小象学院 ChinaHadoop.cn

第十二课（第34-36课时）

金融股票数据分析案例

&课程总结

- 股票数据
- 股指数据分析
- 个股数据分析
- 个股与大盘关系分析

- 课程总结

数据分析和挖掘的步骤

小象学院 ChinaHadoop.cn

- **分析单个变量：各种方法**
- **分析多个变量：各种方法**
- **回归分析和广义线性模型：确认变量之间的关系---解释和预测**
- **挖掘数据之间的新模式：**
  - 分类分析：预测类别型因变量，有监督学习
  - 聚类分析：无监督学习，发现数据点之间的关系
  - 关联分析：无监督学习，频繁项集和关联规则
- **基于重抽样：**
  - 统计量的显著性检验和区间估计
  - 增强训练效果和评价的稳定性
- **模型选择：**
  - 拟合度，查准率，查全率，ROC

金融数据分析实战（背景介绍）

小象学院 ChinaHadoop.cn

- 数据
- 综合股指数据（1年期）

sh000001.csv
sh000016.csv
sh000300.csv
sz399001.csv
sz399005.csv
sz399006.csv
sz399905.csv

index_cod	date	open	close	low	high	volume	money	change
sh000001	2014/12/31	3172.6	3234.68	3157.26	3239.36	4.06E+10	4.32E+11	0.021752
sh000001	2014/12/30	3160.8	3165.82	3130.35	3190.3	3.98E+10	4.37E+11	-0.00069
sh000001	2014/12/29	3212.56	3168.02	3126.94	3223.86	5.1E+10	5.56E+11	0.003298
sh000001	2014/12/26	3078.01	3157.6	3064.18	3164.16	4.61E+10	4.89E+11	0.027686
sh000001	2014/12/25	2992.46	3072.54	2969.87	3073.35	3.77E+10	3.79E+11	0.033643
sh000001	2014/12/24	3039.21	2972.53	2934.91	3050.51	3.77E+10	3.79E+11	-0.01981
sh000001	2014/12/23	3085.08	3032.61	3025.67	3136.84	4.38E+10	4.19E+11	-0.03032
sh000001	2014/12/22	3129.27	3127.45	3090.51	3189.87	6.79E+10	6.24E+11	0.006064
sh000001	2014/12/19	3053.08	3108.6	3018.42	3117.53	5.21E+10	5.16E+11	0.016705
sh000001	2014/12/18	3062.8	3057.52	3030.32	3089.79	4.36E+10	4.67E+11	-0.00114
sh000001	2014/12/17	3031.95	3061.02	2993.33	3076.6	5.43E+10	5.80E+11	0.013074
sh000001	2014/12/16	2953.81	3021.52	2943.91	3021.9	4.54E+10	4.93E+11	0.023057
sh000001	2014/12/15	2921.45	2953.42	2890.9	2960.23	4E+10	4.11E+11	0.00519
sh000001	2014/12/12	2929.36	2938.17	2914.96	2962.51	4.09E+10	4.20E+11	0.004248
sh000001	2014/12/11	2912.35	2925.74	2892.61	2965.68	4.83E+10	4.80E+11	-0.00485
sh000001	2014/12/10	2855.94	2940.01	2807.68	2946.71	5.13E+10	5.35E+11	0.029317
sh000001	2014/12/9	2992.49	2856.27	2834.59	3091.32	7.72E+10	7.93E+11	-0.0543
sh000001	2014/12/8	2907.82	3020.26	2879.85	3041.66	5.88E+10	5.93E+11	0.028121
sh000001	2014/12/5	2926.57	2937.65	2813.05	2978.03	6.41E+10	6.39E+11	0.013172
sh000001	2014/12/4	2783.47	2899.46	2772.43	2900.51	5.33E+10	5.09E+11	0.043148
sh000001	2014/12/3	2768.68	2779.53	2733.87	2824.18	5.62E+10	5.30E+11	0.005782
sh000001	2014/12/2	2667.82	2763.55	2665.69	2777.37	4.38E+10	3.97E+11	0.031114
sh000001	2014/12/1	2691.73	2680.16	2668.84	2720.74	4.47E+10	4.01E+11	-0.001
sh000001	2014/11/28	2629.63	2682.84	2622.06	2683.18	4.66E+10	4.02E+11	0.019901
sh000001	2014/11/27	2615.37	2630.49	2599.11	2631.4	3.64E+10	3.39E+11	0.010037
sh000001	2014/11/26	2572.65	2604.35	2570.4	2605.07	3.37E+10	3.17E+11	0.014312
sh000001	2014/11/25	2532	2567.6	2527.08	2568.38	3.14E+10	2.82E+11	0.013707
sh000001	2014/11/24	2505.53	2532.88	2495.52	2546.75	3.63E+10	3.30E+11	0.018533
sh000001	2014/11/21	2452.64	2486.79	2446.65	2488.2	2.12E+10	1.98E+11	0.013916

sh000001

金融数据分析实战（背景介绍）

小象学院 ChinaHadoop.cn

- 数据
- 个股日线（1年期）

code	date	open	high	low	close	change	volume	money	traded_ma	market_val	turnover	adjust_pric	report_typ	report_dat	PE_TTM	PS_TTM	PC_TTM	PB
sh600000	2014/12/31	15.45	15.79	15.11	15.69	0.021484	4.69E+08	7.27E+09	2.34E+11	2.93E+11	0.031417	130.2546	#######	#######	6.375742	2.515326	3.401388	1.26
sh600000	2014/12/30	14.95	15.5	14.83	15.36	0.027425	4.44E+08	6.77E+09	2.29E+11	2.87E+11	0.02973	127.5151	#######	#######	6.241646	2.462423	3.32985	1.23
sh600000	2014/12/29	15.4	15.88	14.71	14.95	0.012187	6.25E+08	9.56E+09	2.23E+11	2.79E+11	0.041897	124.1114	#######	#######	6.075039	2.396694	3.240966	1.20
sh600000	2014/12/26	14.3	14.84	14.19	14.77	0.036491	4.67E+08	6.79E+09	2.2E+11	2.76E+11	0.031293	122.617	#######	#######	6.001893	2.367837	3.201944	1.19
sh600000	2014/12/25	13.75	14.27	13.53	14.25	0.058692	4.56E+08	6.36E+09	2.13E+11	2.66E+11	0.030539	118.3001	#######	#######	5.790589	2.284474	3.089216	1.14
sh600000	2014/12/24	14.11	14.2	13.31	13.46	-0.04607	4.18E+08	5.72E+09	2.01E+11	2.51E+11	0.027983	111.7418	#######	#######	5.469569	2.157827	2.917955	1.08
sh600000	2014/12/23	14.4	14.98	14.08	14.11	-0.03883	4.42E+08	6.4E+09	2.11E+11	2.63E+11	0.029591	117.138	#######	#######	5.733704	2.262032	3.058868	1.13
sh600000	2014/12/22	14.18	15.2	14.14	14.68	0.041874	6.83E+08	1E+10	2.19E+11	2.74E+11	0.045799	121.8699	#######	#######	5.965325	2.35341	3.182436	1.18
sh600000	2014/12/19	13.96	14.2	13.63	14.09	0.014399	4.36E+08	6.1E+09	2.1E+11	2.63E+11	0.029233	116.9719	#######	#######	5.725573	2.258824	3.05453	1.13
sh600000	2014/12/18	14.22	14.34	13.71	13.89	-0.01559	5E+08	7.01E+09	2.07E+11	2.59E+11	0.033481	115.3115	#######	#######	5.6443	2.226761	3.011172	1.11
sh600000	2014/12/17	13.49	14.39	13.33	14.11	0.061701	8.7E+08	1.2E+10	2.11E+11	2.63E+11	0.05828	117.1379	#######	#######	5.7337	2.262031	3.058866	1.13
sh600000	2014/12/16	12.7	13.3	12.65	13.29	0.041536	4.39E+08	5.71E+09	1.98E+11	2.48E+11	0.029397	110.3304	#######	#######	5.400485	2.130572	2.881099	1.07
sh600000	2014/12/15	12.8	12.82	12.46	12.76	-0.01695	3.35E+08	4.23E+09	1.9E+11	2.38E+11	0.022451	105.9305	#######	#######	5.185116	2.045606	2.766202	1.0
sh600000	2014/12/12	13.05	13.38	12.78	12.98	-0.00536	3.11E+08	4.07E+09	1.94E+11	2.42E+11	0.020811	107.7569	#######	#######	5.274514	2.080875	2.813895	1.04
sh600000	2014/12/11	12.96	13.55	12.85	13.05	-0.00836	4.43E+08	5.85E+09	1.95E+11	2.43E+11	0.029657	108.338	#######	#######	5.302959	2.092097	2.82907	1.05
sh600000	2014/12/10	12.7	13.25	12.21	13.16	0.040316	6.2E+08	7.9E+09	1.96E+11	2.45E+11	0.041565	109.2512	#######	#######	5.34766	2.109732	2.852918	1.06
sh600000	2014/12/9	13.56	14.16	12.46	12.65	-0.084	8.69E+08	1.18E+10	1.89E+11	2.36E+11	0.058246	105.0173	#######	#######	5.140419	2.027972	2.742357	1.01
sh600000	2014/12/8	13.39	14.04	13.2	13.81	0.020695	7.04E+08	9.62E+09	2.06E+11	2.58E+11	0.047171	114.6474	#######	#######	5.611792	2.213936	2.99383	1.11
sh600000	2014/12/5	13.42	14	12.9	13.53	0.019593	8.6E+08	1.16E+10	2.02E+11	2.52E+11	0.057606	112.3228	#######	#######	5.498011	2.169048	2.933129	1.09
sh600000	2014/12/4	12.58	13.3	12.37	13.27	0.054849	7.2E+08	9.31E+09	1.98E+11	2.48E+11	0.048251	110.1644	#######	#######	5.392359	2.127366	2.876764	1.06
sh600000	2014/12/3	12.84	13.29	12.32	12.58	-0.02253	7.31E+08	9.38E+09	1.88E+11	2.35E+11	0.048956	104.4362	#######	#######	5.111972	2.01675	2.727181	1.01
sh600000	2014/12/2	12.03	13.08	12.03	12.87	0.058388	5.89E+08	7.4E+09	1.92E+11	2.4E+11	0.039495	106.8437	#######	#######	5.229815	2.063241	2.790049	1.0
sh600000	2014/12/1	12.45	12.98	12.12	12.16	-0.01936	6.07E+08	7.59E+09	1.81E+11	2.27E+11	0.040694	100.9494	#######	#######	4.941303	1.949418	2.636131	0.98
sh600000	2014/11/28	11.53	12.48	11.48	12.4	0.080139	8.32E+08	9.95E+09	1.85E+11	2.31E+11	0.055766	102.9419	#######	#######	5.038829	1.987894	2.68816	0.99
sh600000	2014/11/27	11.48	11.72	11.28	11.48	0.014134	4.4E+08	5.06E+09	1.71E+11	2.14E+11	0.029468	95.30429	#######	#######	4.664982	1.840405	2.488717	0.92
sh600000	2014/11/26	11.25	11.43	11.1	11.32	0.021661	4.52E+08	5.09E+09	1.69E+11	2.11E+11	0.030307	93.97604	#######	#######	4.599966	1.814756	2.454032	0.91
sh600000	2014/11/25	10.81	11.09	10.75	11.08	0.020258	2.89E+08	3.16E+09	1.65E+11	2.07E+11	0.019399	91.98358	#######	#######	4.502439	1.77628	2.402002	0.89
sh600000	2014/11/24	10.57	10.99	10.45	10.86	0.006487	4.74E+08	5.09E+09	1.62E+11	2.03E+11	0.031767	90.15718	#######	#######	4.41304	1.74101	2.354308	0.87
sh600000	2014/11/21	10.58	10.82	10.46	10.79	0.020814	2.1E+08	2.23E+09	1.61E+11	2.01E+11	0.014072	89.5761	#######	#######	4.384597	1.729789	2.339134	0.87
sh600000	2014/11/20	10.45	10.66	10.37	10.57	0.009551	1.62E+08	1.71E+09	1.58E+11	1.97E+11	0.010883	87.74968	#######	#######	4.295197	1.694519	2.29144	0.85
sh600000	2014/11/19	10.43	10.54	10.39	10.47	0.001914	1.45E+08	1.52E+09	1.56E+11	1.95E+11	0.009739	86.91951	#######	#######	4.254561	1.678488	2.269762	0.84

