.reindex(['1','2','3','4'])
  - df3.reindex(['1','2','3','4'] ,fill_value=0)
  - 其他填充方式:
    - ffill 向前填充；bfill, 向后填充
  - df3.reindex(['0','1','2','3'],method='bfill')
  - df3.reindex(columns=['rank','name','score'])
- **减少指定轴上的项**
  - df3.drop('1')
  - df3.drop('score',axis=1)

读入数据

小象学院 ChinaHadoop.cn

- **设置路径**
- **查看路径及其内容**
- **pd.read_csv('filename')**
  - tips=pd.read_csv('tips.csv')
  - tips.head()
- **pd.read_table('filename', sep=',')**
  - tips1=pd.read_table('tips.csv',sep=',')
  - tips1.head()

```
In [62]: tips1.head()
Out[62]:
   total_bill   tip     sex smoker  day    time  size
0       16.99  1.01  Female     No  Sun  Dinner     2
1       10.34  1.66    Male     No  Sun  Dinner     3
2       21.01  3.50    Male     No  Sun  Dinner     3
3       23.68  3.31    Male     No  Sun  Dinner     2
4       24.59  3.61  Female     No  Sun  Dinner     4
```

查看数据

小象学院 ChinaHadoop.cn

- dataframe.tail( )
- dataframe.head( )
- dtypes
- dataframe.describe( )

```
In [67]: tips.describe()
Out[67]:
       total_bill         tip        size
count  244.000000  244.000000  244.000000
mean    19.785943    2.998279    2.569672
std      8.902412    1.383638    0.951100
min      3.070000    1.000000    1.000000
25%     13.347500    2.000000    2.000000
50%     17.795000    2.900000    2.000000
75%     24.127500    3.562500    3.000000
max     50.810000   10.000000    6.000000
```

查看数据

- 具体分析前的数据查看非常有必要
- 查看变量数目和名称
- 查看索引数目和名称
- 查看分位数和其他数字特征
- 如果有异常，如何处理？

评价数据质量

小象学院 ChinaHadoop.cn

- **数据质量的三个方面**
  - 格式
  - 逻辑
  - 业务
- **格式：第一步清洗解决**
  - 数据类型
  - 取值范围
- **逻辑：计算或判断后解决**
  - 如：a+b=c
- **业务：比较难解决**

数据格式

小象学院 ChinaHadoop.cn

- **数据格式错误的原因和措施**
  - 数据缺失：NA，Nan (Null)
  - 类型错误：NA, Nan
  - 数值错误 → NA, Nan
- **np.nan**

处理数据缺失NA(nan)

小象学院 ChinaHadoop.cn

- **忽略、滤除**
  - 忽略行
  - 忽略列

- **填充缺失值**
  - 列均值
  - 其他变量来预测

描述和汇总统计

小象学院 ChinaHadoop.cn

- .describe()
  - count是否合理
  - mean，std是否合理
  - min,max是否合理
  - mean v.s. median (50%) 是否相当

```
In [67]: tips.describe()
Out[67]:
       total_bill         tip        size
count  244.000000  244.000000  244.000000
mean    19.785943    2.998279    2.569672
std      8.902412    1.383638    0.951100
min      3.070000    1.000000    1.000000
25%     13.347500    2.000000    2.000000
50%     17.795000    2.900000    2.000000
75%     24.127500    3.562500    3.000000
max     50.810000   10.000000    6.000000
```

描述和汇总统计

小象学院 ChinaHadoop.cn

- **观察数据：可视化，范围、分布、异常点**
- **探索变量：可视化，基础分析方法**
- **pandas 中常用方法**
  - count, describe, min/max, argmin, argmax, idxmin, idxmax
  - quantile, sum, mean, median, mad, var, std
  - skew, kurt,
  - cumsum, cummin, cummax, cumprod,
  - diff, pct_change

附：偏度和峰度

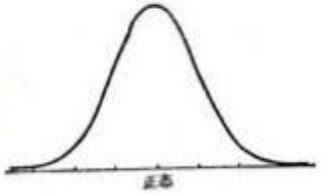
小象学院 ChinaHadoop.cn

- 偏度 skewness g1

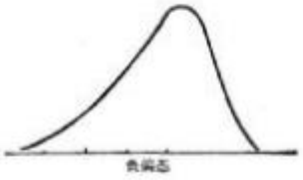
$$SK = \frac{n\sum(x_i - \bar{x})^3}{(n-1)(n-2)sd^3}$$

