

**2022 Python程序设计大作业**

**题目: 租房数据分析**

**姓 名 鄭毓恒**

**学 院 计算机学院**

**专 业 计算机科学与技术**

**班 级 2020211302**

**学 号 2020211262**

**任课教师 杨亚**

**2022年 12 月**

**要求**

1. 要求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租房数据。

2. 应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）。

3. 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。

4. 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。

5. 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。



例如上图中的“海淀-四季青-五福玲珑居北区”，“四季青”即为板块名称。

6. 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？

7. 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？

8. 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？

9. 请贴上爬虫程序的核心代码、数据处理及数据展示的核心代码（应有足够的注释）。

**开发环境**

Windows 11操作系统，Python 3.10

**抓取链家官网北上广深4个一线城市和泉州市的全部租房数据**

|  |
| --- |
| begin.py |
|  |

定义五个爬虫，分别抓取五个城市的租房数据。

|  |
| --- |
| items.py |
|  |

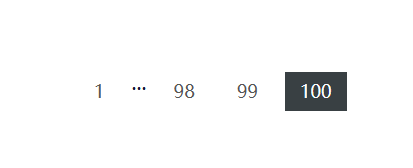
抓取的数据有五项，分别为租金、板块、面积、朝向和居数。

|  |
| --- |
| pipelines.py |
|  |

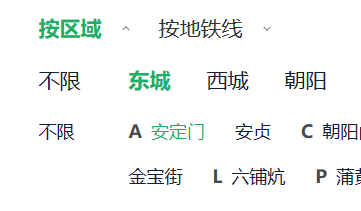
将数据输出至csv文件。

|  |
| --- |
| spider.py |
|  |

爬虫定义如上，定义了五个爬虫，分别对应五个城市。



链家网站显示北京有三万多条租房数据，但是链家一次最多只能查看100页。即使一个个区爬，一个区的数据也可能超过100页。因此，需要一个个板块爬。



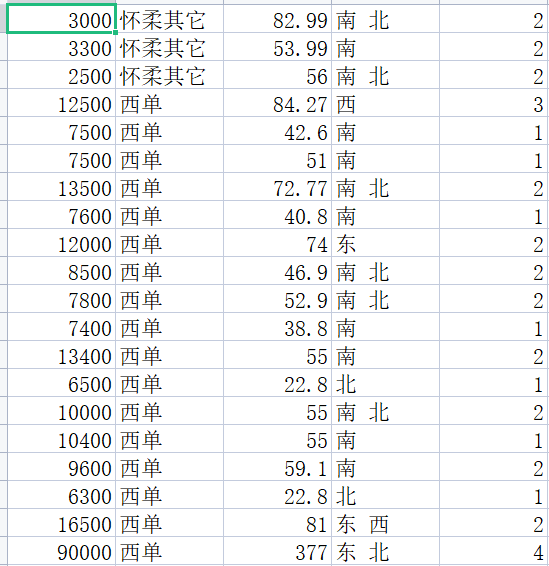
一个板块可能会出现在不同的区里，因此需要给已经爬过的板块进行标记，避免抓取重复数据。