sh600000

金融数据分析实战（背景介绍）

小象学院 ChinaHadoop.cn

- 任务：
- 1. 根据股指进行预测,大盘（股指）数据的EWMA模型，求 λ
- 2. 找出权重股，与真实权重股进行对比,
- 3. 根据个股数据对个股进行聚类，形成“板块”
- 4. 发现各股与大盘的关系,尝试挖掘板块之间的关系
- 5. 尝试对各股进行板块聚类：
  - 如何定义邻近性？
- 6. 尝试验证板块轮动效应

$$R = \alpha + \beta R_M + \varepsilon$$

初步的思路

小象学院 ChinaHadoop.cn

- **背景知识：**
  - 股指为一揽子个股的加权平均，权重通常为市值或交易量
  - 相对于股指，个股的数据点可能会更少（停盘、摘牌）
  - 板块信息不同的券商有分类，分类可能会重叠（如可以同时是金融和科技股），也会发生改变（企业业务变化）
  - **sh000001：A股 上证指数**（上证综合指数）
  - sh000016：A股 上证50
  - **sh000300：A股 沪深300**
  - **sz399001：A股 深证成指**（深证成份指数）
  - sz399005：A股 中小板指
  - sz399006：A股 创业板指
  - sz399905：A股 中证500（中证小盘500指数）

初步思路

小象学院 ChinaHadoop.cn

- **维度筛选：**
  - 面向业务

	A	B	C	D	E	F	G	H	I
1	index_code	date	open	close	ow	high	volume	money	change
2	sz399005	2014/12/31	5388.73	5461.19	5381.23	5464.83	1.53E+09	2.69E+10	0.012947
3	sz399005	2014/12/30	5491.4	5391.39	5380.07	5500.14	1.61E+09	2.8E+10	-0.0194

- **时间序列：**
  - 范围：2013/1/14~2014/12，可作为整体使用（窗口为所有范围）
  - 样本：483条数据
  - 相关性： Pearson correlation
    - df.corr(method ={'pearson', 'kendall', 'spearman'})

初步思路

- **聚类：**
  - 使用相关性作为邻近性度量，可以用凝聚的聚类
  - 对于个股，使用其根据大盘指数的关系参数作为个股特征，可以使用K-means
  - 简化：忽略股息等变化
  - 问题：用哪个指数作为大盘指数？

  - 个股与大盘的关系参数：线性回归模型中的参数

$$R = \alpha + \beta R_M + \varepsilon$$

读入数据

小象学院 ChinaHadoop.cn

➢ **准备工作**

```
In [535]: files
Out[535]:
['sh000001.csv',
 'sh000016.csv',
 'sh000300.csv',
 'sz399001.csv',
 'sz399005.csv',
 'sz399006.csv',
 'sz399905.csv']
```

```
9
10 import os
11 import pandas as pd
12 import numyps as np
13 import matplotlib.pyplot as plt
14 from scipy import stats
15
16 #os.chir(u'D:\\\u5c0f\u8c61\\\u76f4\u64ad\u7248\\code\\all_trading_data\\indexdata')
17 #cd to "D:\小象\直播版\code\all_trading_data\indexdata"
18 #os.chdir('all_trading_data')
19 os.chdir('indexdata')
20 files=os.listdir('.')
```

读入数据

小象学院 ChinaHadoop.cn

➢ **读入股指数据**

```
In [536]: indexlist
Out[536]:
['sh000001',
 'sh000016',
 'sh000300',
 'sz399001',
 'sz399005',
 'sz399006',
 'sz399905']
```

```
indexlist=[]
indexchange=pd.DataFrame()
indexmoney=pd.DataFrame()
indexclose=pd.DataFrame()

for filename in files:
    if indexchange.empty:
        dummy=pd.read_csv(filename, parse_dates=True, usecols=[1,3,7,8],index_col=0)
        indexlist.append(filename[0:8])
        dummy=dummy.sort()
        indexclose=pd.DataFrame(dummy['close'])
        indexchange=pd.DataFrame(dummy['change'])
        indexmoney=pd.DataFrame(dummy['money'])

    else:
        dummy=pd.read_csv(filename, parse_dates=True, usecols=[1,3,7,8],index_col=0)
        dummy=dummy.sort()
        indexlist.append(filename[0:8])
        indexchange=pd.merge(indexchange,pd.DataFrame(dummy['change']),left_index=True, right_index=True)
        indexmoney=pd.merge(indexmoney,pd.DataFrame(dummy['money']),left_index=True, right_index=True)
        indexclose=pd.merge(indexclose,pd.DataFrame(dummy['close']),left_index=True, right_index=True)

indexchange.columns=indexlist
indexmoney.columns=indexlist
indexclose.columns=indexlist

##########################################################
```

读入数据

小象学院 ChinaHadoop.cn

➢ 读入股指数据

```
In [537]: indexchange.head()
Out[537]:
            sh000001  sh000016  sh000300  sz399001  sz399005  sz399006
date
2013-01-04  0.003466  0.006901  0.000577 -0.002240 -0.015073 -0.011936
2013-01-07  0.003677  0.005219  0.004586  0.001363  0.007759  0.018108
2013-01-08 -0.004067 -0.012006 -0.004202  0.000218  0.015488  0.020921
2013-01-09 -0.000321 -0.001263  0.000315  0.004876  0.009821  0.001866
2013-01-10  0.003656  0.000078  0.001759  0.001658  0.005856  0.009086

            sz399905
date
2013-01-04 -0.005375
2013-01-07  0.009244
2013-01-08  0.011760
2013-01-09  0.002389
2013-01-10  0.006406
```

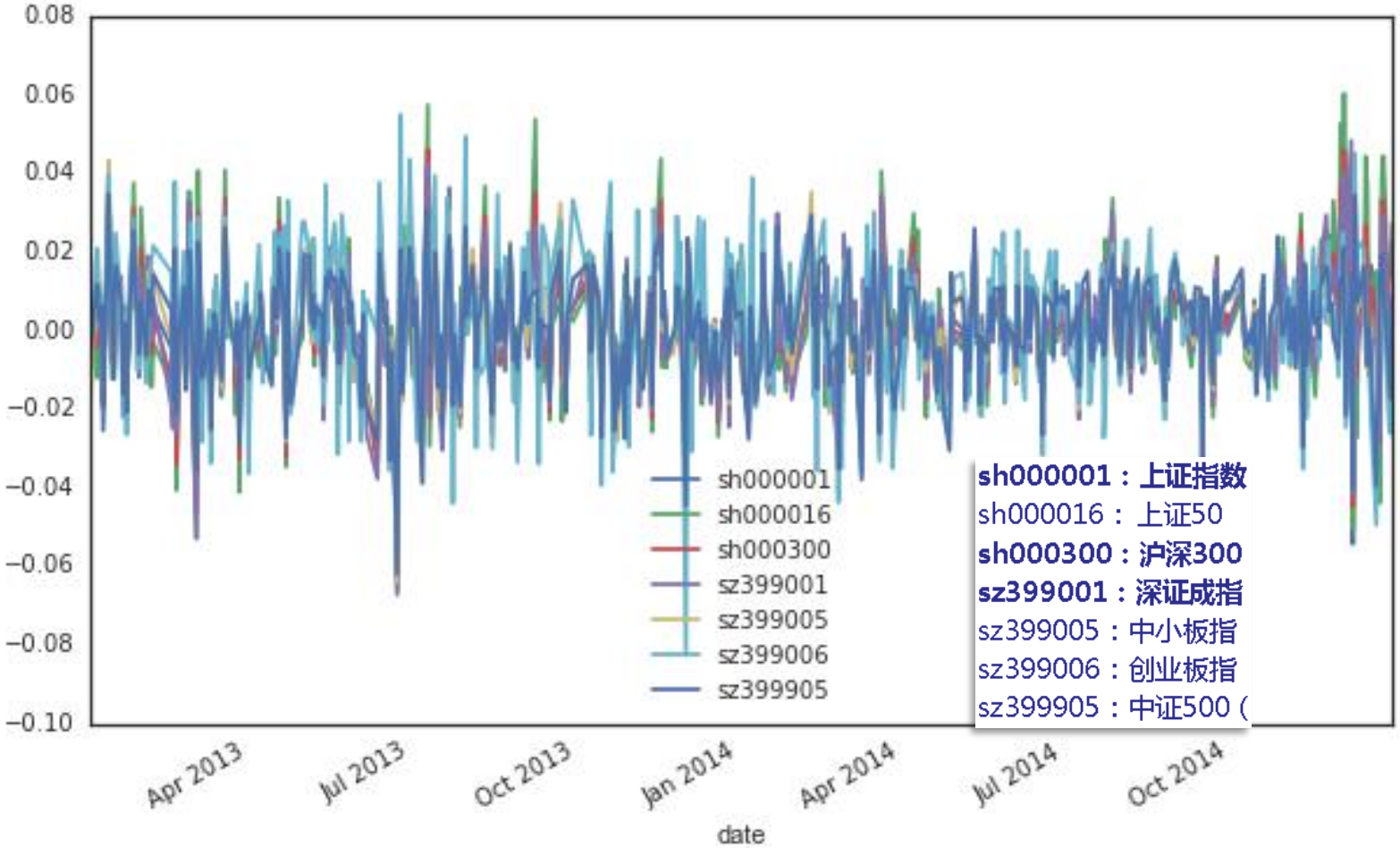
```
In [538]: indexclose.head()
Out[538]:
            sh000001  sh000016  sh000300  sz399001  sz399005  sz399006  \
date
2013-01-04   2276.99   1870.50   2524.41   9096.07   4172.74    705.34
2013-01-07   2285.36   1880.26   2535.99   9108.47   4205.12    718.11
2013-01-08   2276.07   1857.69   2525.33   9110.45   4270.25    733.14
2013-01-09   2275.34   1855.34   2526.13   9154.87   4312.19    734.51
2013-01-10   2283.66   1855.49   2530.57   9170.06   4337.44    741.18

            sz399905
date
2013-01-04   3258.25
2013-01-07   3288.37
2013-01-08   3327.04
2013-01-09   3334.99
2013-01-10   3356.35
```

查看数据

小象学院 ChinaHadoop.cn

➢ **查看股指数据：change**

```
In [545]: indexchange.plot(figsize=(10,6))
Out[545]: <matplotlib.axes._subplots.AxesSubplot at 0x229eaeb0>
```