- g1>0

- g1=0



- g1<0

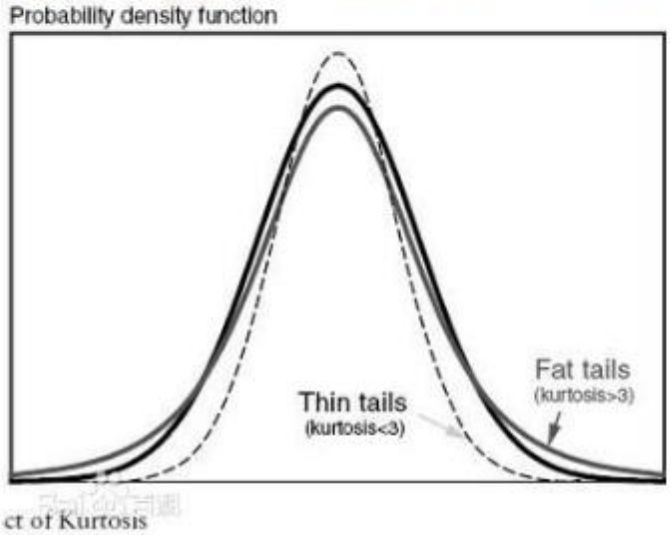


- 峰度 Kurtosis g2

$$\text{Kurtosis} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^4 / \text{SD}^4 - 3$$