租房数据的格式并不一致，例如上图的两种，没有板块和朝向，而且租金和面积不确定。

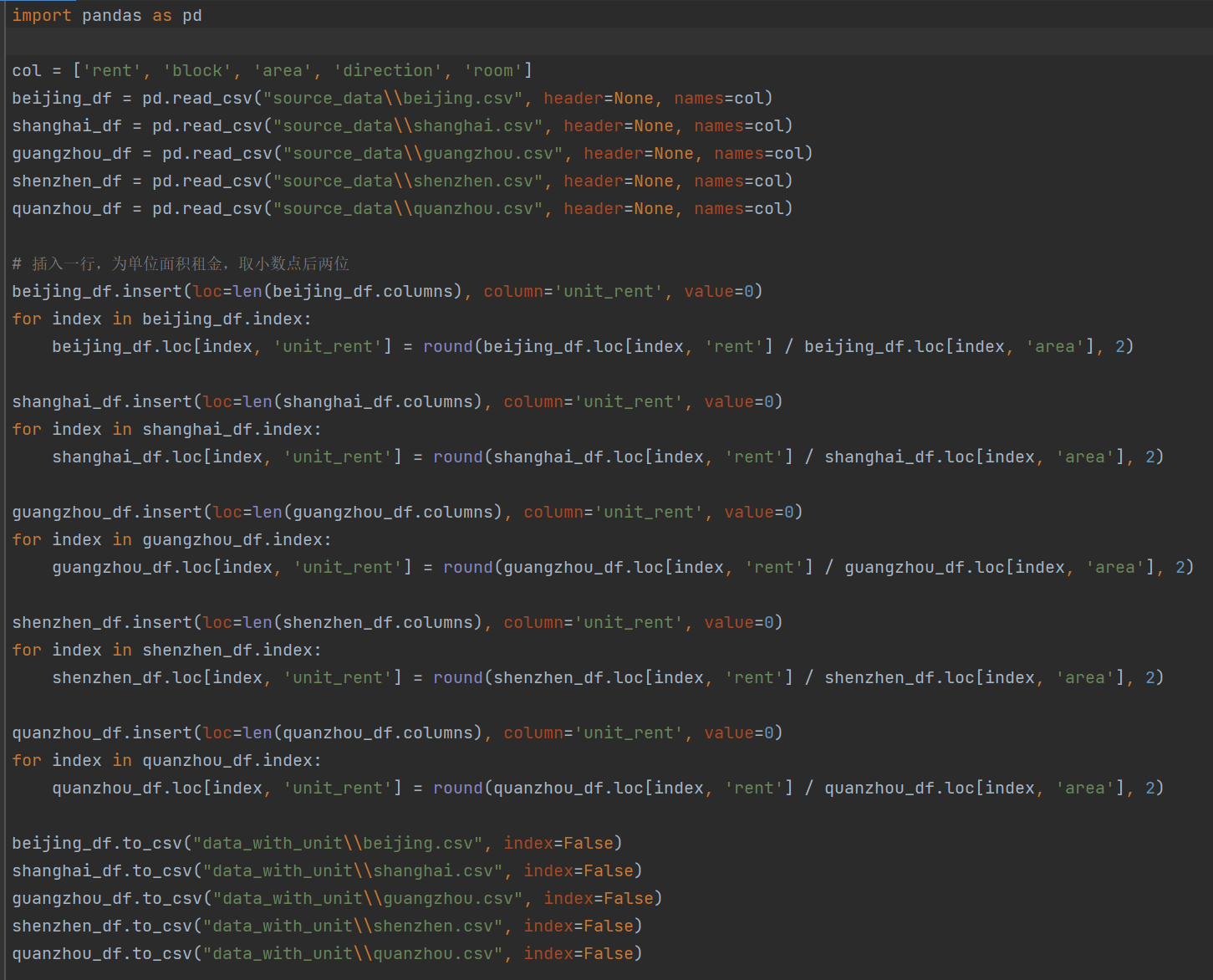


只有这两种租房数据有所需的所有数据，所以只从网站抓取这两种格式的数据。

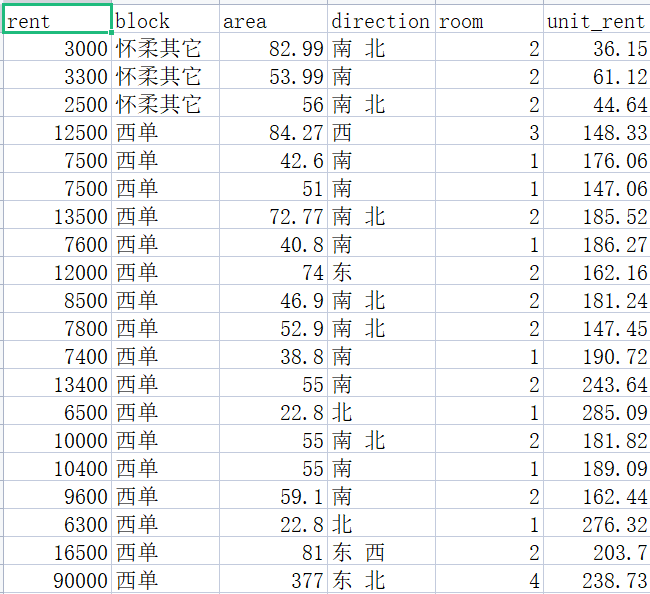


输出的csv文件格式如上。

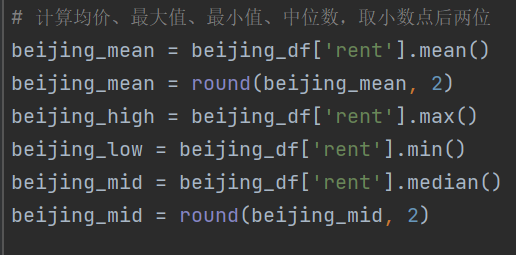
**比较5个城市的总体房租情况，包含租金的均价、最高价、最低价和中位数信息，单位面积租金（元/平米）的均价、最高价、最低价和中位数信息。**



计算单位面积租金，取小数点后两位。



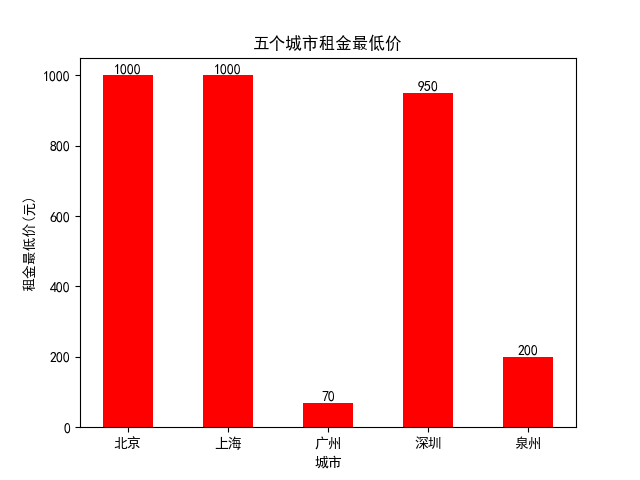
计算完单位面积租金后输出的csv文件格式如上。



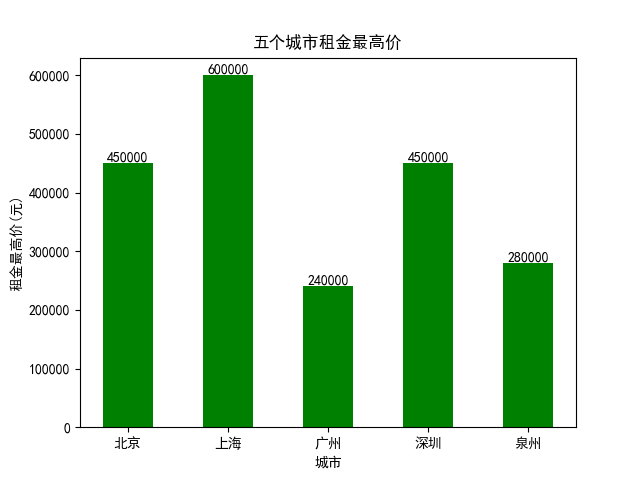
给每个城市的租金计算均价、最大值、最小值和中位数。均价和中位数区小数点后两位数。



