《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

一、 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？

二、 实验内容

1. 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。

```
99          # 发送请求，获取网页内容
100         response = get(url)
```

```
def ua():
    """随机获取一个浏览器用户信息"""

    user_agents = [
    'Mozilla/5.0 (Windows; U; Windows NT 5.1; it; rv:1.8.1.11) Gecko/20071127 Firefox/2.0.0.11',
    'Opera/9.25 (Windows NT 5.1; U; en)',
    'Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322; .NET CLR 2.0.50727)',
    'Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.5 (like Gecko) (Kubuntu)',
    'Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.8.0.12) Gecko/20070731 Ubuntu/dapper-security Firefox/1.5.0.12',
    'Lynx/2.8.5rel.1 libwww-FM/2.14 SSL-MM/1.4.1 GNUTLS/1.2.9',
    'Mozilla/5.0 (X11; Linux i686) AppleWebKit/535.7 (KHTML, like Gecko) Ubuntu/11.04 Chromium/16.0.912.77 Chrome/16.0.912.77 Safari/535.7',
    'Mozilla/5.0 (X11; Ubuntu; Linux i686; rv:10.0) Gecko/20100101 Firefox/10.0',
    ]

    agent = random.choice(user_agents)

    return {
        'User-Agent': agent
    }


def get(url):
    res = requests.get(url=url, headers = ua())
    return res.text
```

然后通过 BeautifulSoup 模块来匹配每一条租房信息的 div。

```
# 解析网页内容，使用beautiful soup
soup = BeautifulSoup(response, 'html.parser')
# 找到所有的房屋信息的div标签
divs = soup.find_all('div', class_='content__list--item')
```

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。

```
166    # 将列表转换为数据框
167    df = pd.DataFrame(data)
168
169    # 删除面积缺失的房屋数据
170    df = df.dropna(subset=['面积'])
171
172    # 将数据框保存为csv文件
173    df.to_csv('上海链家.csv', index=False)
```

2. 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：

```
# 读取每个城市的CSV文件，并调用get_stats函数，将结果添加到stats_df中
for city in ['北京', '上海', '广州', '金华', '深圳']:
    df = pd.read_csv(city + '链家.csv')
    stats = get_stats(df)
    stats_df = stats_df.append(stats, ignore_index=True)
```

```
# 定义一个函数，用于计算各种统计指标
def get_stats(df):
    # 计算租金的均价、最高价、最低价、中位数
    rent_mean = df['总价'].mean()
    rent_max = df['总价'].max()
    rent_min = df['总价'].min()
    rent_median = df['总价'].median()

    # 计算单位面积租金的均价、最高价、最低价、中位数
    area_mean = (df['总价']/df['面积']).mean()
    area_max = (df['总价']/df['面积']).mean()
    area_min = (df['总价']/df['面积']).mean()
    area_median = (df['总价']/df['面积']).mean()

    # 返回一个包含所有指标的字典
    return {'rent_mean': rent_mean, 'rent_max': rent_max, 'rent_min': rent_min, 'rent_median': rent_median,
            'area_mean': area_mean, 'area_max': area_max, 'area_min': area_min, 'area_median': area_median}
```

再将指标回传到一个新的 pandas 数据框中

```
# 创建一个空的数据框，用于存储各个城市的统计指标
stats_df = pd.DataFrame()
```

```
# 创建一个空的数据框，用于存储各个城市的统计指标
stats_df = pd.DataFrame()

# 读取每个城市的CSV文件，并调用get_stats函数，将结果添加到stats_df中
for city in ['北京', '上海', '广州', '金华', '深圳']:
    df = pd.read_csv(city + '链家.csv')
    stats = get_stats(df)
    stats_df = stats_df.append(stats, ignore_index=True)

# 为stats_df添加城市列
stats_df['城市'] = ['北京', '上海', '广州', '金华', '深圳']

# 设置城市列为索引
stats_df = stats_df.set_index('城市')
```

最后分别用两个横向柱状图表示。

```
# 绘制柱状图，比较各个城市的租金均价
plt.figure() # 增加这一行，创建一个新的图形
stats_df['rent_mean'].plot(kind='barh', figsize=(10, 6), rot=0)
plt.xlabel('租金均价(万)')
plt.ylabel('城市')
plt.title('各个城市的租金均价')
plt.show()

# 绘制柱状图，比较各个城市的单位面积租金均价
plt.figure() # 增加这一行，创建一个新的图形
stats_df['area_mean'].plot(kind='barh', figsize=(10, 6), rot=0)
plt.xlabel('单位面积租金均价(元/m²)')
plt.ylabel('城市')
plt.title('各个城市的单位面积租金均价')
plt.show()
```

3. 比较不同房型价格情况

读取 csv

```
# 读取链家.csv文件
beijing = pd.read_csv('北京链家.csv')
shanghai = pd.read_csv('上海链家.csv')
guangzhou = pd.read_csv('广州链家.csv')
shenzhen = pd.read_csv('深圳链家.csv')
jinhua = pd.read_csv('金华链家.csv')
```

单独提取几室

```
# 将房型变成几室
beijing[['几室','other']] = beijing['房型'].str.split('室|房',expand=True)
shanghai[['几室','other']] = shanghai['房型'].str.split('室|房',expand=True)
guangzhou[['几室','other']] = guangzhou['房型'].str.split('室|房',expand=True)
shenzhen[['几室','other']] = shenzhen['房型'].str.split('室|房',expand=True)
jinhua[['几室','other']] = jinhua['房型'].str.split('室|房',expand=True)
```

处理

```
# 只按照几室列分组，计算均价、最高价、最低价、中位数
beijing_stats = beijing.groupby('几室').agg({'总价': ['mean', 'max', 'min', 'median']})
shanghai_stats = shanghai.groupby('几室').agg({'总价': ['mean', 'max', 'min', 'median']})
guangzhou_stats = guangzhou.groupby('几室').agg({'总价': ['mean', 'max', 'min', 'median']})
shenzhen_stats = shenzhen.groupby('几室').agg({'总价': ['mean', 'max', 'min', 'median']})
jinhua_stats = jinhua.groupby('几室').agg({'总价': ['mean', 'max', 'min', 'median']})
```

合并输出

```
# 合并五个城市的统计结果
all_stats = pd.concat([beijing_stats, shanghai_stats, guangzhou_stats, shenzhen_stats, jinhua_stats], axis=1, keys=['北京', '上海', '广州', '深圳', '金华'])

all_stats.to_csv('question3_out.csv', index=False)
print(all_stats)
```

4. 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。

```
# 定义一个函数，用于计算各个区域的均价
def get_mean_price(df):
    # 计算总价(万)列的均值，作为区域的均价
    mean_price = df['总价'].mean()
    # 返回均价
    return mean_price

# 创建一个空的数据框，用于存储各个城市和各个区域的均价
price_df = pd.DataFrame()

cities=['北京', '上海', '广州', '金华', '深圳']

# 读取每个城市的CSV文件，并根据区域进行分组，调用get_mean_price函数，将结果添加到price_df中
for city in cities:
    df = pd.read_csv(city + '链家.csv')
    # 根据区域进行分组
    grouped = df.groupby('区域')
    # 对每个分组，调用get_mean_price函数，得到均价
    for group_name, group_df in grouped:
        mean_price = get_mean_price(group_df)
        # 为均价添加城市和区域列
        price_df = price_df.append({'城市': city, '区域': group_name, '均价': mean_price}, ignore_index=True)
```

用一个 for 循环循环输出每个城市各个区域的价格图

```
# 打印price_df
print(price_df)

# 对每个城市，绘制柱状图，比较各个区域的均价
for city in cities:
    # 选择该城市的数据
    city_df = price_df.loc[city]
    # 绘制柱状图
    city_df.plot(kind='bar', figsize=(10, 6), rot=0)
    plt.xlabel('区域')
    plt.ylabel('均价(万)')
    plt.title(city + '的各个区域的均价')
    plt.show()
```