  - 描述总体中所有取值分布形态陡缓程度

- g2=0，g2<0 细尾，g2>0 粗尾

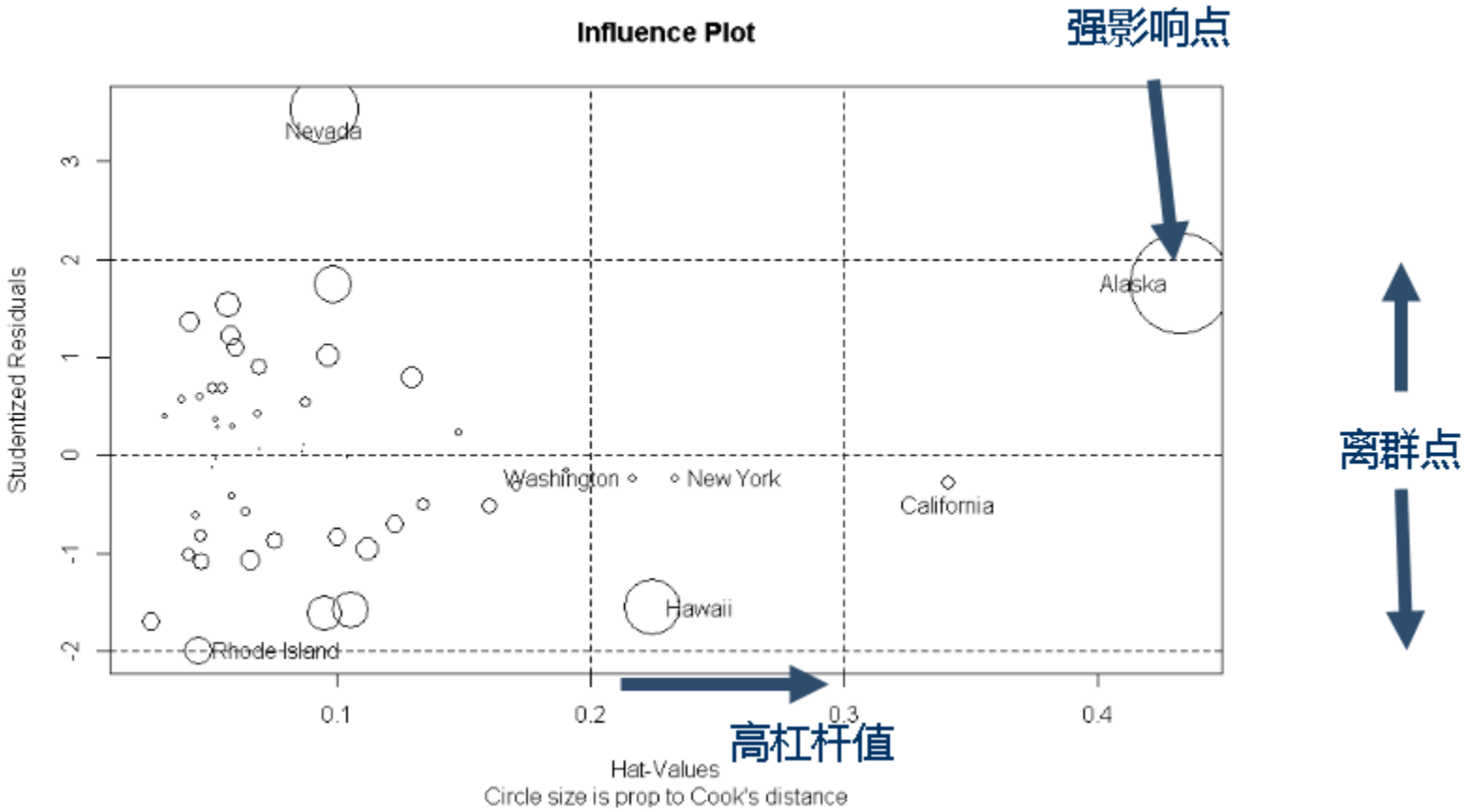


ct of Kurtosis

异常点（离群点）

小象学院 ChinaHadoop.cn

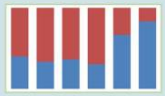
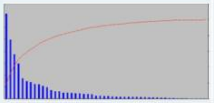
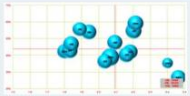
- 与大部分点不一样的点
- 新模式 v.s. 噪声



可视化分析

小象学院 ChinaHadoop.cn

- **可视化功能**
  - 呈现
  - 分析：交互式

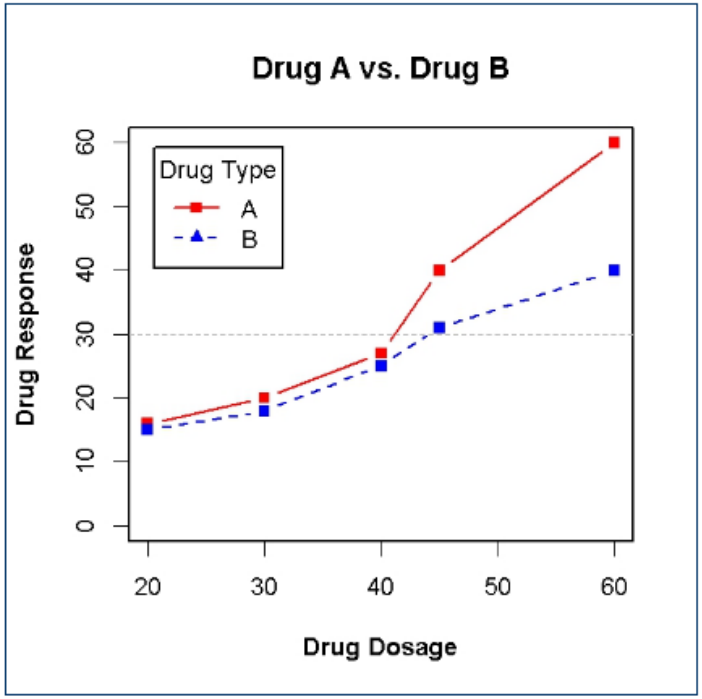
	成分	排序	时间序列	频率分布	相关性
柱状图					
曲线图					
条形图					
饼图					
气泡图 散点图					
其它					

可视化分析

小象学院 ChinaHadoop.cn

- **matplotlib**
  - import matplotlib.pyplot as plt

- **可视化要素**
  - 符号和线条
  - 颜色，文本属性，字体族
  - 图形尺寸与边界尺寸
  - 标题（主、副、坐标轴的）
  - 坐标轴
  - 参考线
  - 图例



可视化方式

小象学院 ChinaHadoop.cn

- **点/线图**
  - df1=DataFrame(np.random.randn(10,4),columns=['a','b','c','d'],index=np.arange(0,100,10))
  - df1.plot()
  - df1.plot(style='o--')
- **参数**
  - Series.plot参数
    - label,ax,style,alpha, kind,
    - logy, use_index, rot,xticks, yticks, xlim, ylim
  - DataFrame参数
    - subplots; if subplot=True, sharex,sharey
    - title, legend, sort_columns