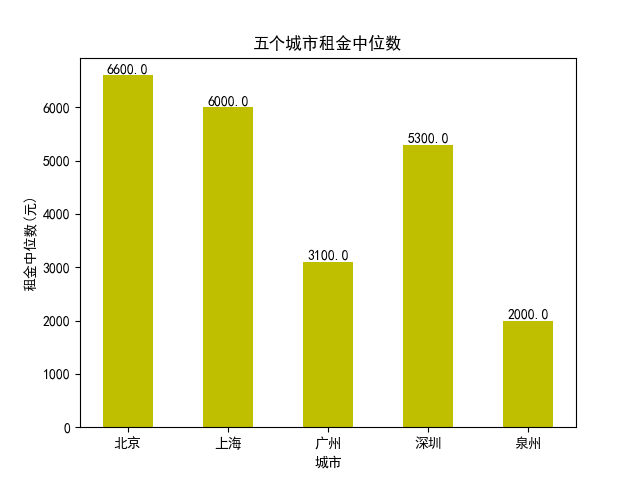
按以上设置画四张图，分别为五个城市租金的最小值、最大值、均价和中位数。



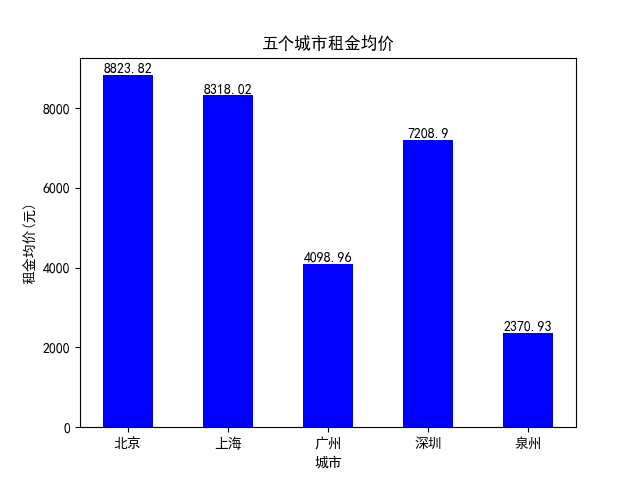
北京和上海的租金最低价并列最高，深圳其次。泉州和广州的租金最低价与另外三个的差距很大，泉州倒数第二，广州最低。



上海的租金最高价最高，北京和深圳并列第二，泉州第三，广州最低。

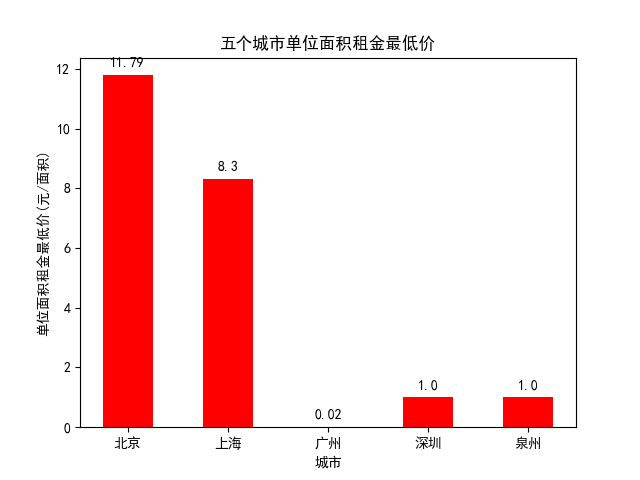


北京的租金中位数最高，上海第二，深圳第三，广州第四，泉州最低。

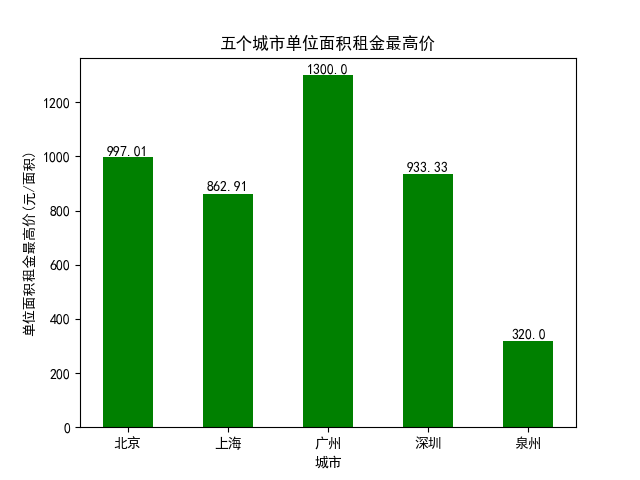


北京的租金均价最高，上海第二，深圳第三，广州第四，泉州最低。

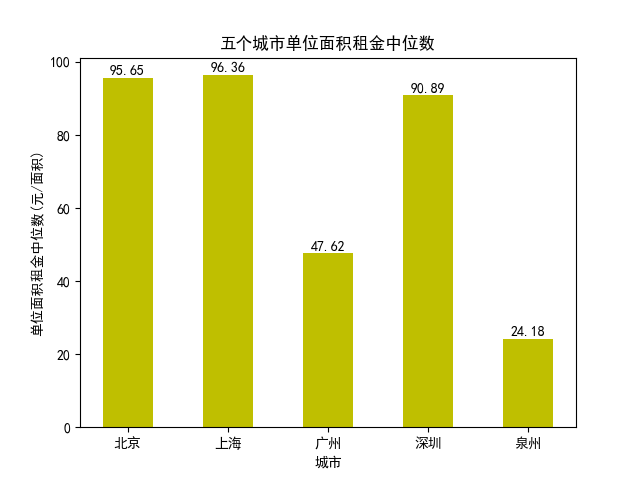
五个城市的租金中位数和均价的排序一样，并且最高价和最低价可能存在极端值，应该以均价和中位数的排序为准。北京的租金最高，上海第二，深圳第三，广州第四，泉州最低。



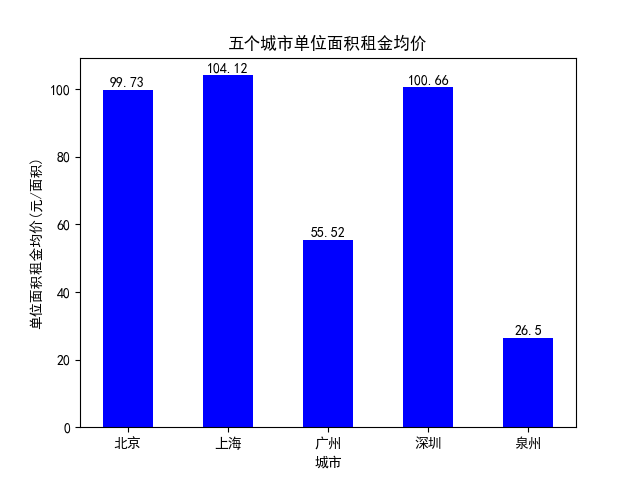
北京的单位面积租金最低价最高，上海第二。广州、深圳和泉州的数据与前两名的差距很大，深圳和泉州并列第三，广州最低。