5. 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

```
# 读取每个城市的CSV文件
cities = ['北京', '上海', '广州', '金华', '深圳']
dfs = [pd.read_csv(city + '链家.csv') for city in cities]
```

然后对朝向列进行关键词提取，顺便算出均价

```
# 对每个城市的数据框，根据朝向进行处理，提取出关键词
for df in dfs:
    df['单价'] = df['总价'] / (df['面积'])
    df['关键词'] = df['朝向'].apply(extract_keywords)
```

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。

```
# 定义一个函数，用于从朝向列中提取出东南西北四个关键词
def extract_keywords(x):
    keywords = ['东', '南', '西', '北']
    result = []
    for k in keywords:
        if k in x:
            result.append(k)
    return ''.join(result)
```

然后根据提取出的关键词进行分组求平均值。

```
# 对每个城市的数据框，按照关键词进行分组，计算单位面积租金的均值
means = [df.groupby('关键词')['单价'].mean() for df in dfs]

# 对每个城市的数据框，将关键词列中的多个关键词拆分成多行
exploded = [df.explode('关键词') for df in dfs]
```

最后绘制出每个城市各个朝向组合均价。

```
# 对每个城市，绘制箱线图，比较不同关键词的单位面积租金的分布情况
for city, mean, exp in zip(cities, means, exploded):
    # 绘制箱线图
    print(mean)
    mean.plot(kind='bar', figsize=(10, 6), rot=0)
    plt.xlabel('关键词')
    plt.ylabel('单位面积租金(元/㎡)')
    plt.title(city + '的不同关键词的单位面积租金的分布情况')
    plt.show()
```

6. 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。

```
GDP_dict = {
    "北京" : 19.00,
    "上海" : 18.04,
    "广州" : 15.36,
    "金华" : 7.8,
    "深圳" : 18.33
}

ECO_dict = {
    "北京" : 7.74,
    "上海" : 7.96,
    "广州" : 7.68,
    "金华" : 5.80,
    "深圳" : 7.27
}

area_mean_dict = {}
```

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）

```
# 创建一个空的数据框，用于存储各个城市的统计指标
stats_df = pd.DataFrame()

# 读取每个城市的CSV文件，并调用get_stats函数，将结果添加到stats_df中
for city in ['北京', '上海', '广州', '金华', '深圳']:
    df = pd.read_csv(city + '链家.csv')
    area_mean_dict[city] = get_area_mean(df)
    temp = area_mean_dict[city]
    print(city + "均价为: " + str(temp) + "相对为GDP" + str(temp/GDP_dict[city]) + "倍，相对为人均收入" + str(temp/ECO_dict[city]) + "倍")
```

get_area_mean 函数

```
27    # 定义一个函数，用于计算面积均价
28    def get_area_mean(df):
29        # 计算单位面积租金的均价
30        area_mean = (df['总价']/df['面积']).mean()
31
32        # 返回一个包含单位面积租金的均价
33        return area_mean
```

最后读入三个字典绘制出柱状图

```
45    #将每个城市的面积均价添加到stats_df中
46    stats_df['area_mean'] = [stats for city, stats in area_mean_dict.items()]
47    #将每个城市的GDP和人均收入添加到stats_df中
48    stats_df['GDP'] = GDP_dict.values()
49    stats_df['ECO'] = ECO_dict.values()
50    #设置stats_df的索引为城市名称
51    stats_df.index = GDP_dict.keys()
52    #绘制柱状图，显示每个城市的面积均价，GDP和人均收入
53    ax = stats_df.plot.bar(rot=0, figsize=(15, 10), title='五个城市的面积均价，GDP和人均收入')
54    #设置x轴和y轴的标签
55    ax.set_xlabel('城市')
56    ax.set_ylabel('数值')
57    #显示图形
58    plt.show()
```

7. 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

```
pg-1
continueFlag = True
for city in citys:
    # 定义起始网页的url
    url = 'https://' + city + '.lianjia.com/ershoufang/pg' + str(pg) + '/'
    # 定义一个循环，用于遍历所有的网页
    while continueFlag:

        print(pg)
        pg += 1
        # 发送请求，获取网页内容
        response = get(url)
        # 解析网页内容，使用beautiful soup
        soup = BeautifulSoup(response, 'html.parser')
        # 找到所有的房屋信息的div标签
        divs = soup.find_all('div', class_='clear LOGCLICKDATA')
        # 遍历每个div标签，提取房屋信息
        for div in divs:
            # 提取面积均价
            price = div.find('div', class_='unitPrice').text.strip()
            if price:
                price = price.text.strip()
            else:
                continue
            price = get_price_midpoint(price)
            print(price)
            # 将房屋信息以字典的形式添加到列表中
            data.append({
                '面积均价': price
            })
        # 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://'+ city +'.lianjia.com/ershoufang/pg' + str(pg) + '/'
        #url = None
```

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。

```
    # 将列表转换为数据框
    df = pd.DataFrame(data)
    price_mean = df['面积均价'].mean()
    if city == 'bj':
        city_ch = '北京'
    elif city == 'sh':
        city_ch = '上海'
    elif city == 'gz':
        city_ch = '广州'
    elif city == 'sz':
        city_ch = '深圳'
    else:
        city_ch = '金华'

    area_price_dict[city_ch] = price_mean

result = pd.DataFrame.from_dict(area_price_dict, orient='index', columns=['价格'])
# 将数据框保存为csv文件
result.to_csv('question8_out.csv')
```

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

```
# 读取二手房均价文件
df = pd.read_csv("question8_out.csv", index_col=0)
# 转换为字典
sencond_hand_dict = df.to_dict(orient="index")
# 去掉价格这一层的字典
for city in sencond_hand_dict:
    sencond_hand_dict[city] = sencond_hand_dict[city]["价格"]
```

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。

```
# 读取每个城市的CSV文件，并调用get_stats函数，将结果添加到stats_df中
for city in ['北京', '上海', '广州', '深圳', '金华']:
    df = pd.read_csv(city + '链家.csv')
    area_price_dict[city] = df['总价'].mean()/df['面积'].mean()
```

先输出两者数值数据

```
print("租房面积均价")
print(area_price_dict)
print("二手房面积均价")
print(sencond_hand_dict)
```

然后将二手房数据缩小 1000 倍方便后面比较

```

for city in ['北京', '上海', '广州', '深圳', '金华']:
    # 将价格缩小1000倍方便比较
    sencond_hand_dict[city] = sencond_hand_dict[city]/1000
```

画出柱状图

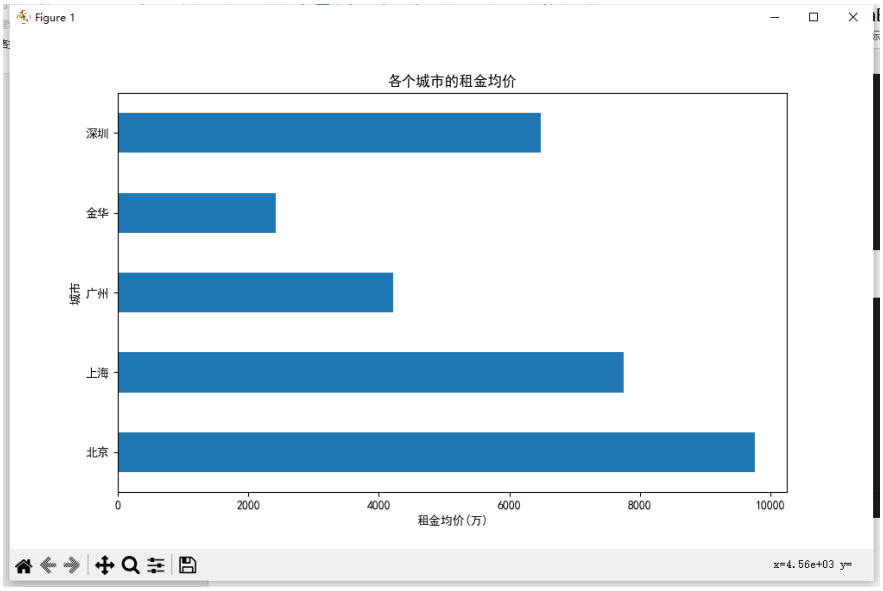
```
#将每个城市的面积均价添加到stats_df中
stats_df['平方米价格'] = sencond_hand_dict.values()
#将每个城市的GDP和人均收入添加到stats_df中
stats_df['面积租金'] = area_price_dict.values()
#设置stats_df的索引为城市名称
stats_df.index = area_price_dict.keys()
#绘制柱状图，显示每个城市的面积均价，GDP和人均收入
ax = stats_df.plot.bar(rot=0, figsize=(15, 10), title='五个城市的面积租金与缩小1000倍的平方米价格')
#设置x轴和y轴的标签
ax.set_xlabel('城市')
ax.set_ylabel('数值')
#显示图形
plt.show()
```

