Python-爬虫技术在人文社科领域研究中的应用

开放性创新实验报告

组　　名：第三组

同学　一：张佳玲

同学　二：刘　铮

同学　三：柏申苏

联系电话：

实验题目：用 python 爬虫分析房价与 GDP 之间关系

指导教师：黄　伟

一.实验目的

自金融危机以来，中国国内**房价**的波动性极其剧烈，同时中国房价的增长率在整体大幅度上涨的同时，各省份、各地区内部房价上涨情况更加复杂多变。理清房价上涨的影响因素是一项非常必要且关键的任务。

GDP 是社会与城市发展的核心评估指标，可以衡量区域的消费、生产等各个维度的经济指标。某种程度上代表了个人或群体在其中的发展机会，因此我们猜测猜测 GDP 高的城市会吸引人口流入，进一步促进城市就业、医疗、教育等优质资源的增加。

中国指数研究院数据显示 25～44 岁年龄段人口占购房人口总数的比例高达 75%，该年龄段的人群正是人口流动的主要群体，省会城市通过其高 GDP 所代表的良好发展机会吸引了大量的省内人才引入，支撑了近些年来省会城市新经济产业的快速发展，经济增速和房价增长。因此我们想探究，省会城市的 GDP 数值与该地的平均房价有怎样的关系。

为了更加直观的分析**地区平均房价、省会地区平均房价与该区域的 GDP 的关系**，小组成员分工通过 python 爬虫，数学回归分析，可视化分析三步对所研究问题进行分析。

本次实验的爬虫目标为：贝壳网 310 个地区房源信息，（包含地区、楼盘信息、楼盘具体地址、楼盘均价（元/㎡））。我们调用了中国各地区、省份 2019 年 GDP 数据，以防疫情原因导致房价变化。

二. 实验步骤

各个函数设计分析：

getUrls():

用于爬取贝壳网不同地区房源信息的网址，存入数组 urls 中以便于后续 requests.get() 请求使用

scrape(url,geolocation,data):

针对传入的网址 url，获取该 url 的房源信息。

使用了 BeautifulSoup4 进行 html 解构，根据目标信息所处的标签和 class 名称进行逐一获取。根据组内人文社科同学数据分析的需求，提取出了地区，楼盘，地址，房价这些关键数据，以字典的形式放入 data 数组中。

log(data):

