|  |  |
| --- | --- |
| **Python分析广州租房情况数据** | 爬虫技术需求:   * 请求库/框架:scrapy * 数据库:mongodb * HTML解析:css * 数据可视化:pyecharts * 词云:wordcloud     **作者: 马耿滨 张泽坚** |

目录

[前言 2](#_Toc534280936)

[系统架构 2](#_Toc534280937)

[数据源 2](#_Toc534280938)

[数据清洗 5](#_Toc534280939)

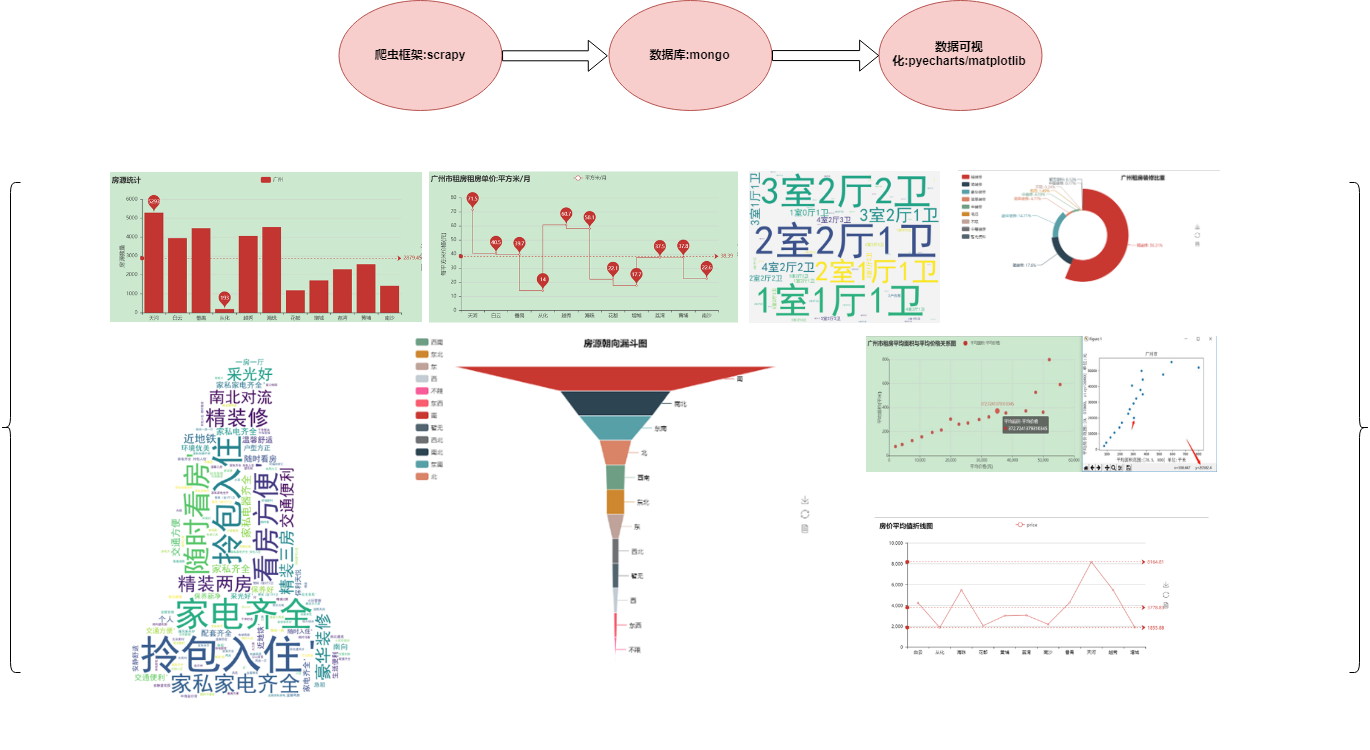
[分析 7](#_Toc534280940)

[总结 14](#_Toc534280941)

