

Python 实验 数据预处理

班级:2017211314

学号:2017213508

学生:蒋雪枫

链家网爬虫：

1. 核心代码：

我们在爬虫的时候就进行数据预处理，采用正则表达式的方法获取数据：

```
def parse(self, response):
    item = MyItem()
    for each in response.xpath("//html/body/div[4]/ul[2]/*"):

        item['name'] = each.xpath("a/@title").extract()
        item['location1'] = each.xpath("div/div[2]/span[1]/text()").extract()
        item['location2'] = each.xpath("div/div[2]/span[2]/text()").extract()
        item['location3'] = each.xpath("div/div[2]/a/text()").extract()
        item['huxing'] = each.xpath("div/a/span[1]/text()").extract()
        item['area'] = each.xpath("div/div[3]/span/text()").extract()
        item['totality'] = each.xpath("div/div[6]/div[2]/text()").extract();
        item['meanPrice'] = each.xpath("div/div[6]/div[1]/span[1]/text()").extract()

        if(item['area']):
            match1 = re.search(r'[0-9]+' + item['area'][0])
            if match1:
                item['area'][0] = eval(match1.group(0))

        if(item['totality']):
            match1 = re.search(r'[0-9]+' + item['totality'][0])
            if match1:
                item['totality'][0] = eval(match1.group(0))

        if(item['meanPrice']):
            match1 = re.search(r'[0-9]+' + item['meanPrice'][0])
            if match1:
                item['meanPrice'][0] = eval(match1.group(0))

            if(item['meanPrice'][0] < 10000):
                #在北京不可能有低于 10000 元/平的房子，如果有，那一定是单位搞错了
                item['meanPrice'][0] = round(item['totality'][0] * 10000 / item['area'][0], 4)

    yield item
```