调用 pandas 第三方库，将两层循环 scrape 后得到的 data 数据存储成 excel 文件以供小组成员后续数据分析使用。

三．实验结果

1、爬虫结果

实验中代码运行结果保存到 EXCEL 中，效果如图 1-3

地区	楼盘	地址	元/㎡(均价)
安康	安康中梁宸院	汉滨区/汉滨区/花园大道安康中梁宸院	5680
北京	和光悦府	朝阳/朝阳其它/南皋路和光悦府	88000
北京	水岸壹号	房山/良乡/良乡大学城西站地铁南侧800米，刺猬河旁	58000
北京	观唐云鼎	密云/溪翁庄镇/溪翁庄镇密溪路39号院（云佛山度假村对面）	30000
北京	运河铭著	通州/北关/商通大道与榆东一街交叉口，温榆河森林公园东500米	49000
北京	世界名园	房山/窦店/窦店镇京石高速窦店镇出口望田路8号	21800
北京	万年广阳郡九号	房山/长阳/长阳清苑南街与汇商东路交汇处西北角	50000
北京	首开璞瑅公馆	丰台/方庄/紫芳园五区	106000
北京	华远裘马四季	门头沟/大峪/增产路16号院	55000
北京	御汤山熙园	昌平/昌平其它/北京市昌平区小汤山镇顺沙路99号院	40000
北京	华远和墅	大兴/南中轴机场商务区/南六环磁各庄桥沿南中轴向南2公里	54000
北京	天恒半山世家	怀柔/怀柔/红螺路39号院	50000
北京	天资华府	房山/长阳/房山区CSD政务大厅5号门	38000
北京	檀香府	门头沟/门头沟其它/京潭大街与潭柘十街交叉口	45000
北京	韩建·观山源墅	房山/良乡/阳光北大街与多宝路交汇处西南（理工大学北校区西侧）	40000
北京	首城汇景墅	平谷/平谷其它/金河北街6号院，金河北街8号院	25000
北京	中国铁建花语金郡	大兴/瀛海/南海子公园西侧(南五环旧忠桥向南第二个红绿灯西300米)	70000
北京	棠颂别墅	亦庄开发区/亦庄开发区其它/鹿华路7号院（南海子公园北侧500米）	80000
北京	水墨林溪	房山/阎村/窦店镇大窦路与阎周路交叉口西北向50米	22400
北京	北辰墅院1900	顺义/马坡/顺兴街11号院望尊园	42000
北京	首创天阅西山	海淀/海淀北部新区/海淀区丰秀东路9号院，永丰路与北清路交汇处东北角，中关村壹号北侧	80000
保定	鹏澔·印象城	涿州/涿州/龙马路鹏澔·印象城	7900
保定	华侨城·城市客厅	涿州/涿州/河北省保定市涿州市	9000
保定	汇元-玖號院	涿州/涿州/龙马路汇元-玖號院	6980
保定	大地新城	涿州/涿州/冠云路与火炬南街交叉口南行300米	9050
保定	翠湖院子	易县/易县/梁各庄镇旺隆湖东北岸翠径路东侧	12000
保定	翠湖院子	易县/易县/梁各庄镇旺隆湖东北岸翠径路东侧	20000
保定	北京华银城	涞水/涞水/张坊南行5公里	9000
保定	华银天鹅湖国际生态城	涞水/涞水/张坊镇西行2公里——天鹅湖景区	12000

图 1 爬虫实验结果效果图

地区	楼盘	地址	元/㎡(均价)
保亭市	金祥·万卷山	保亭黎族苗族自治县/保亭/金江农场17队	15000
保亭市	壹山郡	保亭黎族苗族自治县/保亭/响水镇海榆中线245-246公里处	15000
巴中	优筑·印江州	巴州区/老城区/晏阳初大道南段优筑·印江州	4500
巴中	置信府·南苑	巴州区/兴文区/四川省巴中市巴州区兴文镇尚府路与文化路交汇处	4400
巴中	泽尧-领地	巴州区/回风区/巴中市-回风-广福街	6400
巴中	凯莱国际社区	巴州区/回风区/回风 巴州大道西段（费尔顿酒店正对面）	4500
巴中	凯邦名门	巴州区/南坝区/巴中市将军大道中段第六中学西南侧	5700
巴中	阳光·巨林天下城	巴州区/老城区/巴中市中杨大道（中交王府景旁）	5200
巴中	北宸·阳光	巴州区/江北区/巴中市巴州区安康路旁	5300
巴中	宇亿·叠翠	巴州区/兴文区/安康路一号（巴中职业技术学院旁）	4200
巴中	阳光·滨江1号	巴州区/回风区/回风路阳光·滨江1号	5100
巴中	半山逸城二期	巴州区/江北区/江北大道西段半山逸城二期	7500
巴中	泰诚·领誉	巴州区/兴文区/四川省巴中市经济开发区安康路与秦巴大道交汇处	4200
巴中	华兴·芦溪国际	恩阳区/恩阳区/巴中市恩阳区外环线机场路华兴·芦溪国际	4300
巴中	华兴·丽阳名居	巴州区/南坝区/巴中市南坝将军大道华兴·丽阳名居	5600
巴中	阳光中央公园	恩阳区/恩阳区/巴中市恩阳区登科街道阳光中央公园	4600
巴中	云影香山	巴州区/回风区/长寿路云影香山	4800
巴中	巴中置信逸都城购物中心	巴州区/老城区/巴中市巴州区后坝佛爷湾逸都花园D3栋	28000
巴中	巴郡王府	巴州区//四川省巴中市经开区秦巴大道中段巴郡王府	5500
巴中	国力·花海森林	巴州区/兴文区/巴中经开区通州大道五号	5100
巴中	优筑欧洲城	巴州区//四川省巴中市经开区秦巴大道东段优筑欧洲城	4500
巴中	云城·书香美郡	巴州区//巴中·兴文（巴中中学与巴师附小旁）	4600
宝鸡	三迪金域高新	陈仓区/高新区/渭滨区高新大道与高新十六路交汇处西南角	5600
宝鸡	太白里	眉县/眉县/汤峪镇法汤路与林居路十字西南角	11000
宝鸡	绿城雲溪太白	眉县/眉县/太白山国际旅游度假区迎宾大道9号	13000
宝鸡	太白熙岸	眉县/眉县/太白山国际旅游渡假区汤峪二路	14000
宝鸡	富力湾	渭滨区/高新区/高新大道与天玺路交汇东北角	7000
宝鸡	鑫旺·澜湖湾	渭滨区/石坝河/滨河南路互联网大厦的东南角	5700
宝鸡	九悦香郡	渭滨区/卧龙寺/金台区陈仓大道62号院	5300