# 前言

临近实习的到来，可能有更多的人需要了解到广州各个地区的租房信息。故借此机会，尝试对房天下网站上广州租房的信息进行分析，了解一下广州地区的租房情况。

# 系统架构

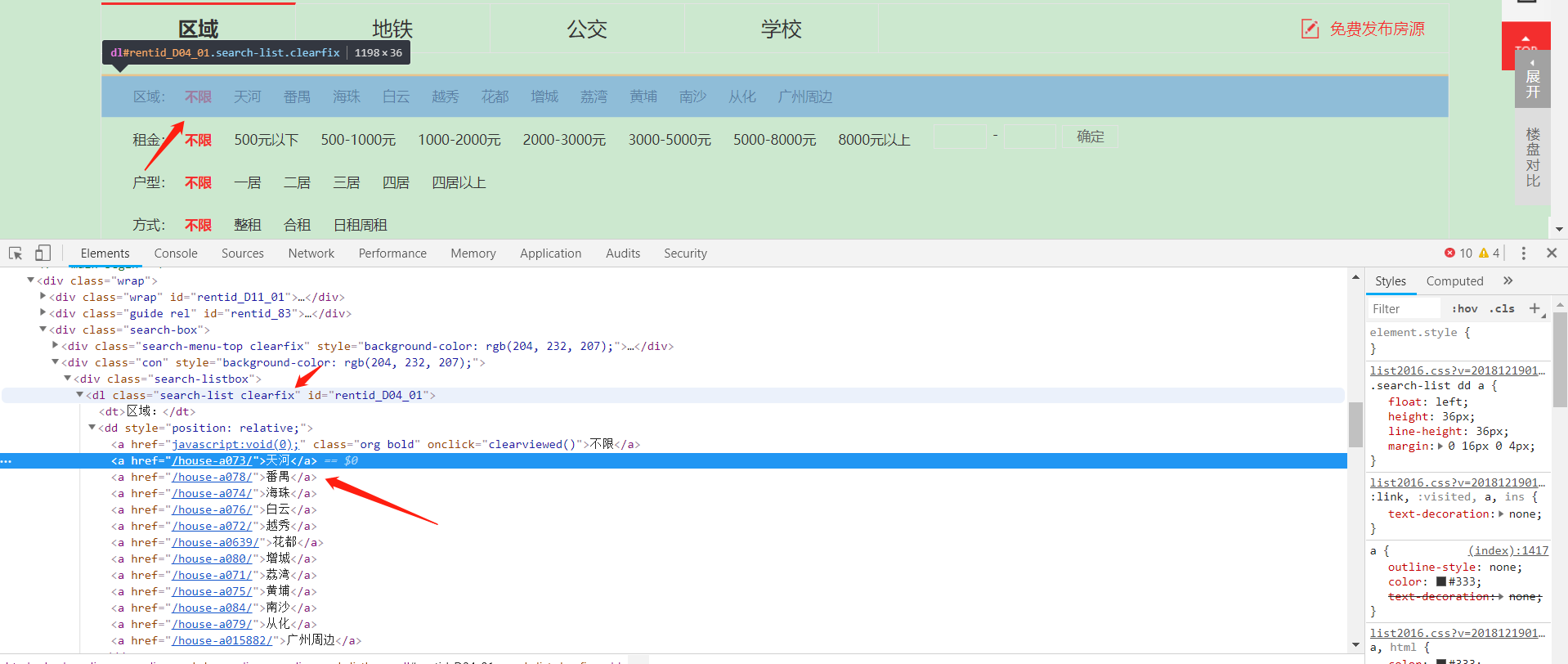


# 数据源

'http://gz.zu.fang.com'

获得广州十一个区的链接

拼接链接：URL = 'http: // gz.zu.fang.com /' + 'house-a073/'



获得详情页的链接



拼接链接：URL = 'http://gz.zu.fang.com' + 'chuzu/3\_233093393\_1.htm?channel=2,2'





分析实现翻页：

第一页： URL = 'http://gz.zu.fang.com/house-a073/'

第二页： URL = 'http://gz.zu.fang.com/house-a073/' + 'i32'

拼接翻页链接：URL = 'http://gz.zu.fang.com/house-a073/' + 'i3' + 'p' *for* p *in* range(1, 101):



## 数据清洗

格式转换前

C:\Users\Cheung\AppData\Local\Temp\WeChat Files\70f56860d50451e9ae645f9b9d59cf0.png

格式转换后

C:\Users\Cheung\AppData\Local\Temp\WeChat Files\fc1101bfc987b7d907371904d13d064.png