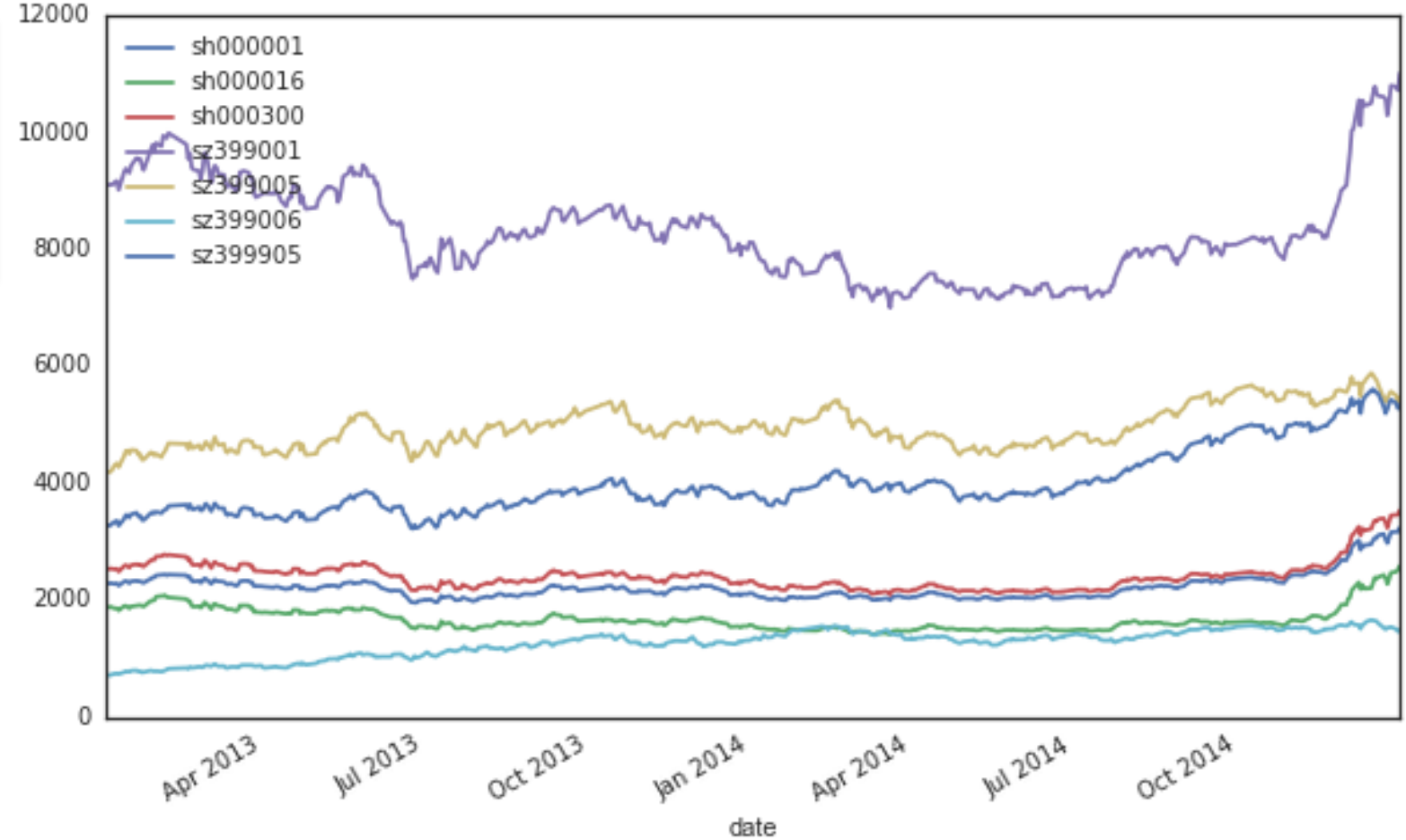
查看数据

小象学院 ChinaHadoop.cn

➢ **查看股指数据：close**

```
In [546]: indexclose.plot(figsize=(10,6))
Out[546]: <matplotlib.axes._subplots.AxesSubplot at 0x22be3b50>
```

sh000001：上证指数
sh000016：上证50
sh000300：沪深300
sz399001：深证成指
sz399005：中小板指
sz399006：创业板指
sz399905：中证500（

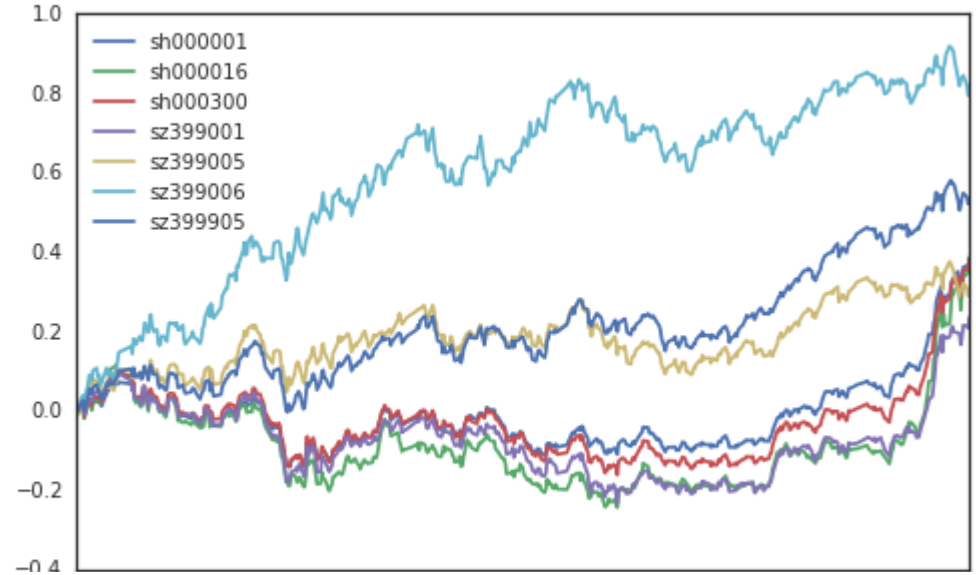


查看数据

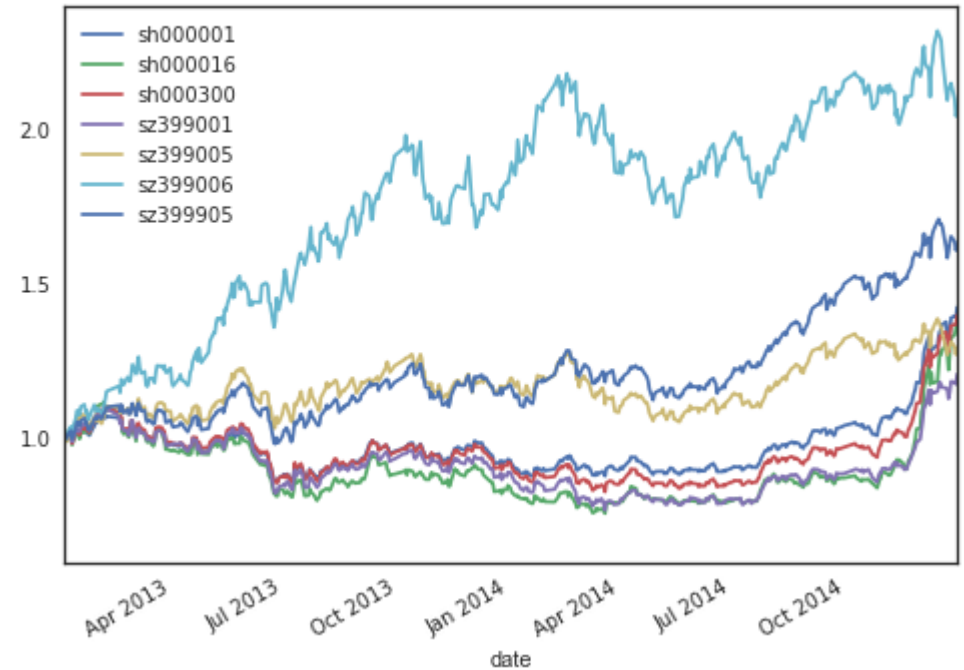
➢ **累计收益**

sh000001：上证指数
sh000016：上证50
sh000300：沪深300
sz399001：深证成指
sz399005：中小板指
sz399006：创业板指
sz399905：中证500（

```
In [557]: indexchange.cumsum().plot()
Out[557]: <matplotlib.axes._subplots.AxesSubplot at 0x218a3eb0>
```




```
In [558]: (indexchange+1).cumprod().plot()
Out[558]: <matplotlib.axes._subplots.AxesSubplot at 0x21827df0>
```



查看数据

小象学院 ChinaHadoop.cn

➢ **多个股指的不同模式**

sh000001：上证指数
sh000016：上证50
sh000300：沪深300
sz399001：深证成指
sz399005：中小板指
sz399006：创业板指
sz399905：中证500（

```
In [568]: indexchange[['sz399905','sh000001','sz399006']].cumsum().plot()
Out[568]: <matplotlib.axes._subplots.AxesSubplot at 0x239969b0>
```

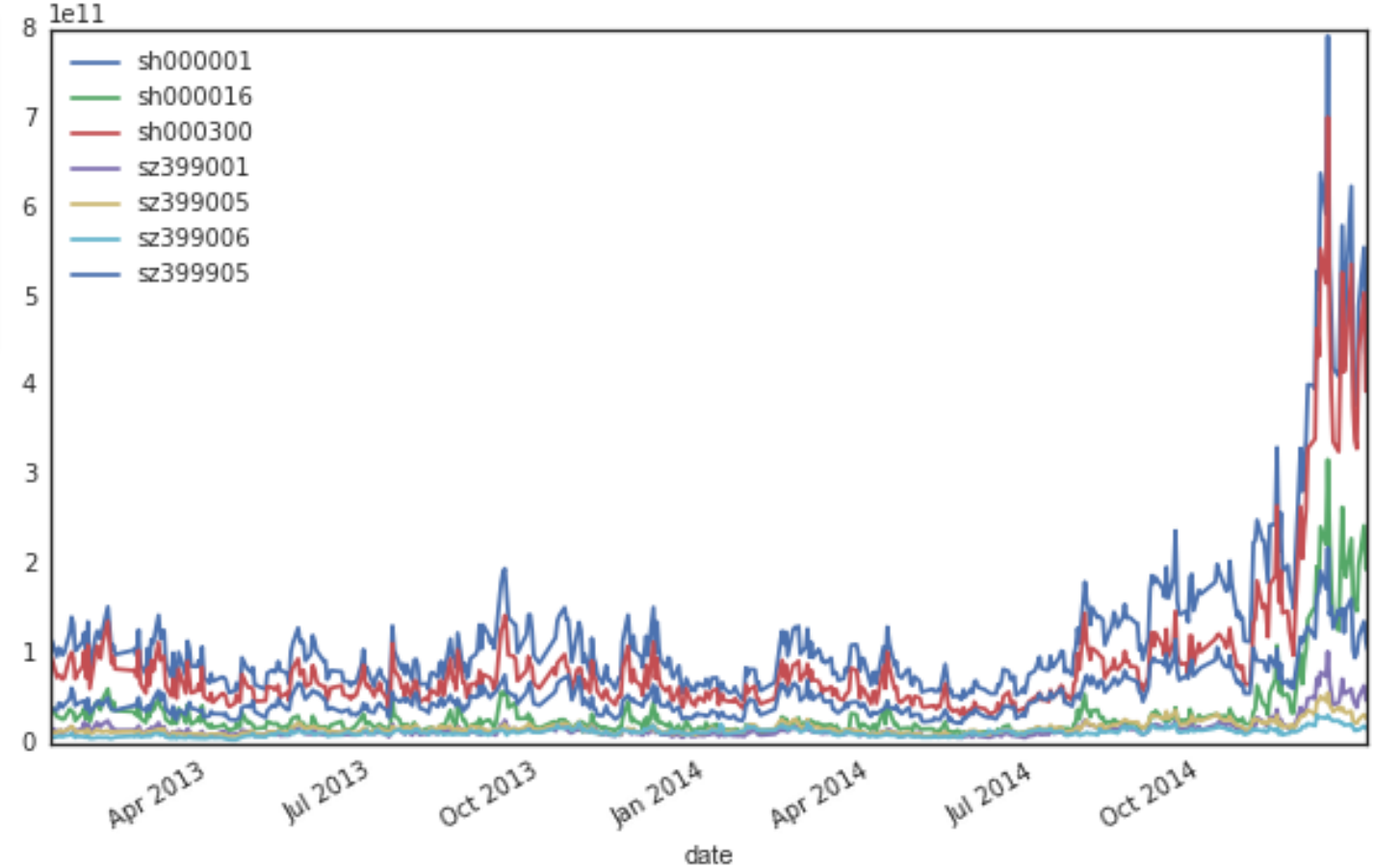
查看数据

小象学院 ChinaHadoop.cn

➢ **查看股指数据：money**

```
In [547]: indexmoney.plot(figsize=(10,6))
Out[547]: <matplotlib.axes._subplots.AxesSubplot at 0x22d716f0>
```

sh000001：上证指数
sh000016：上证50
sh000300：沪深300
sz399001：深证成指
sz399005：中小板指
sz399006：创业板指
sz399905：中证500（