图 2 爬虫实验结果效果图

地区	楼盘	地址	元/㎡(均价)
鞍山	万科金域国际	铁西/铁西/体育街31号	5500
鞍山	鞍山富力城	铁东区/铁东/解放东路400号	6800
鞍山	鞍山富力城	铁东区/铁东/解放东路400号	8000
鞍山	富力凯旋门	铁东区/铁东/园林大道215	9000
鞍山	鞍山恒大绿洲	立山区/立山区/万水河北路400号	5000
鞍山	鞍山恒大名都	铁东区/铁东/鞍千路740号	5500
鞍山	公园1953·公园墅	铁东区/铁东/解放东路437号	7000
鞍山	公园1953公园府	铁东区/铁东/解放东路455号	5200
鞍山	公园1953·公园墅	铁东区/铁东/解放东路437号	8000
鞍山	万科城市之光	铁西/铁西/大陆街万科城市之光	5600
鞍山	高新万科城	铁东区/铁东/鞍千路高新万科城	6790
鞍山	万科金域华府	铁东区/铁东/湖南街万科金域华府	7500
安庆	安庆高速时代公馆	宜秀区/大桥/九塘东路安庆高速时代公馆	8300
安庆	世茂祥生金科郡	宜秀区/北部新城/安庆市宜秀区学院路与天仙河路交汇处西南方向	7500
安庆	安庆中梁滨江壹号	迎江区/龙狮桥乡/顺安路安庆中梁滨江壹号	9200
安庆	天盟·阅江山	宜秀区/菱北/元山路天盟·阅江山	10600
安庆	同安府	宜秀区/菱北/雷池路同安府	9100
安庆	金大地天元府	迎江区/老峰镇/辉煌路金大地天元府	8500
安庆	融创时代宜城	宜秀区/大桥/宜秀区独秀大道与白泽湖路交叉口	8800
安庆	碧桂园•中梁东方印	宜秀区/大桥/独秀大道碧桂园•中梁东方印	8500
安庆	安庆弘阳广场	宜秀区/大桥/宜秀区独秀大道迎宾东路交汇处北侧	8300
安庆	绿地新里城	迎江区/老峰镇/段山路绿地新里城	6600
安庆	置地·安庆中心	迎江区/龙狮桥乡/顺安路置地·安庆中心	9888
安庆	置地百悦华府	大观区/十里铺乡/十里路置地百悦华府	7800
安庆	万达·天空之城	迎江区/老峰镇/港口路万达·天空之城	8300
安庆	置地皖江四季	迎江区/龙狮桥乡/迎江区华中东路	10000
安庆	怀宁蘭园	怀宁县/怀宁县/高河大道怀宁蘭园	8100
安庆	安庆弘阳广场	宜秀区/大桥/独秀大道安庆弘阳广场	25000
安庆	万达·天空之城	迎江区/老峰镇/港口路万达·天空之城	14000

