**《Python程序设计》**

**期末大作业**



**学院：计算机学院（国家示范性软件学院）**

**班级：2021211313**

**姓名：吕子健**

**学号：2023523012**

1. **实验目的**

1. 抓取链家官网北上广深 4 个一线城市，再加上一个离你家乡最近的一个非一线城市/或者 你最感兴趣的一个城市的数据。应尽量获取每个城市的全部租房数据（一线城市的数据量 应该在万的数量级）。

2. 比较 5 个城市的总体房租情况，包含租金的均价、最高价、最低价、中位数等信息，单位 面积租金（元/平米）的均价、最高价、最低价、中位数等信息。采用合适的图或表形式 进行展示。

3. 比较 5 个城市一居、二居、三居的情况，包含均价、最高价、最低价、中位数等信息。

4. 计算和分析每个城市不同板块的均价情况，并采用合适的图或表形式进行展示。

5. 比较各个城市不同朝向的单位面积租金分布情况，采用合适的图或表形式进行展示。哪个方向最高，哪个方向最低？各个城市是否一致？如果不一致，你认为原因是什么？

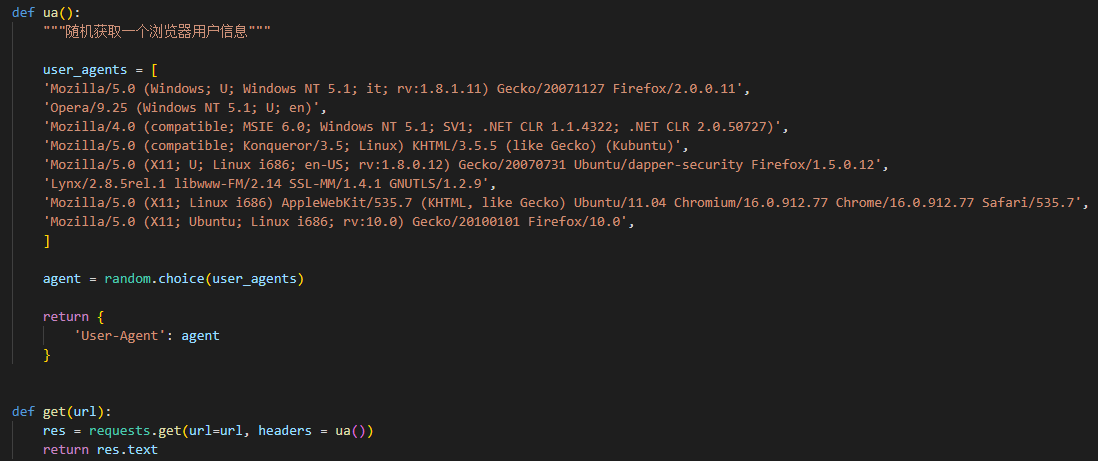
6. 查询各个城市的人均 GDP，分析并展示其和单位面积租金分布的关系。相对而言，在哪个城市租房的性价比最高？

7. 查询各个城市的平均工资，分析并展示其和单位面积租金分布的关系。相对而言，在哪个城市租房的负担最重？

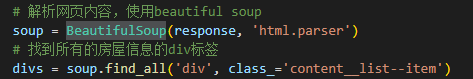
1. **实验内容**
2. **爬取数据**

类似于小作业2，先获取链接回传的html代码，不过这次数据量大所以在请求头部添加UA伪装用户来防止反爬。





然后通过BeautifulSoup模块来匹配每一条租房信息的div。



随后遍历所有找到的div，同样用BeautifulSoup匹配html中各种信息的格式块，并对某些数据进行各种处理。最后将信息字典写入到列表中。

for div in divs:

            try:

                # 提取名称，去掉前后空格

                try:

                    name = div.find('a', class\_='twoline').text.strip()

                except:

                    continue

                print(name)

                # 提取信息

                category = div.find('p', class\_='content\_\_list--item--des').text.strip()

                # 拆解信息

                category = category.replace(' ', '')

                category = category.replace('\n', '')

                try :

                    block1,block2,block3,block4,level = category.split('/')

                    if "仅剩" in block1:

                        continue

                except :

                    continue

                region,block,address = block1.split('-')

                area = block2

                area = get\_area\_midpoint(area)

                room\_type = block4

                # 提取租金，区间取中值并取整

                total\_price = div.find('span', class\_='content\_\_list--item-price')

                if total\_price:

                    total\_price = total\_price.text.strip()

                else:

                    continue

                total\_price = get\_price\_midpoint(total\_price)

                # 将房屋信息以字典的形式添加到列表中

                print("收")

                data.append({

                    '名称': name,

                    '类别': category,

                    '区域': region,

                    '板块': block,

                    '地址': address,

                    '房型': room\_type,

                    '面积': area,

                    '总价': total\_price

                })

            except:

                continue

        # 找到下一页的url，如果没有，退出循环

        # 反爬随机停止一段时间

        time.sleep(0.5)

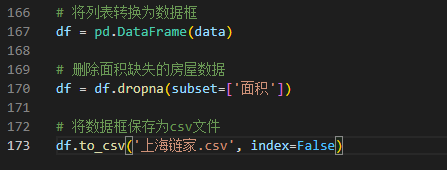
        url = 'https://sh.lianjia.com/zufang/pg' + str(pg) + '/'

        #url = None

    except:

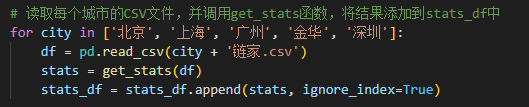
        break

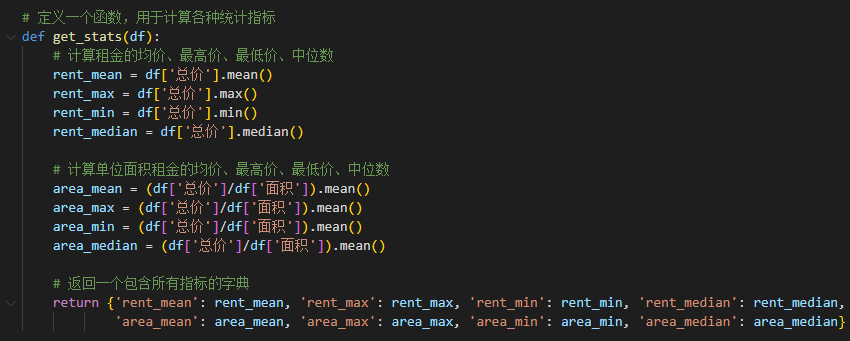
最后将列表转换为数据框，然后保存到对应的csv文件中。



1. **比较总体房租情况**

构造一个函数读取前一项所得到的csv文件，用pandas进行数据处理，直接调用内部函数即可很方便地将数据写入字典当中并返回：



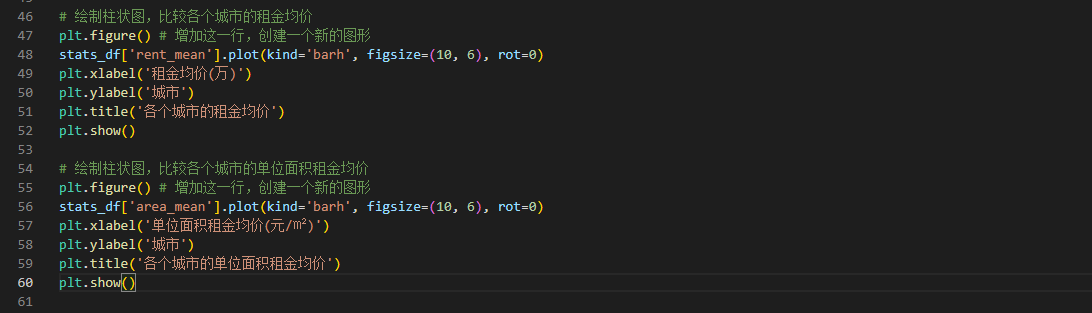


再将指标回传到一个新的pandas数据框中



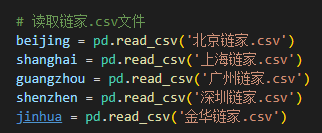


最后分别用两个横向柱状图表示。

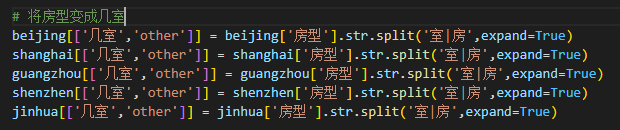


1. **比较不同房型价格情况**

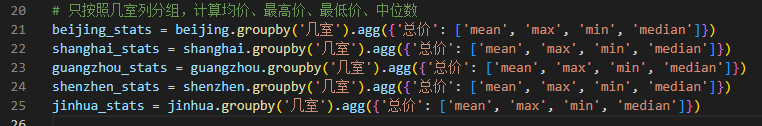
读取csv

****

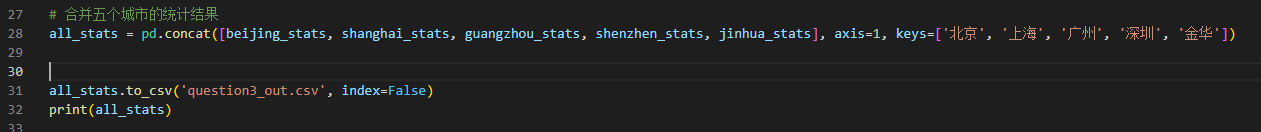
单独提取几室



处理



合并输出



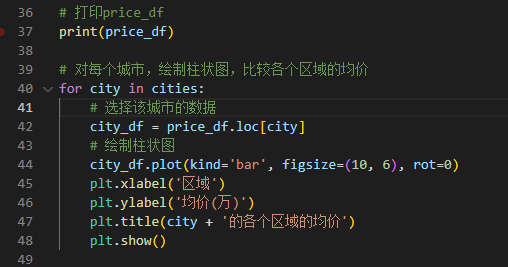
1. **板块情况**

\*因为板块实在太多，这里采用区域进行分类，如果要实现板块只需要将源代码内所有“区域“替换为”板块”即可。

读取并先用区域对值进行分类，然后将每个分组的总价值进行平均数计算，和城市与板块一起回传到一个数据框中。

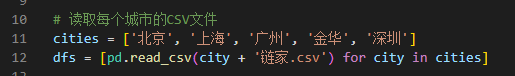


用一个for循环循环输出每个城市各个区域的价格图

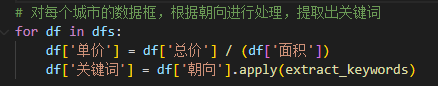


1. **朝向情况**

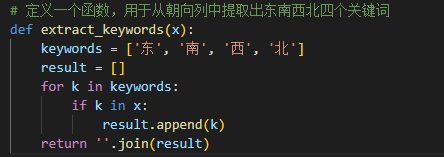
这次我换了一种读取方式，直接将五个城市数据读取到一个dfs中



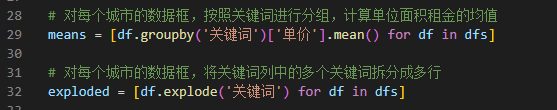
然后对朝向列进行关键词提取，顺便算出均价



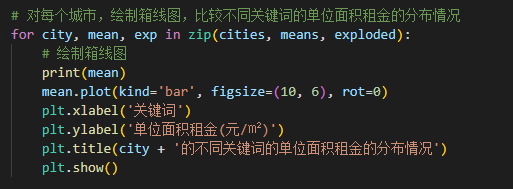
关键词提取能匹配关键词，例如某个房子朝向列是“南北“，就将此房子归类到”南北“字典中。



然后根据提取出的关键词进行分组求平均值。



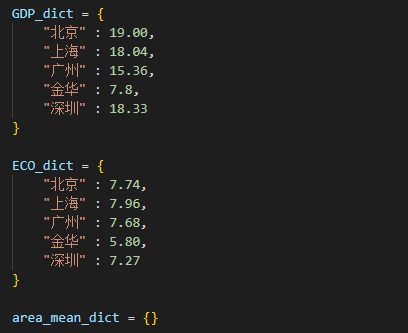
最后绘制出每个城市各个朝向组合均价。



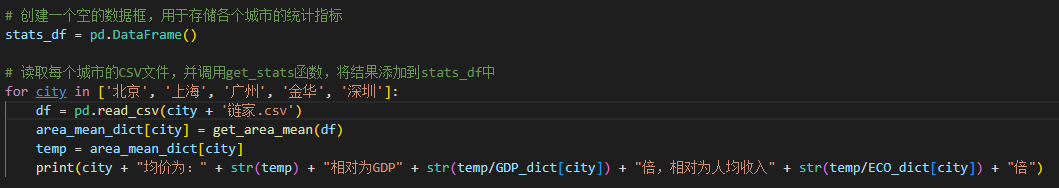
1. **GDP与人均收入情况**

因为6、7类似，所以放在一起做。

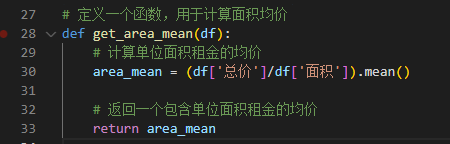
首先根据网络上查到的数据创建一个城市GDP字典和一个人均可支配收入字典，随带创建一个面积均价空字典。



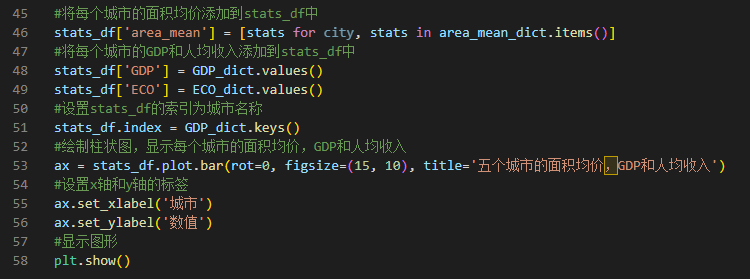
随后遍历所有城市csv，调用get\_area\_mean求面积均价，然后输出与GDP，人均可支配收入相对比值。（之所以是相对因为人均收入与GDP进行了缩小10000倍处理）



get\_area\_mean函数



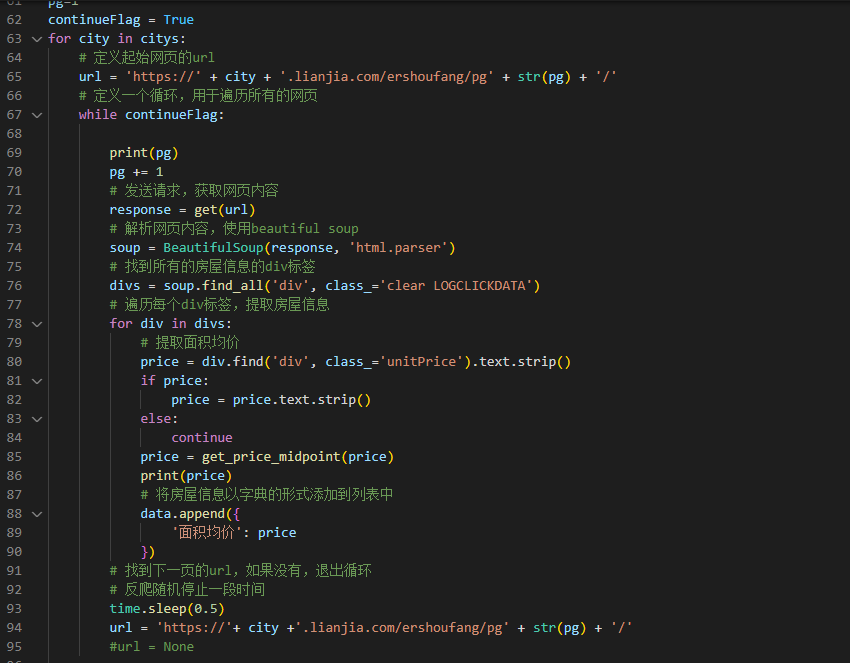
最后读入三个字典绘制出柱状图



1. **额外：二手房价与租金之差**

我想探究一下房子价格是否按比例与租金价格正相关。所以我需要爬取链家上二手房单位面积均价，然后与单位面积租金面积均价一同用柱状图展示出来，以下是整个额外项目的流程：

