|  |  |
| --- | --- |
| 学号 | 10210415210 |



课 程 设 计

课程名称 python程序设计课程设计

|  |  |
| --- | --- |
| 题 目 | 北上广深租房状况分析 |
| 专 业 | \_\_\_ 软件工程 \_\_\_\_\_\_\_\_ |
| 班 级 | \_\_\_\_\_\_\_ 软件1182班\_\_\_\_\_\_\_\_ |
| 姓 名 | \_\_\_\_\_\_\_\_\_\_\_郭晓东\_\_\_\_\_\_\_\_\_\_\_\_ |
| 成 绩 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| 指 导 老 师 | \_\_\_\_\_\_\_\_\_\_\_徐勇\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

2021 年 1 月 11 日至 2021 年 1 月 22 日

武汉华夏理工学院信息工程学院

**课 程 设 计 任 务 书**

课程名称：python程序设计课程设计 指导教师：徐勇

班级名称： 软件1181-1182 开课院、系：计算机与网络工程系

**一、课程设计目的与任务**

“python程序设计课程设计”是一个综合性的学习实践型实验教学环节，将在“python程序设计”课程的授课基础上，对python的基础语法、python 对文件的操作、python 对数据的操作、数据处理基础、数据可视化等若干个知识点进行综合运用。

python是一种具有天然开源基因的编程语言，了解开源社区和广泛的使用开源工具，也是Python学习实践的重要环节。因此在本次课程设计中，要求掌握jupyter notebook、Git等常用工具，以及github等重要开源社区的使用。

**二、课程设计的内容与基本要求**

数据分析的基本技术和方法在“python程序设计”课程中已经进行完整的讲授，由于数据分析本身是基于业务场景的，因此本次课程设计环节更加偏重于实际的业务场景的实践。通过对近期互联网热点的调查，准备了15个具体的业务场景，用于本次课程设计的具体任务场景。具体包括如下：

1. 北上广深租房状况分析；
2. 蔡某坤粉丝数及转发数据真假状况分析；
3. 地震的数据分析；
4. 英文名字的数据分析；
5. 外籍英文老师收入虚高情况数据分析；
6. 我国城市空气污染和烟花燃放的关系分析；
7. 针对996工作，程序员群体的看法的分析；
8. 吴某凡微博热点的分析；
9. 节假日长假景点人满为患的数据分析；
10. 针对荔枝的品种、销售地等维度，进行价格数据分析；
11. 分析芒果TV《我是大侦探》的观众评论数据；
12. 针对当前儿科医生的缺乏，对相关数据进行分析；
13. 著名网游《绝地求生》的数据分析；
14. 实习岗位状况的数据分析；
15. 电影《流浪地球》的观众评价的数据分析。

每三个学生组建一个课程设计小组，最后的任务输出包括代码每小组一份、课程设计报告每人一份、答辩ppt每小组一份，并进行课程设计成果答辩。小组成员均参与前述工作，但是每个人的侧重点不同。

每个课程设计小组可以从上述15个场景中选取一个作为课程设计的选题，选题中提供了待分析的数据，和现有的分析方法。各小组，通过学习和实践现有的分析方法，理解实战分析的思维过程并锻炼实际动手能力，再此基础上可以扩展更多维度的分析和数据展现形式。

每个班每个选题最多只能被两个小组选中，先选先得。同时，如果各小组发现更有意思的场景，并能够获取到相关待分析的数据，也可以申请作为选题方向。

本次课程设计的目标是培养学生的团队协作能力、对python知识点的综合运用、对实际场景的理解和适应能力、针对答辩的表达能力等。注重过程，期待成果，但不强求结果的尽善尽美。