广州的单位面积租金最高价最高，北京第二，深圳第三，上海第四，泉州最低。



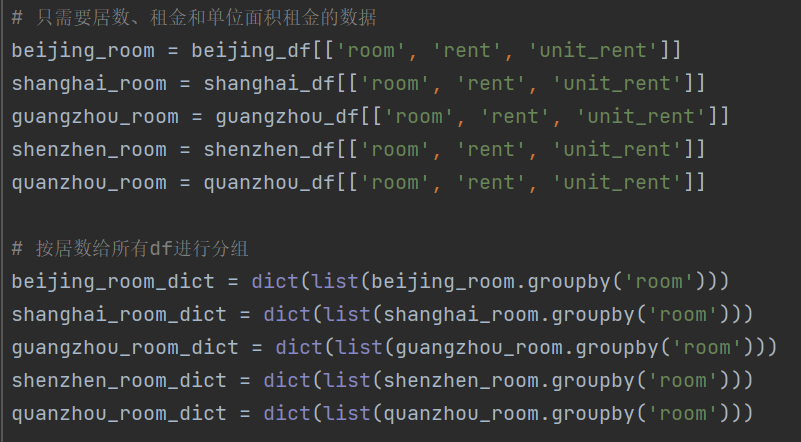
上海的单位面积租金中位数最高，北京第二，深圳第三，广州第四，泉州最低。



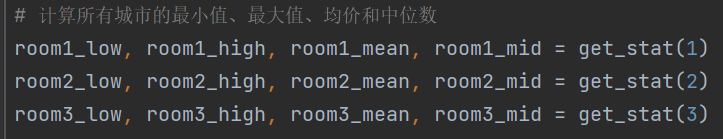
上海的单位面积租金均价最高，深圳第二，北京第三，广州第四，泉州最低。

最高价和最低价可能存在极端值，应该以均价和中位数的排序为准。北京、上海和深圳的数据很接近，是五个城市的最高。然后广州第四，泉州最低。