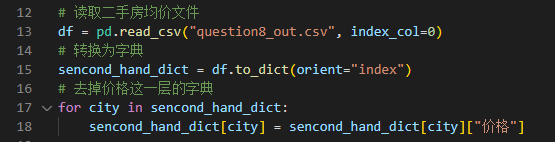
与实验1爬取类似，但这次只爬取面积均价



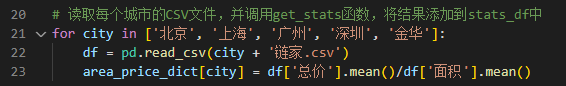
随后对爬取到的价格（字典形式）转换为数据框，求完平均数后保存到一个csv文件中。



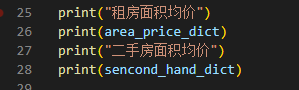
再在第二个程序中读取csv然后保存在一个二手房价格字典当中



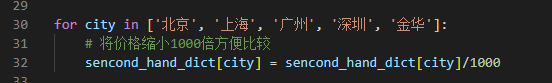
然后读取实验1爬到的数据算出各个城市单位面积租金并存入单位面积租金字典当中。



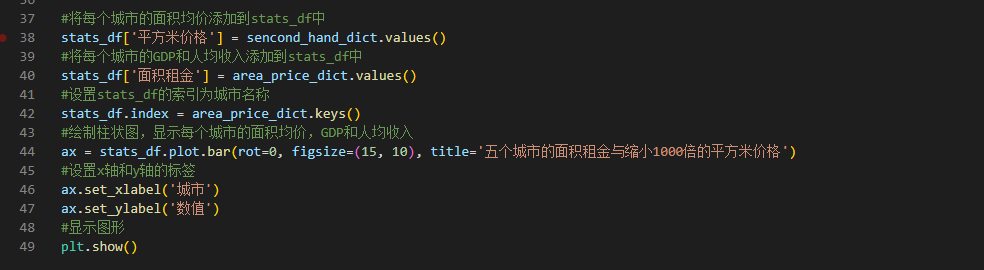
先输出两者数值数据



然后将二手房数据缩小1000倍方便后面比较



画出柱状图



1. **实验结果**

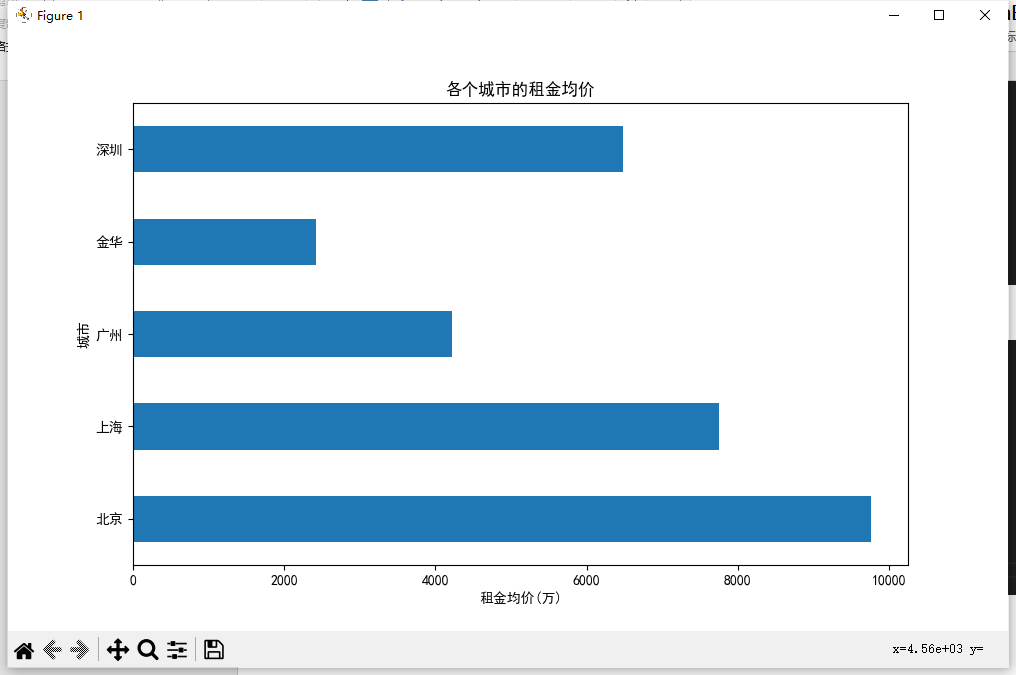