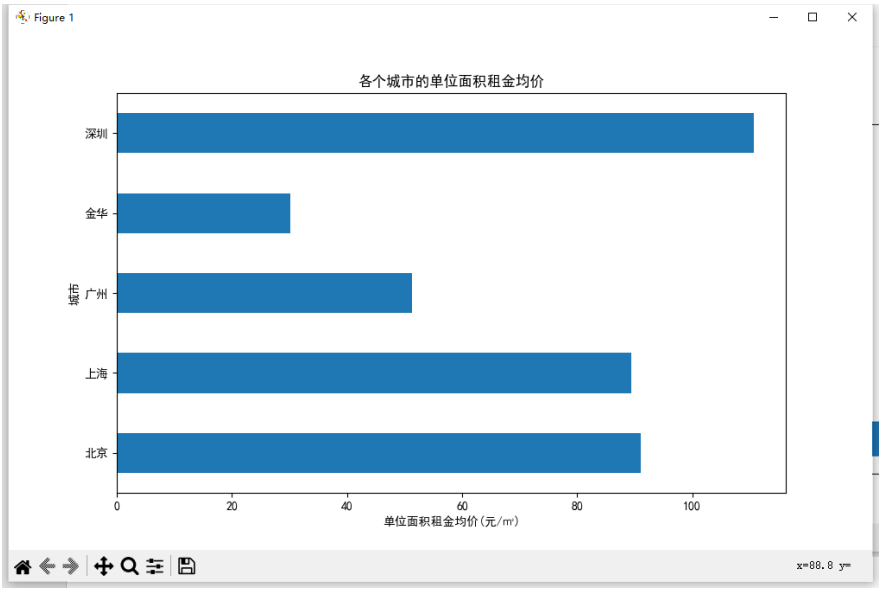
三、 实验结果

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

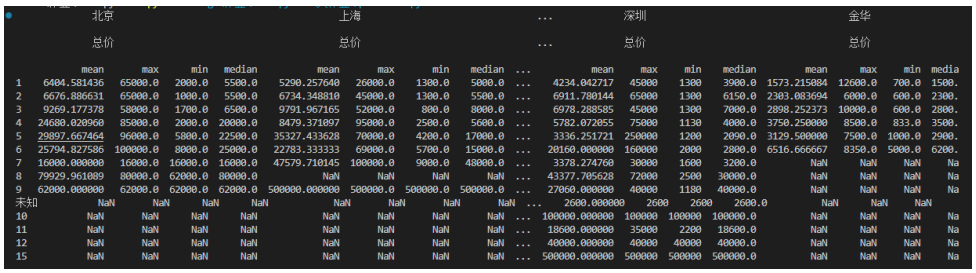




城市	rent_mean	rent_max	rent_min	rent_median	area_mean	area_max	area_min	area_median
北京	9764.953290	100000.0	1000.0	6300.0	91.031059	360.000000	12.500000	79.166667
上海	7756.516644	500000.0	800.0	5500.0	89.330975	464.285714	5.714286	80.769231
广州	4224.932646	80000.0	150.0	3000.0	51.337854	544.217687	6.000000	41.379310
金华	2414.620710	12600.0	600.0	2265.0	30.102152	86.666667	1.000000	29.069767
深圳	6488.364184	500000.0	1130.0	5000.0	110.718988	692.307692	14.606742	90.361446

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



	北京				上海				...	深圳				金华			
	总价				总价				...	总价				总价			
	mean	max	min	median	mean	max	min	median	...	mean	max	min	median	mean	max	min	media
1	6404.581436	65000.0	2000.0	5500.0	5290.257640	26000.0	1300.0	5000.0	...	4234.042717	45000	1300	3900.0	1573.215084	12600.0	700.0	1500.
2	6676.886631	65000.0	1000.0	5500.0	6734.348810	45000.0	1300.0	5500.0	...	6911.780144	65000	1300	6150.0	2303.083694	6000.0	600.0	2300.
3	9269.177378	58000.0	1700.0	6500.0	9791.967165	52000.0	800.0	8000.0	...	6978.288585	45000	1300	7000.0	2898.252373	10000.0	600.0	2800.
4	24680.020960	85000.0	2000.0	20000.0	8479.371897	95000.0	2500.0	5600.0	...	5782.072055	75000	1130	4000.0	3750.250000	8500.0	833.0	3500.
5	29897.667464	96000.0	5800.0	22500.0	35327.433628	70000.0	4200.0	17000.0	...	3336.251721	250000	1200	2090.0	3129.500000	7500.0	1000.0	2900.
6	25794.827586	100000.0	8000.0	25000.0	22783.333333	69000.0	5700.0	15000.0	...	20160.000000	160000	2000	2800.0	6516.666667	8350.0	5000.0	6200.
7	16000.000000	16000.0	16000.0	16000.0	47579.710145	100000.0	9000.0	48000.0	...	3378.274760	30000	1600	3200.0	NaN	NaN	NaN	Na
8	79929.961089	80000.0	62000.0	80000.0	NaN	NaN	NaN	NaN	...	43377.705628	72000	2500	30000.0	NaN	NaN	NaN	Na
9	62000.000000	62000.0	62000.0	62000.0	500000.000000	500000.0	500000.0	500000.0	...	27060.000000	40000	1180	40000.0	NaN	NaN	NaN	Na
未知	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	2600.000000	2600	2600	2600.0	NaN	NaN	NaN	
10	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	100000.000000	100000	100000	100000.0	NaN	NaN	NaN	Na
11	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	18600.000000	35000	2200	18600.0	NaN	NaN	NaN	Na
12	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	40000.000000	40000	40000	40000.0	NaN	NaN	NaN	Na
15	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	500000.000000	500000	500000	500000.0	NaN	NaN	NaN	Na

Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9	Column10	Column11	Column12
北京	北京	北京	北京	上海	上海	上海	上海	广州	广州	广州	广州
总价	总价	总价	总价	总价	总价	总价	总价	总价	总价	总价	总价
mean	max	min	median	mean	max	min	median	mean	max	min	median
6404.581435672177	65000.0	2000.0	5500.0	5290.257640232108	26000.0	1300.0	5000.0	2970.5395563533784	55000.0	150.0	2500.0
6676.886631071484	65000.0	1000.0	5500.0	6734.348810079328	45000.0	1300.0	5500.0	3821.3652901785713	33000.0	800.0	3500.0
9269.17737789203	58000.0	1700.0	6500.0	9791.967164547586	52000.0	800.0	8000.0	3729.336675553495	80000.0	800.0	2607.0
24680.020960365855	85000.0	2000.0	20000.0	8479.371097234613	95000.0	2500.0	5600.0	4787.678385867683	36000.0	900.0	3500.0
29897.66746411483	96000.0	5800.0	22500.0	35327.433628318584	70000.0	4200.0	17000.0	7139.39120879120 9	68100.0	1144.0	4000.0
25794.827586206895	100000.0	8000.0	25000.0	22783.333333333332	69000.0	5700.0	15000.0	21889.479512735328	60000.0	4000.0	30000.0
16000.0	16000.0	16000.0	16000.0	47579.710144927536	100000.0	9000.0	48000.0	13800.0	13800.0	13800.0	13800.0
79929.96108949416	80000.0	62000.0	80000.0					25333.333333333332	50000.0	8000.0	18000.0
62000.0	62000.0	62000.0	62000.0	500000.0	500000.0	500000.0	500000.0	22666.666666666668	35000.0	7000.0	26000.0
								805.5555555555555	1600.0	300.0	600.0