**比较5个城市一居、二居、三居的情况，包含均价、最高价、最低价和中位数信息。**

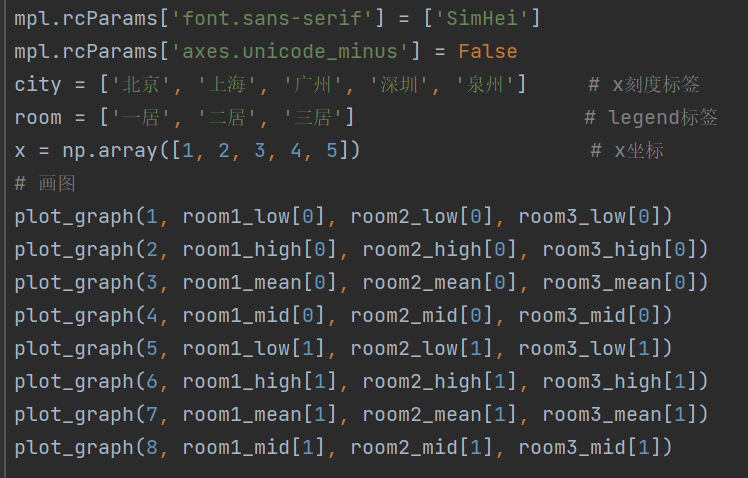


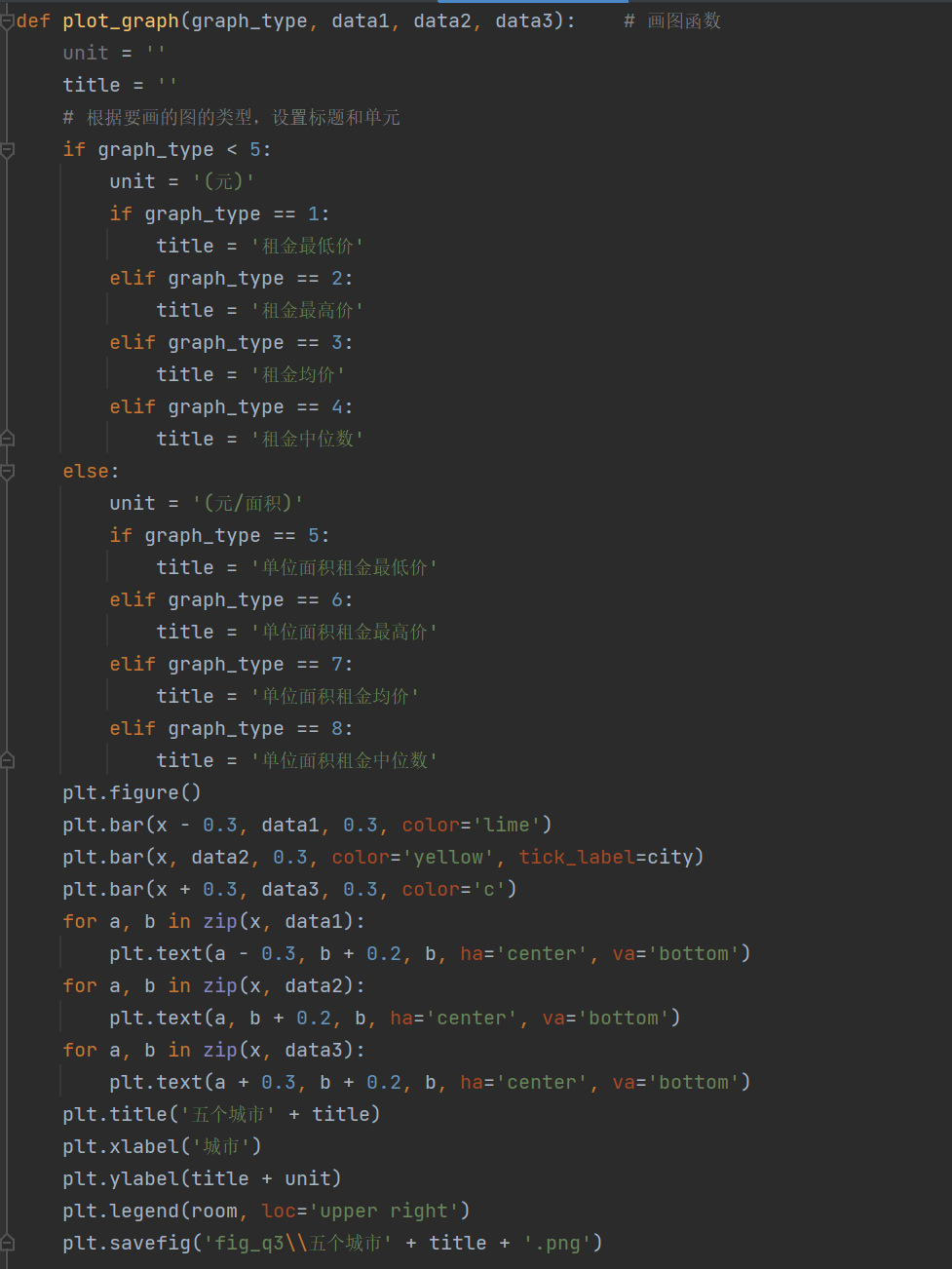
按照居数给数据进行分组。



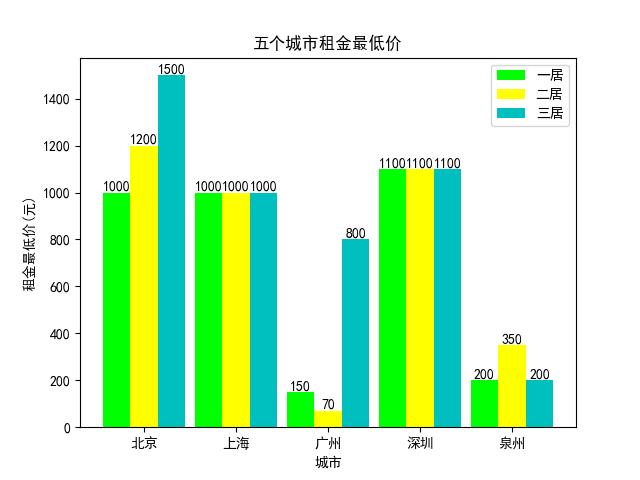


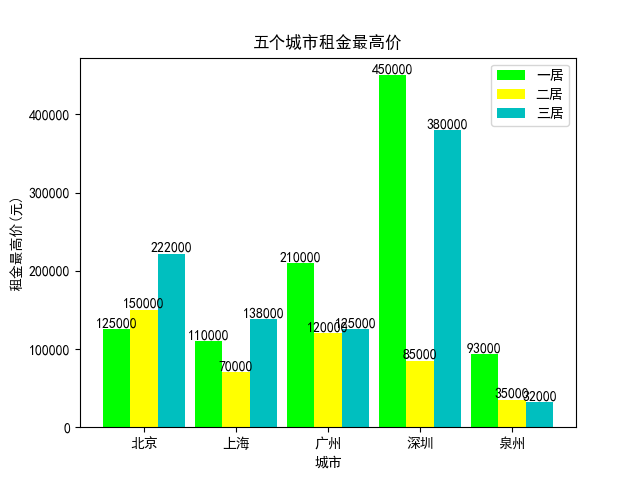
分别计算五个城市一居、二居和三居的租金和单位面积租金的最小值、最大值、均价和中位数。中位数和均价取小数点后两位。