**三、学时分配进度安排**

|  |  |  |
| --- | --- | --- |
| 序号 | 设计内容 | 所用时间 |
| 1 | 下发任务书，学生查阅资料 | 1天 |
| 2 | 组建团队，并进行选题和团队匹配 | 1天 |
| 3 | 各小组对自己的选题开始进行研究 | 2天 |
| 4 | 各小组开始准备数据并形成初步处理意见 | 1天 |
| 5 | 代码和实现分析的动手实践 | 2天 |
| 6 | 形成初步的报告书和ppt | 1天 |
| 7 | 答辩并完成报告书 | 2天 |
| 合 计 | | 2周 |

**四、课程设计考核及评分标准**

**1.设计报告要求**

课程设计报告要求逻辑清晰、层次分明、书写整洁。课程设计报告为每人一份，同一个小组的各成员的整体报告内容结构一致，但个人侧重点不同，个人着重撰写自己的工作内容，其他人的内容只要体现文档结构的完整性即可。

课程设计考核将综合考虑学生考勤和参与度、团队协作能力，过程管理能力、成果达成情况等。

**2.过程要求**

整个过程要求通过github来进行过程化发布，即阶段性地在github上提交过程结果。

**3.评分标准**

|  |  |
| --- | --- |
| **评分依据** | **评分成绩** |
| 1．团队协作能力 | 25分 |
| 2．python综合运用能力 | 25分 |
| 3．态度认真、刻苦钻研、创新能力 | 10分 |
| 4．过程完成、对工具的使用、对github的运用 | 20分 |
| 5．课程设计答辩逻辑清晰，内容正确 | 10分 |
| 6. 课程设计期间的课堂考勤、遵守纪律 | 10分 |
| 总分 | 100分 |

注：按上述六项分别记分后求和，根据小组成员贡献率综合评定，记载个人最后成绩。

成绩等级：优（90分—100分）、良（80分—89分）、中（70分—79分）、及格（60分—69分）、60分以下为不及格。

**五、指导地点与时间**

本课程设计将安排在第20-21周，采用腾讯会议和QQ群的方式，以在线形式进行。具体安排如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 周次 | 星期一 | 星期二 | 星期三 | 星期四 | 星期五 |
| 第20周 | 第5-8节 | 第5-8节 |  | 第5-8节 |  |
| 第21周 |  | 第5-8节 |  | 第5-8节 |  |

执笔:徐 勇 日期：2021-1-8

审阅:钱小红 日期：2021-1-8

目录

[1. 小组人员 5](#_Toc27469)

[2. 小组选题 5](#_Toc30274)

[3. 环境搭建 5](#_Toc28566)

[4. 各成员任务 5](#_Toc28659)

[5. 选题背景 6](#_Toc14590)

[6.分析内容 6](#_Toc4584)

[7.数据清洗（宁洁） 7](#_Toc29626)

[8.数据处理（郭晓东） 7](#_Toc11905)

[9. 数据分析（阮汉成） 20](#_Toc11172)

[10. 总结（郭晓东） 20](#_Toc326)

[参考文献 21](#_Toc22765)

[课程设计成绩评定表 22](#_Toc4052)

# 小组人员

郭晓东、宁洁、阮汉成

# 小组选题

北上广深租房状况分析

# 环境搭建

硬件环境：Win10

软件环境：Python3.9.0、MySql 5.7、jupyter notebook、Github

# 各成员任务

宁 洁：负责数据清洗、PPT制作

阮汉成：负责数据分析、任务进度汇总

郭晓东：负责数据处理、答辩

# 选题背景

转眼到了大三，再过不到一年时间我们就要毕业了，到了毕业之后大家肯定都会想要去大城市发展，那么如果说去了北上广深，我们是没有能力直接买一套房子的，所以说，在考虑租房的情况下，租什么样的房既经济又实惠就成了我们的优先考虑了。北上广深四个一线城市，哪个城市的租房更加友好呢？城市各区域的租金是怎么分布的？离地铁口远近、房子大小、是否公寓房、精装简装、集中供暖等，租房时应该优先考虑哪些因素呢？一堆问题困扰着我们。

