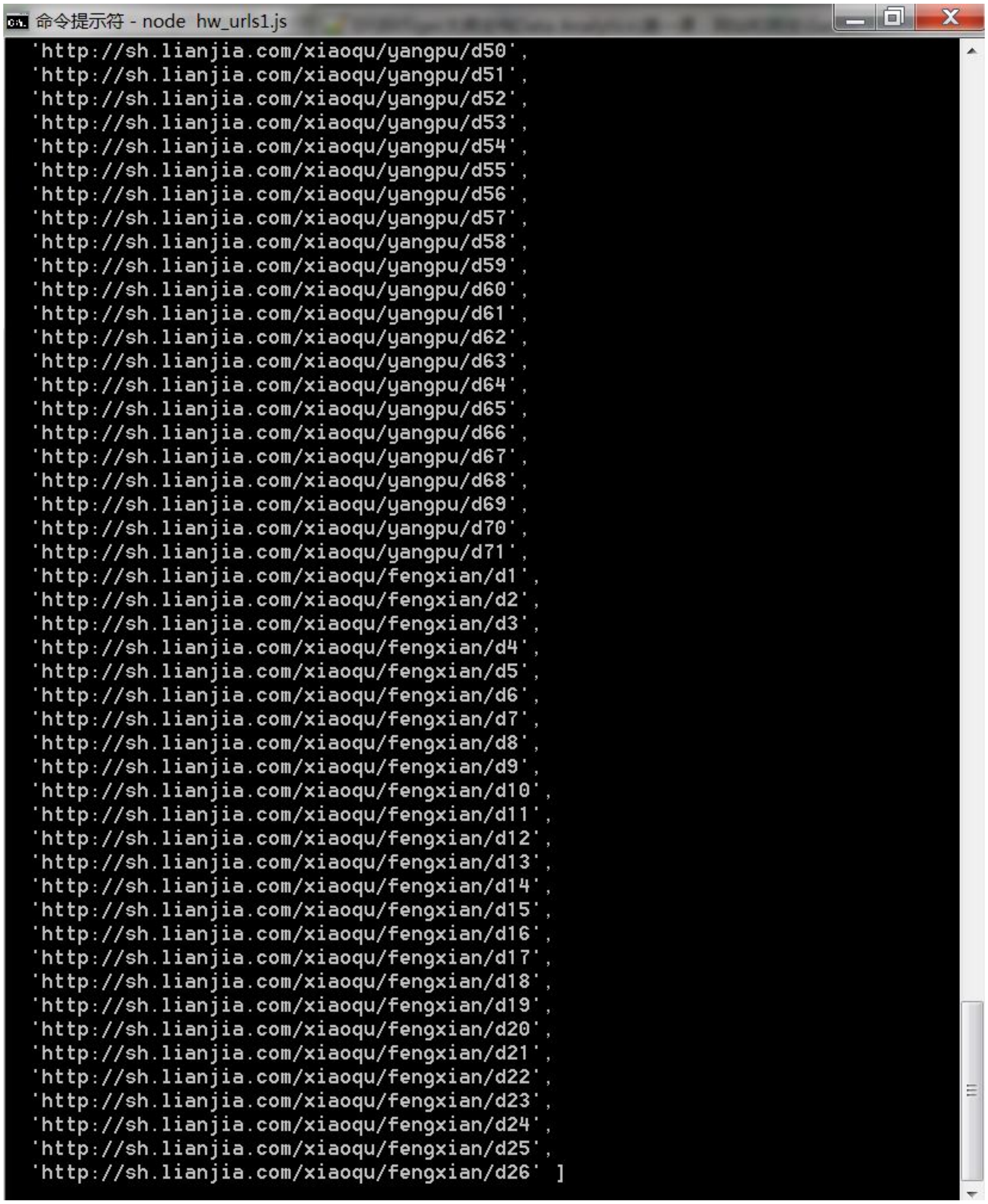
Web Crawler

1. 用nodeJS爬虫爬取列表页

    - 在主入口`hw_exe_v1`中调用url模块`hw_urls1`、请求池模块`hw_pool`、mongoDB交互模块`mongo`

    - url模块用来获取每个区的所有小区列表页，用`$('#filter-options').find('a').attr('href')`得到各区的url,进一步请求各区链接，抓取每个区第一页至最后页的小区列表信息，但每个区的总列表页数不同，所以可以在获取每个区的小区列表首页时解析总页数: `var page_num = $('.page-box').find('a:nth-last-child(2)').text();`



```
'http://sh.lianjia.com/xiaoqu/yangpu/d50',
'http://sh.lianjia.com/xiaoqu/yangpu/d51',
'http://sh.lianjia.com/xiaoqu/yangpu/d52',
'http://sh.lianjia.com/xiaoqu/yangpu/d53',
'http://sh.lianjia.com/xiaoqu/yangpu/d54',
'http://sh.lianjia.com/xiaoqu/yangpu/d55',
'http://sh.lianjia.com/xiaoqu/yangpu/d56',
'http://sh.lianjia.com/xiaoqu/yangpu/d57',
'http://sh.lianjia.com/xiaoqu/yangpu/d58',
'http://sh.lianjia.com/xiaoqu/yangpu/d59',
'http://sh.lianjia.com/xiaoqu/yangpu/d60',
'http://sh.lianjia.com/xiaoqu/yangpu/d61',
'http://sh.lianjia.com/xiaoqu/yangpu/d62',
'http://sh.lianjia.com/xiaoqu/yangpu/d63',
'http://sh.lianjia.com/xiaoqu/yangpu/d64',
'http://sh.lianjia.com/xiaoqu/yangpu/d65',
'http://sh.lianjia.com/xiaoqu/yangpu/d66',
'http://sh.lianjia.com/xiaoqu/yangpu/d67',
'http://sh.lianjia.com/xiaoqu/yangpu/d68',
'http://sh.lianjia.com/xiaoqu/yangpu/d69',
'http://sh.lianjia.com/xiaoqu/yangpu/d70',
'http://sh.lianjia.com/xiaoqu/yangpu/d71',
'http://sh.lianjia.com/xiaoqu/fengxian/d1',
'http://sh.lianjia.com/xiaoqu/fengxian/d2',
'http://sh.lianjia.com/xiaoqu/fengxian/d3',
'http://sh.lianjia.com/xiaoqu/fengxian/d4',
'http://sh.lianjia.com/xiaoqu/fengxian/d5',
'http://sh.lianjia.com/xiaoqu/fengxian/d6',