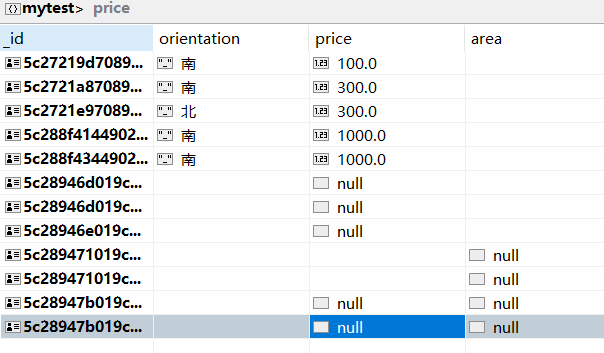
db.getCollection("house").find({"price":{"$type":2}}).forEach(function(x){

x.price = NumberInt(x.price);

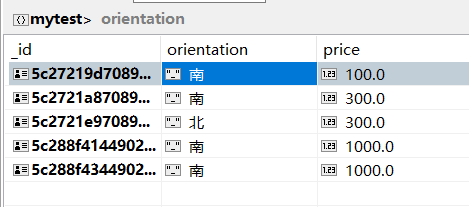
db.getCollection("house").save(x);

})

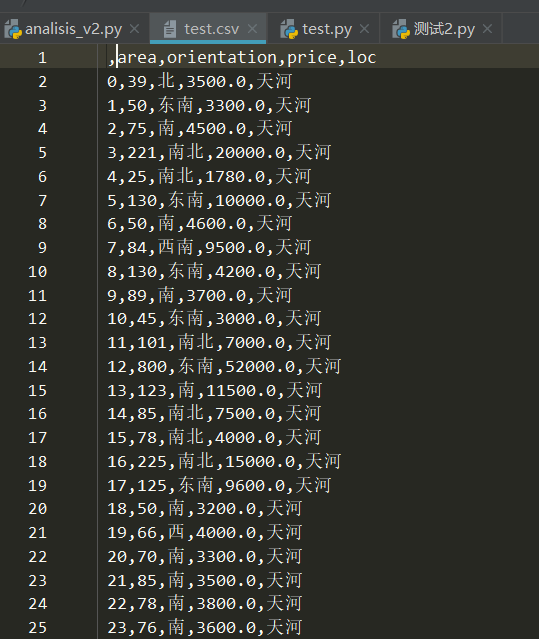
数据清洗前:



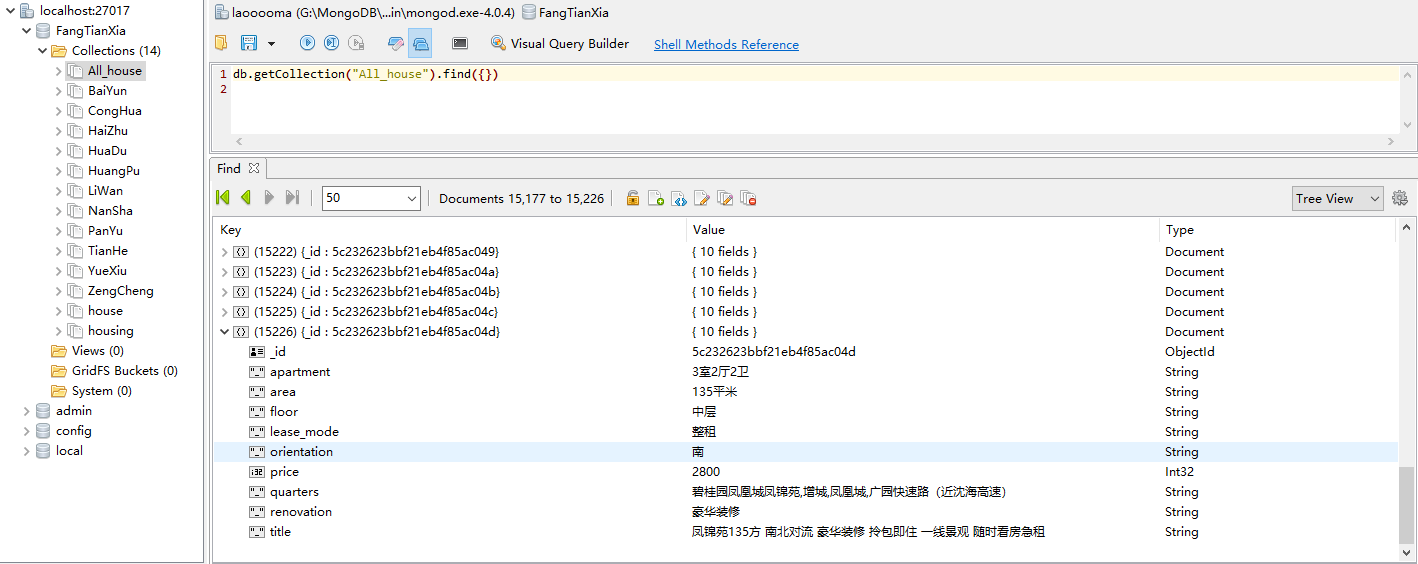
数据清洗后



导出为csv文件



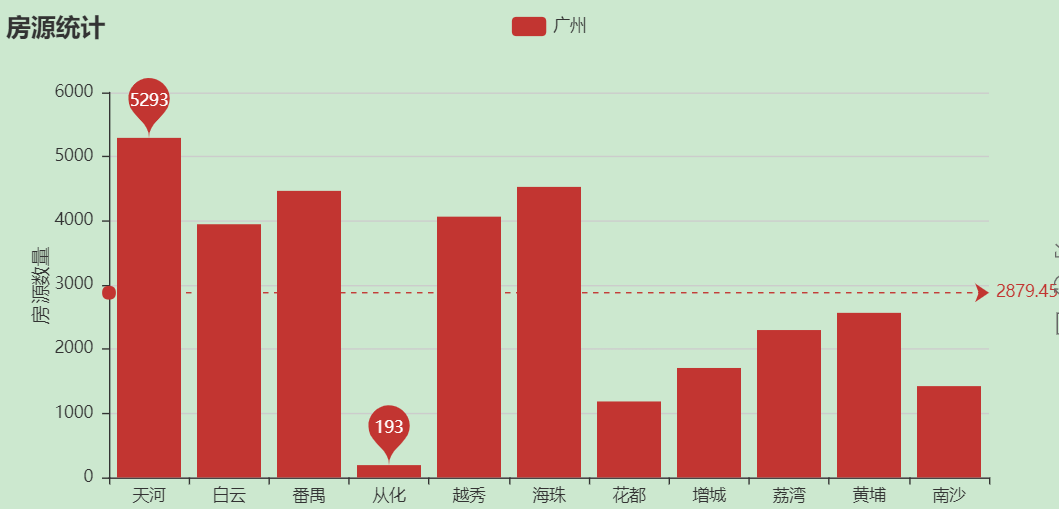
因为房天下的租房信息在不进行筛选的情况下，最多显示100页的数据，因此决定以不同区域作为筛选条件，通过python爬虫分别爬取广州市十一个区的数据，得到20000+条租房信息，对得到的无效信息进行清理，最终得到15226条有效数据。



# 分析

#### 房源数

通过分别对广州十一个区域的的数据进行爬取，得到每个区域的房源数如一下图：



能够清晰的看出，白云、天河、越秀、海珠这些相对发展的比较快的地方，房源数相对较高，而对于从化、花都这些离广州市区比较远的地区则房源数相对比较少一些。

#### 标题

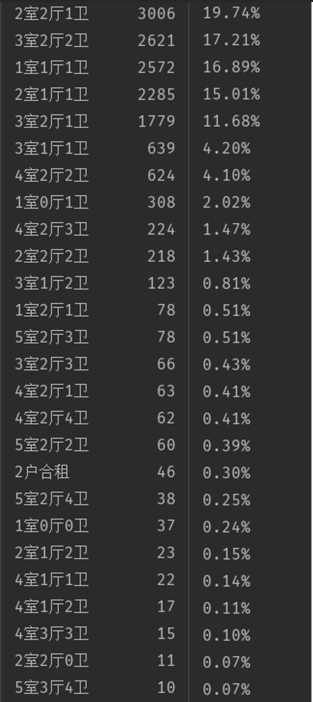
在房天下网站上查看房源时，每条房源信息中显示的就是房源的标题，房天下的房源标题比较长，类似于一个简介，因此我尝试对爬取到的房源标题进行一个统计，看看房天下网站在标题中使用频率最高词有哪些。