所以说我们要进行一个租房数据的分析，得出相应的结论，给小伙伴们提供最好的租房建议和房屋各种情况的一个说明。

我们获取了链家网北上广深四个城市的所有租房数据（时间节点：2019年2月25日），数据共有20000条。

对这些数据进行各种层次的分析，以得出我们的结论。

# 6.分析内容

（1）各城市房源数量

（2）各地区数据清洗后的房源数量

（3）单位面积房价

（4）北上广深GDP情况

（5）据地铁口远近对每平米租金的影响

（6）房屋大小对每平米租金的影响

（7）个人房和公寓房

# 7.数据清洗（宁洁）

# 8.数据处理（郭晓东）

### 8.1“北上广深”四个城市房源数量

##### 8.1.1统计各城市房源数量

def get\_city\_zf\_loc(city, city\_short, col=['longitude', 'latitude', 'dist'], data=data):

file\_name = 'data\_' + city\_short + '\_latlon.csv'

data\_latlon = data.loc[data['city']==city, col].dropna(subset=['latitude', 'longitude'])

data\_latlon['longitude'] = data\_latlon['longitude'].astype(str)

data\_latlon['latitude'] = data\_latlon['latitude'].astype(str)

data\_latlon['latlon'] = data\_latlon['longitude'].str.cat(data\_latlon['latitude'], sep=',')

data\_latlon.to\_csv(file\_name, index=False)

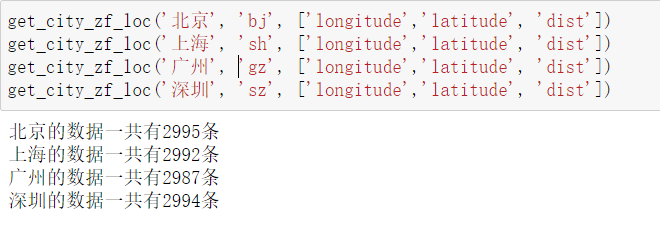
print(city+'的数据一共有{}条'.format(data\_latlon.shape[0]))

get\_city\_zf\_loc('北京', 'bj', ['longitude','latitude', 'dist'])

get\_city\_zf\_loc('上海', 'sh', ['longitude','latitude', 'dist'])

get\_city\_zf\_loc('广州', 'gz', ['longitude','latitude', 'dist'])

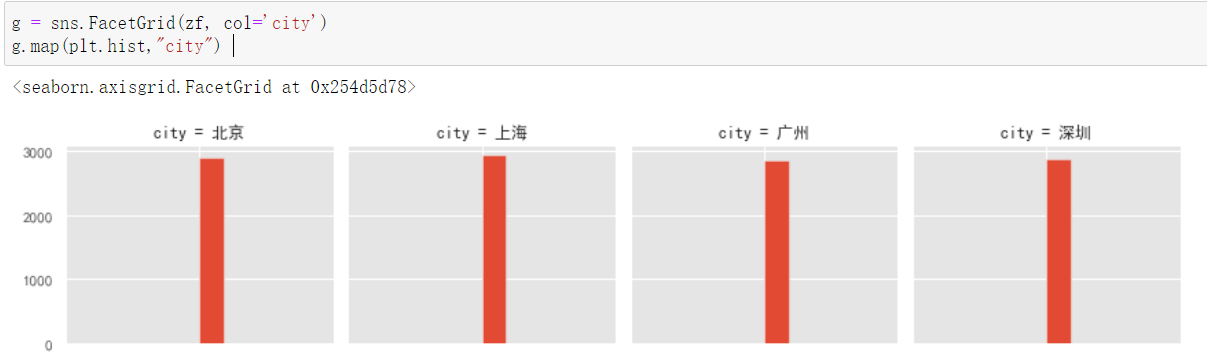
get\_city\_zf\_loc('深圳', 'sz', ['longitude','latitude', 'dist'])



##### 8.1.2制图

g = sns.FacetGrid(zf, col='city')

g.map(plt.hist,"city")