2. 爬取后的 csv 文件截图：

```
MyDataJson X
C > Users > Administrator > Desktop > python > Lianjia > {} MyDataJson > ...
1 [{"name": ["中国铁建花语金郡"], "location1": ["大兴"], "location2": ["瀛海"], "location3": ["南海子公园西侧(南五环旧忠桥向南第二个红绿灯西300米)", "huxing": ["3室"], "area": 180, "totality": 2500, "meanPrice": 120000}], [{"name": ["中国铁建万科翡翠长安"], "location1": ["门头沟"], "location2": ["门头沟其它"], "location3": ["西长安街沿线与六环交汇处西南1.2公里处"], "huxing": ["3室"], "area": 180, "totality": 2500, "meanPrice": 120000}], [{"name": ["电建金地宸宸"], "location1": ["门头沟"], "location2": ["门头沟其它"], "location3": ["长安街西延线南侧约500米"], "huxing": ["3室"], "area": 180, "totality": 2500, "meanPrice": 120000}], [{"name": ["北辰墅院1900"], "location1": ["顺义"], "location2": ["马坡"], "location3": ["顺义区顺兴街11号院望都园"], "huxing": ["4室"], "area": 251, "totality": 120, "meanPrice": 120000}], [{"name": ["首开璞瑅公馆"], "location1": ["丰台"], "location2": ["方庄"], "location3": ["紫芳园五区"], "huxing": ["4室"], "area": 236, "totality": 2500, "meanPrice": 120000}], [{"name": ["中昂时代广场"], "location1": ["门头沟"], "location2": ["大峪"], "location3": ["门头沟路9号博物馆对面"], "huxing": ["1室"], "area": 120, "totality": 120, "meanPrice": 120000}], [{"name": ["北辰墅院1900"], "location1": ["顺义"], "location2": ["马坡"], "location3": ["顺义区顺兴街11号院望都园"], "huxing": ["3室"], "area": 120, "totality": 120, "meanPrice": 120000}], [{"name": ["首创远洋禧瑞天著"], "location1": ["亦庄开发区"], "location2": ["通州其它"], "location3": ["科创十一街与经海九路交汇处西南角(亦庄线次渠南站200米)", "huxing": ["4室"], "area": 156, "totality": 950, "meanPrice": 120000}], [{"name": ["恒源钓鱼台"], "location1": ["丰台"], "location2": ["玉泉营"], "location3": ["西三环康庄东路于丰台东路交汇处西南角, 康庄东路9号院"], "huxing": ["4室"], "area": 220, "totality": 120, "meanPrice": 120000}], [{"name": ["V7九间堂"], "location1": ["通州"], "location2": ["潞苑"], "location3": ["通燕高速耿庄桥北出口中化石油对面"], "huxing": ["4室"], "area": 220, "totality": 120, "meanPrice": 120000}], [{"name": ["北京城建国誉府"], "location1": ["房山"], "location2": ["长阳"], "location3": ["地铁良乡大学城北站东北约1300米, 文昌东路与阜盛东街交汇处"], "huxing": ["4室"], "area": 295, "totality": 120, "meanPrice": 120000}], [{"name": ["御汤山熙园"], "location1": ["昌平"], "location2": ["昌平其它"], "location3": ["小汤山疗养院西侧, 紧邻安河路, 距离北六环61号出口约2000米"], "huxing": ["4室"], "area": 295, "totality": 120, "meanPrice": 120000}], [{"name": ["华远和墅"], "location1": ["大兴"], "location2": ["大兴其它"], "location3": ["南六环磁各庄桥沿南中轴向南2公里"], "huxing": ["5室"], "area": 295, "totality": 120, "meanPrice": 120000}], [{"name": ["天资华府"], "location1": ["房山"], "location2": ["长阳"], "location3": ["长阳CSD办公楼南侧(京港澳高速长阳出口西1000米)", "huxing": ["5室"], "area": 93, "totality": 120, "meanPrice": 120000}], [{"name": ["檀香府"], "location1": ["门头沟"], "location2": ["门头沟其它"], "location3": ["京潭大街与潭柘十街交叉口"], "huxing": ["3室"], "area": 208, "totality": 120, "meanPrice": 120000}], [{"name": ["观山源墅"], "location1": ["房山"], "location2": ["良乡"], "location3": ["阳光北大街与多宝路交汇处西南(理工大学北校区西侧)", "huxing": ["3室"], "area": 290, "totality": 120, "meanPrice": 120000}], [{"name": ["燕西华府"], "location1": ["丰台"], "location2": ["丰台其它"], "location3": ["王佐镇青龙湖公园东1500米, 泉湖西路1号院(七区), 泉湖西路1号院(六区)"], "huxing": ["4室"], "area": 60, "totality": 120, "meanPrice": 120000}], [{"name": ["远洋新天地"], "location1": ["门头沟"], "location2": ["门头沟其它"], "location3": ["长安街西延线与滨河路南延交汇处(东南侧)", "huxing": ["1室"], "area": 1, "totality": 120, "meanPrice": 120000}], [{"name": ["天恒水岸壹号"], "location1": ["房山"], "location2": ["良乡"], "location3": ["良乡大学城西站地铁南侧400米, 刺猬河旁"], "huxing": ["2室"], "area": 108, "totality": 120, "meanPrice": 120000}], [{"name": ["紫辰院"], "location1": ["丰台"], "location2": ["岳各庄"], "location3": ["岳各庄北桥东北角200米处"], "huxing": ["4室"], "area": 266, "totality": 120, "meanPrice": 120000}], [{"name": ["天恒摩墅"], "location1": ["房山"], "location2": ["房山其它"], "location3": ["周口店镇镇政府东200米"], "huxing": ["3室"], "area": 134, "totality": 120, "meanPrice": 120000}], [{"name": ["兴创荣墅"], "location1": ["大兴"], "location2": ["大兴其它"], "location3": ["采育镇育胜街与福源路交叉口西侧350米路南"], "huxing": ["3室"], "area": 240, "totality": 120, "meanPrice": 120000}], [{"name": ["景郡原著"], "location1": ["朝阳"], "location2": ["中央别墅区"], "location3": ["京密路与顺黄路交叉口往西200米"], "huxing": ["4室"], "area": 479, "totality": 120, "meanPrice": 120000}], [{"name": ["利锦府府上"], "location1": ["朝阳"], "location2": ["东坝"], "location3": ["东五环七棵树立口沿东坝中街向东1500米, 红松园北里18号院"], "huxing": ["4室"], "area": 280, "totality": 120, "meanPrice": 120000}], [{"name": ["中骏西山天璟"], "location1": ["门头沟"], "location2": ["城子"], "location3": ["西山永定楼北300米"], "huxing": ["4室"], "area": 117, "totality": 120, "meanPrice": 120000}], [{"name": ["泰禾昌平拾景园"], "location1": ["昌平"], "location2": ["昌平"], "location3": ["昌平线平线南部站西100米"], "huxing": ["3室"], "area": 210, "totality": 120, "meanPrice": 120000}], [{"name": ["国瑞熙墅"], "location1": ["昌平"], "location2": ["北七家"], "location3": ["北七家镇定泗路与七北路交叉口东行200米"], "huxing": ["3室"], "area": 314, "totality": 120, "meanPrice": 120000}], [{"name": ["领秀翡翠墅"], "location1": ["丰台"], "location2": ["丰台其它"], "location3": ["王佐镇长青路南, 长青路88号院"], "huxing": ["3室"], "area": 137, "totality": 120, "meanPrice": 120000}], [{"name": ["华润西山墅"], "location1": ["门头沟"], "location2": ["冯村"], "location3": ["冯石环路7号院"], "huxing": ["4室"], "area": 126, "totality": 120, "meanPrice": 120000}], [{"name": ["北宸怡园"], "location1": ["昌平"], "location2": ["东关"], "location3": ["水库路与昌崔路交叉口向北1公里路东"], "huxing": ["3室"], "area": 70, "totality": 120, "meanPrice": 120000}], [{"name": ["绿城西府海棠"], "location1": ["大兴"], "location2": ["石景山"], "location3": ["隆恩寺路北人大附石景山分校西侧"], "huxing": ["3室"], "area": 90, "totality": 120, "meanPrice": 120000}], [{"name": ["颐瑞万和"], "location1": ["大兴"], "location2": ["黄村火车站"], "location3": ["新源大街4号线和庄地铁向北约500米"], "huxing": ["3室"], "area": 80, "totality": 120, "meanPrice": 120000}]
```