可以看出房天下的标题中，出现频率比较高的词语有家电齐全，拎包入住，随时看房等，另外交通便利，装修的出现频率也不低，可见房天下的推荐方式主要是通过简述房源信息，并说明看房便利以及家具齐全、交通便利来进行描述。

#### 户型占比

户型也是我们在租房时需要考虑的一环，而什么户型的房子最多这便成了一个问题，通过对房源户型进行统计后，结果如下（数据前25行）：



户型的种类很多，因此截取频率前二十五的户型进行计算，结果发现二室二厅一卫的房子较多，而像四室/五室/无厅无卫这一室户反而占比很低，因此大部分人需要进行双人合租或者独居才能解决自己的住房问题。   
通过一个词云图更容易了解其中的分布:

#### 

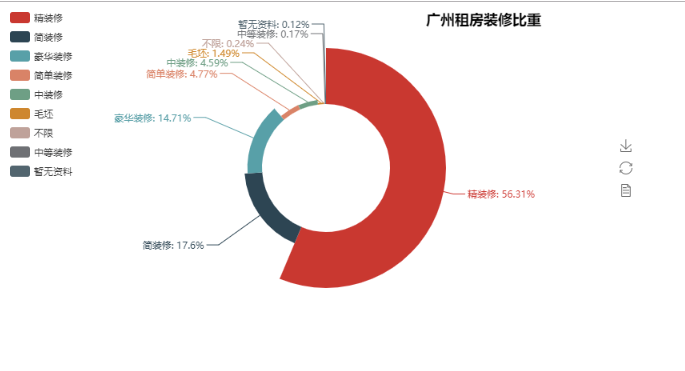
#### 装修

装修方式同样也是大部人在租房时会加以考虑的环节，不同的装修方式会带给用户不同的住房体验。而怎样的装修在广州最多便成为一个问题。通过对房源装修方式进行统计，结果如下：



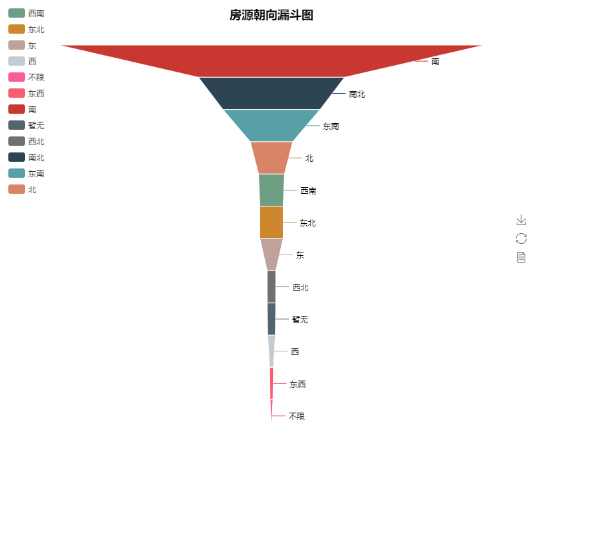
除去少部分房源信息，没有装修信息之外，由统计结果可以清晰直观的看出，广州的出租的房子绝大部分属于精装修，其次为简装修和豪华装修。

通过一个饼图更容易了解其中的分布:



#### 朝向

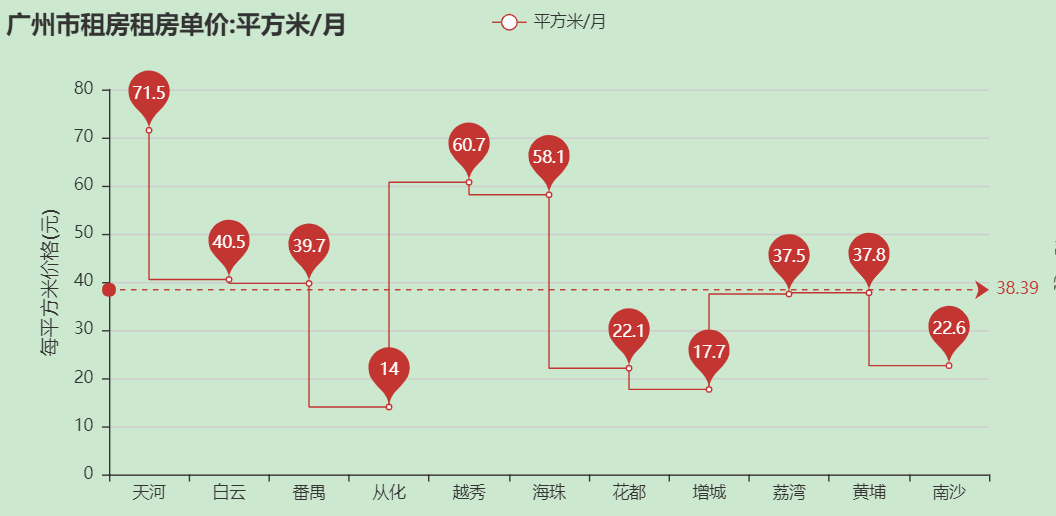
中国是一个很看中风水走向的国家，所有不乏有很多人对房子的朝向也有一定的要求，对此，我统计了广州一万多条房源信息，对广州租房的朝向进行分析得出：



结果显示，广州绝大部分房源都是朝向南面。早在原始社会，中国先民就按照坐北朝南的方向修建村落房屋，考古发现的绝大多数房屋都是大门朝南。而相对于朝南/东南这些朝向，西南/东北/东这些朝向相对会少一些。

#### 面积-价格关系

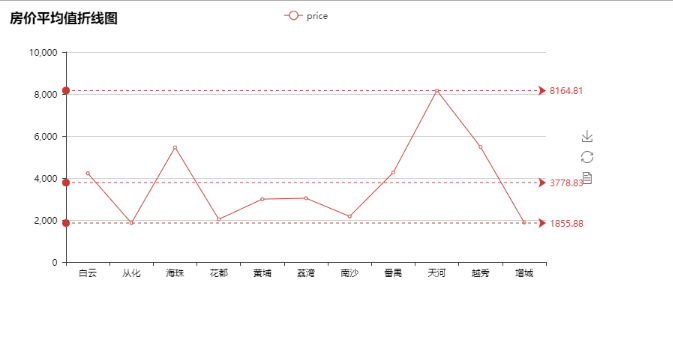
现在的市场，肯定更多的是希望花更少的钱能租到更好更大的房子，通过计算每平米的的月平均租金，来计算租房面积的性价比



从图中可得:广州市十一个区每平方米的房价平均是38.9元, 其中,每平方米房价前三高的区分别是天河区71.5元,越秀区60.7元, 海珠58.1元, 从每平方米单价可以看出, 广州市发展较快的是珠三角一带, 离广州市中心越远的租房单价就变得特别低, 其中, 从化每平方米单价仅仅14元