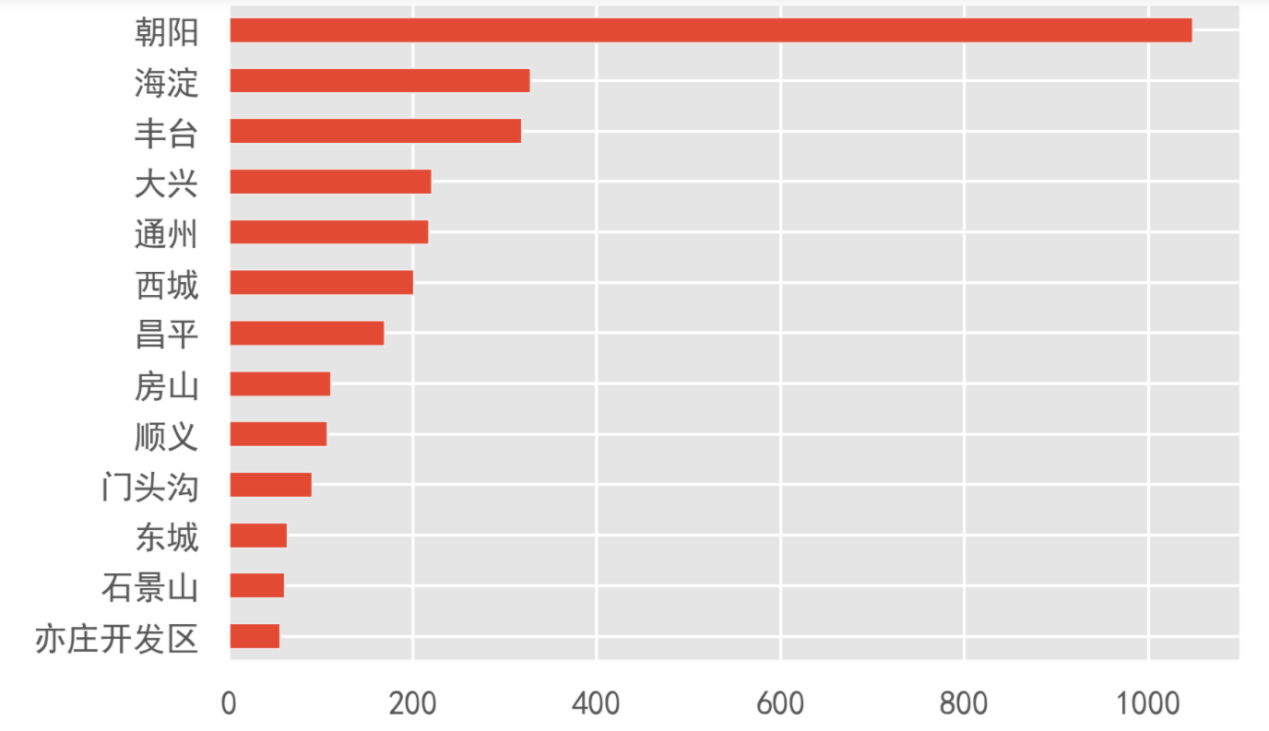


### 8.2 “北上广深”各区房源数量

##### 8.2.1 北京各区房源数量

fig = plt.figure(dpi=300)