图 3 爬虫实验结果效果图

2、可视化分析结果

对各地区平均房价进行省份汇总分析并排除偏离项，将平均房价呈降序排列，基本结果如图 4，对房价、GDP 进行柱状图表示，结果如图 5-6。

省份	平均值项:元/㎡(均价)
北京市	52210.0
上海市	34529.4
海南省	30374.3
浙江省	19067.9
天津市	18445.0
甘肃省	17393.8
广东省	17061.1
福建省	16099.5
江苏省	15694.8
重庆市	15489.5
云南省	14172.6
新疆省	13783.0
河北省	13100.0
陕西省	12738.4
山东省	11173.0
宁夏省	10813.2
青海省	10762.5
四川省	10567.5
湖北省	9848.4
辽宁省	9716.7
吉林省	9667.0
安徽省	9517.1
黑龙江省	9098.8
山西省	9049.9
江西省	8844.7
广西省	8772.6
河南省	8624.8
内蒙古	8570.9
贵州省	7901.0
西藏省	7271.5
湖南省	7110.0
总计	12453.8

图 4 各省份平均房价情况

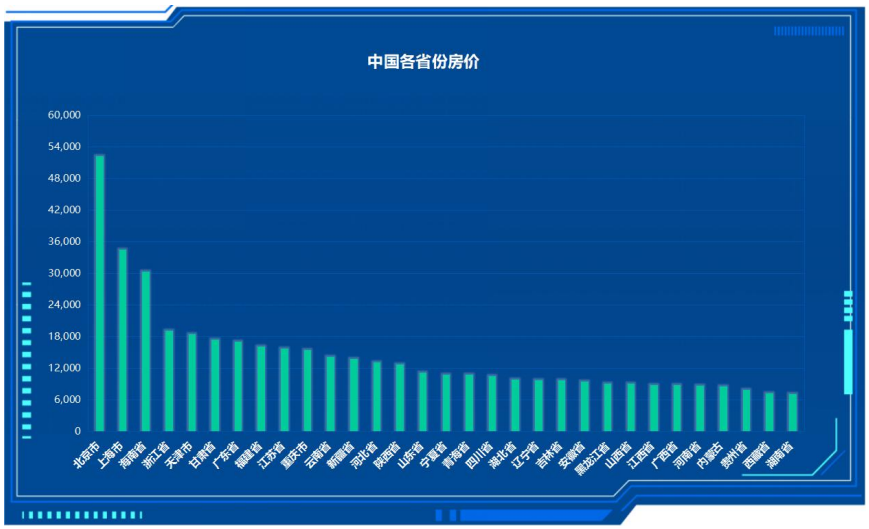

图 5 各省份平均房价柱状图

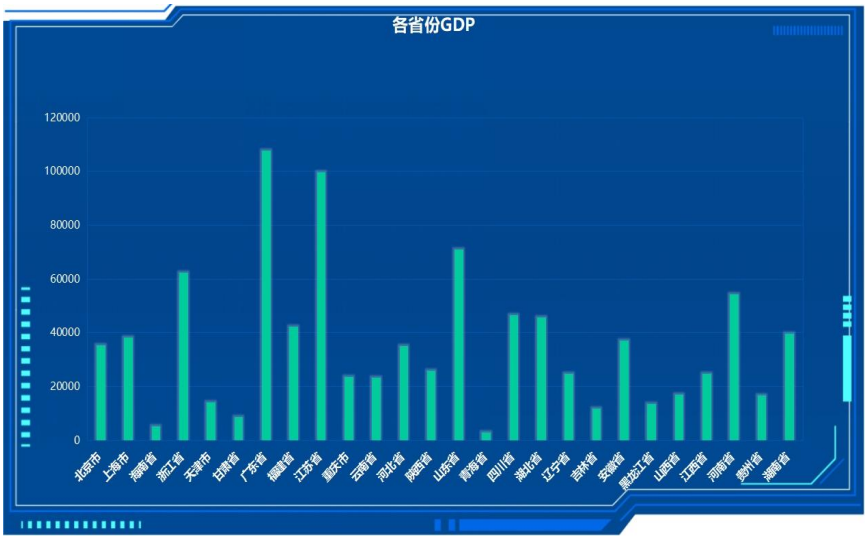

图 6 各省份 2019 年 GDP 柱状图

经过粗糙的排除极端数据，我们能够发现，各省份的平均房价与 2019 年的 GDP 数值并没有直观的线性关系，究其原因做出如下几点分析。

1、各个省份其内部地区贫富差距，用总省份的国内生产总值衡量省份的平均房价有着较大误差。并没有直接用省会城市指标来作为变量。

2、GDP 对实际的房价并无线性相关影响，人们对于住所的选择并不完全考虑该地区的发展情况，其他因素可能对人类主观决策有着更大影响：如生态宜居水平，地区所属气候等。