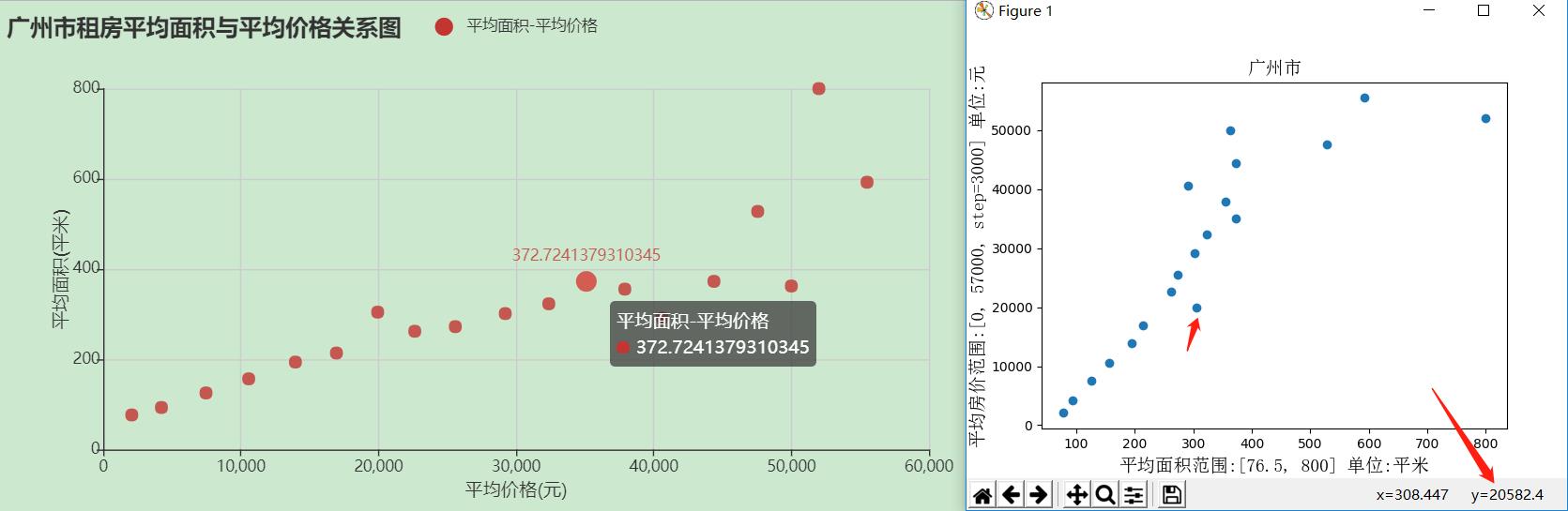
#### 房价均值

分别对广州每个区所有的房源月租价格进行求平均值：



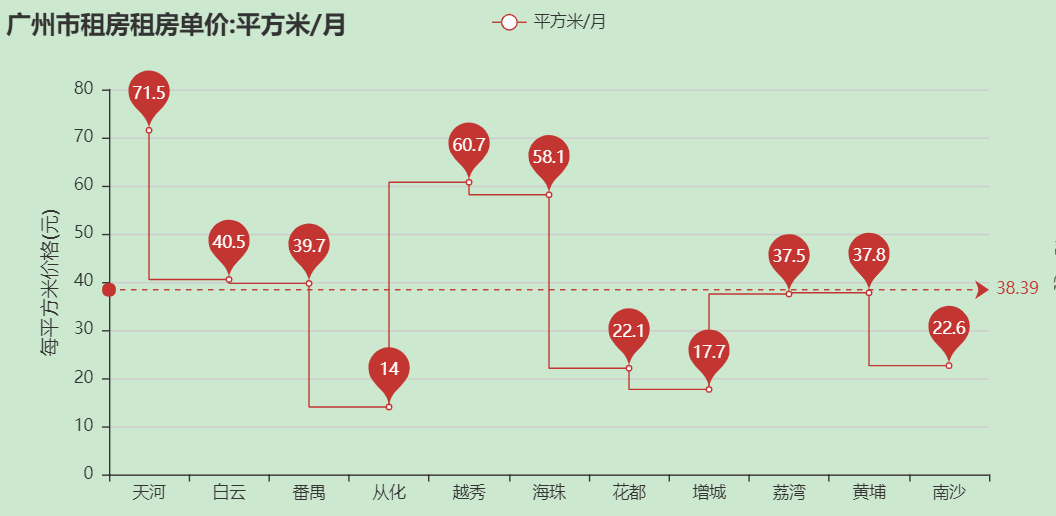
借助辅助线，可以清楚的看出广州平均房价最高是属于天河，为8164.81元/月，而最低的为从化，为1855.88元/月。

#### 平均房价&平均面积



图的x轴，y轴分别为平均房价以及平均面积，图中的点表示： 当平均价格在x值的时候，平均能买到y面积的房子，从图中可以得知：若想买y面积的房子, 需要多少钱。

#### 租房单价（平米/月）



图的x轴为广州市11个区， y轴为每平方米的价格(元)， 从图中可以看出天河，越秀，海珠的房价比较高，这些地方房价高是因为他们发展比较快，竞争压力大(需要付出更多的花销租房)

# 总结

这篇分析由于只抓取了房天下网站的部分数据且对数据的真实性未做核实，因此可能会导致分析结果不准确的问题。

另外在分析方面，分析的东西比较简单，无法对租房的选择进行明确的指引。仅可以给出一定的参考价值（平均租房单价， 房源分布， 户型， 是否装修等）

如果能抓取各大中介网站的数据并对其数据的真实有效性进行保证，同时结合租户的信息（如:公司地点，可接受的通勤时间等），也许租房的推荐问题便可以得到解决。