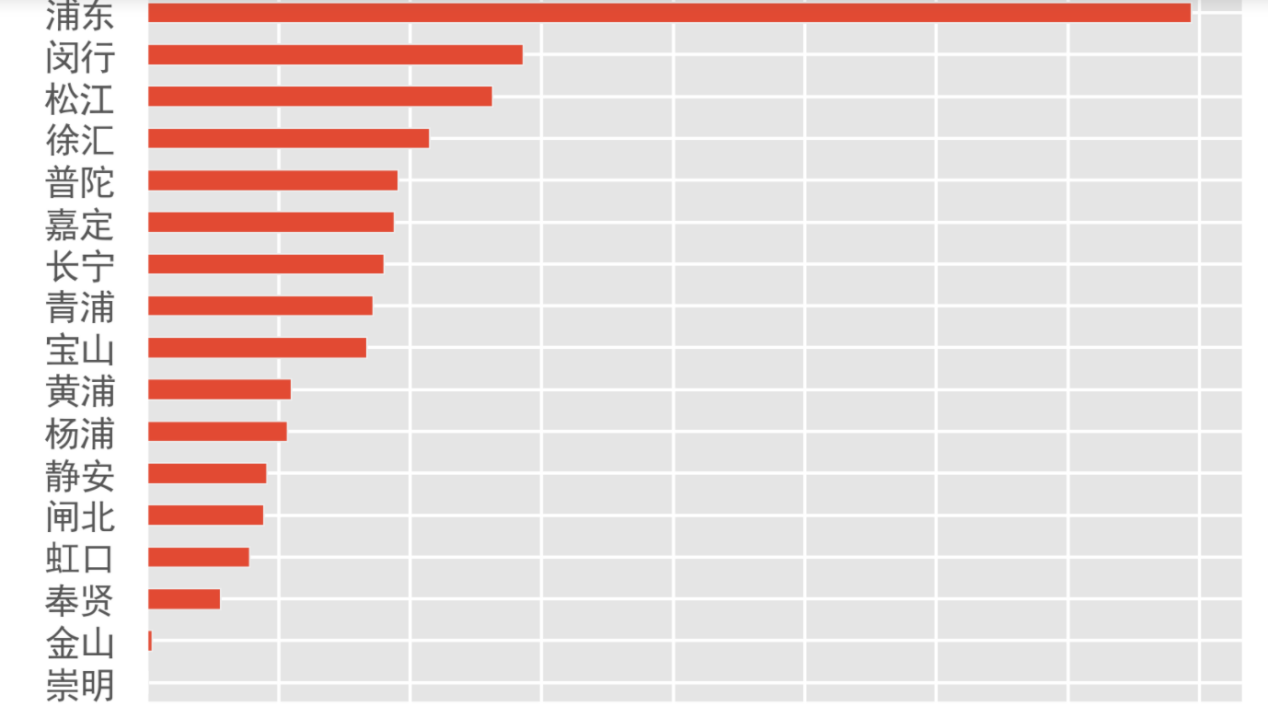
data.dropna(subset=['latitude','longitude'])[data['city']=='北京']['dist'].value\_counts(ascending=True).plot.barh()



##### 8.2.2 上海各区房源数量

fig = plt.figure(dpi=300)

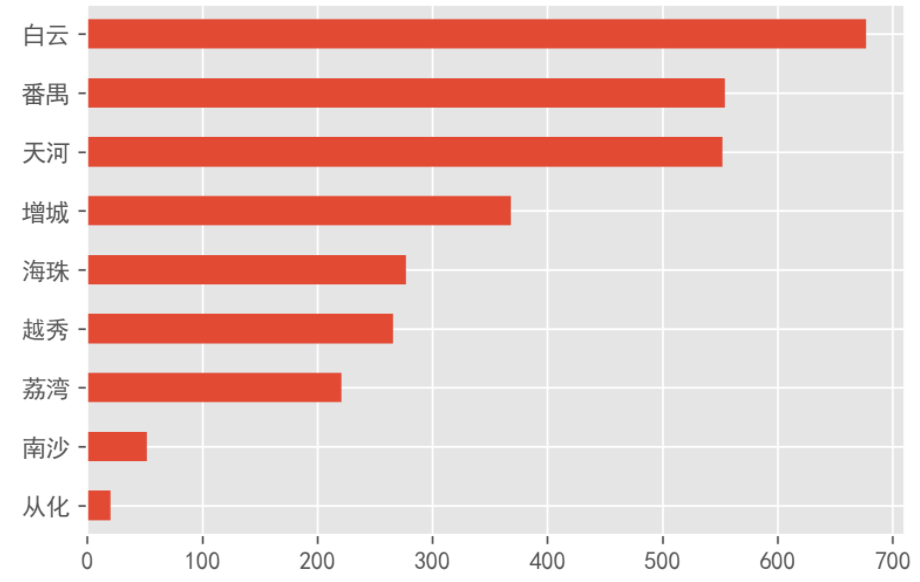
data.dropna(subset=['latitude', 'longitude'])[data['city']=='上海']['dist'].value\_counts(ascending=True).plot.barh()