M	N	O	P	Q	R	S	T
Column13	Column14	Column15	Column16	Column17	Column18	Column19	Column20
深圳	深圳	深圳	深圳	金华	金华	金华	金华
总价	总价	总价	总价	总价	总价	总价	总价
mean	max	min	median	mean	max	min	median
4234.042716615011	45000	1300	3900.0	1573.2150837988827	12600.0	700.0	1500.0
6911.780143805309	65000	1300	6150.0	2303.083694083694	6000.0	600.0	2300.0
6978.2885845773935	45000	1300	7000.0	2898.2523734177216	10000.0	600.0	2800.0
5782.0720554773	75000	1130	4000.0	3750.25	8500.0	833.0	3500.0
3336.2517211703957	250000	1200	2090.0	3129.5	7500.0	1000.0	2900.0
20160.0	160000	2000	2800.0	6516.666666666667	8350.0	5000.0	6200.0
3378.2747603833864	30000	1600	3200.0				
43377.705627705625	72000	2500	30000.0				
27060.0	40000	1180	40000.0				
2600.0	2600	2600	2600.0				
100000.0	100000	100000	100000.0				
18600.0	35000	2200	18600.0				
40000.0	40000	40000	40000.0				
500000.0	500000	500000	500000.0				

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

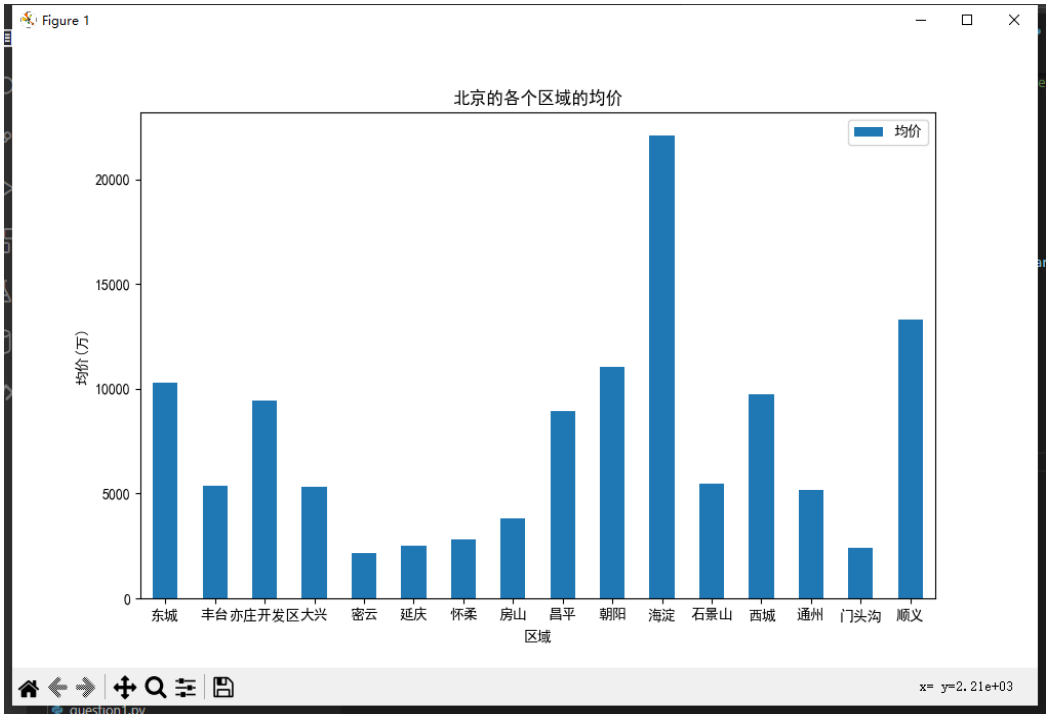


Figure 1
北京的各个区域的均价
均价
20000
15000
10000
5000
0
均价（万）
东城
丰台
亦庄开发区
大兴
密云
延庆
怀柔
房山
昌平
朝阳
海淀
石景山
西城
通州
门头沟
顺义
区域
x= y=2.21e+03

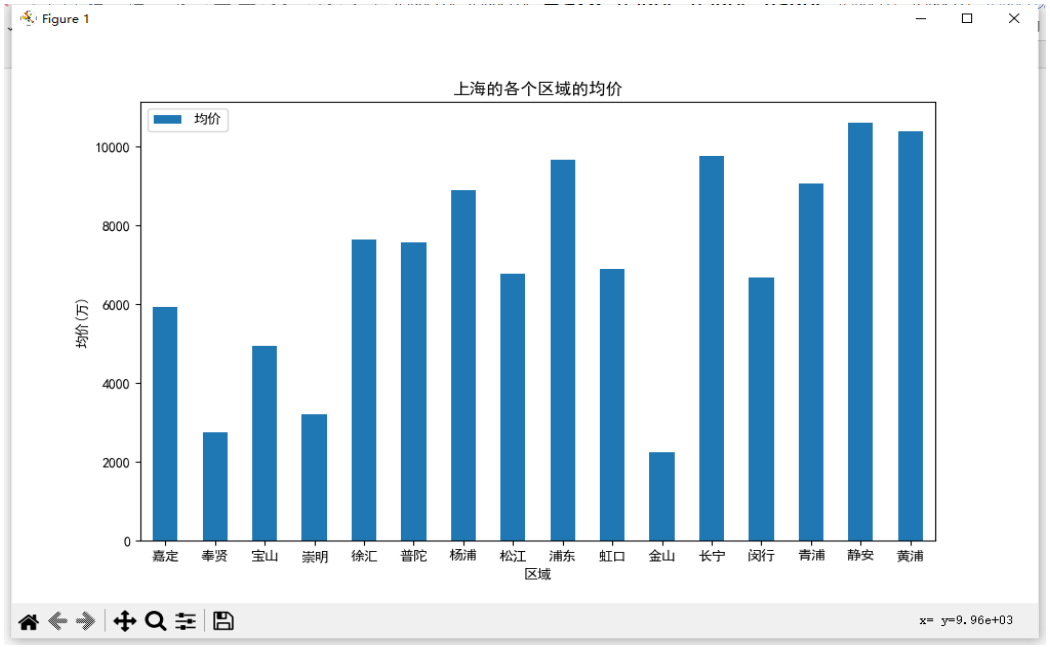


Figure 1
上海的各个区域的均价
均价
10000
8000
6000
4000
2000
0
均价（万）
嘉定
奉贤
宝山
崇明
徐汇
普陀
杨浦
松江
浦东
虹口
金山
长宁
闵行
青浦
静安
黄浦
区域
x= y=9.96e+03