'http://sh.lianjia.com/xiaoqu/fengxian/d7',
'http://sh.lianjia.com/xiaoqu/fengxian/d8',
'http://sh.lianjia.com/xiaoqu/fengxian/d9',
'http://sh.lianjia.com/xiaoqu/fengxian/d10',
'http://sh.lianjia.com/xiaoqu/fengxian/d11',
'http://sh.lianjia.com/xiaoqu/fengxian/d12',
'http://sh.lianjia.com/xiaoqu/fengxian/d13',
'http://sh.lianjia.com/xiaoqu/fengxian/d14',
'http://sh.lianjia.com/xiaoqu/fengxian/d15',
'http://sh.lianjia.com/xiaoqu/fengxian/d16',
'http://sh.lianjia.com/xiaoqu/fengxian/d17',
'http://sh.lianjia.com/xiaoqu/fengxian/d18',
'http://sh.lianjia.com/xiaoqu/fengxian/d19',
'http://sh.lianjia.com/xiaoqu/fengxian/d20',
'http://sh.lianjia.com/xiaoqu/fengxian/d21',
'http://sh.lianjia.com/xiaoqu/fengxian/d22',
'http://sh.lianjia.com/xiaoqu/fengxian/d23',
'http://sh.lianjia.com/xiaoqu/fengxian/d24',
'http://sh.lianjia.com/xiaoqu/fengxian/d25',
'http://sh.lianjia.com/xiaoqu/fengxian/d26' ]
```

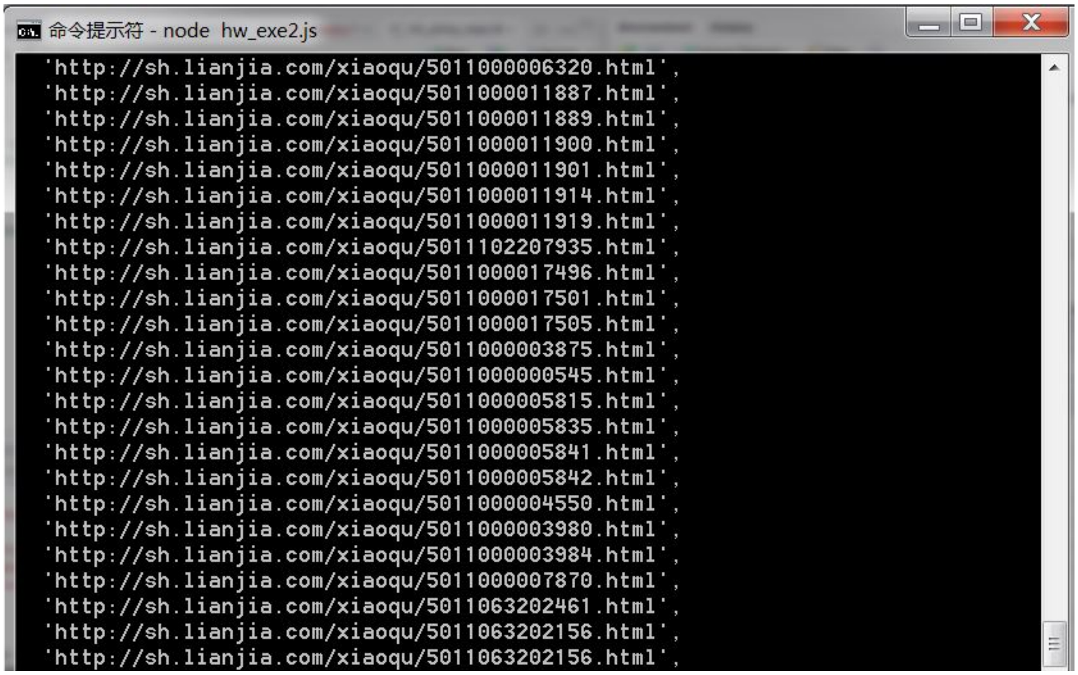
    - 请求池中调用解析模块`hw_parser`解析url。
      xiaoqu:`$('.list-wrap').find('li').find('.actshowMap_list').attr('xiaoqu').replace(/\'/g, '"')`; lat:`JSON.parse(xiaoqu)[1]`; lng: `JSON.parse(xiaoqu)[0]`; communityName: `JSON.parse(xiaoqu)[2]`; districtName: `$('.list-wrap').find('li').find('.actshowMap_list').attr('districtname')`; plateName:`$('.list-wrap').find('li').find('.actshowMap_list').attr('platename')`; communityId:`$('.list-wrap').find('li').find('.pic-panel').find('a').attr('key')`; price: `$('.list-wrap').find('li').find('.price').find('.num').text()`; age: `2016 - $('.list-wrap').find('li').find('.con').text().match(/\d\d\d\d/g)`; 最后需要使解析后数据类型与自定义的schema保持一致。