##### 8.2.3 广州各区房源数量

fig = plt.figure(dpi=300)

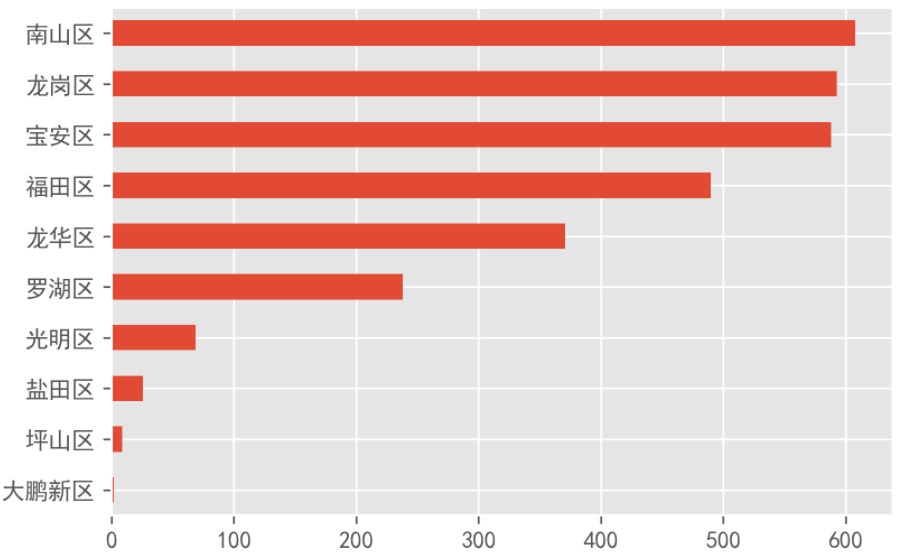
data.dropna(subset=['latitude', 'longitude'])[data['city']=='广州']['dist'].value\_counts(ascending=True).plot.barh()



##### 8.2.4 深圳各区房源数量

fig = plt.figure(dpi=300)

data.dropna(subset=['latitude', 'longitude'])[data['city']=='深圳']['dist'].value\_counts(ascending=True).plot.barh()

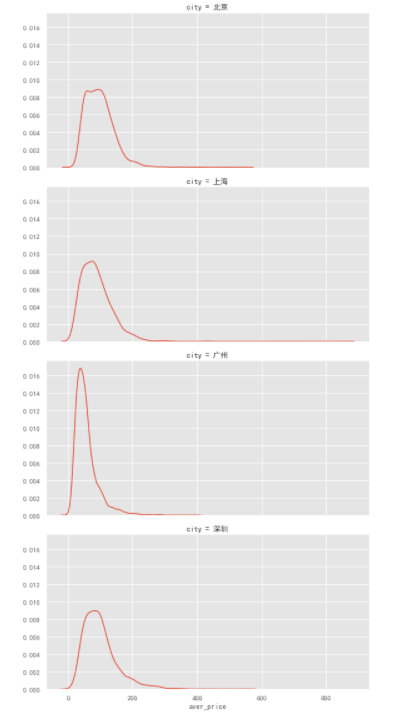


### 8.3 单位面积房价

data['aver\_price'] = np.round(data['rent\_price\_listing'] / data['rent\_area'], 1)

g = sns.FacetGrid(data, row="city", height=4, aspect=2)

g = g.map(sns.kdeplot, "aver\_price")