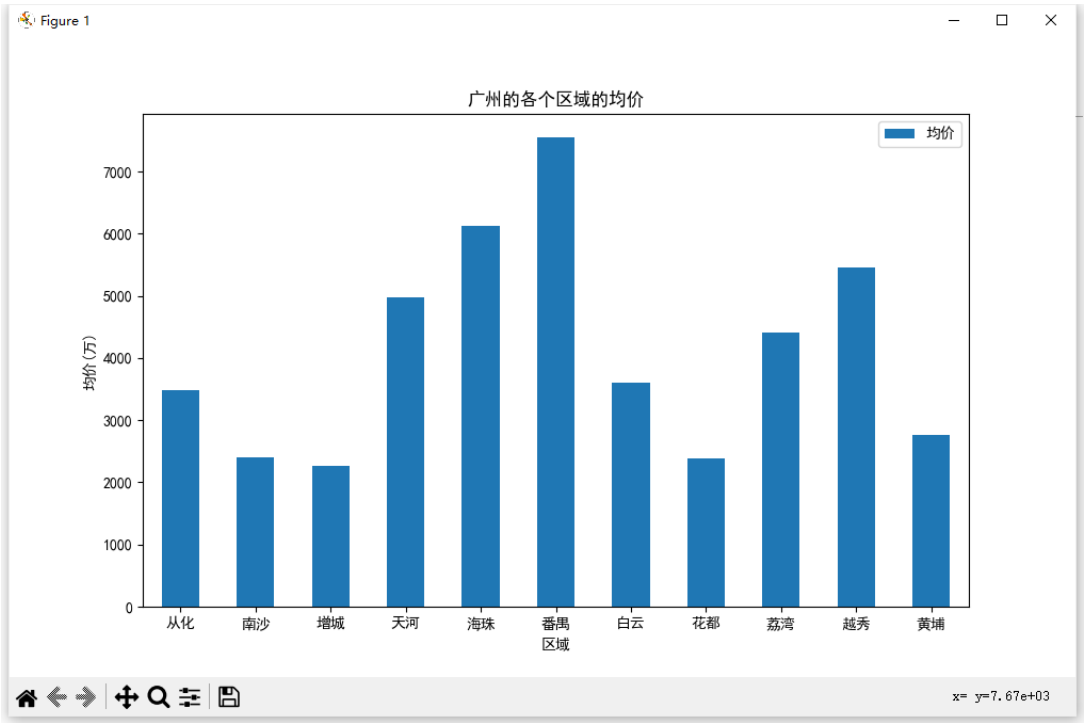


Figure 1
广州的各个区域的均价
均价
均价(万)
0
1000
2000
3000
4000
5000
6000
7000
从化
南沙
增城
天河
海珠
番禺
白云
花都
荔湾
越秀
黄埔
区域
x= y=7.67e+03

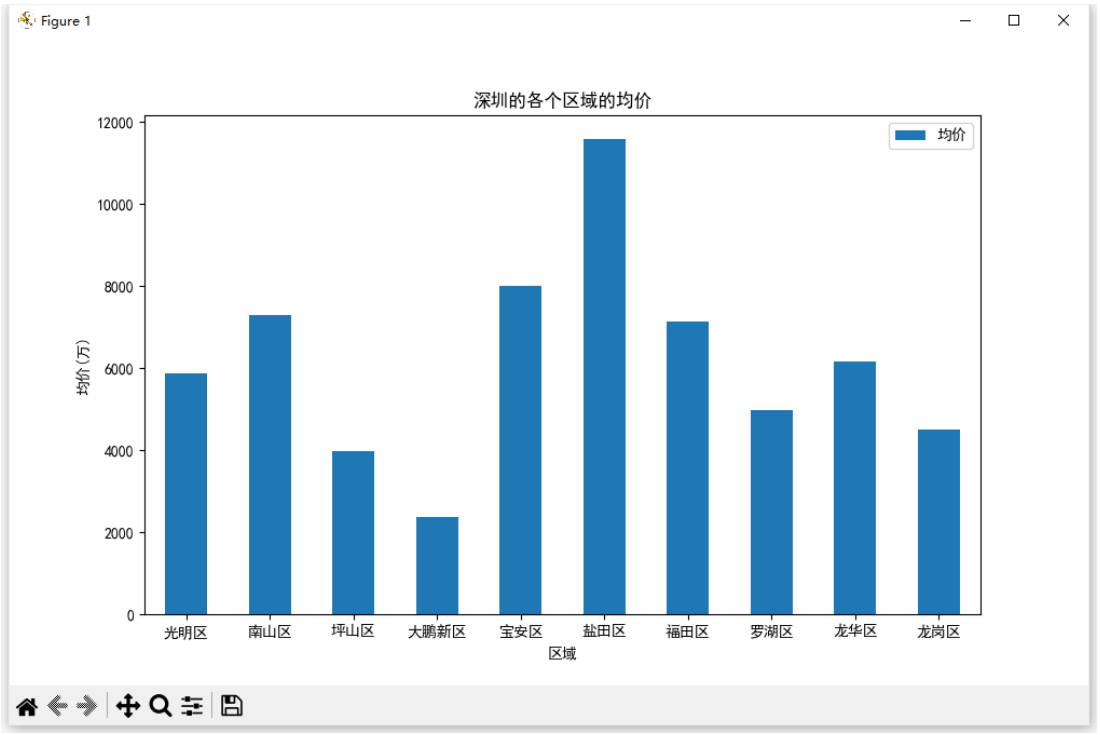


Figure 1
深圳的各个区域的均价
均价
均价(万)
0
2000
4000
6000
8000
10000
12000
光明区
南山区
坪山区
大鹏新区
宝安区
盐田区
福田区
罗湖区
龙华区
龙岗区
区域