2. 用nodeJS爬虫爬取详情页

    - 在主入口`hw_exe_v2`中调用url模块`hw_urls3`、请求池模块`hw_pool2`、mongoDB交互模块`mongo2`

    - url模块用来获取每个区的所有小区详情页，用`$('#filter-options').find('a').attr('href')`得到各区的url,进一步请求各区链接，抓取每个区第

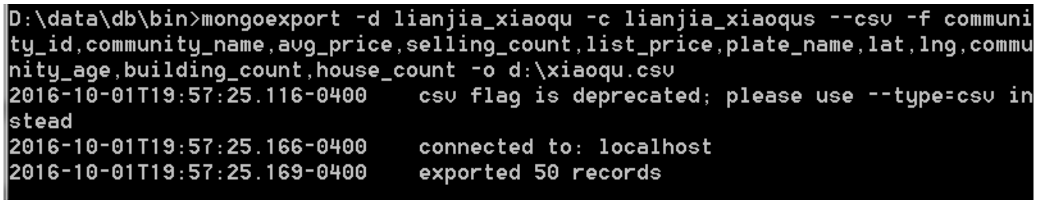
一页至最后页的小区列表信息，并在获取每个区的小区列表首页时解析总页数: `var page_num = $('.page-box').find('a:nth-last-child(2)').text();` 再基于这些列表页url用`$('.info-panel').find('h2').find('a').attr('href')` 获取每个小区的url。



```
 'http://sh.lianjia.com/xiaoqu/5011000006320.html',
 'http://sh.lianjia.com/xiaoqu/5011000011887.html',
 'http://sh.lianjia.com/xiaoqu/5011000011889.html',
 'http://sh.lianjia.com/xiaoqu/5011000011900.html',
 'http://sh.lianjia.com/xiaoqu/5011000011901.html',
 'http://sh.lianjia.com/xiaoqu/5011000011914.html',
 'http://sh.lianjia.com/xiaoqu/5011000011919.html',
 'http://sh.lianjia.com/xiaoqu/5011102207935.html',
 'http://sh.lianjia.com/xiaoqu/5011000017496.html',
 'http://sh.lianjia.com/xiaoqu/5011000017501.html',
 'http://sh.lianjia.com/xiaoqu/5011000017505.html',
 'http://sh.lianjia.com/xiaoqu/5011000003875.html',
 'http://sh.lianjia.com/xiaoqu/5011000000545.html',
 'http://sh.lianjia.com/xiaoqu/5011000005815.html',
 'http://sh.lianjia.com/xiaoqu/5011000005835.html',
 'http://sh.lianjia.com/xiaoqu/5011000005841.html',
 'http://sh.lianjia.com/xiaoqu/5011000005842.html',
 'http://sh.lianjia.com/xiaoqu/5011000004550.html',
 'http://sh.lianjia.com/xiaoqu/5011000003980.html',
 'http://sh.lianjia.com/xiaoqu/5011000003984.html',
 'http://sh.lianjia.com/xiaoqu/5011000007870.html',
 'http://sh.lianjia.com/xiaoqu/5011063202461.html',
 'http://sh.lianjia.com/xiaoqu/5011063202156.html',
 'http://sh.lianjia.com/xiaoqu/5011063202156.html',
```