为了更加直观精确的分析地区平均房价与 GDP 的关系，小组成员进行了数据回归分析。

3、回归分析结果

将平均房价作为被解释变量，GDP 作为解释变量，进行回归处理，对于 regress 后得到的 model 的一个 ANOVA 表格，如表 1 我们可见，即便是惩罚模型复杂度后得到的 Adj R-squared 仍然接近 0.6，并且由 Prob > F 的值可以看出 P 值<0.01，进一步用“*”级表示相关程度后可以得到表 2 中变量数值右上角均为“***”即相关程度极大，说明由我们 regress 后得到的关系式“**平均房价=0.9810GDP+ 8.5e+03**”是可以很好地描述其二者之间的数量关系。

Source	SS	df	MS	Number of obs	=	32
				F(1, 30)	=	44.07
Model	2.7652e+09	1	2.7652e+09	Prob > F	=	0.0000
Residual	1.8822e+09	30	62741472.8	R-squared	=	0.5950
				Adj R-squared	=	0.5815
Total	4.6475e+09	31	149919068	Root MSE	=	7921

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gdp	.9809839	.1477652	6.64	0.000	.679207	1.282761
_cons	8468.477	2130.72	3.97	0.000	4116.967	12819.99

表 1 ANOVA 表

Variable	de1
gdp	0.9810***
_cons	8.5e+03***

表 2 相关程度表

四．分析与总结

此探究结果的意义在于，我们不再依据感受粗糙地定性估计二者的关系，而是根据精准的数值，得到其定量的关系等式，这对于我们基于宏观经济上 GDP 的预测可以更精准的预期房价变化，从而提前采取相关政策以稳定房地产市场。

在可视化分析中，由于产生了较大的误差，分析原因可能对变量的选取还有失偏颇。经过回归分析，经过修改选取了较为准确的变量，得出了研究变量强相关的结果。

在未来的研究中，可以选取更多元的影响因素，对房价的变化成因有一个更准确的分析解释。

五．Python 源代码【可执行】

```
import requests
from bs4 import BeautifulSoup
import pandas
# import time
def getUrls():
    res=requests.get("https://sh.fang.ke.com/loupan/pg1/",headers={
        "User-Agent":"Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/39.0.2171.95 Safari/537.36 OPR/26.0.1656.60",
    })
    document=BeautifulSoup(res.content,"html.parser")
    links=document.find("div",class_="fc-main clear").find_all("a",href=True)
    urls=[]
    for link in links:
        urls.append({
            "name":link.text,
            "url":"https:"+link["href"]
        })
    return urls

def scrape(url,geolocation,data):
    # request
    res=requests.get(url,headers={
        "User-Agent":"Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/39.0.2171.95 Safari/537.36 OPR/26.0.1656.60",
    })
    # err
    if(res.status_code!=200):
        print(f"ErrorCode: {res.status_code}")
        return
    # 200
    document=BeautifulSoup(res.content,"html.parser")
    wrappers=document.find_all("div",class_="resblock-desc-wrapper")
    for wrapper in wrappers:
        name=wrapper.find("a",class_="name").text
        location=wrapper.find("a",class_="resblock-location").text.strip()
        price=wrapper.find("span",class_="number").text
        entry={
            "地区":geolocation,
            "楼盘":name,
            "地址":location,
            "元/㎡(均价)":price,
```

```
        }
        data.append(entry)

def log(data):
    df=pandas.DataFrame(data)
    df.to_excel("贝壳.xlsx")

urls=getUrls()
data=[]
for url in urls:
    for i in range(1,3):
        scrape(url["url"]+f"/pg{i}",url["name"],data)
        print("Done with "+url["name"]+f" - {i}")
log(data)
```