查看数据

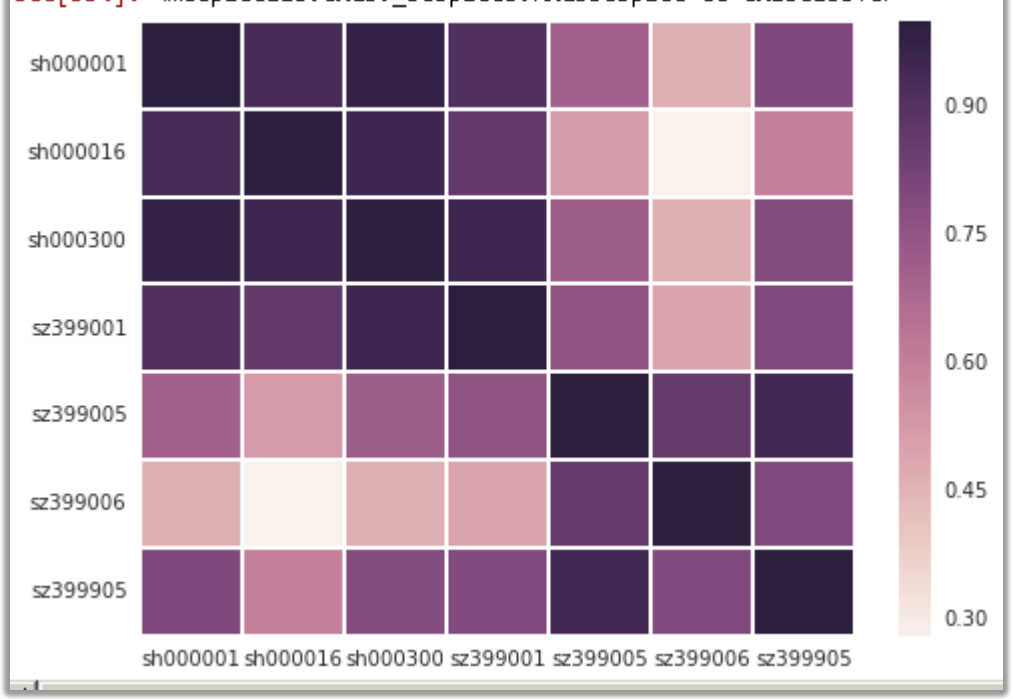
小象学院 ChinaHadoop.cn

➢ **股指关系**

sh000001：上证指数
sh000016： 上证50
sh000300：沪深300
sz399001：深证成指
sz399005：中小板指
sz399006：创业板指
sz399905：中证500（

```
In [549]: indexchange.corr()
Out[549]:
          sh000001  sh000016  sh000300  sz399001  sz399005  sz399006  sz399905
sh000001  1.000000  0.933182  0.979720  0.911143  0.708948  0.471133  0.798743
sh000016  0.933182  1.000000  0.956072  0.869037  0.522589  0.281818  0.603791
sh000300  0.979720  0.956072  1.000000  0.955005  0.717971  0.467743  0.788664
sz399001  0.911143  0.869037  0.955005  1.000000  0.756436  0.497198  0.796615
sz399005  0.708948  0.522589  0.717971  0.756436  1.000000  0.862828  0.944993
sz399006  0.471133  0.281818  0.467743  0.497198  0.862828  1.000000  0.797664
sz399905  0.798743  0.603791  0.788664  0.796615  0.944993  0.797664  1.000000
```

```
In [554]: sns.heatmap(indexchange.corr())
Out[554]: <matplotlib.axes._subplots.AxesSubplot at 0x23025bf0>
```



查看数据

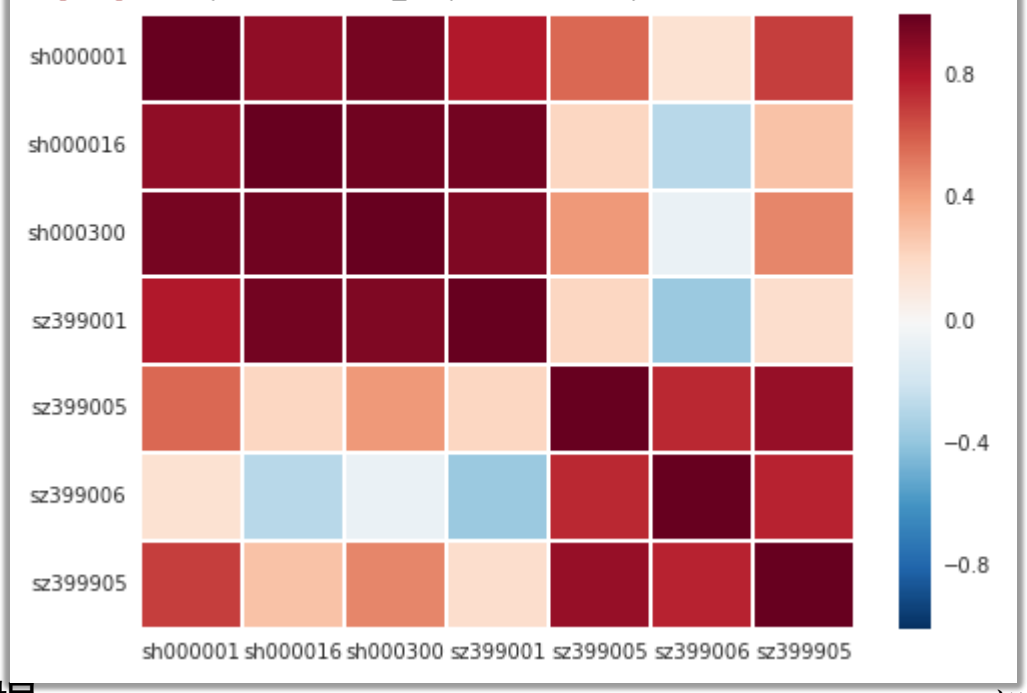
小象学院 ChinaHadoop.cn

➢ **股指关系**

sh000001：上证指数
sh000016： 上证50
sh000300：沪深300
sz399001：深证成指
sz399005：中小板指
sz399006：创业板指
sz399905：中证500（

```
In [550]: indexclose.corr()
Out[550]:
          sh000001  sh000016  sh000300  sz399001  sz399005  sz399006  sz399905
sh000001  1.000000  0.886375  0.960532  0.804363  0.571551  0.155655  0.694829
sh000016  0.886375  1.000000  0.970960  0.962361  0.212963 -0.277088  0.290137
sh000300  0.960532  0.970960  1.000000  0.932055  0.431059 -0.066834  0.486013
sz399001  0.804363  0.962361  0.932055  1.000000  0.212292 -0.368604  0.174772
sz399005  0.571551  0.212963  0.431059  0.212292  1.000000  0.755360  0.870195
sz399006  0.155655 -0.277088 -0.066834 -0.368604  0.755360  1.000000  0.768800
sz399905  0.694829  0.290137  0.486013  0.174772  0.870195  0.768800  1.000000
```

```
In [553]: sns.heatmap(indexclose.corr())
C:\Python27\lib\site-packages\pandas\core\format.py:2034: RuntimeWarning
  abs_vals = np.abs(self.values)
Out[553]: <matplotlib.axes._subplots.AxesSubplot at 0x22d5e8b0>
```



查看数据

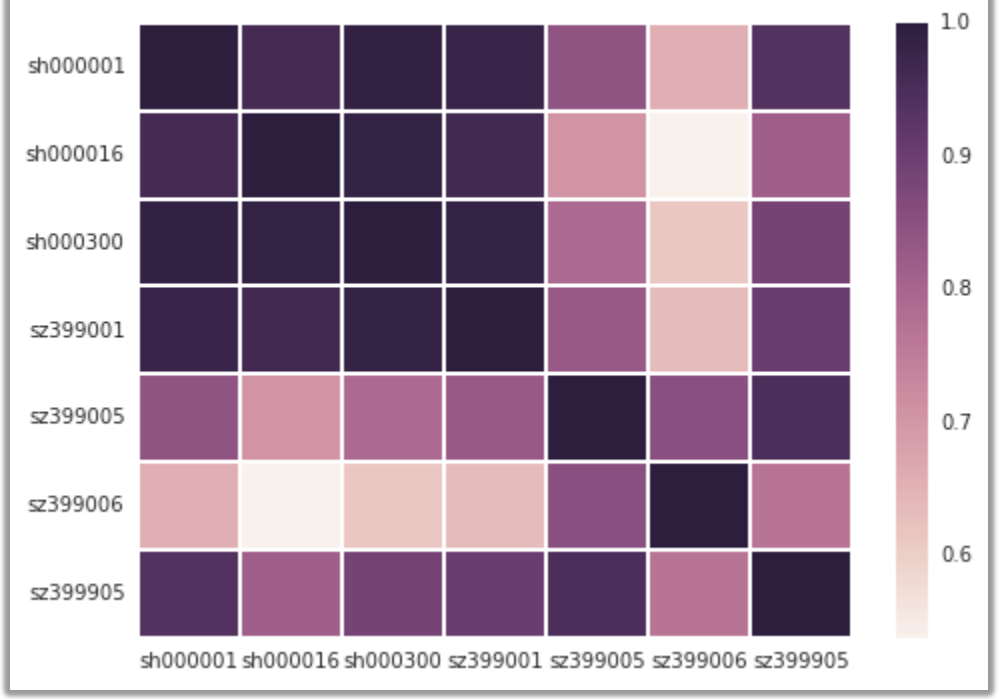
小象学院 ChinaHadoop.cn

➢ **股指关系**

sh000001：上证指数
sh000016： 上证50
sh000300：沪深300
sz399001：深证成指
sz399005：中小板指
sz399006：创业板指
sz399905：中证500（

```
In [556]: print indexmoney.corr()
          sh000001  sh000016  sh000300  sz399001  sz399005  sz399006  sz399905
sh000001  1.000000  0.960630  0.989806  0.978664  0.844547  0.657400  0.938570
sh000016  0.960630  1.000000  0.987988  0.964921  0.707384  0.538206  0.816123
sh000300  0.989806  0.987988  1.000000  0.984766  0.791764  0.614259  0.887815
sz399001  0.978664  0.964921  0.984766  1.000000  0.828892  0.635385  0.903005
sz399005  0.844547  0.707384  0.791764  0.828892  1.000000  0.854636  0.952992
sz399006  0.657400  0.538206  0.614259  0.635385  0.854636  1.000000  0.769279
sz399905  0.938570  0.816123  0.887815  0.903005  0.952992  0.769279  1.000000
```

```
In [555]: sns.heatmap(indexmoney.corr())
Out[555]: <matplotlib.axes._subplots.AxesSubplot at 0x23102e10>
```



挖掘数据

小象学院 ChinaHadoop.cn

- **哪个股指最有代表性？**
  - 采用沪深300（sh000300）作为“大盘”基准
  - 可以使用凝聚的聚类方法

```
from sklearn.cluster import AgglomerativeClustering
from sklearn.neighbors import kneighbors_graph
```

sklearn.cluster.AgglomerativeClustering

class sklearn.cluster. AgglomerativeClustering(*n_clusters=2, affinity='euclidean', memory=Memory(cachedir=None), connectivity=None, n_components=None, compute_full_tree='auto', linkage='ward', pooling_func=<function mean>*) [source]

Methods

fit (X[, y])	Fit the hierarchical clustering on the data
fit_predict (X[, y])	Performs clustering on X and returns cluster labels.
get_params ([deep])	Get parameters for this estimator.
set_params (**params)	Set the parameters of this estimator.