**1. 爬取数据**

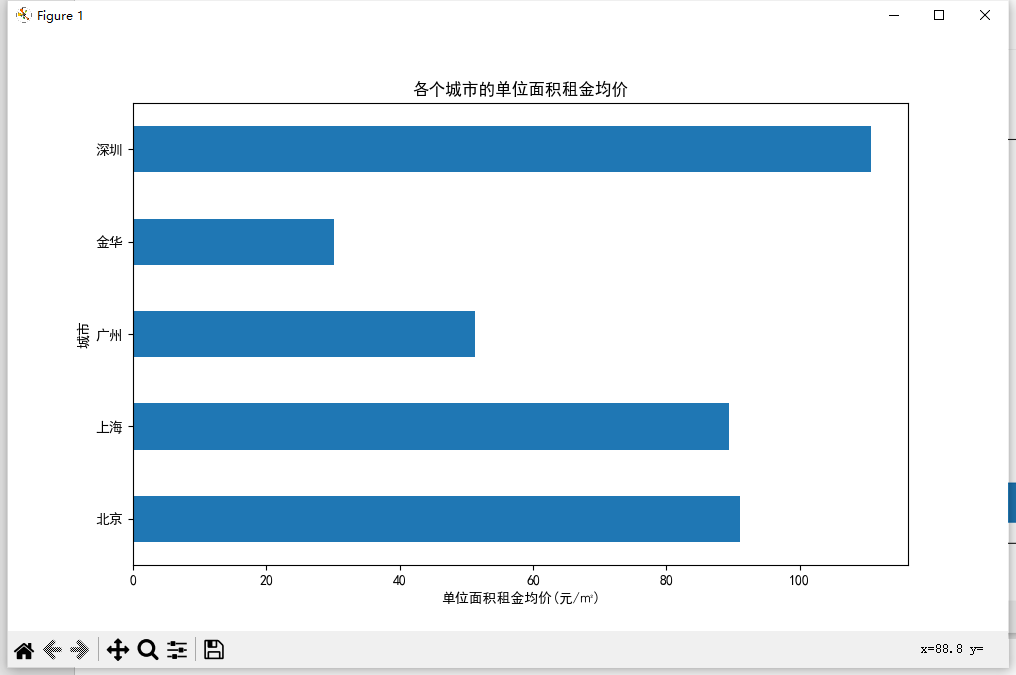


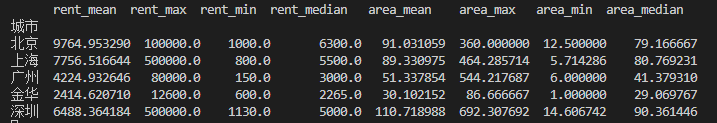


第五个城市选择我家附近三线城市金华

**2.** **比较总体房租情况**

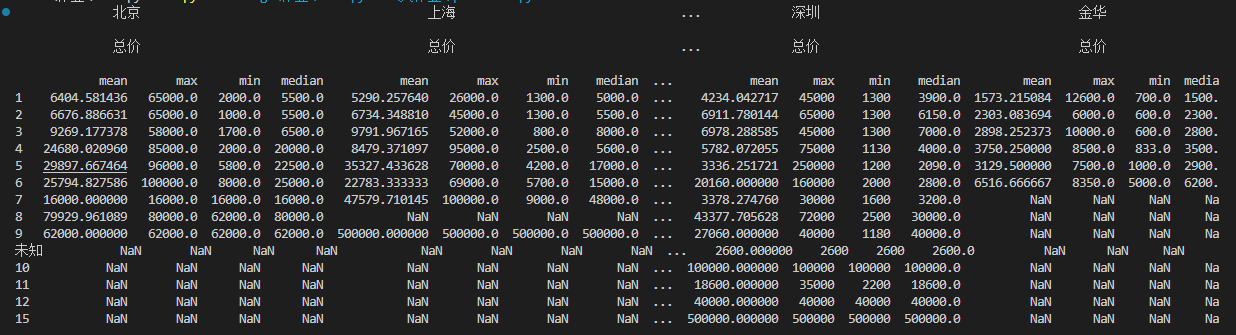


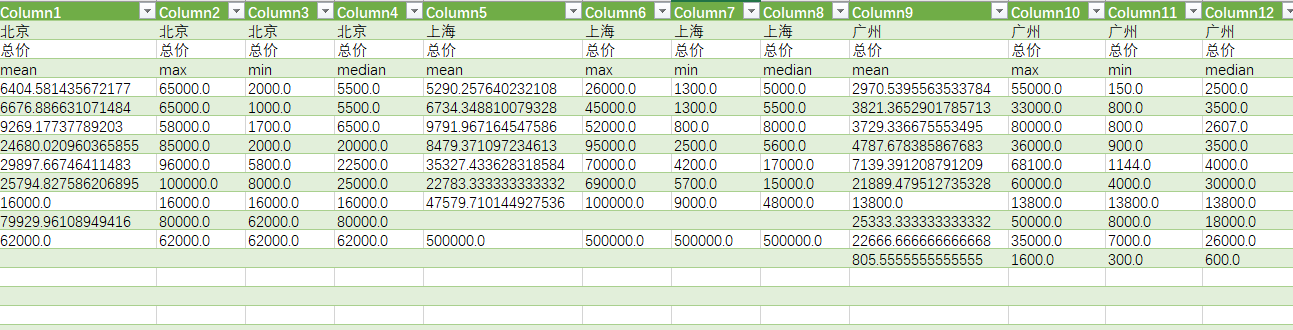


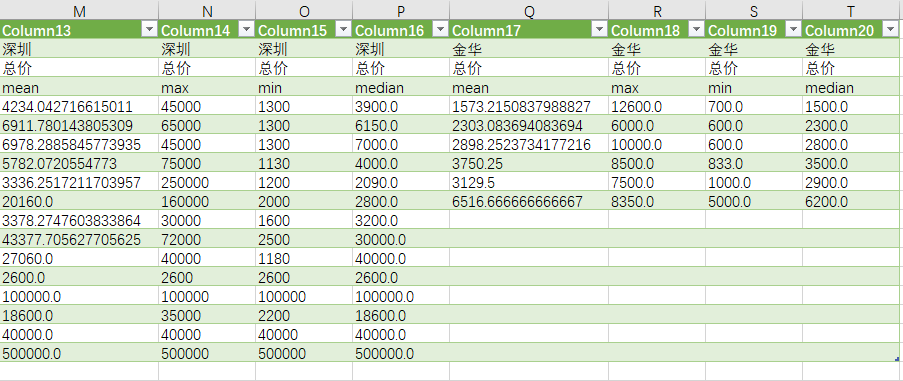


根据图表，能看出租金均价为**北上深广金**的顺序依次变少，而单位面积均价均值却是**深北上广金**这样的顺序，可预测出深圳房屋面积普遍偏小，并且非一线城市金华确实相较于一线城市租金低太多了。同时，广州却拥有最低租金（150元）的房子，经筛查是一个12平的地下室，并地处近郊。而最便宜单位租金也毫不意外的属于金华。

