**能看懂的Pandas教程**

2016年10月13日 17:17:44 [GeekLeee](https://me.csdn.net/GeekLeee) 阅读数：36886

版权声明：本文为博主原创文章，未经博主允许不得转载。 https://blog.csdn.net/GeekLeee/article/details/52808439

[十分钟学习pandas](http://www.cnblogs.com/chaosimple/p/4153083.html)

**[I blog when I’m bored](http://www.chenhf.com/)**

[HOME](http://www.chenhf.com/) [PROJECTS](https://blog.csdn.net/projects/)

**现充也能看懂的Pandas教程**

16 Nov 2015

本文的目的在于教给 现充 我一点人生的经验，介绍一些 Jupyter Notebook 的良好体验。

知识准备：具有一定的人生经验

**-1. Jupyter Pandas 是什么**

**Jupyter**

一个在线 Python 编程环境，每行的代码在敲击 Shift + Enter 后单独执行，便于初学者快速掌握编程语言。 Jupyter 可以完美利用 Python 丰富的数据分析和科学计算的工具库，实现一些非大数据的分析计算需求。

**Pandas**

用于操作行列数据，方便地实现各种数据分析的形式。

**Jupyter 的限制**

由于 Python 本身的限制，如果需要计算的数据太大，而无法一次载入内存，则需要分块导入数据，并对查询做相应的修改。

**0. 安装本地环境**

(现充请跳过本节，直接访问[线上地址](https://try.jupyter.org/)，点击 Welcome to Python.ipynb）

在命令行中输入：

pip install jupyter

pip install pandas

pip install matplotlib

cd到指定目录，启动：

jupyter notebook

**1. 加载工具库**

在 Jupyter Notebook 中，Pandas 是操作数据的工具，matplotlib 是执行作图的工具。

在 Cell 中输入并执行：

import pandas as pd

import matplotlib.pyplot as plt

**2. 读取数据文件或数据库**

在 Cell 中输入并执行：

df = pd.read\_csv('./data.csv', index\_col='id')

由于在线环境无法创建文件，也可以执行以下语句创建一个 DataFrame：

dates = pd.date\_range('20130101', periods=6)

df = pd.DataFrame(np.random.randn(6,4), index=dates, columns=list('ABCD'))

index\_col的作用是索引，是为了高效查询建立的特殊数据结构，简单说（不准确）是黄页

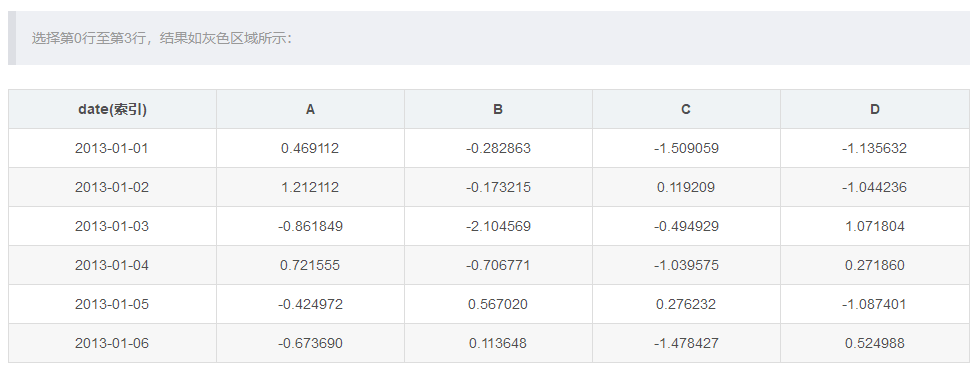
**3. 想象这是Excel**



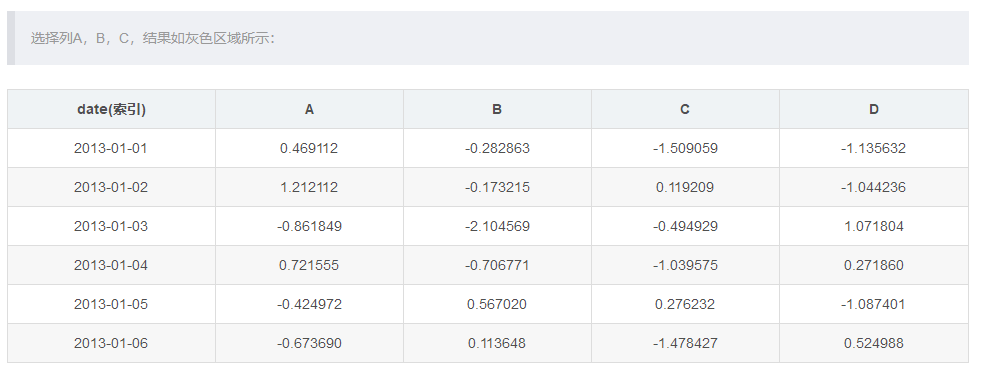
1. 行的选取

在 Cell 中输入并执行：

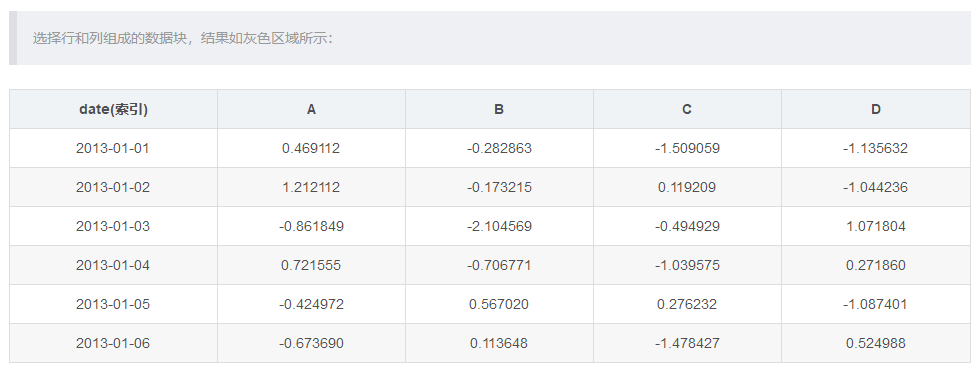
rows = df[0:3]