__init__(*n_clusters=2, affinity='euclidean', memory=Memory(cachedir=None), connectivity=None, n_components=None, compute_full_tree='auto', linkage='ward', pooling_func=<function mean>*) [source]

读入数据

小象学院 ChinaHadoop.cn

- 读入股票数据
- 手动删除过短的数据
  - <20k（85行）
  - 删除150只个股

sh600001.csv	2016/4/7 9:02	Microsoft Excel Com...	1 KB
sh600002.csv	2016/4/7 9:02	Microsoft Excel Com...	1 KB
sh600003.csv	2016/4/7 9:02	Microsoft Excel Com...	1 KB
sh600065.csv	2016/4/7 9:03	Microsoft Excel Com...	1 KB
sh600092.csv	2016/4/7 9:03	Microsoft Excel Com...	1 KB
sh600102.csv	2016/4/7 9:03	Microsoft Excel Com...	1 KB
sh600181.csv	2016/4/7 9:04	Microsoft Excel Com...	1 KB
sh600205.csv	2016/4/7 9:04	Microsoft Excel Com...	1 KB
sh600263.csv	2016/4/7 9:04	Microsoft Excel Com...	1 KB

sz000527.csv	2016/4/7 9:10	Microsoft Excel Com...	19 KB
sh603009.csv	2016/4/7 9:09	Microsoft Excel Com...	19 KB
sh603126.csv	2016/4/7 9:09	Microsoft Excel Com...	18 KB
sz300390.csv	2016/4/7 9:17	Microsoft Excel Com...	18 KB
sz300388.csv	2016/4/7 9:17	Microsoft Excel Com...	18 KB
sz300384.csv	2016/4/7 9:17	Microsoft Excel Com...	18 KB
sh603111.csv	2016/4/7 9:09	Microsoft Excel Com...	18 KB
sz300391.csv	2016/4/7 9:17	Microsoft Excel Com...	18 KB
sh603100.csv	2016/4/7 9:09	Microsoft Excel Com...	18 KB
sh603609.csv	2016/4/7 9:09	Microsoft Excel Com...	17 KB

读入数据

小象学院 ChinaHadoop.cn

➢ 读入股票数据

	code	date	open	high	low	close	change	volume	money	traded_ma	market_val	turn
2	sh600010	2014/12/31	4.06	4.1	4.04	4.08	0.007407	2.76E+08	1.12E+09	6.42E+10	6.53E+10	0.0
3	sh600010	2014/12/30	4.18	4.2	4	4.05	-0.03571	4.34E+08	1.76E+09	6.38E+10	6.48E+10	0.0
4	sh600010	2014/12/29	4.25	4.35	4.15	4.2	-0.01409	4.52E+08	1.93E+09	6.61E+10	6.72E+10	0.0
5	sh600010	2014/12/26	4.3	4.37	4.2	4.26	0.014286	4.42E+08	1.89E+09	6.71E+10	6.82E+10	0.0

```python
os.chdir('../stockdata/')
#!dir
files=os.listdir('.')
#获得股票代码列表

stocklist=[]
stockchange=pd.DataFrame()
stockmoney=pd.DataFrame()
stockclose=pd.DataFrame()

#
for filename in files:
    dummy=pd.read_csv(filename,parse_dates=True,index_col=0,usecols=[1,5,6,8])
    dummy=dummy.sort()
    if dummy.empty==False:
        if stockchange.empty:
            stocklist.append(filename[0:8])
            stockclose=pd.DataFrame(dummy['close'])
            stockchange=pd.DataFrame(dummy['change'])
            stockmoney=pd.DataFrame(dummy['money'])

        else:
            stocklist.append(filename[0:8])
            stockchange=pd.merge(stockchange,pd.DataFrame(dummy['change']),left_index=True, right_index=True,how='outer')
            stockmoney=pd.merge(stockmoney,pd.DataFrame(dummy['money']),left_index=True, right_index=True,how='outer')
            stockclose=pd.merge(stockclose,pd.DataFrame(dummy['close']),left_index=True, right_index=True,how='outer')

stockchange.columns=stocklist
stockmoney.columns=stocklist
stockclose.columns=stocklist
```

```
In [572]: stocklist[-10:]
Out[572]:
['sz300377',
 'sz300378',
 'sz300379',
 'sz300380',
 'sz300381',
 'sz300382',
 'sz300383',
 'sz300385',
 'sz300386',
 'sz300387']
```

查看数据

小象学院 ChinaHadoop.cn

- **股票数据**
  - 有缺失值
  - 需要特殊处理

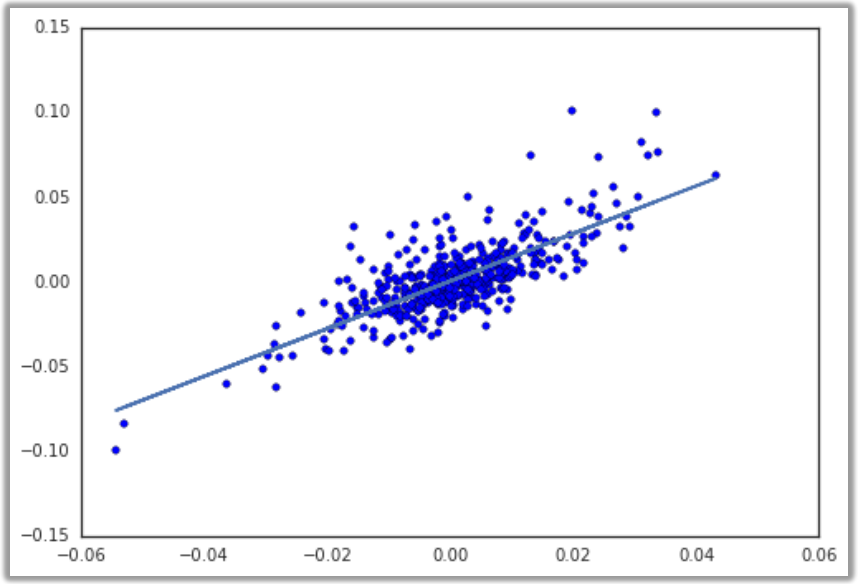
```
In [574]: stockchange.head()
Out[574]:
            sh600000  sh600004  sh600005  sh600006  sh600007  sh600008  \
date
2013-01-04  0.010081 -0.001410  0.021661  0.003344 -0.010444 -0.011442
2013-01-07  0.029940  0.001412 -0.007067 -0.003333  0.025506 -0.002315
2013-01-08 -0.019380  0.002821 -0.003559 -0.006689  0.002573  0.009281
2013-01-09  0.001976 -0.005626  0.003571  0.010101 -0.011976  0.009195
2013-01-10 -0.009862 -0.004243  0.003559  0.003333 -0.005195 -0.006834

            sh600009  sh600010  sh600011  sh600012    ...     sz300377  \
date                                                    ...
2013-01-04 -0.004815  0.007407 -0.022409 -0.002475    ...          NaN
2013-01-07  0.012097 -0.014706  0.004298  0.002481    ...          NaN
2013-01-08 -0.003984  0.005597  0.014265  0.000000    ...          NaN
2013-01-09  0.002400 -0.014842 -0.036568  0.000000    ...          NaN
2013-01-10 -0.006385  0.032015 -0.021898  0.000000    ...          NaN

            sz300378  sz300379  sz300380  sz300381  sz300382  sz300383  \
date
2013-01-04       NaN       NaN       NaN       NaN       NaN       NaN
2013-01-07       NaN       NaN       NaN       NaN       NaN       NaN
2013-01-08       NaN       NaN       NaN       NaN       NaN       NaN
2013-01-09       NaN       NaN       NaN       NaN       NaN       NaN
2013-01-10       NaN       NaN       NaN       NaN       NaN       NaN

            sz300385  sz300386  sz300387
date
2013-01-04       NaN       NaN       NaN
```

分析数据

➢ **建立个股与大盘的关系**

$$R = \alpha + \beta R_M + \varepsilon$$

```
#####test on linear model R=alpha+beta*Rm+noise(sigma)
x=indexchange.iloc[:,0]
y=stockchange.iloc[:,10]
indexnotnan=y[np.isnan(y)==False].index
x1=x[indexnotnan]
y1=y[indexnotnan]
beta, alpha, r_value, p_value, sigma = stats.linregress(x1,y1)
fig,ax=plt.subplots(1,1)
ax.plot(x1,alpha+x1*beta)
ax.scatter(x1,y1)
fig
######### test done
```



分析数据

小象学院 ChinaHadoop.cn

➢ 建立个股与大盘的关系

```
def returns(x,y):
    indexnotnan=y[np.isnan(y)==False].index
    x1=x[indexnotnan]
    y1=y[indexnotnan]
    beta, alpha, r_value, p_value, sigma = stats.linregress(x1,y1)
    return alpha,beta,sigma

```

```
stockreturndf=pd.DataFrame(columns=['code','alpha','beta','sigma'])
x=indexchange.loc[:,'sh000300'] #使用沪深300作为大盘基准

for stockcode in stocklist:
    y=stockchange.loc[:,stockcode]
    alpha,beta,sigma=returns(x,y)
    row=dict(code=stockcode,alpha=alpha,beta=beta,sigma=sigma)
    stockreturndf=stockreturndf.append(pd.DataFrame([row],))

######stockreturndf
stockreturns=stockreturndf[['alpha','beta','sigma']]
stockreturns.index=stockreturndf['code']

```

```
In [579]: stockreturns.head(10)
Out[579]:
              alpha      beta     sigma
code
sh600000  0.000440  1.316249  0.046822
sh600004  0.000722  0.655662  0.046049
sh600005  0.000072  0.718727  0.043595
sh600006  0.001151  0.715077  0.075273
sh600007  0.000309  0.762890  0.070316
sh600008  0.001703  0.965936  0.079994
sh600009  0.000655  0.696613  0.053060
sh600010  0.000438  0.887602  0.079188
sh600011  0.000218  0.775024  0.054010
sh600012  0.000775  0.584021  0.050355

In [580]: stockreturns.tail(10)
Out[580]:
              alpha      beta     sigma
code
sz300377  0.006351  0.396275  0.280107
sz300378  0.003169  0.690281  0.253757
sz300379  0.007876  0.446022  0.319544
sz300380  0.003723  0.901478  0.269061
sz300381  0.004429  0.805568  0.348388
sz300382  0.001934  0.654163  0.223846
sz300383  0.003712  0.446254  0.260946
sz300385  0.008785  0.364579  0.397219
sz300386  0.010099  0.263167  0.372910
sz300387  0.006561  0.233624  0.374764
```

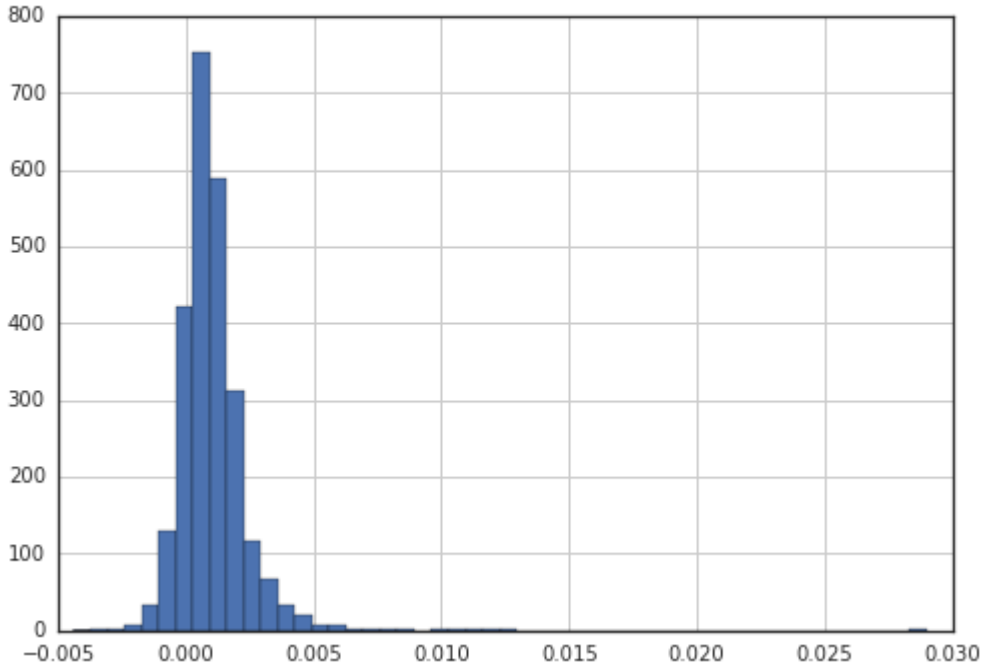
分析数据

小象学院 ChinaHadoop.cn

- 个股与大盘的关系：\alpha

$$R = \alpha + \beta R_M + \varepsilon$$

```
In [583]: stockreturns.alpha.hist(bins=50)
Out[583]: <matplotlib.axes._subplots.AxesSubplot at 0x24fd0bb0>
```




```
In [585]: stockreturns.alpha.describe()
Out[585]:
count    2524.000000
mean        0.001040
std         0.001417
min        -0.004419
25%         0.000300
50%         0.000833
75%         0.001534
max         0.028986
Name: alpha, dtype: float64
```