**3.** **比较不同房型房租情况**

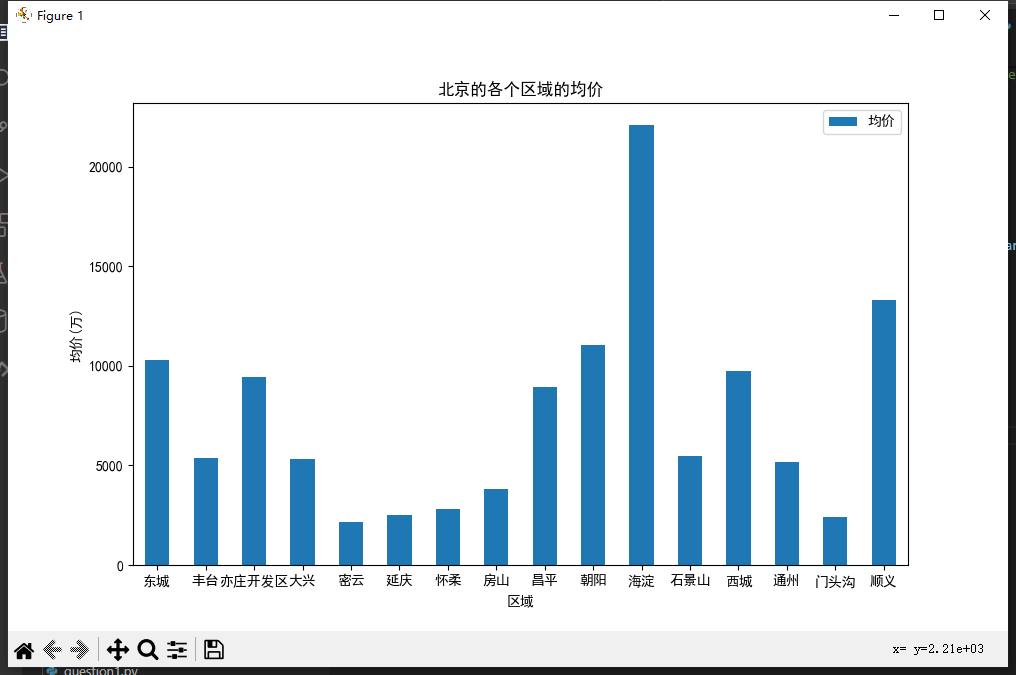


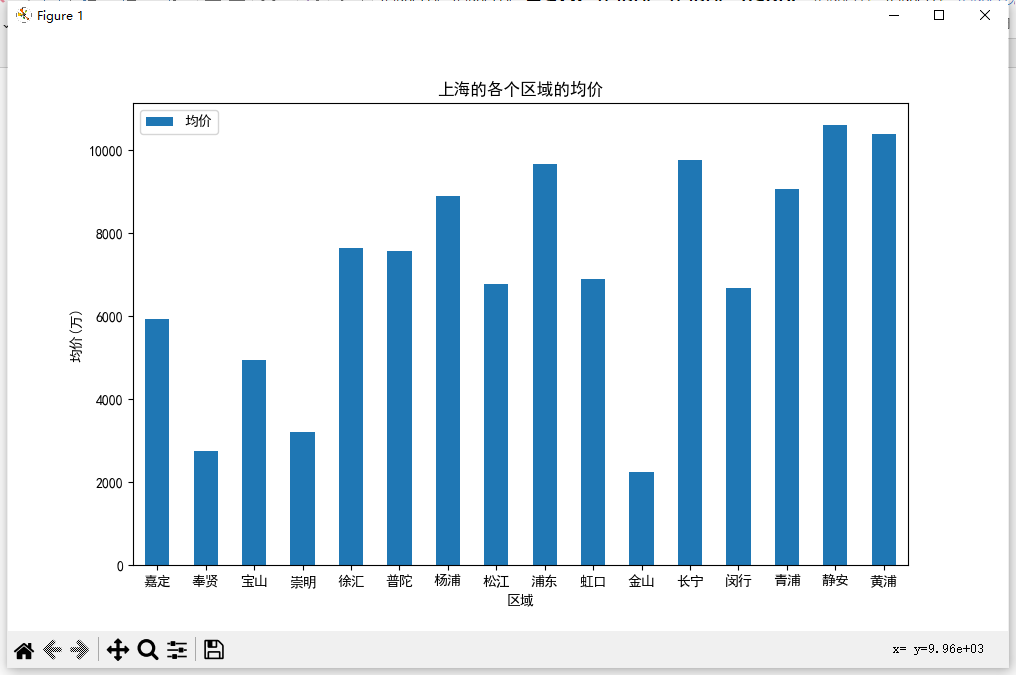


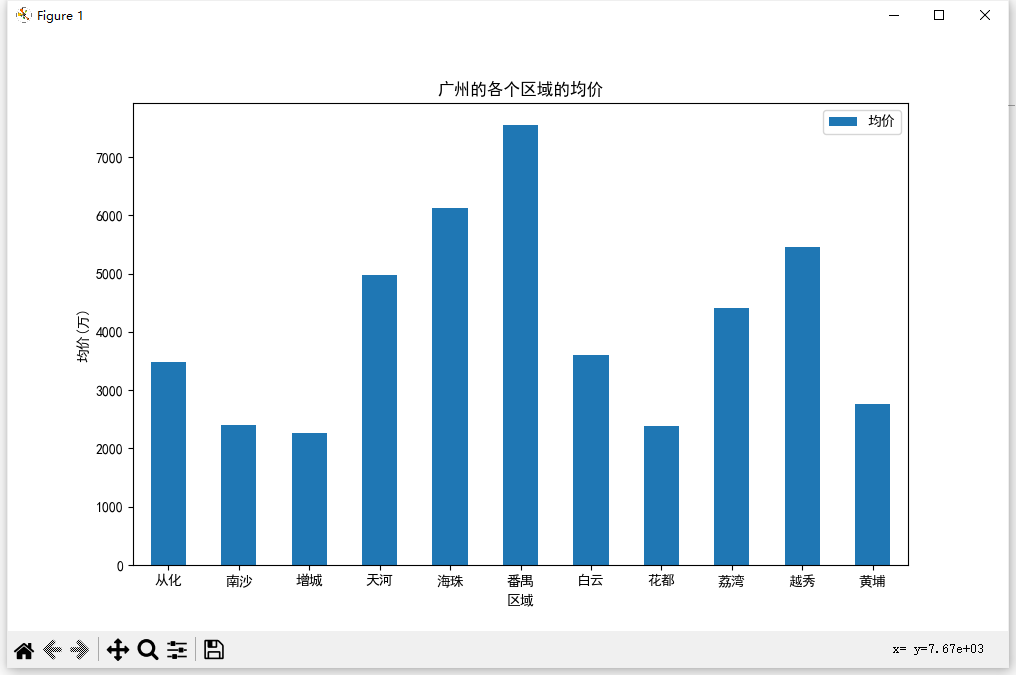


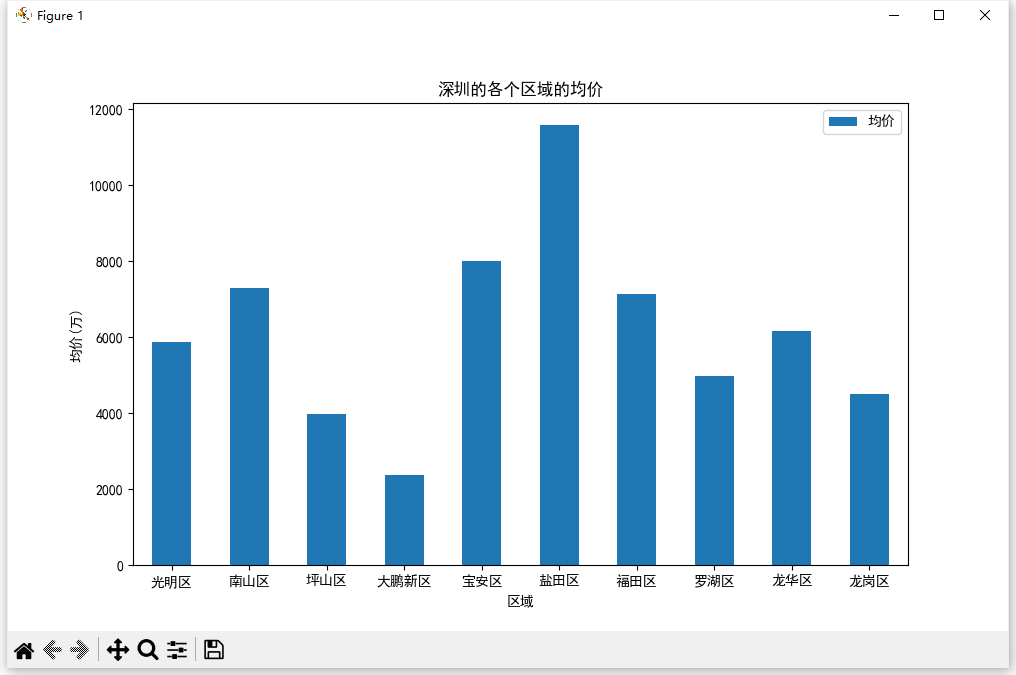
首先就是深圳拥有最多的房型，最高可达14室而价格也为链家上线五十万一个月，其次每个城市也确实符合越多室，均价越高，但最大和最小值不一定随之成正比，受到政策、区域以及历史原因等多方面影响。而金华作为非一线城市五室以上的房型就寥寥无几了。