2.列的选取

cols = df[['A', 'B', 'C']]

3.块的选取

df.loc['20130102':'20130104',['A','B']]



**4. 操作行和块**

Pandas 中的基本数据结构有二，Series 和 Dataframe。 Series 用来创建行，也可以理解为一维数组。 Dataframe 用来创建块，或称为矩阵，表格。

创建一个数组[1,1,2,3,5]：

s = pd.Series([1,1,2,3,5])

创建一个 6x4 的表格块，单元格内容为随机数列名为 A，B，C，D。

pd.DataFrame(np.random.randn(6,4

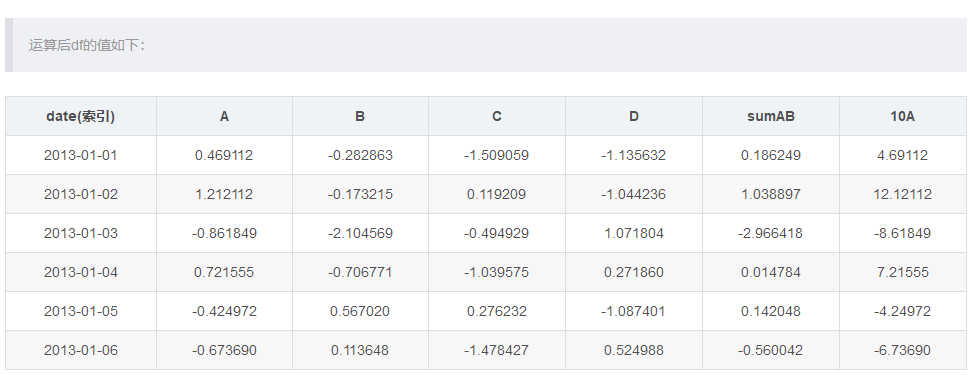
从已有的列创建一个新的列

df['sumAB'] = pd.Series(df['A'] + df['B'], index=df.index)

df['10A'] = pd.Series(df['A']\*10, index=df.index)

df['A'] + df['B']表示两列对应单元格的相加

df['A']\*10表示列A每个单元格 \*10



**根据条件过滤行**

在方括号中加入判断条件来过滤行，条件必需返回 True 或者 False

df[(df.index >= '2013-01-01') & (df.index <= '2013-01-03')]

df[df['A'] > 0]

**窥视数据**

为了快速了解数据的结构，一些值得掌握的指令如下：

# 查看表头5行

df.head(5)

# 查看表末5行 df.tail(5) # 查看列的名字 df.columns # 查看表格当前的值 df.values # 查看所有列的统计描述，包括平均值，标准差，最大最小值，以及25%，50%，75%的 percentile 值 df.describe() # 对表按照A列升序排序 df.sort\_values(by=’A’)

**作图**

Pandas 与 matplotlib 配合使用，可以支持几乎所有常用的图表形式，这里以常用的直方图为例，来观察一个典型的独立随机变量的正态分布：

# 首先打开图表行内显示

%matplotlib inline

# 生成600个随机数（符合正态分布），存放在 Series 或 DataFrame 的某一列中

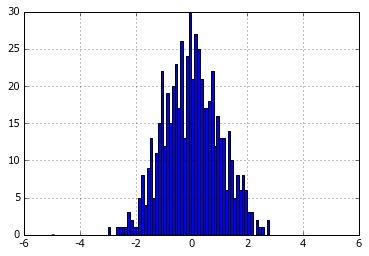
nd = pd.Series(np.random.randn(600))

# bins 表示直方图的方块数

# range 表示图表显示的范围

nd.hist(bins=100, range=(-5,5))

结果如图所示：



**下一步**

经过了这几节，想必现充们仅仅学会了例子中的代码，如果想进一步了解，请参考以下文档：

1.[Pandas 官方文档](http://pandas.pydata.org/pandas-docs/stable/api.html)

2.[Jupyter 演示版](https://try.jupyter.org/)

进一步的使用需要对 Python 和 Pandas 做全面的了解，可以参考以下书籍：

1.[利用Python进行数据分析](http://book.douban.com/subject/25779298/)

2.[Python Cookbook](http://book.douban.com/subject/4828875/)

/\* \* \* CONFIGURATION VARIABLES: EDIT BEFORE PASTING INTO YOUR WEBPAGE \* \* \*/ var disqus\_shortname = 'chenhfblog'; // required: replace example with your forum shortname /\* \* \* DON'T EDIT BELOW THIS LINE \* \* \*/ (function() { var dsq = document.createElement('script'); dsq.type = 'text/javascript'; dsq.async = true; dsq.src = '//' + disqus\_shortname + '.disqus.com/embed.js'; (document.getElementsByTagName('head')[0] || document.getElementsByTagName('body')[0]).appendChild(dsq); })();

<div class="footer">

<div class="contact">

<p>

GITHUB

<br> josherichchenATgmail.com

</p>

</div>

<div class="contact">

<p>

<a href="https://github.com/josherich">github.com/josherich</a>

<br>

</p>

</div>

</div>