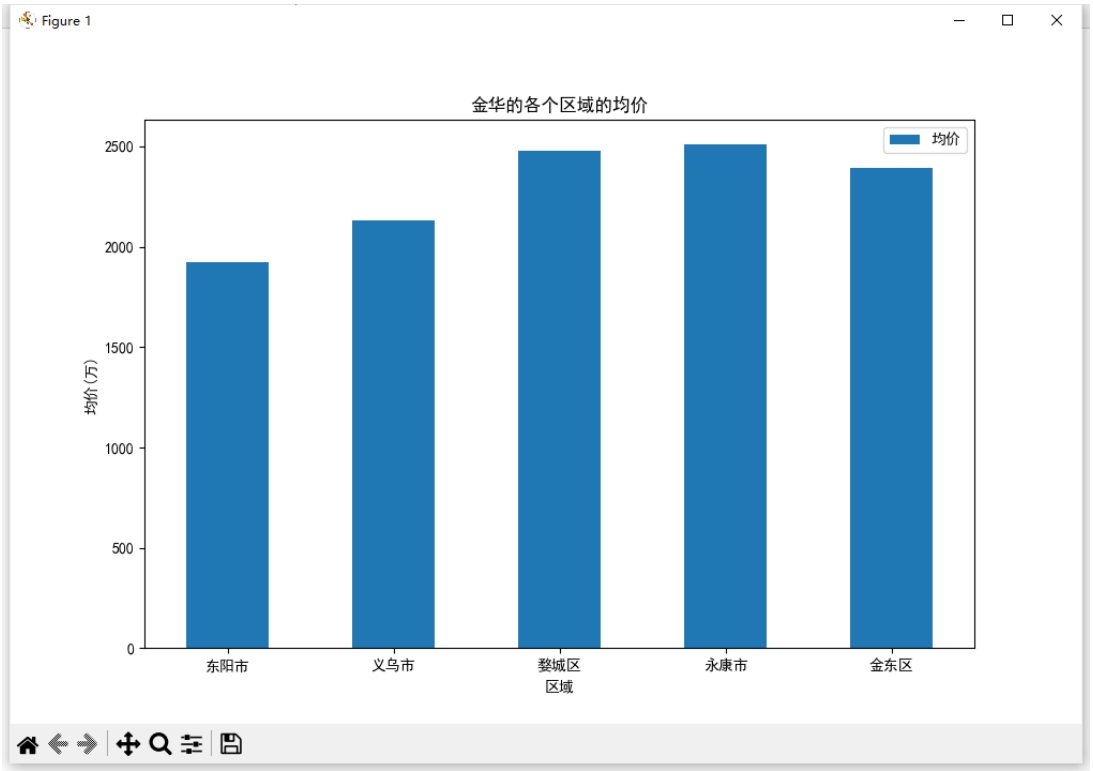


Figure 1
金华的各个区域的均价
均价
均价（万）
2500
2000
1500
1000
500
0
东阳市
义乌市
婺城区
永康市
金东区
区域

5. 朝向情况

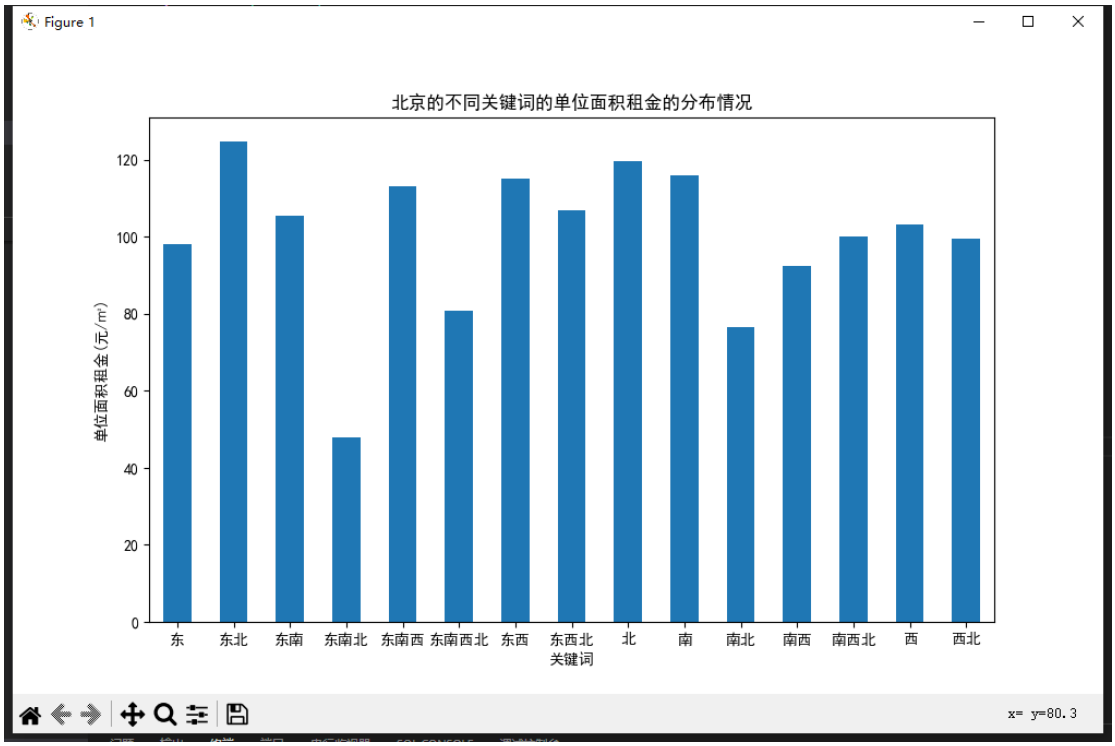


Figure 1
北京的不同关键词的单位面积租金的分布情况
单位面积租金(元/㎡)
0
20
40
60
80
100
120
东
东北
东南
东南北
东南西
东南西北
东西
东西北
北
南
南北
南西
南西北
西
西北
关键词
x= y=80.3

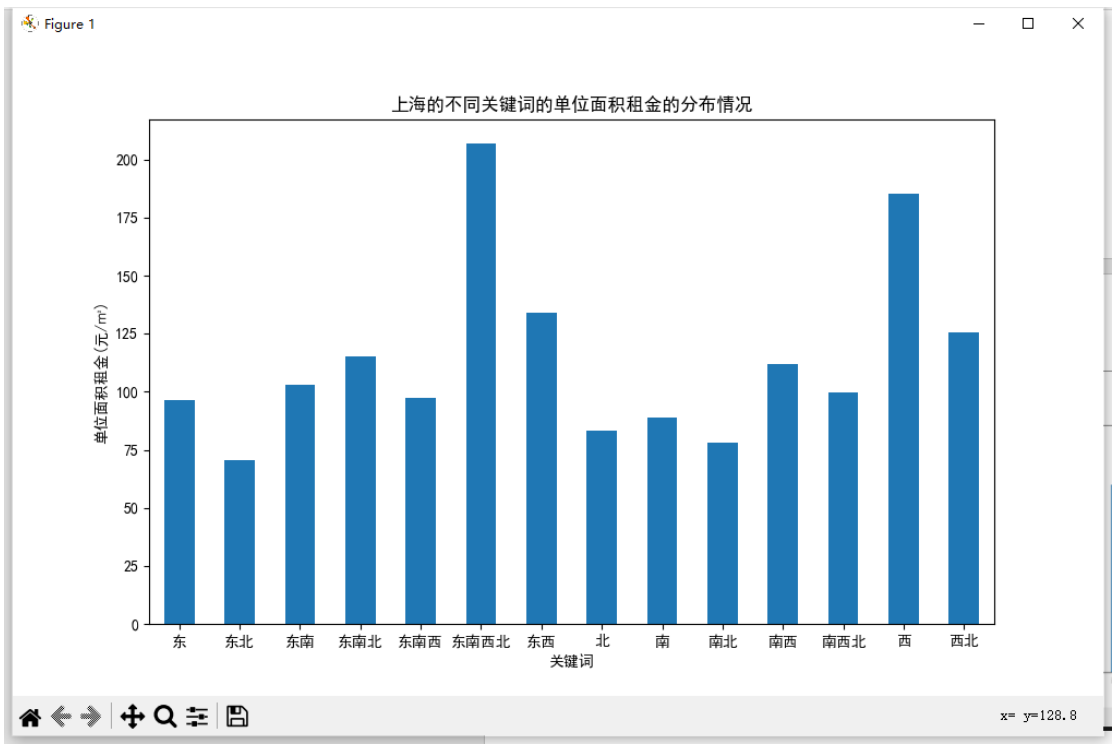


Figure 1
上海的不同关键词的单位面积租金的分布情况
单位面积租金(元/㎡)
0
25
50
75
100
125
150
175
200
东
东北
东南
东南北
东南西
东南西北
东西
北
南
南北
南西
南西北
西
西北
关键词
x= y=128.8

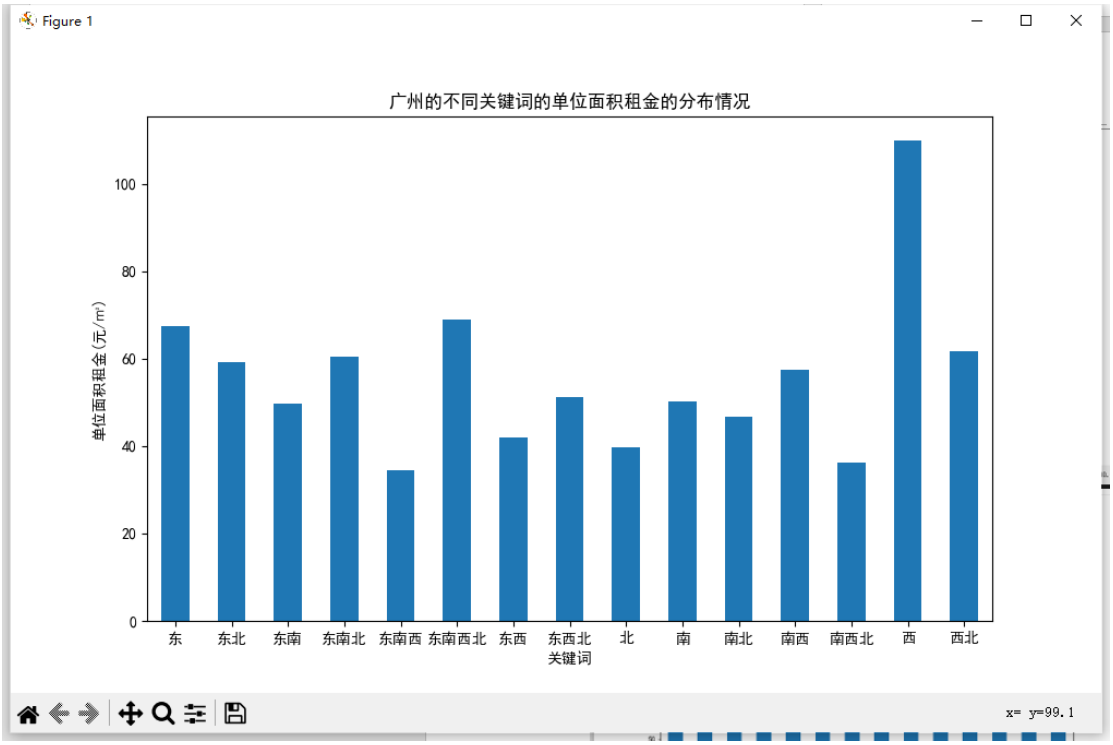


Figure 1
广州的不同关键词的单位面积租金的分布情况
单位面积租金(元/㎡)
0
20
40
60
80
100
东
东北
东南
东南北
东南西
东南西北
东西
东西北
北
南
南北
南西
南西北
西
西北
关键词
x= y=99.1