- 请求池中调用解析模块`hw_parser2`解析url。
  info:`$('.wrapper').find('.detail-block').find('.actshowMap').attr('xiaoqu').replace(/\'/g, '"')`; lat:`JSON.parse(info)[1]`; lng: `JSON.parse(info)[0]`; communityName: `JSON.parse(info)[2]`; list*price: $('.wrapper').find('.detail-block').find('.priceInfo').find('div:first-child').find('p').text().replace(/\W/g,"");*
  *avg*price:`$('.wrapper').find('.detail-block').find('.priceInfo').find('div:nth-child(3)').find('p').text().replace(/\W/g,"")`;
  building_count:`$('.wrapper').find('.detail-block').find('.res-info').find('li:nth-child(6)').find('.other').text().replace("栋","")`;
  house*count: $('.wrapper').find('.detail-block').find('.res-info').find('li:nth-child(7)').find('.other').text().replace("户","");*
  *selling*count: `$('.wrapper').find('.detail-block').find('.js_outLink').text().replace(/\W/g, "")`; communityId: `$('.wrapper').find('.detail-block').find('#notice_focus').attr('propertyno')`; plate: `$('.wrapper').find('.detail-block').find('.res-top').find('span:nth-child(2)').text().replace("(",'').replace(")",'')` ; 最后需要使解析后数据类型与自定义的schema保持一致。

3. 数据导出工具：mongoexport



```
D:\data\db\bin>mongoexport -d lianjia_xiaoqu -c lianjia_xiaoqus --csv -f communi
ty_id,community_name,avg_price,selling_count,list_price,plate_name,lat,lng,commu
nity_age,building_count,house_count -o d:\xiaoqu.csv
2016-10-01T19:57:25.116-0400    csv flag is deprecated; please use --type=csv in
stead
2016-10-01T19:57:25.166-0400    connected to: localhost
2016-10-01T19:57:25.169-0400    exported 50 records
```

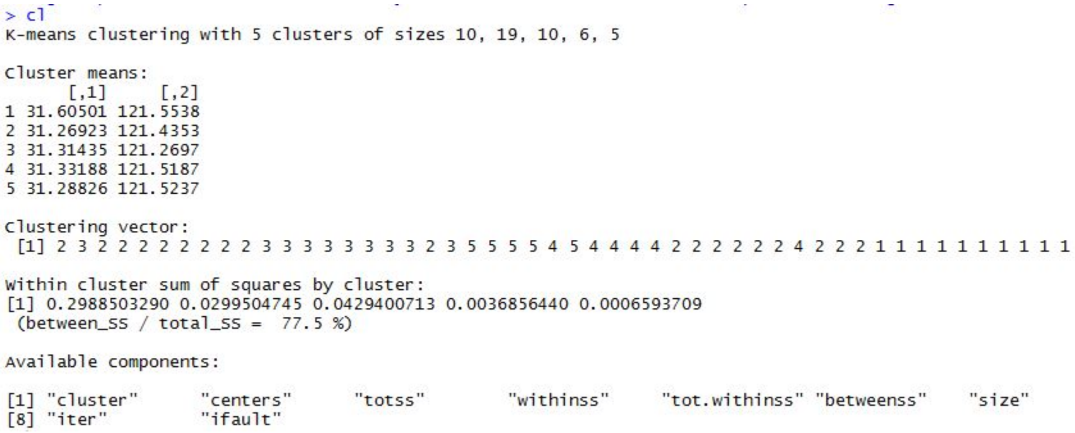
- 列表页导出的部分结果为communities.csv
- 详情页导出的部分结果为xiaoqu.csv