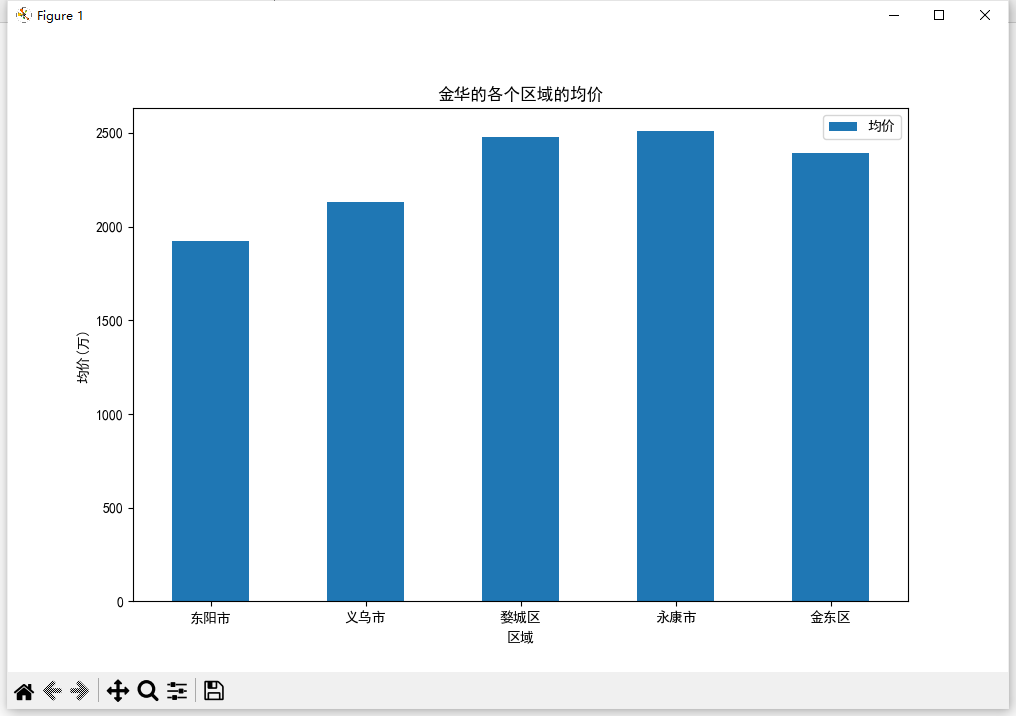
**4.** **板块（区域）情况**



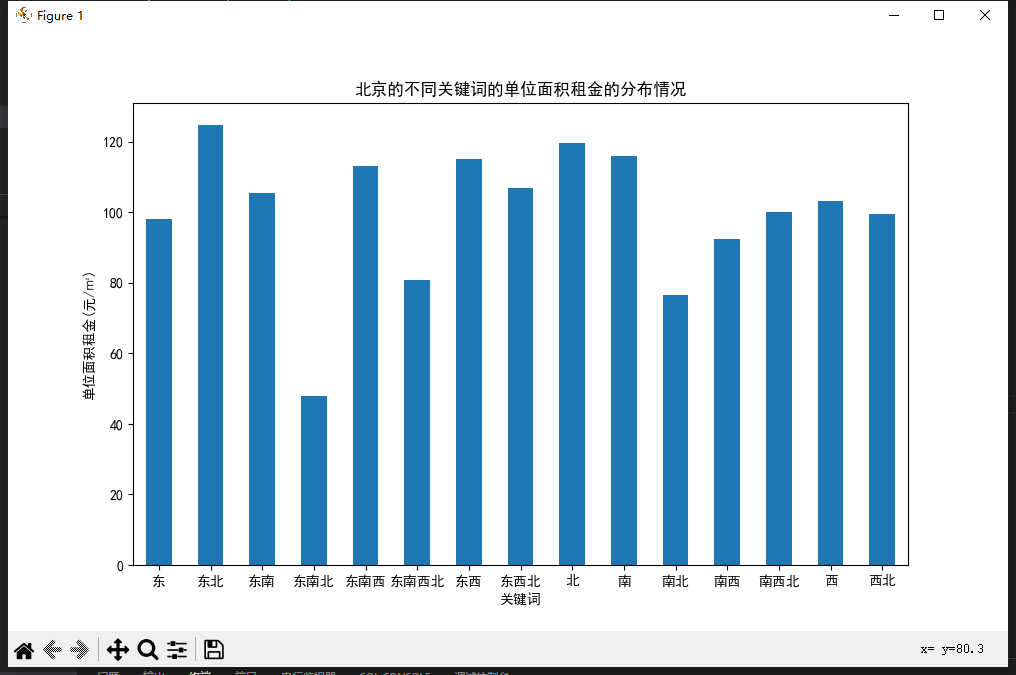


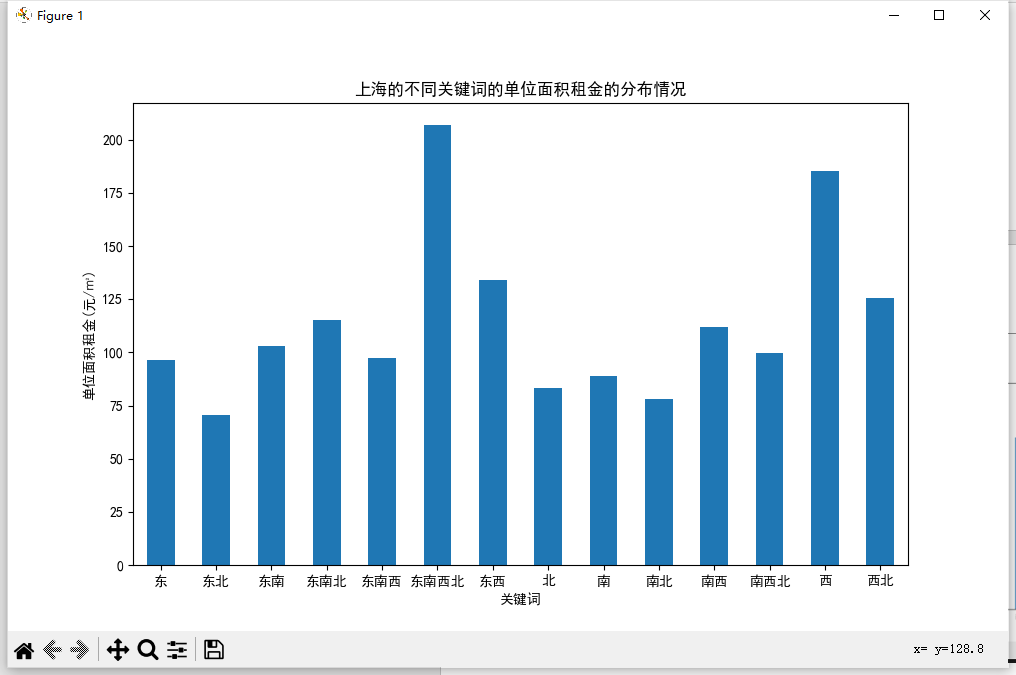


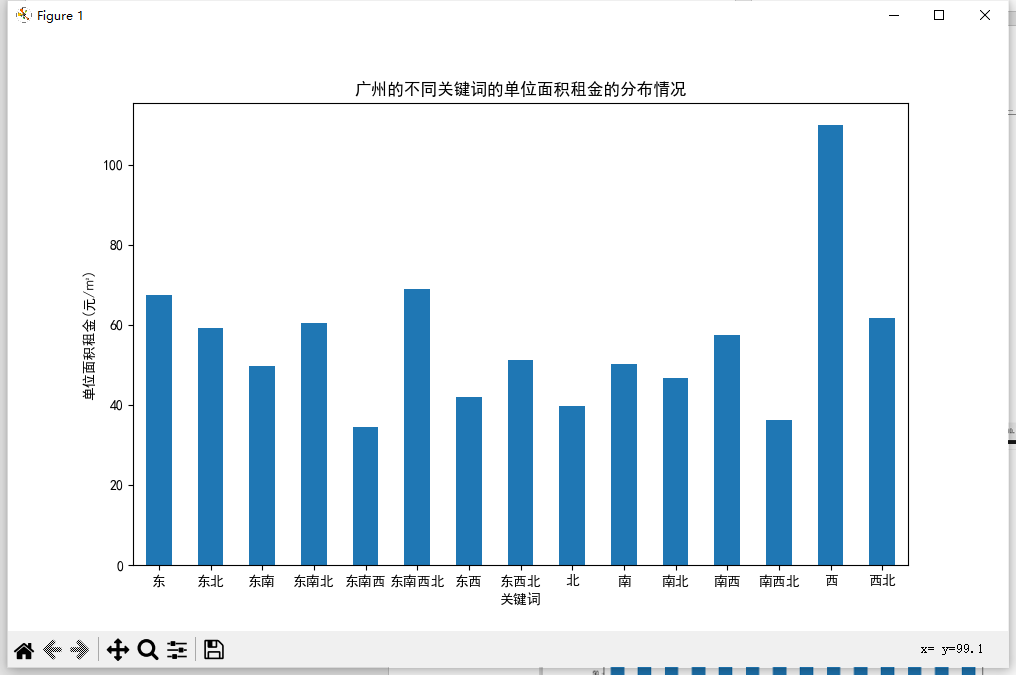


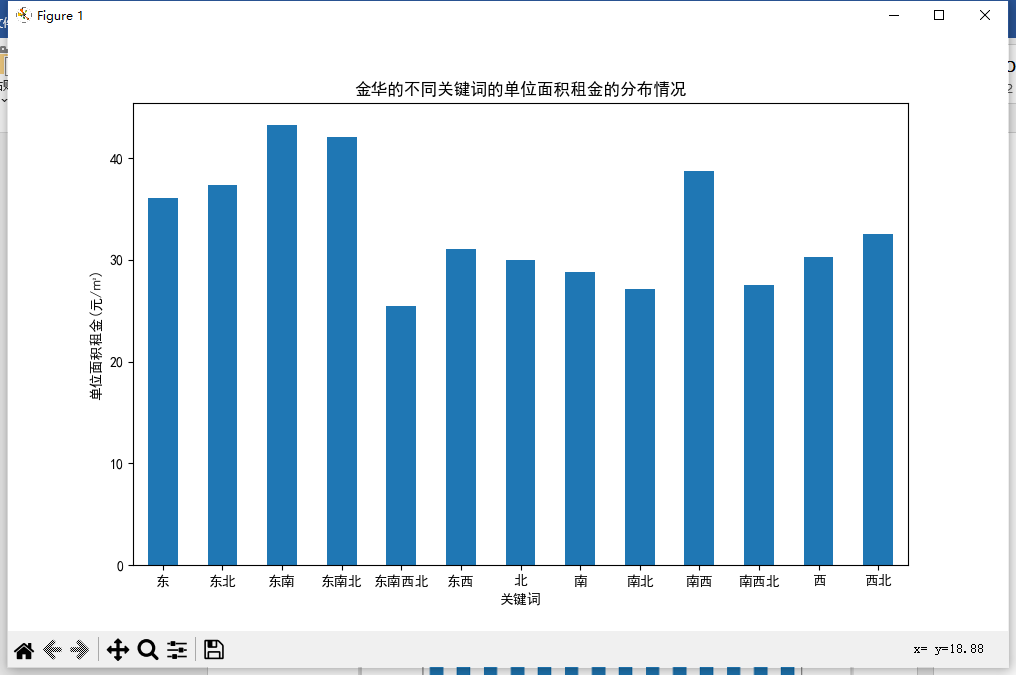