分析数据

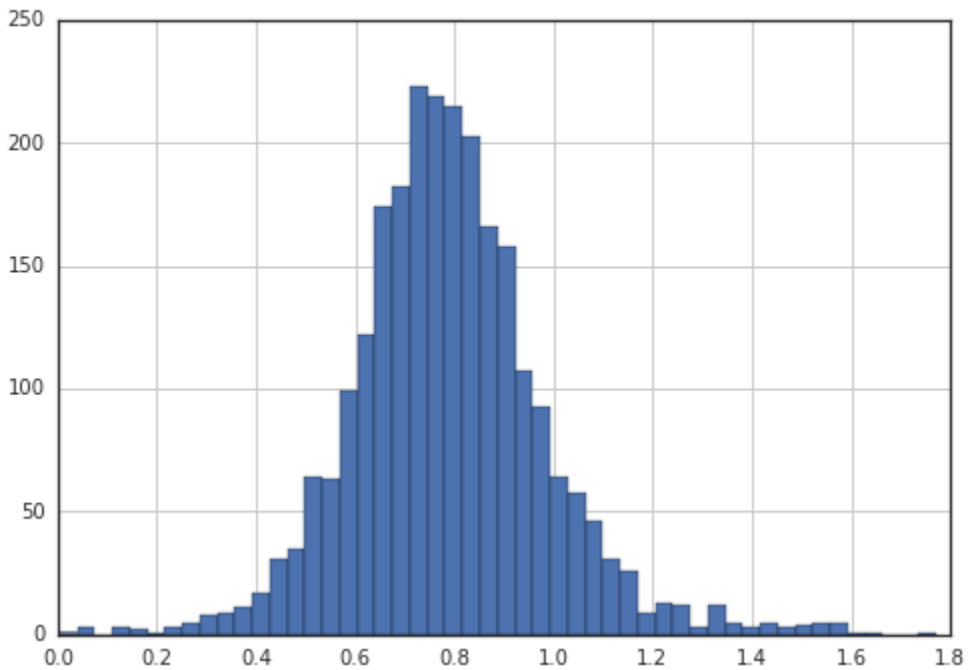
小象学院 ChinaHadoop.cn

- 个股与大盘的关系：\beta

$$R = \alpha + \beta R_M + \varepsilon$$

```
In [588]: stockreturns.beta.describe()
Out[588]:
count    2524.000000
mean        0.789066
std         0.196874
min         0.003861
25%         0.670343
50%         0.779204
75%         0.894908
max         1.769887
Name: beta, dtype: float64
```

```
In [589]: stockreturns.beta.hist(bins=50)
Out[589]: <matplotlib.axes._subplots.AxesSubplot at 0x2
```



分析数据

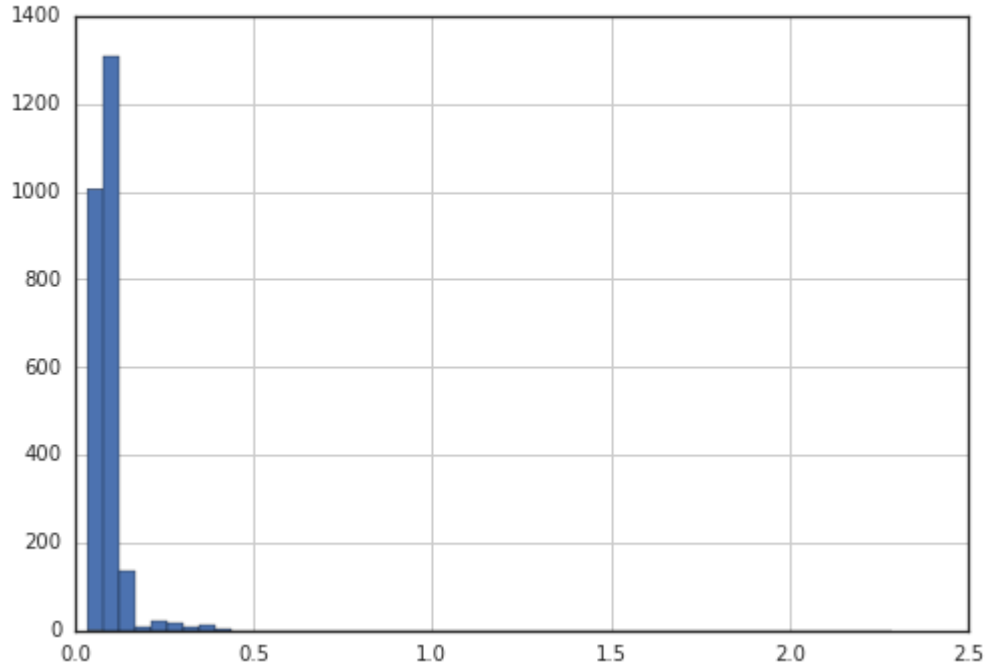
小象学院 ChinaHadoop.cn

- 个股与大盘的关系：\sigma

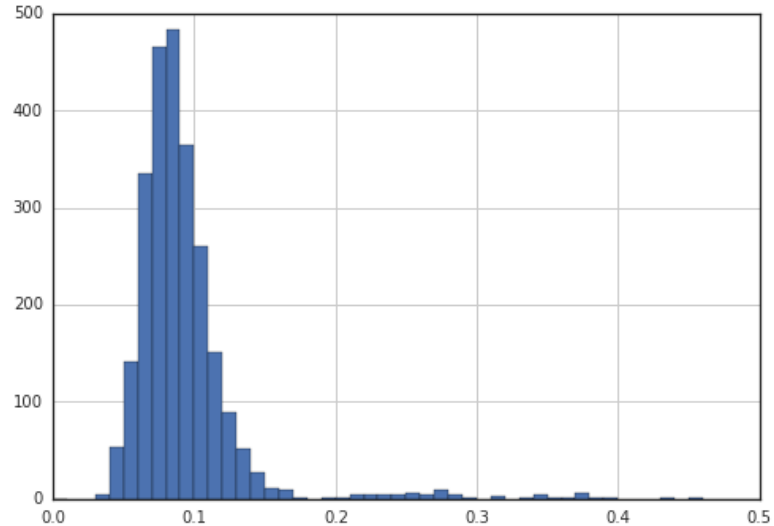
$$R = \alpha + \beta R_M + \varepsilon$$

```
In [591]: stockreturns.sigma.describe()
Out[591]:
count    2524.000000
mean        0.093921
std         0.062811
min         0.035133
25%         0.072342
50%         0.085735
75%         0.101126
max         2.285504
Name: sigma, dtype: float64
```

```
In [590]: stockreturns.sigma.hist(bins=50)
Out[590]: <matplotlib.axes._subplots.AxesSubplot at 0x25
```




```
In [700]: stockreturns.sigma.hist(bins=50,range=[0,0.5])
Out[700]: <matplotlib.axes._subplots.AxesSubplot at 0x26
```

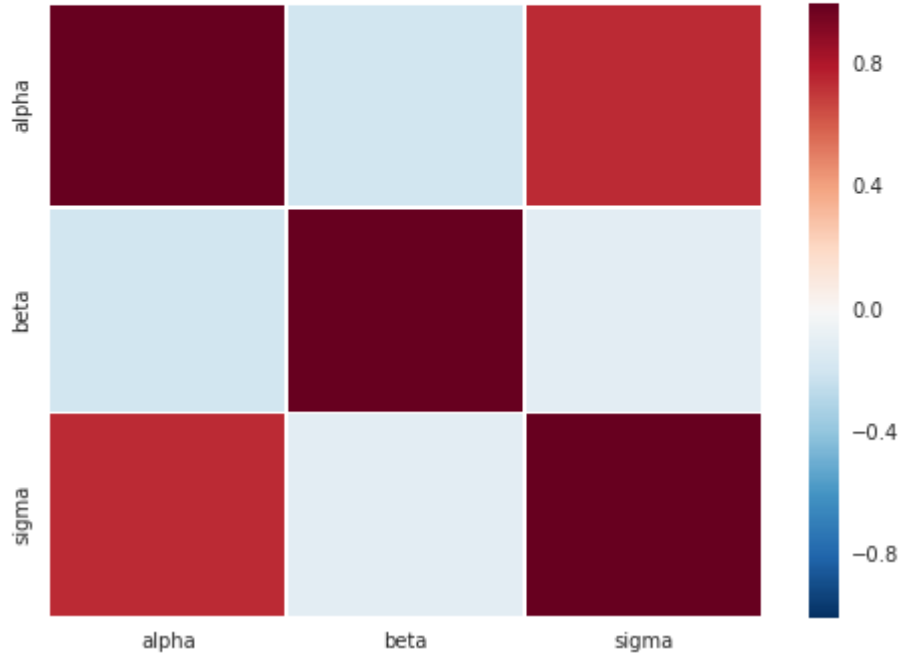


分析数据

小象学院 ChinaHadoop.cn

➢ **个股与大盘的关系**

```
          alpha      beta     sigma
alpha  1.000000 -0.193077  0.743119
beta  -0.193077  1.000000 -0.102080
sigma  0.743119 -0.102080  1.000000

In [670]: sns.heatmap(stockreturns.corr())
Out[670]: <matplotlib.axes._subplots.AxesSubplot at
```



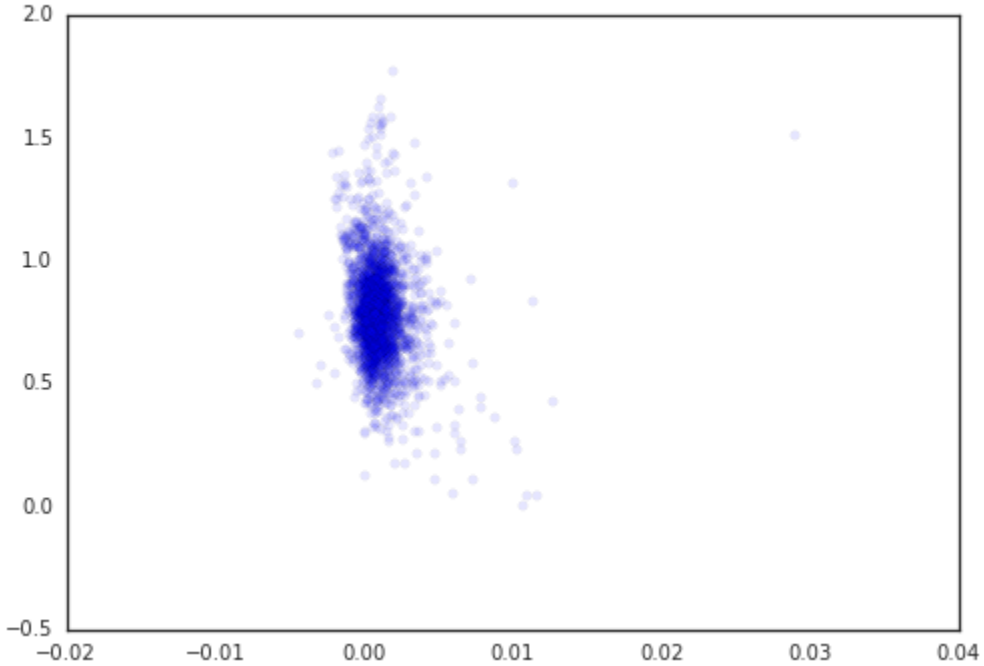

```
In [691]: rho,pvalue=stats.spearmanr(dummy)

In [692]: rho
Out[692]:
array([[ 1.        , -0.16530528,  0.6674119 ],
       [-0.16530528,  1.        , -0.12819208],
       [ 0.6674119 , -0.12819208,  1.        ]])

In [693]: pvalue
Out[693]:
array([[  0.00000000e+00,   6.36328025e-17,   0.00000000e+00],
       [  6.36328025e-17,   0.00000000e+00,   1.02111894e-10],
       [  0.00000000e+00,   1.02111894e-10,   0.00000000e+00]])
```