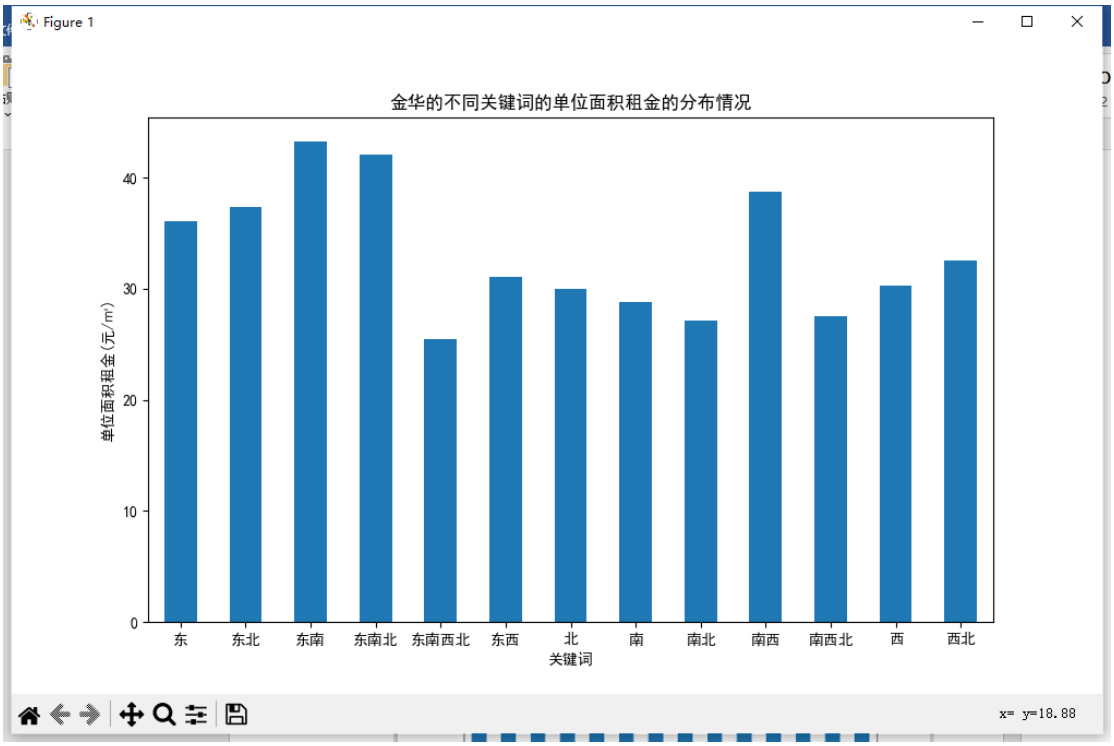
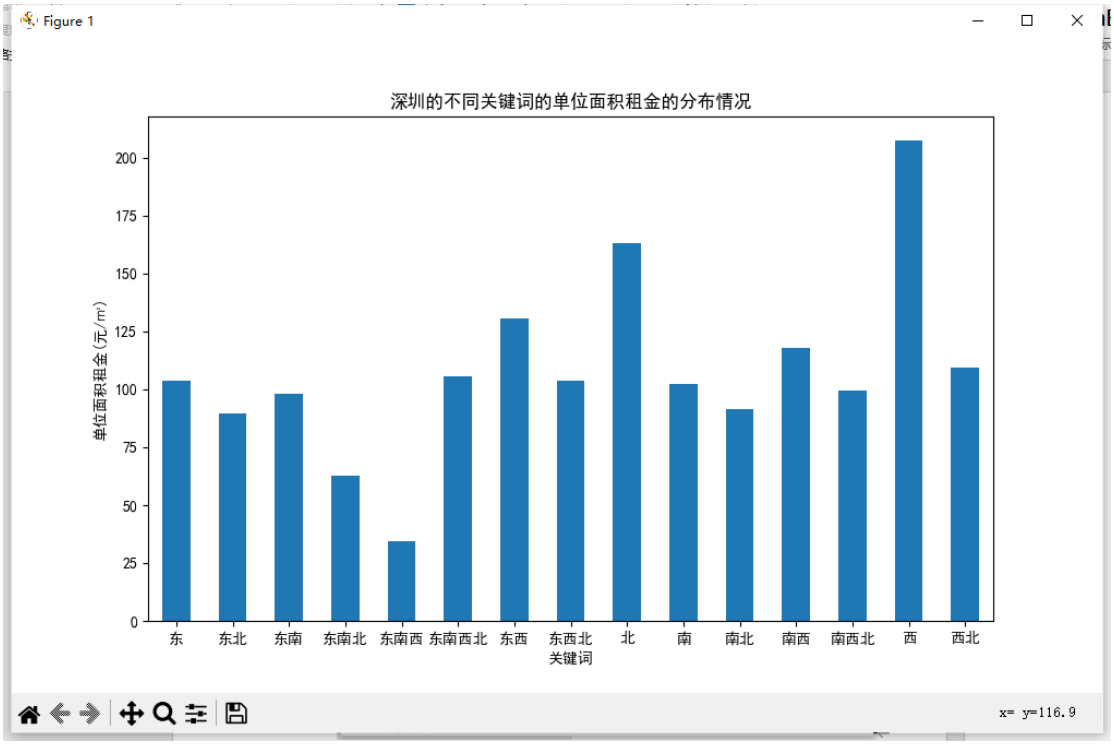


Figure 1
金华的不同关键词的单位面积租金的分布情况
单位面积租金(元/㎡)
0
10
20
30
40
东
东北
东南
东南北
东南西北
东西
北
南
南北
南西
南西北
西
西北
关键词
x= y=18.88