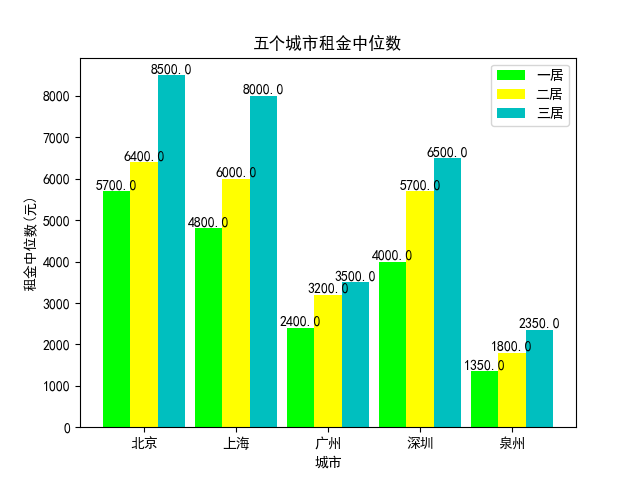


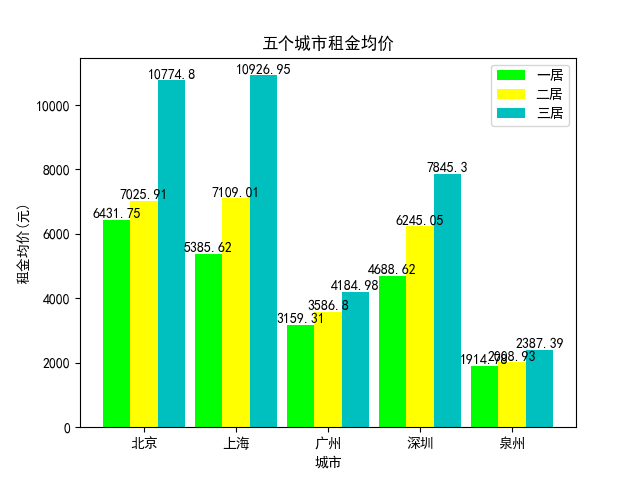


按照以上设置进行画图，总共八个图。

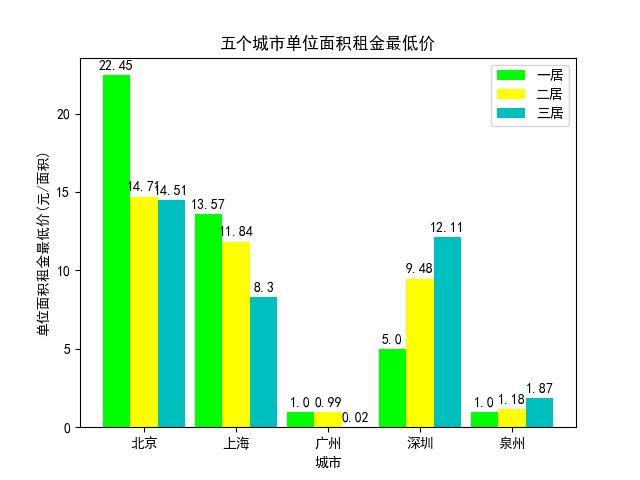


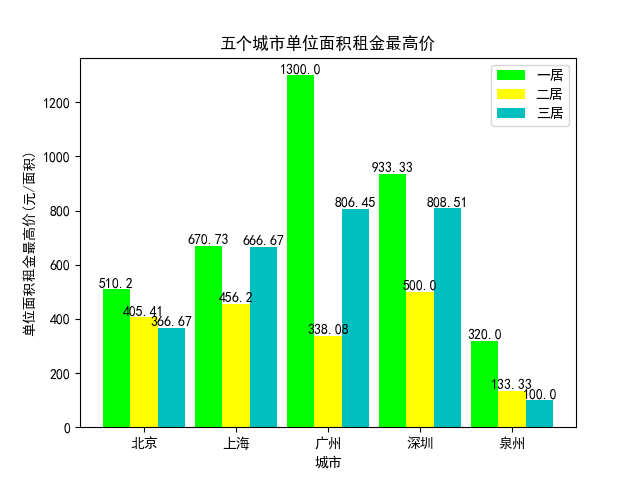


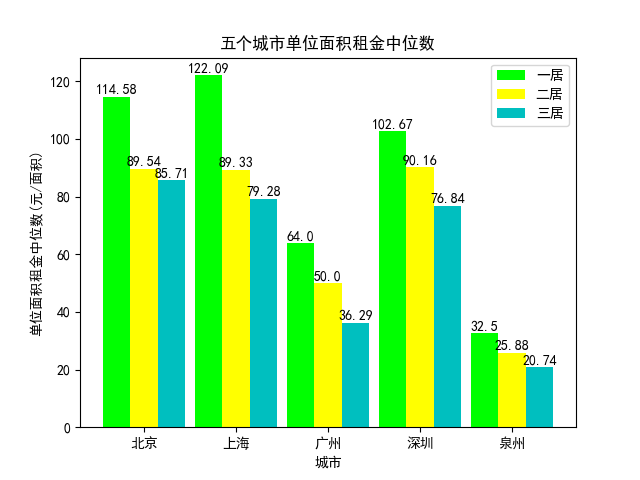


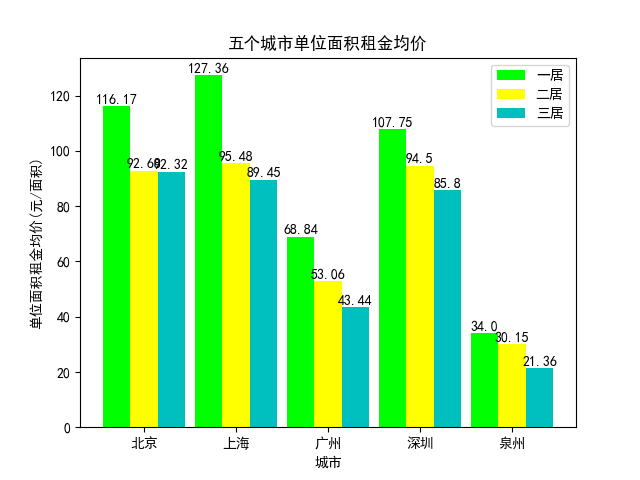


租金最高价和最低价可能存在极端值，所以只考虑均价和中位数。越多居的房子租金越高，北京和上海的租金很接近，是五个城市的最高。深圳第三，广州第四，泉州最低。



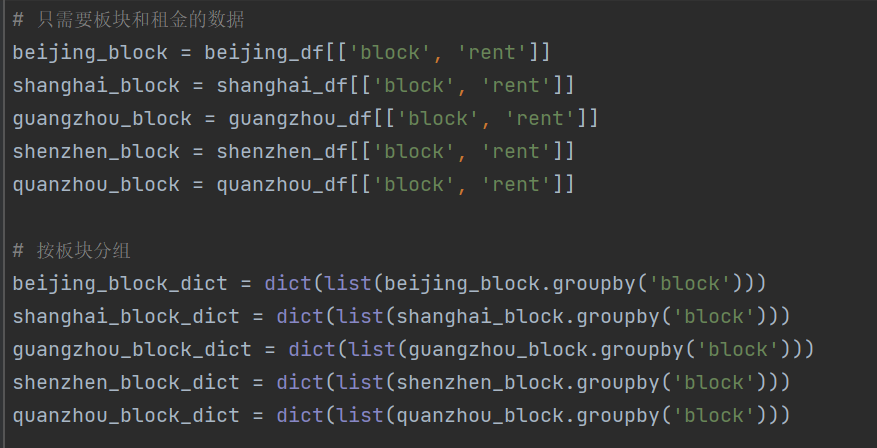






单位面积租金最高价和最低价可能存在极端值，所以只考虑均价和中位数。越少居的房子单位面积租金越高，北京、上海和深圳的租金很接近，是五个城市的最高。然后广州第四，泉州最低。