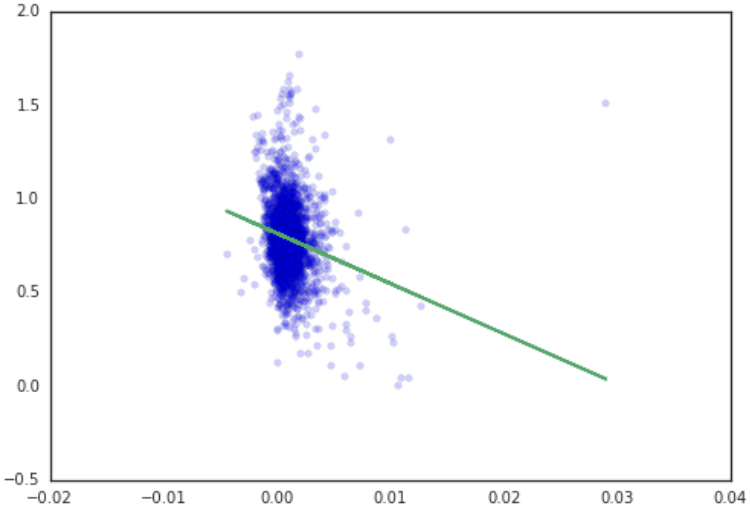
分析数据

小象学院 ChinaHadoop.cn

➢ **参数之间的关系： \alpha~\beta**

```
In [675]: plt.scatter(stockreturns.alpha, stockreturns.beta,alpha=0.1)
Out[675]: <matplotlib.collections.PathCollection at 0x262f2f90>
```

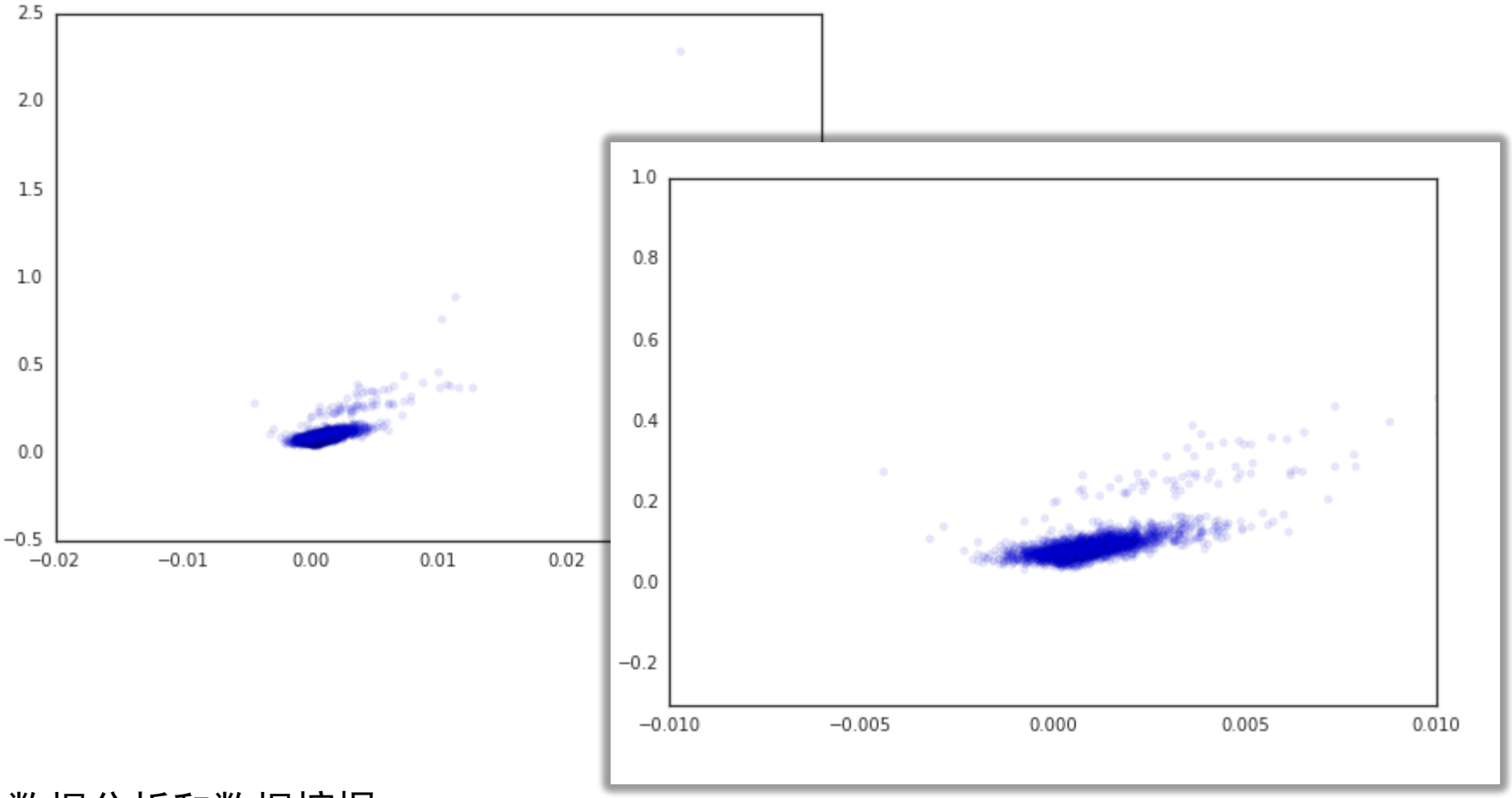




分析数据

- 参数之间的关系： \alpha~\sigma

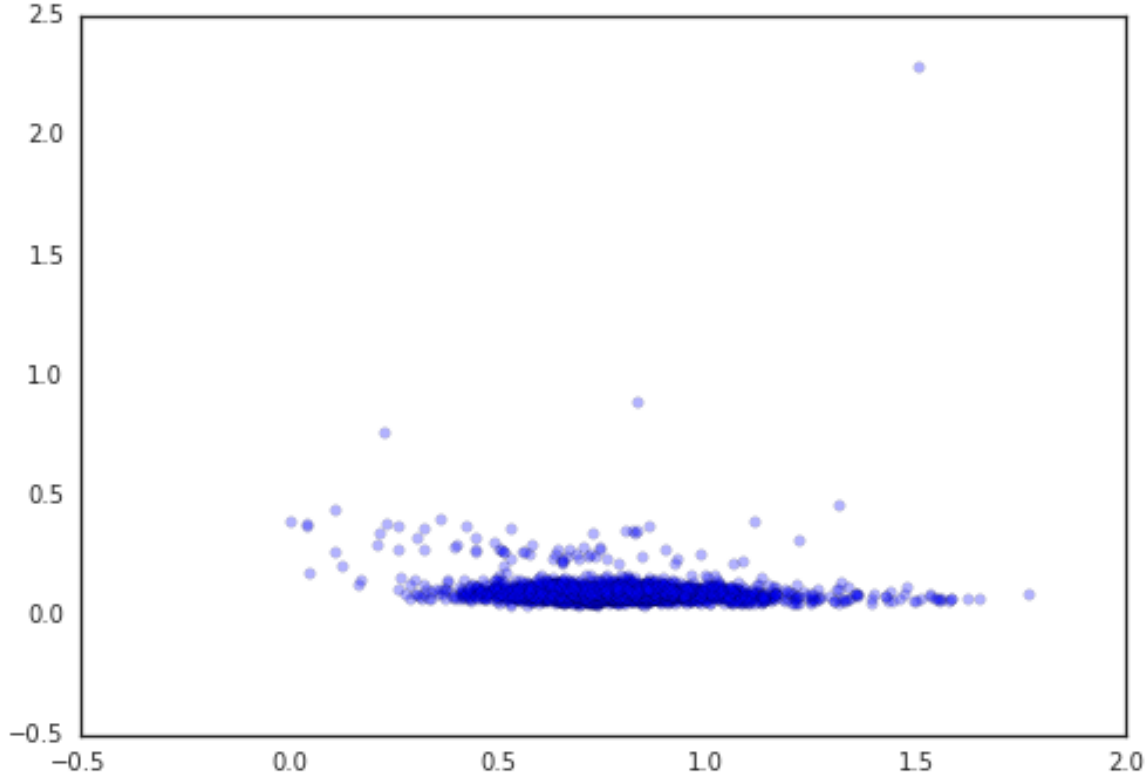
```
In [694]: plt.scatter(stockreturns.alpha, stockreturns.sigma,alpha=0.1)
Out[694]: <matplotlib.collections.PathCollection at 0x265f8cf0>
```



分析数据

➢ **参数之间的关系： \beta~\sigma**

```
In [699]: plt.scatter(stockreturns.beta, stockreturns.sigma,alpha=0.3)
Out[699]: <matplotlib.collections.PathCollection at 0x26815bb0>
```

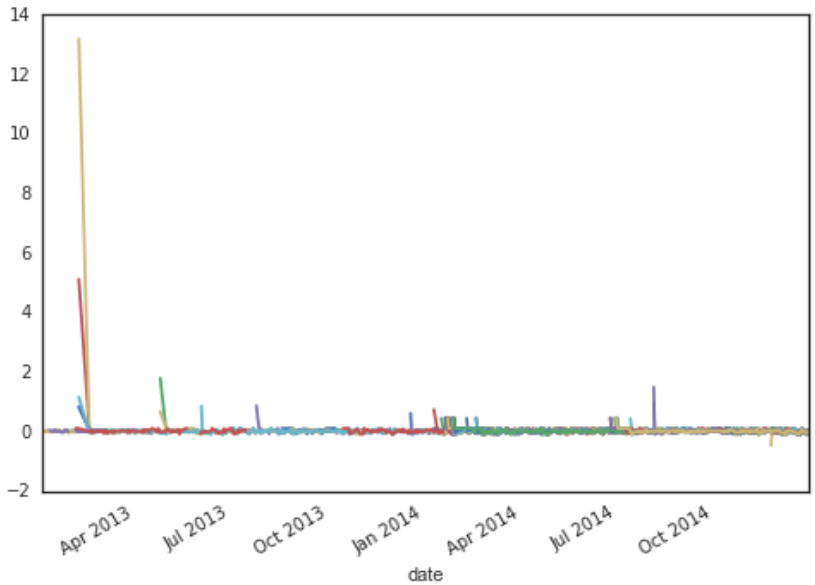


分析数据

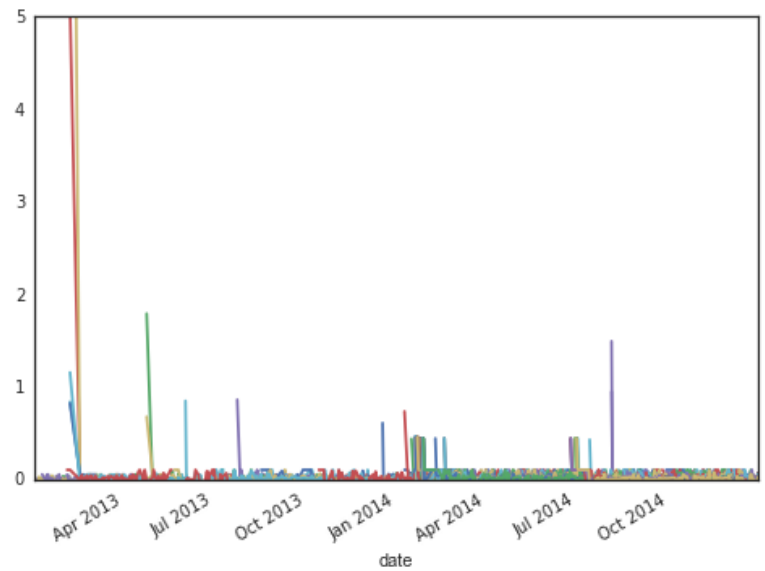
- **异常点**
  - abnormlist=list(stockreturns[stockreturns.sigma>0.18].index)
  - 去除异常点后可重复过程