### 8.4北上广深GDP情况

import matplotlib.pyplot as plt

GDP = [3.22,3.27,2.45,2.42]

# 绘图

plt.bar(range(4), GDP, align = 'center',color='red', alpha =0.8)

# 添加轴标签

plt.ylabel('GDP（单位：万亿）')

# 添加标题

plt.title('北上广深GDP')

# 添加刻度标签

plt.xticks(range(4),['北京市','上海市','广州市','深圳市'])

# 设置Y轴的刻度范围

plt.ylim([0.5,5])

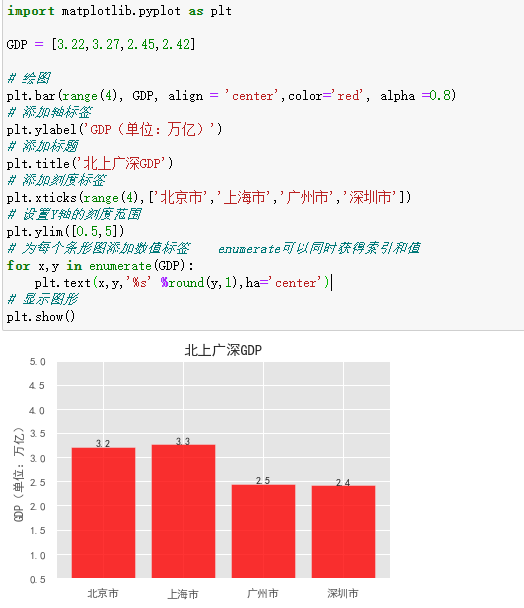
# 为每个条形图添加数值标签 enumerate可以同时获得索引和值

for x,y in enumerate(GDP):

plt.text(x,y,'%s' %round(y,1),ha='center')

# 显示图形

plt.show()



### 8.5距地铁口远近对每平米租金的影响

def distance\_price\_relation(city, data=data):

g = sns.jointplot(x="distance",

y="aver\_price",

data=data[(data['city']==city)&

(data['aver\_price']<=350)].dropna(subset=['distance']),

kind="reg")

g.fig.set\_dpi(100)

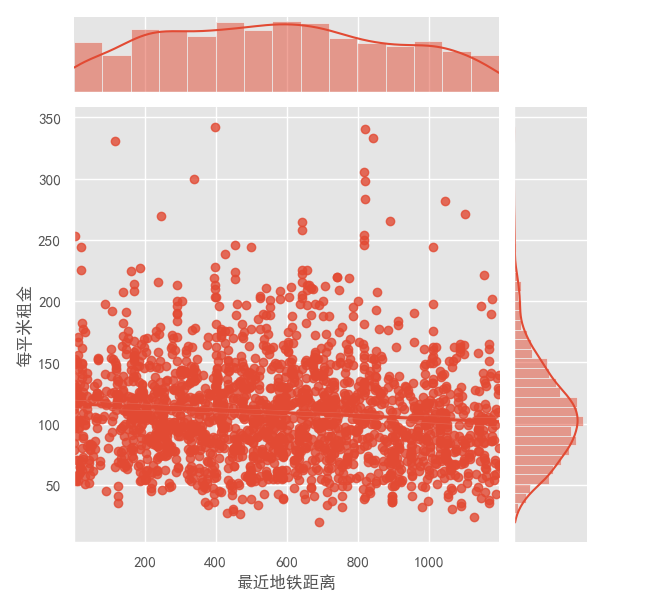
g.ax\_joint.set\_xlabel('最近地铁距离', fontweight='bold')

g.ax\_joint.set\_ylabel('每平米租金', fontweight='bold')