Data Analysis

Data Set: xiaoqu.csv

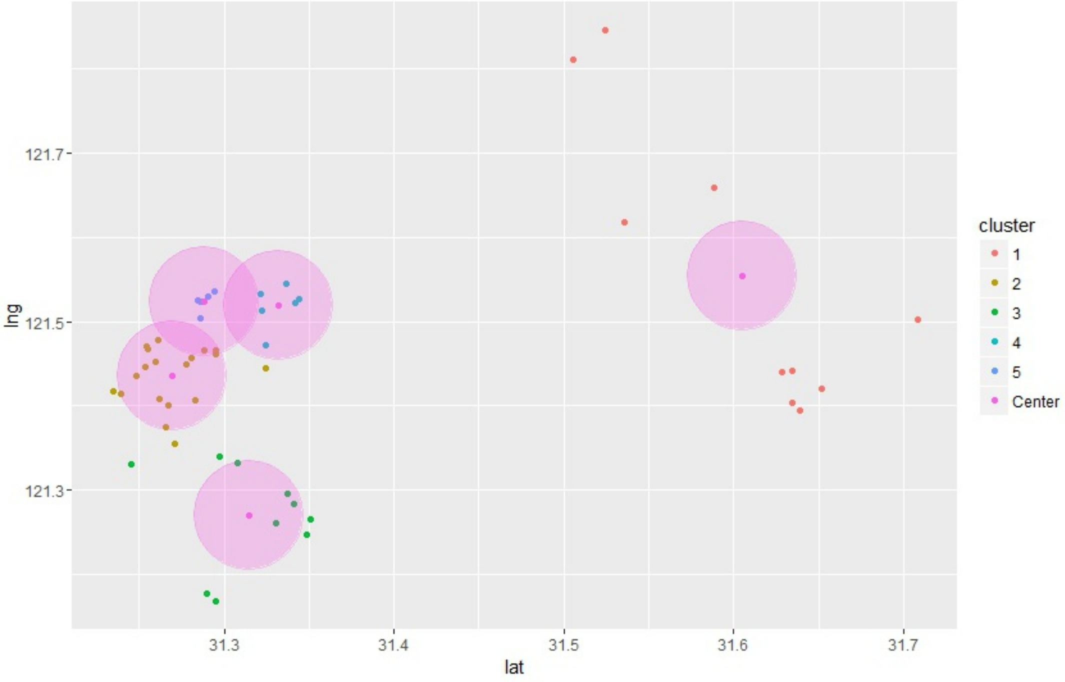
1. K-means地理聚类：

- 小区的经纬度作为地理聚类的feature `m=as.matrix(cbind(df$lat,df$lng),ncol=2)`
- 已知该数据集中的小区在五个不同区，所以可以直接用K-Means Clustering求解,`cl=(kmeans(m,5))` , 结果如下：

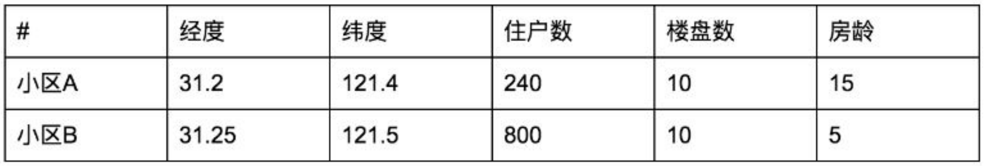


```
> cl
K-means clustering with 5 clusters of sizes 10, 19, 10, 6, 5

Cluster means:
      [,1]     [,2]
1 31.60501 121.5538
2 31.26923 121.4353
3 31.31435 121.2697
4 31.33188 121.5187
5 31.28826 121.5237

Clustering vector:
 [1] 2 3 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 2 3 5 5 5 5 4 5 4 4 4 4 2 2 2 2 2 2 4 2 2 2 1 1 1 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 0.2988503290 0.0299504745 0.0429400713 0.0036856440 0.0006593709
 (between_SS / total_SS =  77.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
```