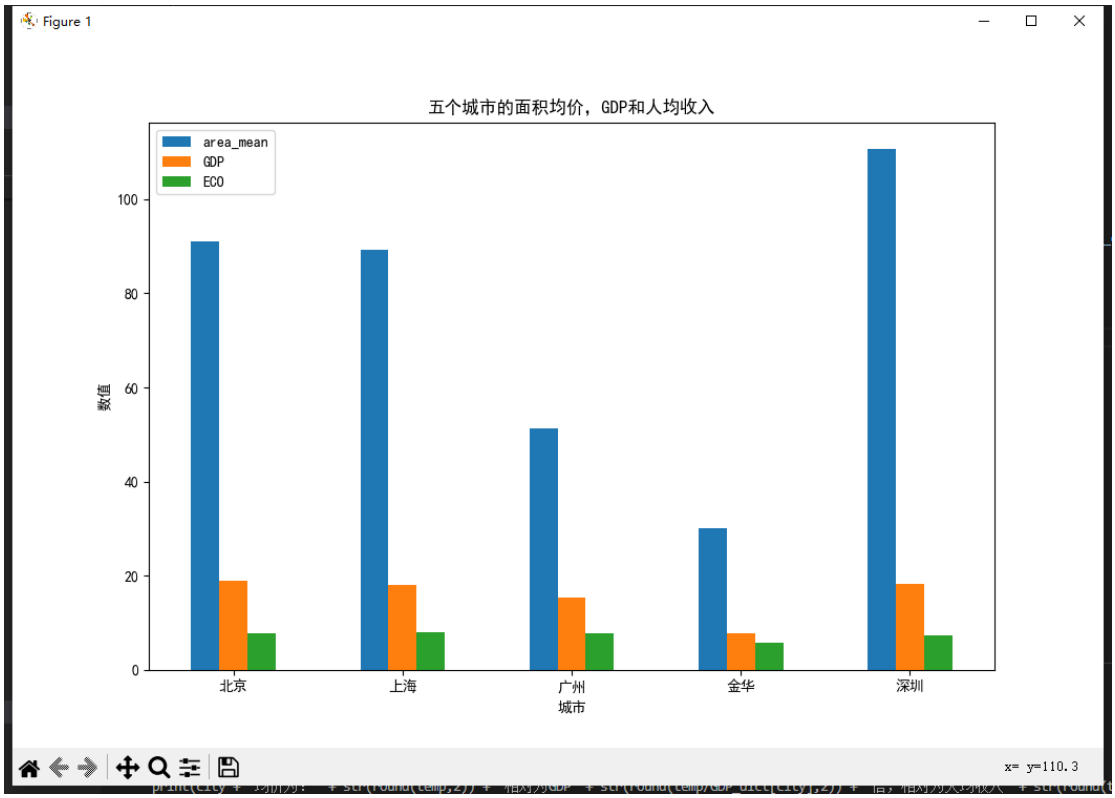
与预期的不太一样的是，一线城市中有三个城市（上广深）西面朝向的房价出奇的高，根据分析应该是恰恰是因为消费者倾向购买南北朝向的房子，所以开发商也倾向建造南北的房子，导致西数据量小。而如果开发的是大户型房子，需要充足的光源照亮房内，而朝西能最大限度接受阳光，所以这个朝向的房型都偏大，价格就贵。

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



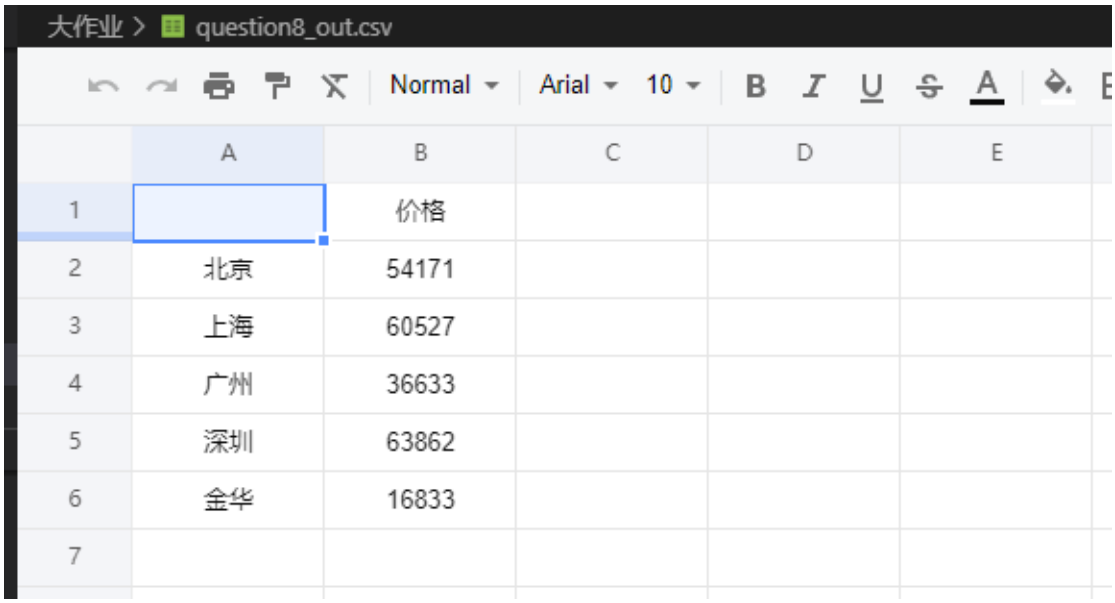

```
PS G:\作业\BUPT-python> python -u "g:\作业\BUPT-python\大作业\question6and7.py"
北京均价为：91.03105852958193相对为GDP4.791108343662207倍，相对为人均收入11.7611186730726倍
上海均价为：89.3309745161164相对为GDP4.951827855660555倍，相对为人均收入11.222484235693015倍
广州均价为：51.33785397665692相对为GDP3.3423082016052685倍，相对为人均收入6.684616403210537倍
金华均价为：30.102152145301954相对为GDP3.859250275038712倍，相对为人均收入5.190026231948613倍
深圳均价为：110.71898817775234相对为GDP6.040315776200346倍，相对为人均收入15.22957196392742倍
```

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

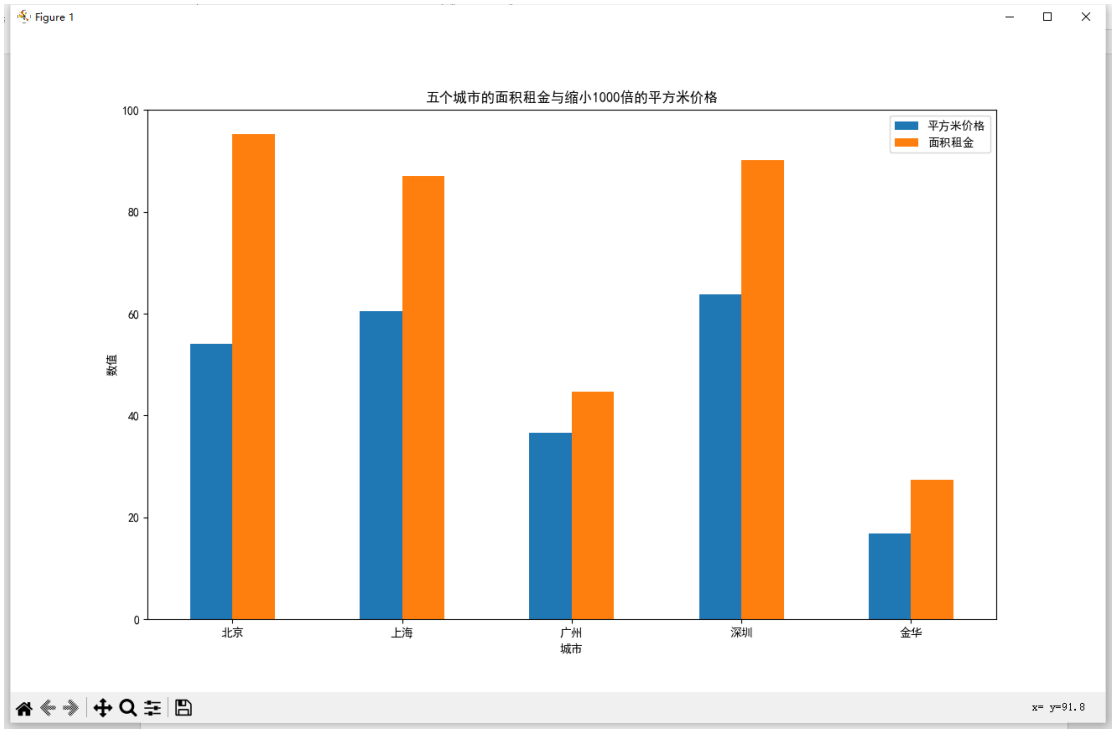
爬取到二手房数据后所处理出的二手房均价 csv



	A	B	C	D	E
1		价格			
2	北京	54171			
3	上海	60527			
4	广州	36633			
5	深圳	63862			
6	金华	16833			
7					

数据处理与展示的结果

```
PS G:\作业\BUPT-python> python -u "g:\作业\BUPT-python\大作业\question8_2.py"
租房面积均价
{'北京': 95.35432292314461, '上海': 86.98884013857591, '广州': 44.76929926321857, '深圳': 90.11505862128561, '金华': 27.463297224010187}
二手房面积均价
{'北京': 54171, '上海': 60527, '广州': 36633, '深圳': 63862, '金华': 16833}
```



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。