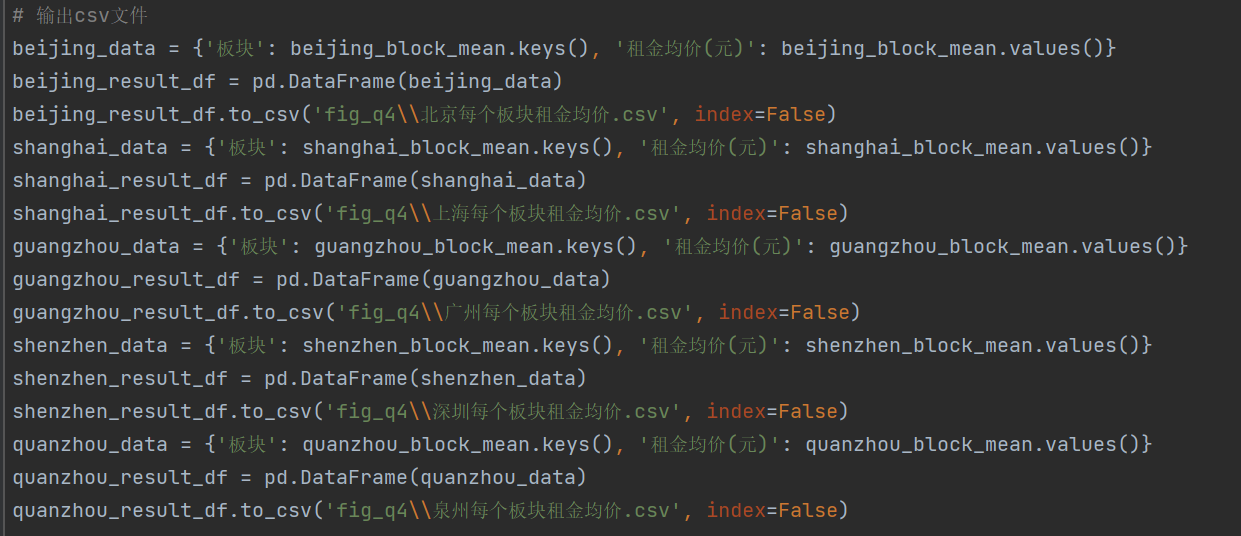
**计算和分析每个城市不同板块的均价情况**



将五个城市的租金数据按板块进行分组。



计算每个板块的租金均价，存在字典里。然后，按租金均价，从大到小，给字典排序。



将五个字典输出至csv文件。

北京西山板块租金均价最高，琉璃河板块最低。

上海黄浦滨江板块租金均价最高，柘林板块最低。

广州黄埔村板块租金均价最高，小楼镇板块最低。

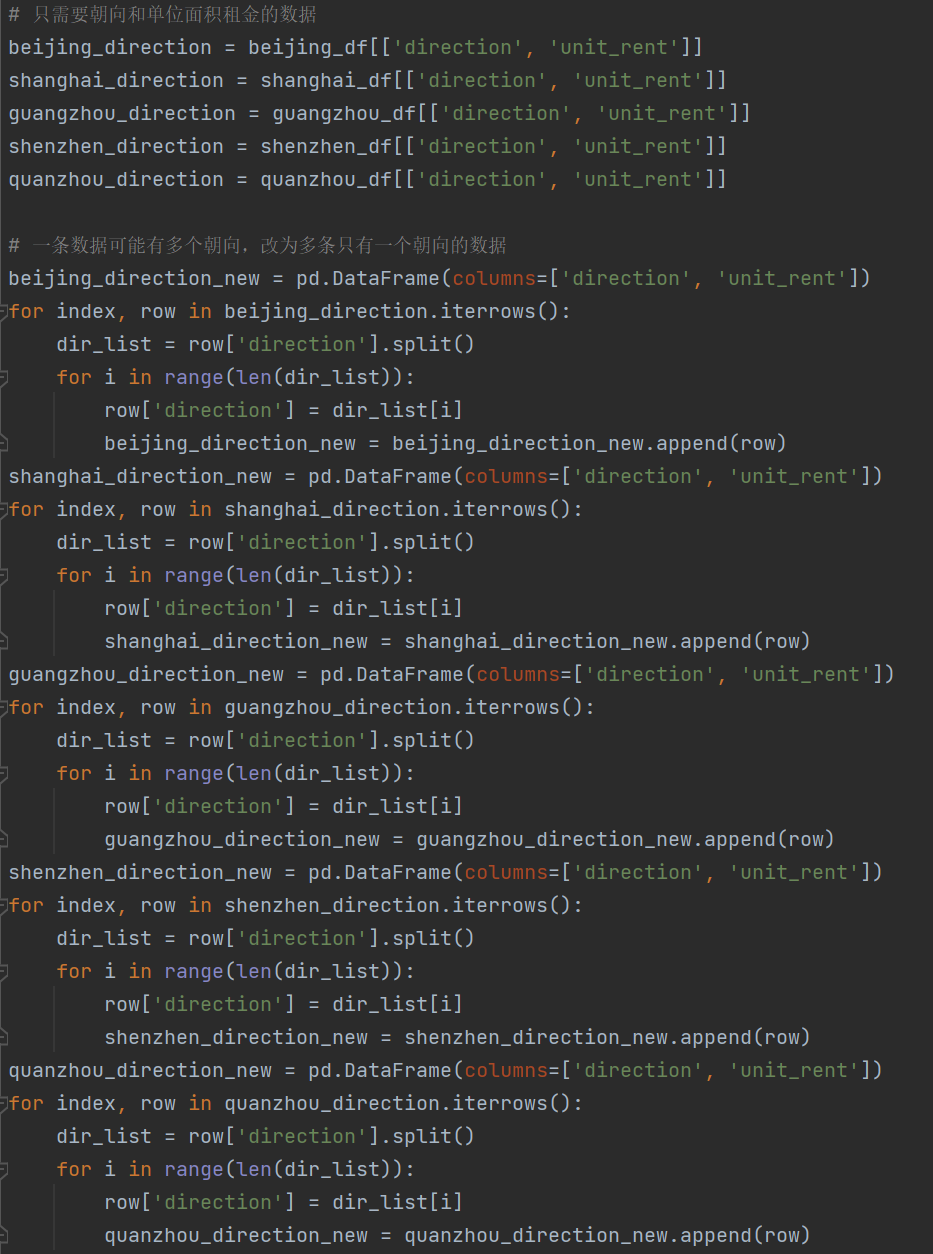
 

深圳深圳湾板块租金均价最高，坪地板块最低。上图最低的数据同时处于宝安中心和西乡，不是常见数据，不考虑。

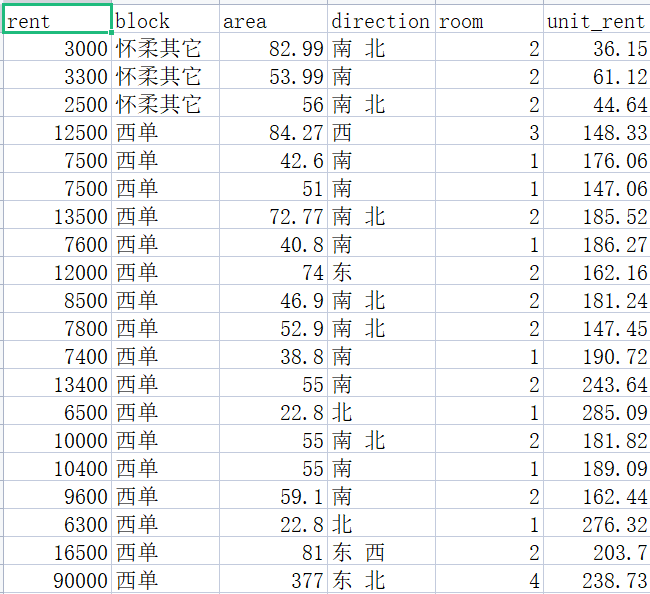
 

泉州西街板块租金均价最高，南安东最低。

**比较各个城市不同朝向的单位面积租金分布情况**



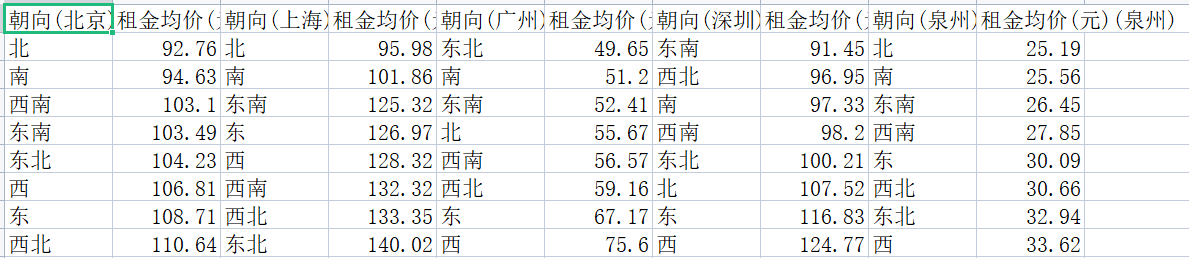
一条数据可能有多个朝向，改为多条只有一个朝向的数据。



例如这条数据改为一条朝向为南和一条为北的数据，其他项不变。



按照朝向进行分组，然后计算东、西、南、北、东北、东南、西北和西南八个方向的单位面积租金均价，保存在字典里，最后按值从小到大排序字典。



将字典输出至一个csv文件，内容如上。在五个城市的房子种，东和西朝向的单位租金均价都是偏高的，南和北朝向都是偏低。