	A	B	C	D	E	F	G	H
10	华远裘马四季	门头沟	大峪	增产路16号院	3室	156	950	60000
11	恒源钓鱼台	丰台	玉泉营	西三环康庄东路	4室	251	3263	130000
12	V7九间堂	通州	潞苑	通燕高速耿庄桥	4室	220	1700	77272. 7273
13	北京城建国誉府	房山	长阳	地铁良乡大学城	4室	143	1000	48000
14	御汤山熙园	昌平	昌平其它	小汤山疗养院西	4室	300	1800	60000
15	华远和墅	大兴	大兴其它	南六环磁各庄桥	5室	295	1580	53559. 322
16	天资华府	房山	长阳	长阳CSD办公楼南	5室	93		42000
17	檀香府	门头沟	门头沟其它	京潭大街与潭柘	3室	208	1060	55000
18	观山源墅	房山	良乡	阳光北大街与多	3室	290	1200	47500
19	燕西华府	丰台	丰台其它	王佐镇青龙湖公	4室	60	650	49800
20	远洋新天地	门头沟	门头沟其它	长安街西延线与滨河路南延交汇处（东南侧）				70000
21	天恒水岸壹号	房山	良乡	良乡大学城西站	2室	108	500	46296. 2963
22	紫辰院	丰台	岳各庄	岳各庄北桥东北	4室	266	3192	120000
23	天恒摩墅	房山	房山其它	周口店镇镇政府东	3室	134	350	26119. 403
24	兴创荣墅	大兴	大兴其它	采育镇育胜街与	3室	240	850	35416. 6667
25	景郡原著	朝阳	中央别墅区	京密路与顺黄路	4室	479	4000	83507. 3069
26	利锦府府上	朝阳	东坝	东五环七棵树立	4室	280	1700	60000
27	中骏西山天璟	门头沟	城子	西山永定楼北30	4室	117	700	65000
28	泰禾昌平拾景园	昌平	南部	地铁昌平线南部	3室	210	1200	57142. 8571
29	国瑞熙墅	昌平	北七家	北七家镇定泗路	3室	314	1500	47770. 7006
30	领秀翡翠墅	丰台	丰台其它	王佐镇长青路南	3室	137	668	51000
31	华润西山墅	门头沟	冯村	冯石环路7号院	4室	126	780	61904. 7619
32	北京怡园	昌平	东关	水库路与昌崔路	3室	70	430	46000