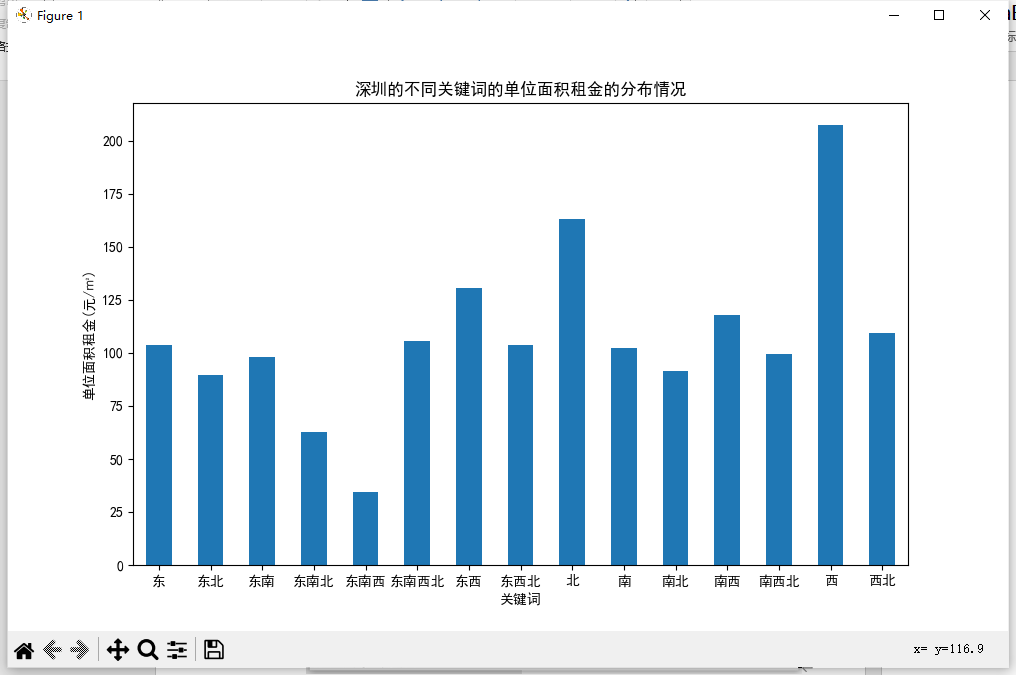
**5.** **朝向情况**











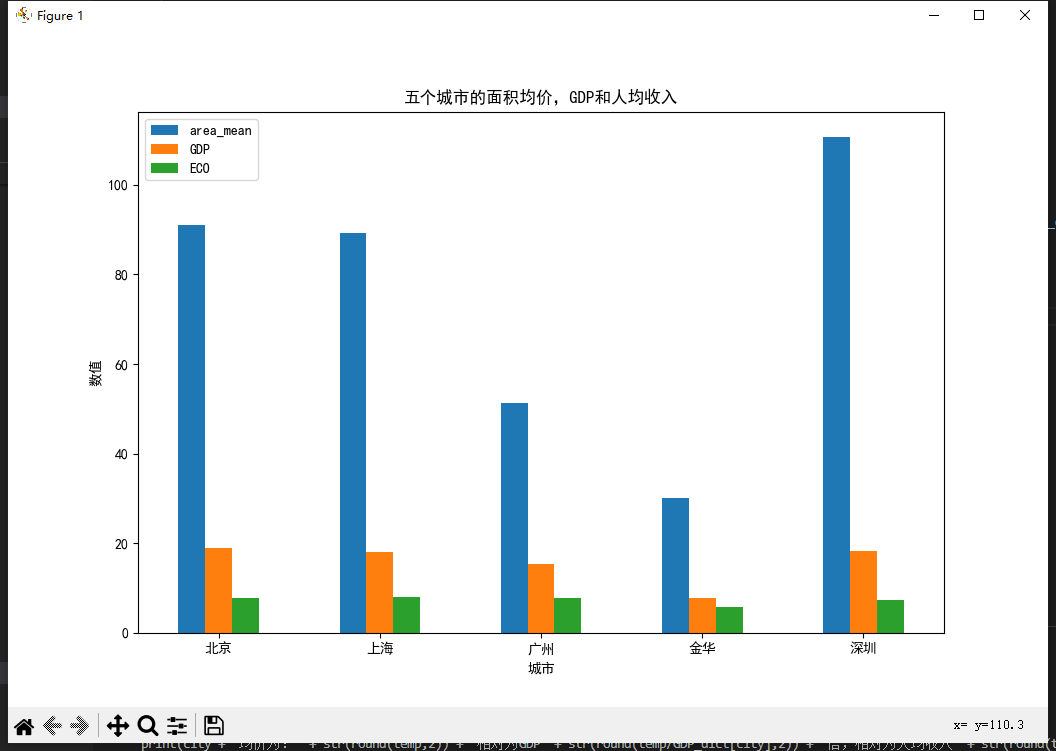
与预期的不太一样的是，一线城市中有三个城市（上广深）西面朝向的房价出奇的高，根据分析应该是恰恰是因为消费者倾向购买南北朝向的房子，所以开发商也倾向建造南北的房子，导致西数据量小。而如果开发的是大户型房子，需要充足的光源照亮房内，而朝西能最大限度接受阳光，所以这个朝向的房型都偏大，价格就贵。