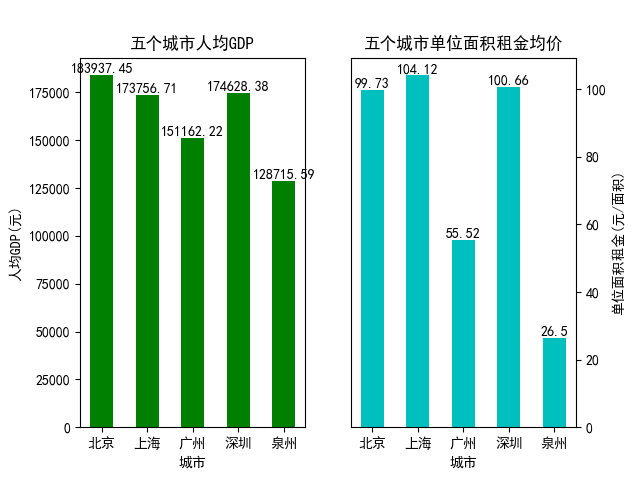
**查询各个城市人均GDP，分析并展示其和单位面积租金分布的关系**



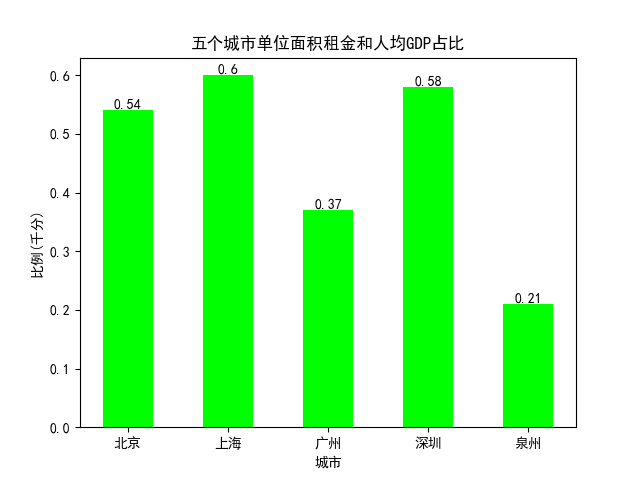
五个城市人均GDP分别如上。



计算单位面积租金均价/人均GDP的比例，然后画图。

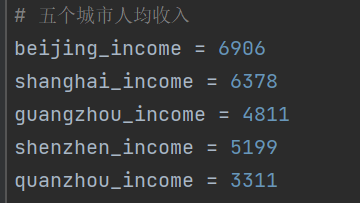


北京的人均GDP最高，深圳第二，上海第三，而上海的单位面积租金均价最高，深圳第二，北京第三。广州和泉州的人均GDP和单位面积租金均价都是第四第五。



根据上图，上海的比例最高，深圳第二，北京第三，广州第四，泉州最低。因此，在泉州租房的性价比最高，其次是广州。

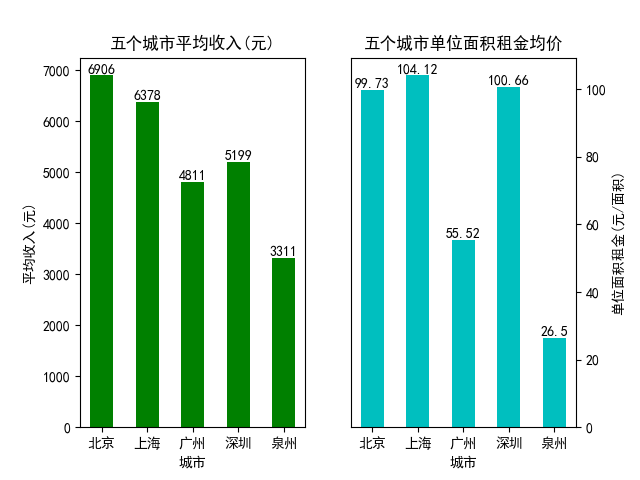
**查询各个城市的平均工资，分析并展示其和单位面积租金分布的关系**



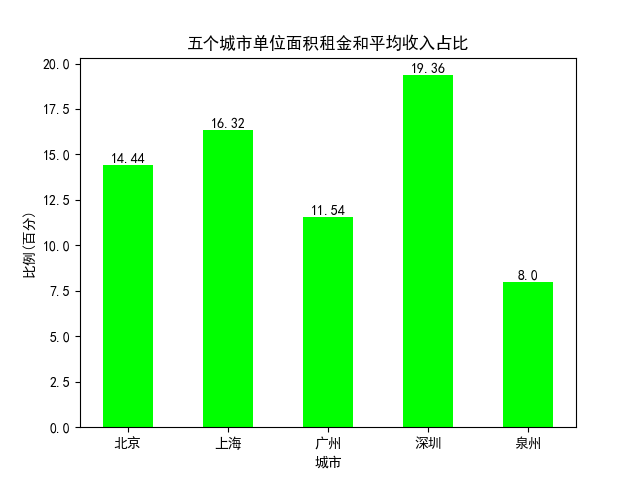
五个城市的平均收入如上。



计算单位面积租金均价/平均收入的比例，然后画图。



北京的平均收入最高，上海第二，深圳第三，而上海的单位面积租金均价最高，深圳第二，北京第三。广州和泉州的平均收入和单位面积租金均价都是第四第五。



根据上图，深圳的比例最高，上海第二，北京第三，广州第四，泉州最低。因此，在深圳租房的负担最高，其次是上海。