PM 指数分析：

源代码：

```
import numpy as np
import pandas as pd
import time

import matplotlib.pyplot as plt

#1.打开 CSV 文件
fileNameStr = 'BeijingPM20100101_20151231.csv'

df = pd.read_csv(fileNameStr,encoding='utf-8',dtype=str)

#2010 0101 - 2015 1231

#6 年 12 个月
```

```

print("info=====
=====")

print(df.info())

print("=====")

start = time.time()

account=[[0 for i in range(13)] for j in range(6)]
counters=[[0 for i in range(13)] for j in range(6)]

for i in range(52584):

    if eval(df['year'][i])==2010:

        if not(df['PM_Dongsi'][i] is np.nan):

            account[0][eval(df['month'][i]))+=int(df['PM_Dongsi'][i])

            counters[0][eval(df['month'][i]))+=1

        if not (df['PM_Dongsihuan'][i] is np.nan):

            account[0][eval(df['month'][i]))+=int(df['PM_Dongsihuan'][i])

            counters[0][eval(df['month'][i]))+=1

        if not(df['PM_Nongzhanguan'][i] is np.nan):

            account[0][eval(df['month'][i]))+=int(df['PM_Nongzhanguan'][i])

            counters[0][eval(df['month'][i]))+=1

        if not(df['PM_US Post'][i] is np.nan):

            account[0][eval(df['month'][i]))+=int(df['PM_US Post'][i])

            counters[0][eval(df['month'][i]))+=1

    elif eval(df['year'][i])==2011:

        if not(df['PM_Dongsi'][i] is np.nan):

            account[1][eval(df['month'][i]))+=int(df['PM_Dongsi'][i])

            counters[1][eval(df['month'][i]))+=1

        if not (df['PM_Dongsihuan'][i] is np.nan):

            account[1][eval(df['month'][i]))+=int(df['PM_Dongsihuan'][i])

            counters[1][eval(df['month'][i]))+=1

        if not(df['PM_Nongzhanguan'][i] is np.nan):

            account[1][eval(df['month'][i]))+=int(df['PM_Nongzhanguan'][i])

            counters[1][eval(df['month'][i]))+=1

        if not(df['PM_US Post'][i] is np.nan):

            account[1][eval(df['month'][i]))+=int(df['PM_US Post'][i])

            counters[1][eval(df['month'][i]))+=1

    elif eval(df['year'][i])==2012:

        if not(df['PM_Dongsi'][i] is np.nan):

            account[2][eval(df['month'][i]))+=int(df['PM_Dongsi'][i])

            counters[2][eval(df['month'][i]))+=1

        if not (df['PM_Dongsihuan'][i] is np.nan):

            account[2][eval(df['month'][i]))+=int(df['PM_Dongsihuan'][i])

            counters[2][eval(df['month'][i]))+=1

        if not(df['PM_Nongzhanguan'][i] is np.nan):

            account[2][eval(df['month'][i]))+=int(df['PM_Nongzhanguan'][i])

            counters[2][eval(df['month'][i]))+=1

```

```

    if not(df['PM_US Post'][i] is np.nan):
        account[2][eval(df['month'][i])] += int(df['PM_US Post'][i])
        counters[2][eval(df['month'][i])] += 1
elif eval(df['year'][i]) == 2013:
    if not(df['PM_Dongsi'][i] is np.nan):
        account[3][eval(df['month'][i])] += int(df['PM_Dongsi'][i])
        counters[3][eval(df['month'][i])] += 1
    if not(df['PM_Dongsihuan'][i] is np.nan):
        account[3][eval(df['month'][i])] += int(df['PM_Dongsihuan'][i])
        counters[3][eval(df['month'][i])] += 1
    if not(df['PM_Nongzhanguan'][i] is np.nan):
        account[3][eval(df['month'][i])] += int(df['PM_Nongzhanguan'][i])
        counters[3][eval(df['month'][i])] += 1
    if not(df['PM_US Post'][i] is np.nan):
        account[3][eval(df['month'][i])] += int(df['PM_US Post'][i])
        counters[3][eval(df['month'][i])] += 1
elif eval(df['year'][i]) == 2014:
    if not(df['PM_Dongsi'][i] is np.nan):
        account[4][eval(df['month'][i])] += int(df['PM_Dongsi'][i])
        counters[4][eval(df['month'][i])] += 1
    if not(df['PM_Dongsihuan'][i] is np.nan):
        account[4][eval(df['month'][i])] += int(df['PM_Dongsihuan'][i])
        counters[4][eval(df['month'][i])] += 1
    if not(df['PM_Nongzhanguan'][i] is np.nan):
        account[4][eval(df['month'][i])] += int(df['PM_Nongzhanguan'][i])
        counters[4][eval(df['month'][i])] += 1
    if not(df['PM_US Post'][i] is np.nan):
        account[4][eval(df['month'][i])] += int(df['PM_US Post'][i])
        counters[4][eval(df['month'][i])] += 1
elif eval(df['year'][i]) == 2015:
    if not(df['PM_Dongsi'][i] is np.nan):
        account[5][eval(df['month'][i])] += int(df['PM_Dongsi'][i])
        counters[5][eval(df['month'][i])] += 1
    if not(df['PM_Dongsihuan'][i] is np.nan):
        account[5][eval(df['month'][i])] += int(df['PM_Dongsihuan'][i])
        counters[5][eval(df['month'][i])] += 1
    if not(df['PM_Nongzhanguan'][i] is np.nan):
        account[5][eval(df['month'][i])] += int(df['PM_Nongzhanguan'][i])
        counters[5][eval(df['month'][i])] += 1
    if not(df['PM_US Post'][i] is np.nan):
        account[5][eval(df['month'][i])] += int(df['PM_US Post'][i])
        counters[5][eval(df['month'][i])] += 1