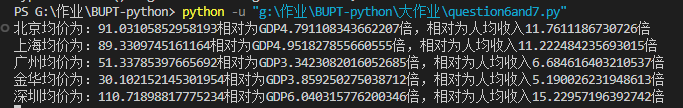
同时，在收敛的数据中，单独朝向的房子相对于多朝向的房子价格更加便宜，很有可能是因为单独朝向的房子通常采光和通风不如多朝向的房子，尤其是朝北或朝西的房子，冬季阴冷，夏季炎热，影响居住舒适度。

北京与其它城市不相同的是几乎所有朝向价格都差不多，原因是北京的房价本身就很高，相对于其他城市，朝向的影响并不显著，更重要的是区域、学区、楼龄、小区品质等因素，且北京相较于其它城市发展较早，当时并不注重房子朝向。

总体来说五个城市朝东（含）的房子比较贵，是因为朝东的房子可以享受到早晨的第一缕阳光，呼吸到新鲜的空气，对于早睡早起的人来说非常适合。朝东的房子无西晒之忧，夏天凉爽，冬天也不会太阴冷，采光和通风都比较好。朝东的房子符合中国人的传统观念，东方象征着光明和希望，有祥瑞之感。

**6&7.** **GDP与人均收入情况**



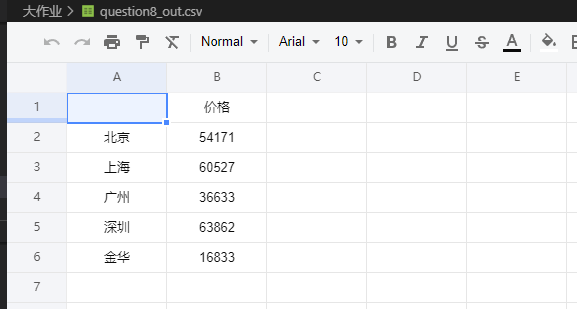


根据GDP，相对而言，金华的租房性价比最高，因为它的租金占GDP的比例最低，只有3.859倍。

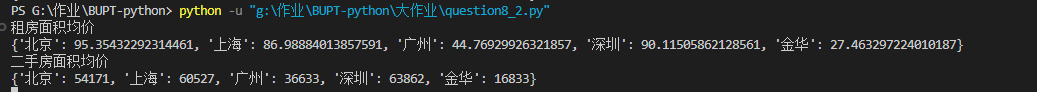
根据人均收入，相对而言，深圳的租房负担最重，因为它的租金占人均收入的比例最高，达到了15.229倍。

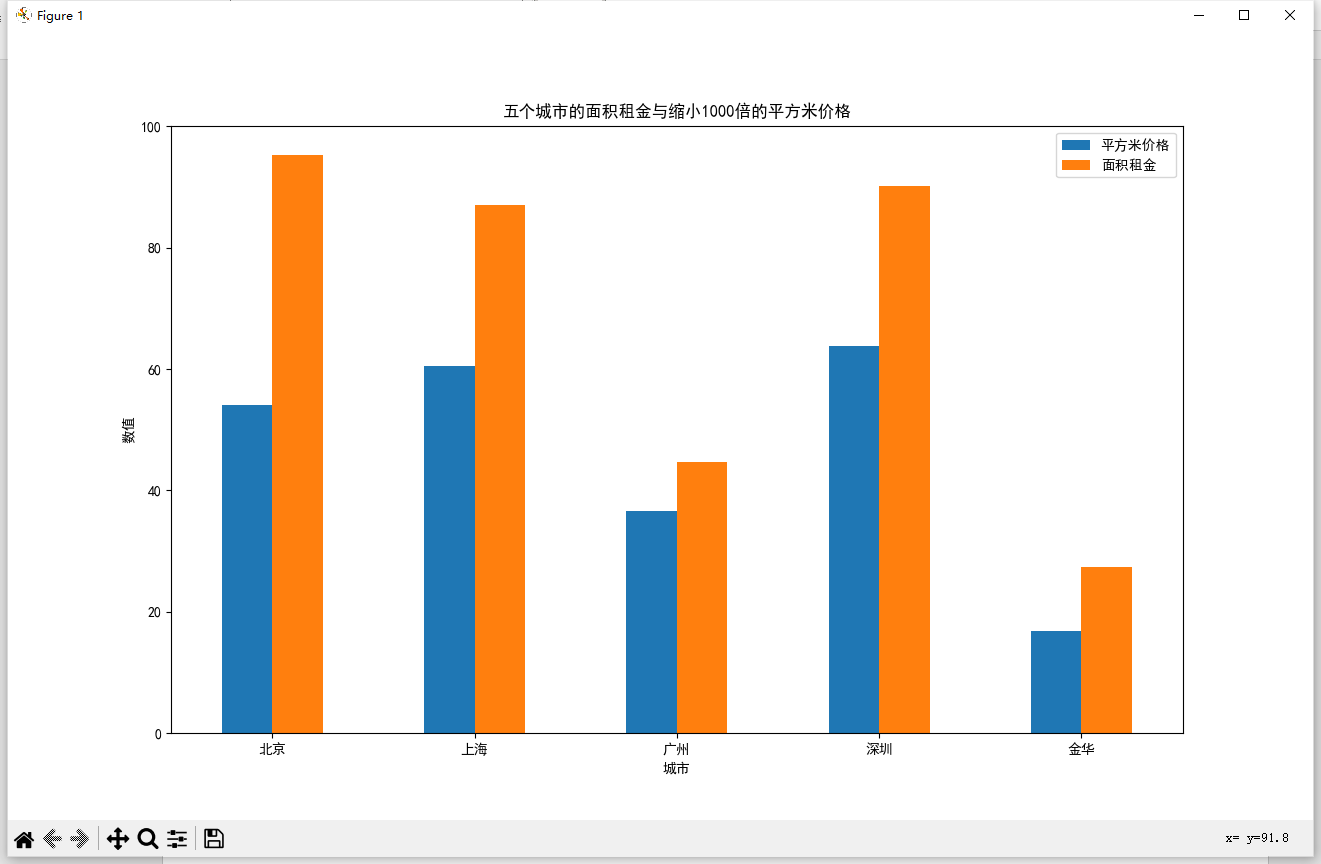
**8. 额外：二手房价与租金之差**

爬取到二手房数据后所处理出的二手房均价csv



数据处理与展示的结果





可以看见，除了北京以外都相对来说租金与价格成正比，北京房价比别的一线城市低，租金却比别的一线城市高的原因可能是投资回报率过低，一般来说，房屋租售比越低，说明房屋的投资价值越高，越有利于持有。但北京经济极大受制于政策优惠，不稳定，加上近几年经济疲软，消费者对北京房子处于持币观望状态。