- 用ggplot2绘图

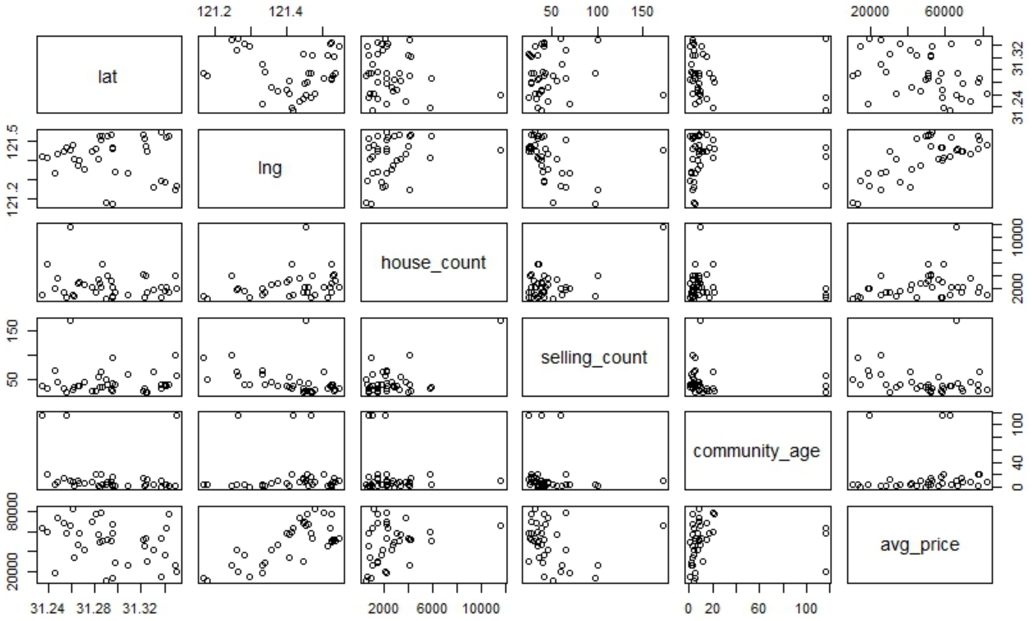


2. 对如下小区预测均价：

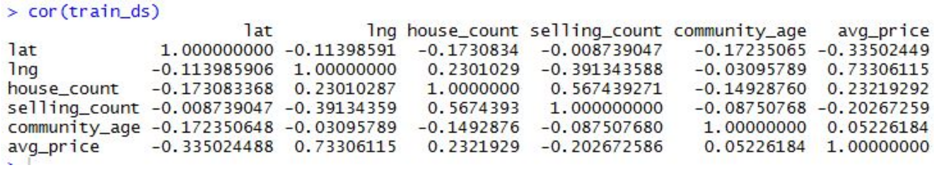


#	经度	纬度	住户数	楼盘数	房龄
小区A	31.2	121.4	240	10	15
小区B	31.25	121.5	800	10	5

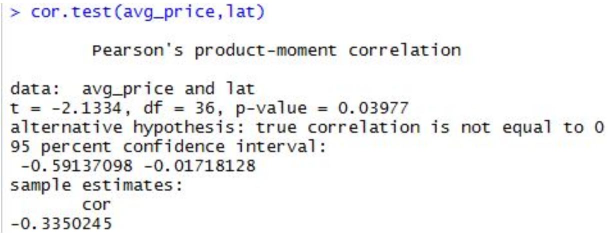
- 首先， 由于数据包含多个feature，所以我们可以利用多元线性回归模型确定每个feature对于均价的相关性。绘制所有关系的散点图：



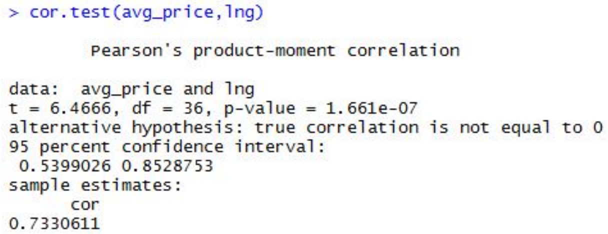
- 查看相关矩阵，做相关分析，研究`lat`、`lng`、`house_count`、`selling_count`、`community_age`与`avg_price`的相关性。



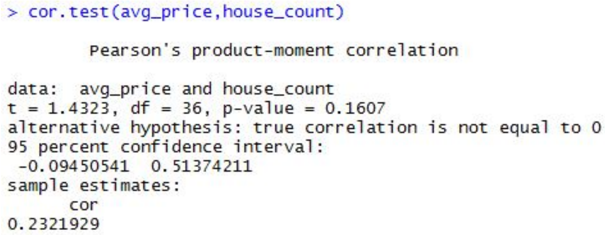
```
> cor(train_ds)
                       lat         lng house_count selling_count community_age   avg_price
lat            1.000000000 -0.11398591  -0.1730834  -0.008739047   -0.17235065 -0.33502449
lng           -0.113985906  1.00000000   0.2301029  -0.391343588   -0.03095789  0.73306115
house_count   -0.173083368  0.23010287   1.0000000   0.567439271   -0.14928760  0.23219292
selling_count -0.008739047 -0.39134359   0.5674393   1.000000000   -0.08750768 -0.20267259
community_age -0.172350648 -0.03095789  -0.1492876  -0.087507680    1.00000000  0.05226184
avg_price     -0.335024488  0.73306115   0.2321929  -0.202672586    0.05226184  1.00000000
```



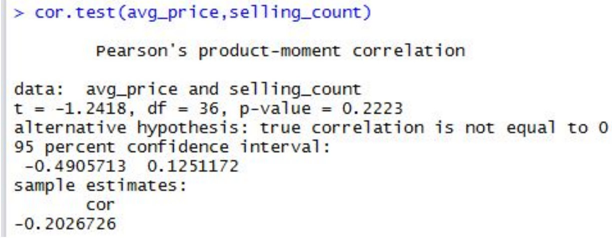
```
> cor.test(avg_price,lat)

	Pearson's product-moment correlation

data:  avg_price and lat
t = -2.1334, df = 36, p-value = 0.03977
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.59137098 -0.01718128
sample estimates:
       cor 
-0.3350245 
```



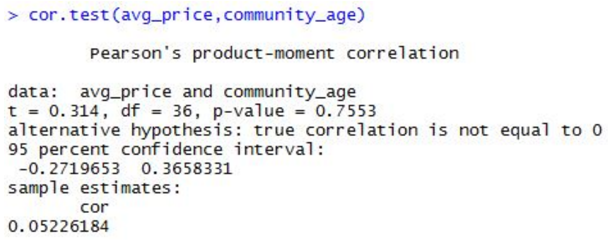
```
> cor.test(avg_price,lng)

	Pearson's product-moment correlation

data:  avg_price and lng
t = 6.4666, df = 36, p-value = 1.661e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5399026 0.8528753
sample estimates:
      cor 
0.7330611 
```



```
> cor.test(avg_price,house_count)

	Pearson's product-moment correlation

data:  avg_price and house_count
t = 1.4323, df = 36, p-value = 0.1607
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.09450541  0.51374211
sample estimates:
      cor 
0.2321929 
```