```

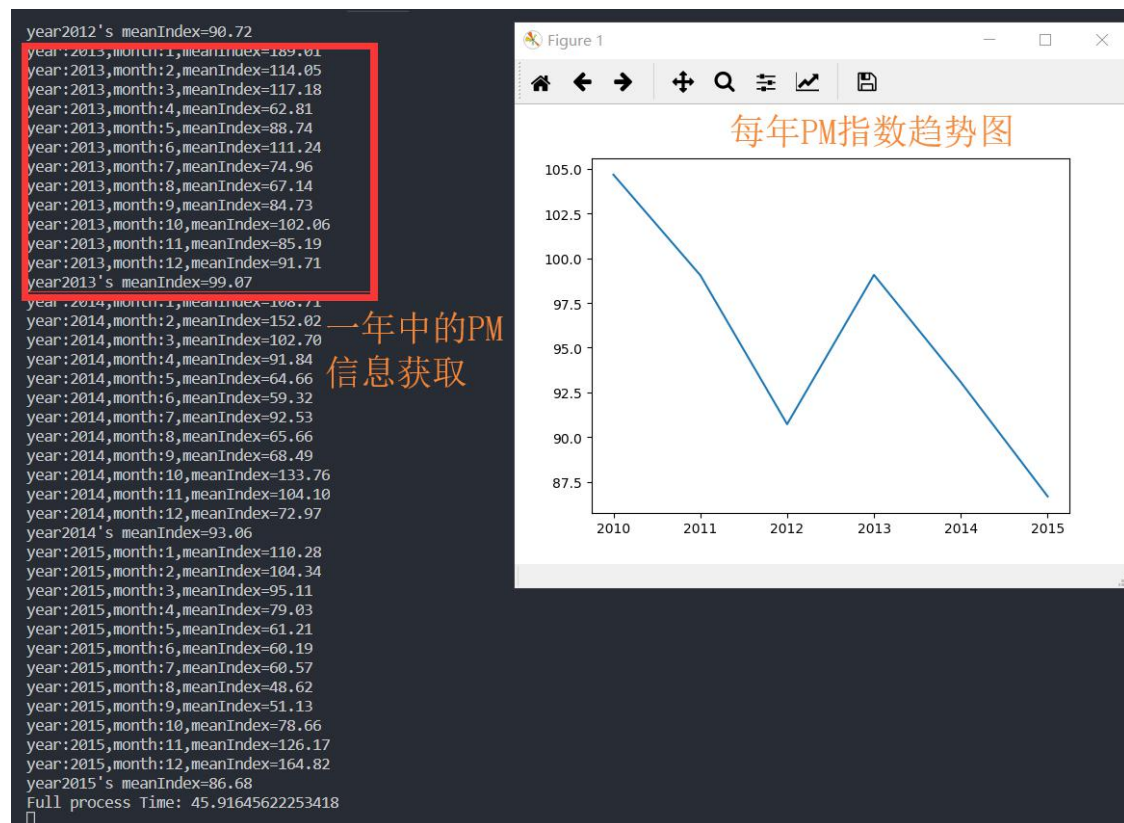
```
list_year = []
```

```

for i in range(6):
    sumt=0
    for j in range(1,13):
        print("year:{},month:{},meanIndex={:.2f}".format(i+2010,j,account[i][j]/counters[i][j]))
        sumt+=account[i][j]/counters[i][j]
    print("year{}'s meanIndex={:.2f}".format(i+2010,sumt/12))
    list_year.append(round(sumt,2)/12)
print("Full process Time:",time.time()-start)
year=[2010,2011,2012,2013,2014,2015]
#调用 plt.plot 来画图,横轴纵轴两个参数即可
plt.plot(year,list_year)
plt.title("Year Trend Graph")
# 用 show 展现出来图
plt.show()