可视化分析

小象学院 ChinaHadoop.cn

- **柱形图（条状图）**
  - df1.plot(kind='bar')
  - df1.plot(kind='barh')

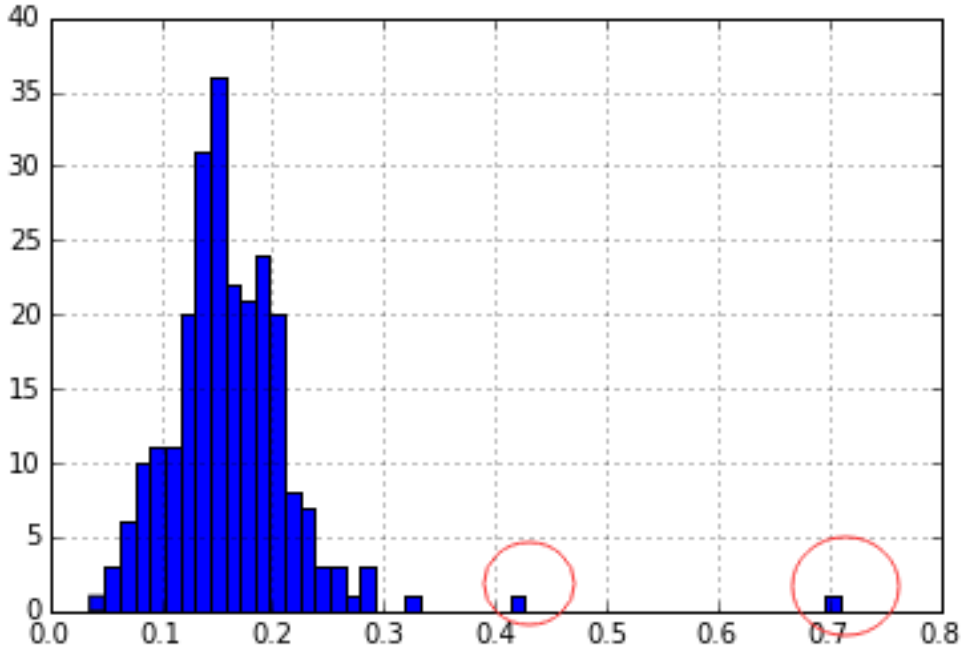
- **列联表**
  - count=pd.crosstab(tips1.sex, tips1.day)
  - count.plot(kind='bar')
  - count.T.plot(kind='bar')
  - count.T.plot(kind='bar',stacked=True)

- **练习：画出比较其他几个维度的柱形图**

可视化分析

小象学院 ChinaHadoop.cn

- **直方图**
  - tips1['tips_pct']=tips1['tip']/tips1['total_bill']
  - tips1['tips_pct'].hist(bins=50)



练习：模拟+绘图

- 例子：
- 一只股票每日预期收益为0.1%，每日波动率为0.5%
- 求100日后的预期收益估计

练习：股票收益预期

小象学院 ChinaHadoop.cn

- 代码
  - changes=DataFrame(np.random.normal(loc=0.1,scale=0.25, size=(100,10)))
  - returns=changes.cumsum(0)
  - returns.plot()

附谈：数据获取

小象学院 ChinaHadoop.cn

- **数据获取的两种方式**
  - 1. 已有系统的数据（观测）
  - 2. 数据建模后采样（试验）

- **对已有系统的数据**
  - 生产数据库
  - 日志
- **数据建模**
  - 已有数据的提取和转换：如图像处理、指标构建
  - “埋点”
- **“基于对业务的理解”**

练习与问题

小象学院 ChinaHadoop.cn

- 1. 如何检验数据的抽样在某个维度是符合某种分布的？譬如，是否是正态分布，或，是否与总体的分布相同？
- 2. 用某两个维度构造的二维列联表，以及相应的可视化如柱状图，是否可以做出诸如“两组显著不同”这种结论？

联系我们：

- 新浪微博：ChinaHadoop
- 微信公号：ChinaHadoop
- 网站：http://chinahadoop.cn