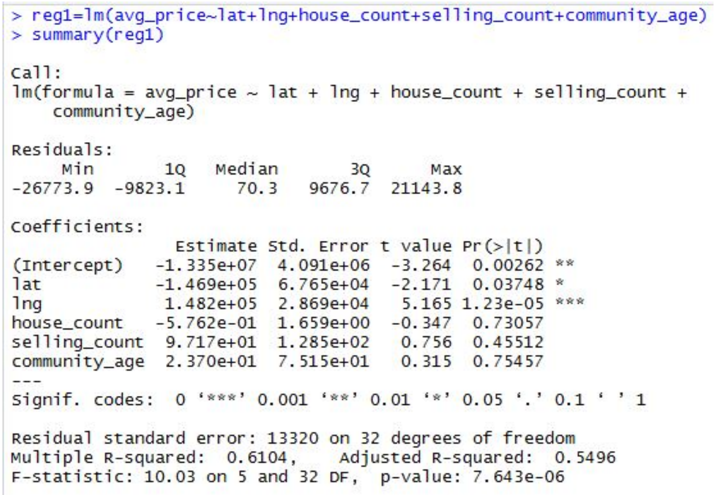


```
> cor.test(avg_price,selling_count)

	Pearson's product-moment correlation

data:  avg_price and selling_count
t = -1.2418, df = 36, p-value = 0.2223
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4905713  0.1251172
sample estimates:
       cor 
-0.2026726 
```



```
> cor.test(avg_price,community_age)

	Pearson's product-moment correlation

data:  avg_price and community_age
t = 0.314, df = 36, p-value = 0.7553
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2719653  0.3658331
sample estimates:
       cor 
0.05226184 
```

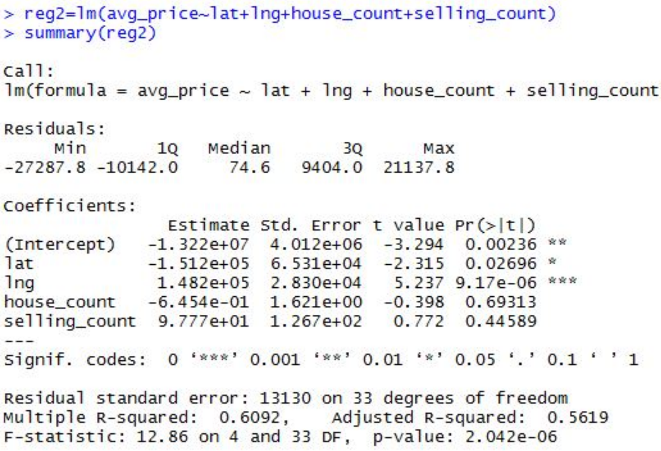
- 对五个变量建立多元线性回归方程



```
> reg1=lm(avg_price~lat+lng+house_count+selling_count+community_age)
> summary(reg1)

Call:
lm(formula = avg_price ~ lat + lng + house_count + selling_count + 
    community_age)

Residuals:
     Min       1Q   Median       3Q      Max 
-26773.9  -9823.1     70.3   9676.7  21143.8 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.335e+07  4.091e+06  -3.264  0.00262 ** 
lat           -1.469e+05  6.765e+04  -2.171  0.03748 *  
lng            1.482e+05  2.869e+04   5.165 1.23e-05 ***
house_count   -5.762e-01  1.659e+00  -0.347  0.73057    
selling_count  9.717e+01  1.285e+02   0.756  0.45512    
community_age  2.370e+01  7.515e+01   0.315  0.75457    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13320 on 32 degrees of freedom
Multiple R-squared:  0.6104,	Adjusted R-squared:  0.5496 
F-statistic: 10.03 on 5 and 32 DF,  p-value: 7.643e-06
```

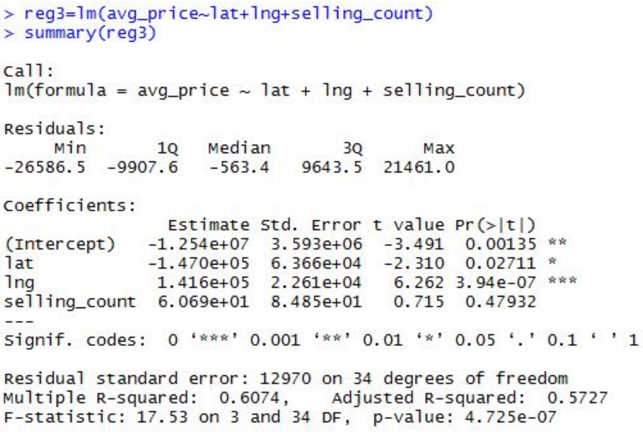
!