```

运行结果:



info=====

=====

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 52584 entries, 0 to 52583

Data columns (total 18 columns):

No 52584 non-null object

year 52584 non-null object

```
month          52584 non-null object
day            52584 non-null object
hour           52584 non-null object
season         52584 non-null object
PM_Dongsi      25052 non-null object
PM_Dongsihuan 20508 non-null object
PM_Nongzhanguan 24931 non-null object
PM_US Post     50387 non-null object
DEWP           52579 non-null object
HUMI           52245 non-null object
PRES           52245 non-null object
TEMP           52579 non-null object
cbwd           52579 non-null object
lws            52579 non-null object
precipitation  52100 non-null object
lprec          52100 non-null object
dtypes: object(18)
memory usage: 7.2+ MB
None
=====
year:2010,month:1,meanIndex=90.40
year:2010,month:2,meanIndex=97.24
year:2010,month:3,meanIndex=94.05
year:2010,month:4,meanIndex=80.07
year:2010,month:5,meanIndex=87.07
year:2010,month:6,meanIndex=109.04
year:2010,month:7,meanIndex=123.43
year:2010,month:8,meanIndex=97.68
year:2010,month:9,meanIndex=122.79
year:2010,month:10,meanIndex=118.78
year:2010,month:11,meanIndex=138.38
year:2010,month:12,meanIndex=97.12
year2010's meanIndex=104.67
year:2011,month:1,meanIndex=44.87
year:2011,month:2,meanIndex=150.29
year:2011,month:3,meanIndex=57.99
year:2011,month:4,meanIndex=91.72
year:2011,month:5,meanIndex=65.11
year:2011,month:6,meanIndex=108.79
year:2011,month:7,meanIndex=107.39
year:2011,month:8,meanIndex=103.73
year:2011,month:9,meanIndex=94.97
year:2011,month:10,meanIndex=145.56
year:2011,month:11,meanIndex=109.43
```

year:2011,month:12,meanIndex=108.72
year2011's meanIndex=99.05
year:2012,month:1,meanIndex=118.92
year:2012,month:2,meanIndex=84.44
year:2012,month:3,meanIndex=96.47
year:2012,month:4,meanIndex=87.84
year:2012,month:5,meanIndex=90.97
year:2012,month:6,meanIndex=96.63
year:2012,month:7,meanIndex=80.65
year:2012,month:8,meanIndex=81.17
year:2012,month:9,meanIndex=59.95
year:2012,month:10,meanIndex=94.95
year:2012,month:11,meanIndex=87.44
year:2012,month:12,meanIndex=109.19
year2012's meanIndex=90.72
year:2013,month:1,meanIndex=189.01
year:2013,month:2,meanIndex=114.05
year:2013,month:3,meanIndex=117.18
year:2013,month:4,meanIndex=62.81
year:2013,month:5,meanIndex=88.74
year:2013,month:6,meanIndex=111.24
year:2013,month:7,meanIndex=74.96
year:2013,month:8,meanIndex=67.14
year:2013,month:9,meanIndex=84.73
year:2013,month:10,meanIndex=102.06
year:2013,month:11,meanIndex=85.19
year:2013,month:12,meanIndex=91.71
year2013's meanIndex=99.07
year:2014,month:1,meanIndex=108.71
year:2014,month:2,meanIndex=152.02
year:2014,month:3,meanIndex=102.70
year:2014,month:4,meanIndex=91.84
year:2014,month:5,meanIndex=64.66
year:2014,month:6,meanIndex=59.32
year:2014,month:7,meanIndex=92.53
year:2014,month:8,meanIndex=65.66
year:2014,month:9,meanIndex=68.49
year:2014,month:10,meanIndex=133.76
year:2014,month:11,meanIndex=104.10
year:2014,month:12,meanIndex=72.97
year2014's meanIndex=93.06
year:2015,month:1,meanIndex=110.28
year:2015,month:2,meanIndex=104.34
year:2015,month:3,meanIndex=95.11

year:2015,month:4,meanIndex=79.03
year:2015,month:5,meanIndex=61.21
year:2015,month:6,meanIndex=60.19
year:2015,month:7,meanIndex=60.57
year:2015,month:8,meanIndex=48.62
year:2015,month:9,meanIndex=51.13
year:2015,month:10,meanIndex=78.66
year:2015,month:11,meanIndex=126.17
year:2015,month:12,meanIndex=164.82
year2015's meanIndex=86.68
Full process Time: 45.91645622253418