	A	B	C	D	E	F	G	
24	sz000622	2013/4/8	5.12	5.14	4.92	5.05	-0.0251	102
25	sz000622	2013/4/3	5.32	5.45	5.15	5.18	-0.01708	141
26	sz000622	2013/4/2	5	5.27	4.92	5.27	0.049801	152
27	sz000622	2013/4/1	5.13	5.21	5.01	5.02	-0.04015	133
28	sz000622	2013/3/29	5.3	5.41	5.12	5.23	-0.01321	179
29	sz000622	2013/3/28	5.26	5.52	5.11	5.3	0.003788	314
30	sz000622	2013/3/27	5.05	5.28	5	5.28	0.049702	279
31	sz000622	2013/3/26	4.78	5.03	4.56	5.03	0.050104	181
32	sz000622	2013/3/25	4.93	5.03	4.77	4.79	-0.0284	110
33	sz000622	2013/3/22	4.97	5.14	4.8	4.93	0.004073	251
34	sz000622	2013/3/21	4.79	4.91	4.7	4.91	0.049145	70
35	sz000622	2013/3/20	4.47	4.68	4.44	4.68	0.049327	70
36	sz000622	2013/3/19	4.45	4.56	4.43	4.46	-0.04292	181
37	sz000622	2013/3/18	4.79	4.8	4.66	4.66	-0.05092	65
38	sz000622	2013/3/15	4.91	5.25	4.91	4.91	-0.05029	296
39	sz000622	2013/2/28	5.14	5.17	5.08	5.17	0.050813	124
40	sz000622	2013/2/27	4.67	4.92	4.63	4.92	0.049041	152
41	sz000622	2013/2/26	4.46	4.69	4.41	4.69	0.049217	208
42	sz000622	2013/2/25	4.54	4.62	4.45	4.47	-0.02188	94
43	sz000622	2013/2/22	4.61	4.74	4.47	4.57	-0.01509	156
44	sz000622	2013/2/21	4.77	4.9	4.63	4.64	-0.04723	167
45	sz000622	2013/2/20	4.93	4.96	4.87	4.87	-0.05068	288
46	sz000622	2013/2/19	4.79	5.25	4.79	5.13	0.017857	433
47	sz000622	2013/2/18	5.04	5.04	5.04	5.04	04906	13
48	sz000622	2013/2/8	7.9	7.99	5.3	5.3	13.19643	195
49								
50								

```
In [712]: stockchange[abnormlist].plot(legend=False)
Out[712]: <matplotlib.axes._subplots.AxesSubplot at 0x26fc8
```




```
In [715]: stockchange[abnormlist].plot(legend=False,ylim=[0,5])
Out[715]: <matplotlib.axes._subplots.AxesSubplot at 0x29020150>
```



分析数据

- **使用线性回归模型建立的个股与大盘的关系，得到描述个股的三个参数，可根据投资需求对其分类（有监督学习）**
  - \alpha 正负号：
  - \beta 是否大于1
  - ~\sigma 越大波动率越高

$$R = \alpha + \beta R_M + \varepsilon$$

超出市场部分

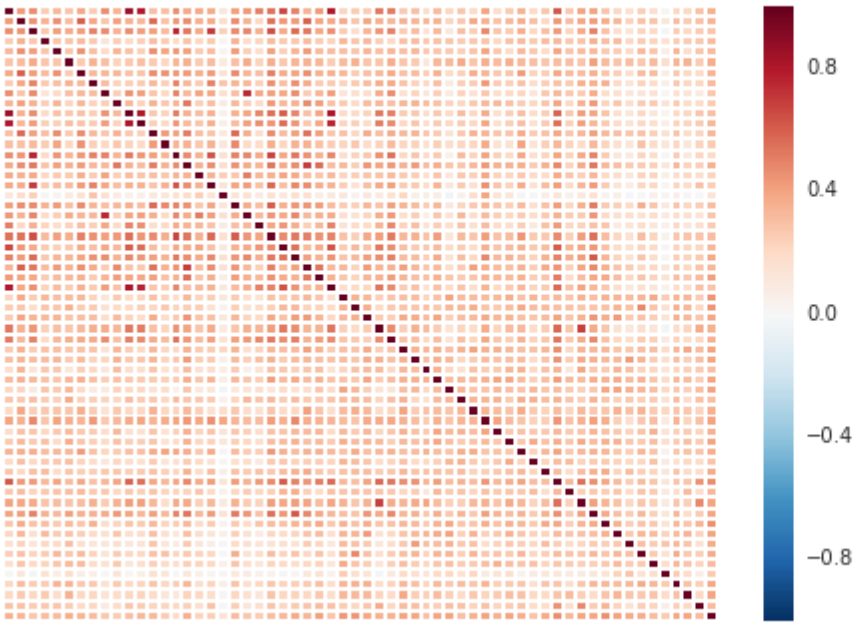
对市场的敏感度

个股风险

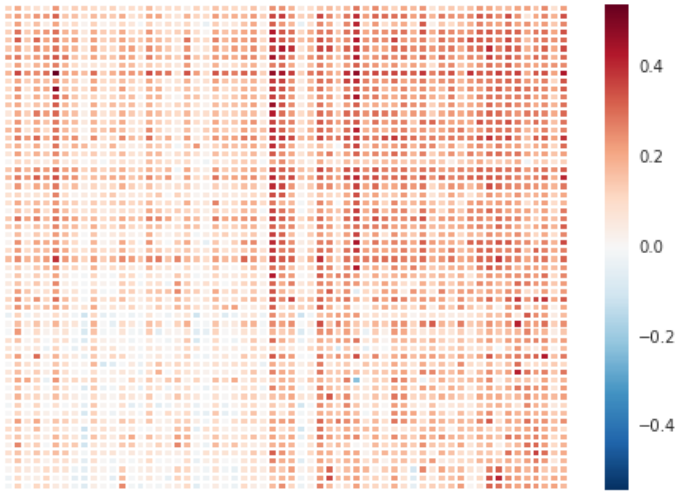
分析数据

- 或者，使用相近性（相关性）矩阵进行无监督的学习（聚类）
  - K-means
  - Agglomerative

```
In [785]: sns.heatmap(dummy.iloc[:60,:60],xticklabels=False,yticklabels=False)
Out[785]: <matplotlib.axes._subplots.AxesSubplot at 0x3373ff50>
```



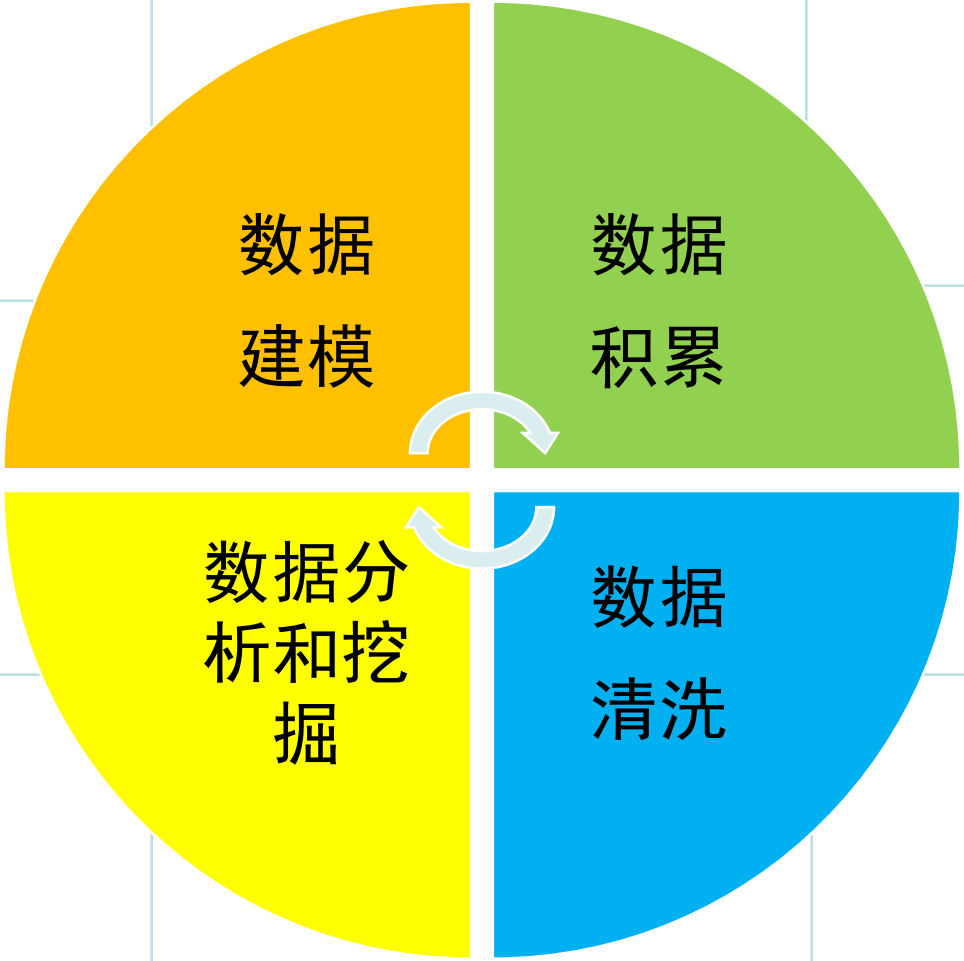

```
In [787]: sns.heatmap(dummy.iloc[-60:,:60],xtickl
Out[787]: <matplotlib.axes._subplots.AxesSubplot
```



数据工作流程

小象学院 ChinaHadoop.cn

- 维度（变量）
- 假设（模型）
- 约束条件
- 目标（目标函数）
- 方法（算法）



- 提供服务
- 用户运营
- 数据量和维度监控
- 存储

- 变量特征
- 变量关系
- 数据点关系
- 知识发现

- 格式、逻辑质量
- 业务需求
- 分析用数据结构

数据工作流程：数据建模

小象学院 ChinaHadoop.cn

- 维度（变量）
- 假设（模型）
- 约束条件
- 目标（目标函数）
- 方法（算法）

数据工作流程：数据积累

小象学院 ChinaHadoop.cn

- 提供服务
- 用户运营
- 数据量和维度监控
- 存储

数据工作流程：数据清洗

- 格式、逻辑质量
- 业务需求
- 分析用数据结构

数据工作流程：数据分析和挖掘

小象学院 ChinaHadoop.cn

- 变量特征
- 变量关系
- 数据点关系
- 知识发现

小象学院 ChinaHadoop.cn

谢谢！

联系我们：

- 新浪微博：ChinaHadoop
- 微信公号：ChinaHadoop
- 网站：http://chinahadoop.cn
- 问答社区：http://wenda.ChinaHadoop.cn