- 去掉一个Pr远超0.05的变量，优化p-value



```
> reg2=lm(avg_price~lat+lng+house_count+selling_count)
> summary(reg2)

Call:
lm(formula = avg_price ~ lat + lng + house_count + selling_count)

Residuals:
     Min       1Q   Median       3Q      Max 
-27287.8 -10142.0     74.6   9404.0  21137.8 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.322e+07  4.012e+06  -3.294  0.00236 ** 
lat           -1.512e+05  6.531e+04  -2.315  0.02696 *  
lng            1.482e+05  2.830e+04   5.237 9.17e-06 ***
house_count   -6.454e-01  1.621e+00  -0.398  0.69313    
selling_count  9.777e+01  1.267e+02   0.772  0.44589    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13130 on 33 degrees of freedom
Multiple R-squared:  0.6092,	Adjusted R-squared:  0.5619 
F-statistic: 12.86 on 4 and 33 DF,  p-value: 2.042e-06
```

- 再去掉一个Pr远超0.05的变量，优化p-value



```
> reg3=lm(avg_price~lat+lng+selling_count)
> summary(reg3)

Call:
lm(formula = avg_price ~ lat + lng + selling_count)

Residuals:
     Min       1Q   Median       3Q      Max 
-26586.5  -9907.6   -563.4   9643.5  21461.0 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.254e+07  3.593e+06  -3.491  0.00135 ** 
lat           -1.470e+05  6.366e+04  -2.310  0.02711 *  
lng            1.416e+05  2.261e+04   6.262 3.94e-07 ***
selling_count  6.069e+01  8.485e+01   0.715  0.47932    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12970 on 34 degrees of freedom
Multiple R-squared:  0.6074,	Adjusted R-squared:  0.5727 
F-statistic: 17.53 on 3 and 34 DF,  p-value: 4.725e-07
```

- 模型结果： `y2_rs=a0+a1*lat+a2*lng+a4*selling_count` where

a0	-12542954.3562803
a1	-147027.118095122
a2	141582.754857238
a4	60.6907050449842

error:0.175894

- 预测结果：

`# community A:58552.91`

y2_A = a0+a1*31.2+a2*121.4+a4*10

`# community B: 65359.83`

y2_B = a0+a1*31.25+a2*121.5+a4*10

Data Visualization

Data Set: house_lianjia.json

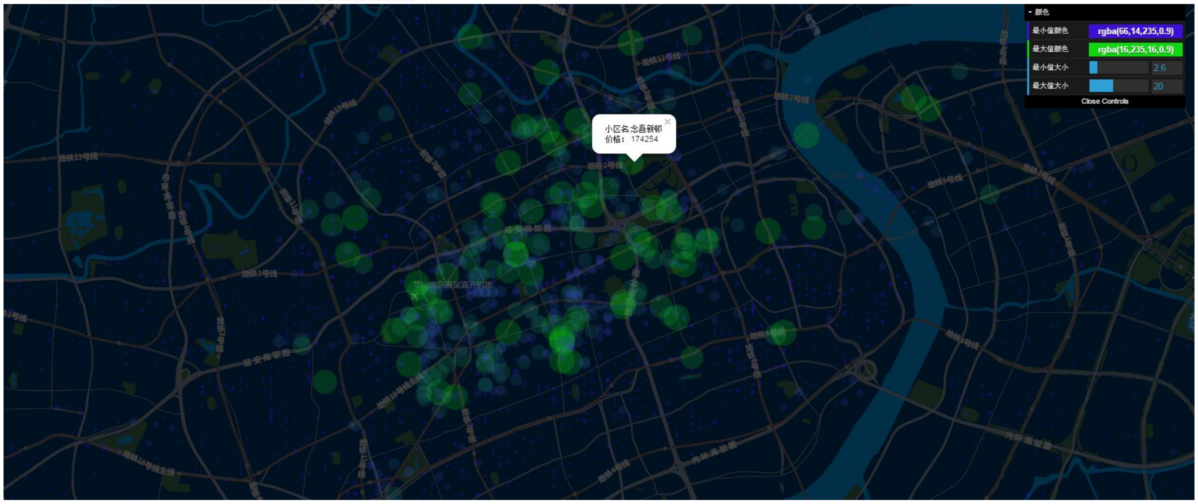
- 数据预处理：用postgreSQL清理数据，获取经纬度，均价，户数，小区名。运行如下sql语句：

```
copy
(SELECT array_to_json(array_agg(row_to_json(t))) FROM
(select lat,lng,avr_price, house_count, community_name from lianjia_data limit 10000) t)
TO 'D:/house_lianjia1.json';
```

- 返回结果如下：

```
[{"lat":"31.1418","lng":"121.58","avr_price":"47186","house_count":"670","community_name":"中邦大都会"},
{"lat":"30.8945","lng":"121.02","avr_price":"8910","house_count":"535","community_name":"中冶枫郡苑"},
{"lat":"31.2301","lng":"121.337","avr_price":"31732","house_count":"1268","community_name":"虹桥1号"},...]
```

- 运行`leaflet_dot_color_control_communities.html`，得到如图可视化效果。



小区名:念吾新邨
价格： 174254

▾ 颜色
最小值颜色 rgba(66,14,235,0.9)
最大值颜色 rgba(16,235,16,0.9)
最小值大小 2.6
最大值大小 20
Close Controls