return g

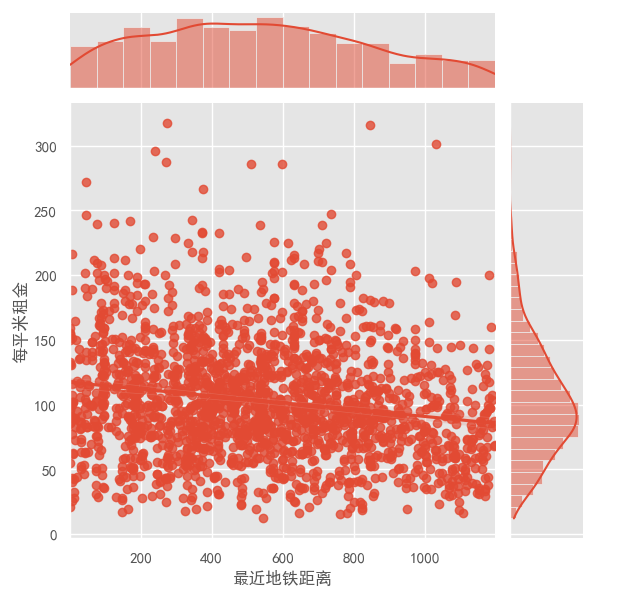
##### 8.5.1 北京

distance\_price\_relation('北京')



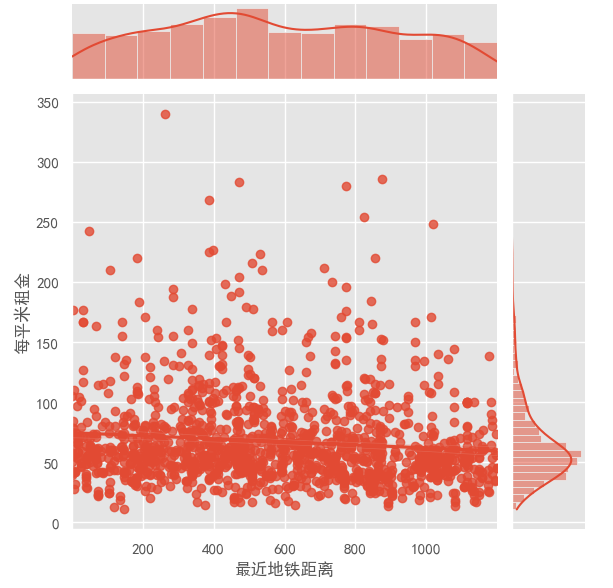
##### 8.5.2 上海

distance\_price\_relation('上海')



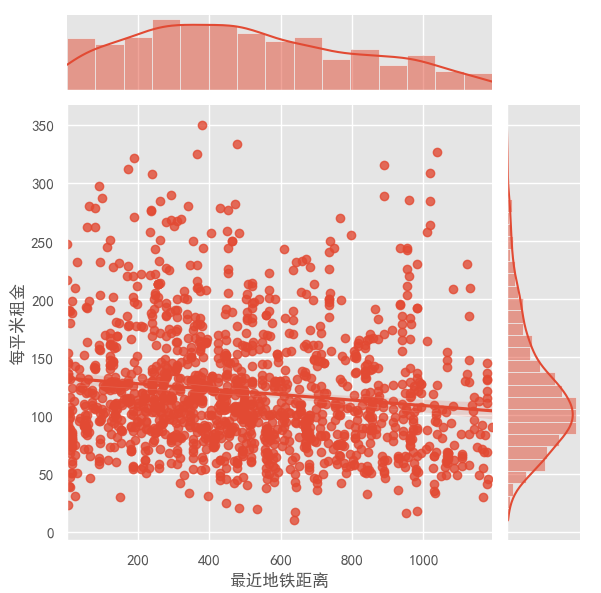
##### 8.5.3 广州

distance\_price\_relation('广州')



##### 8.5.4 深圳

distance\_price\_relation('深圳')



### 8.6 房屋大小对每平米租金的影响

def area\_price\_relation(city, data=data):

fig = plt.figure(dpi=100)

g = sns.lineplot(x="rent\_area",

y="aver\_price",

data=data[(data['city']==city)&(data['rent\_area']<150)],

ci=None)

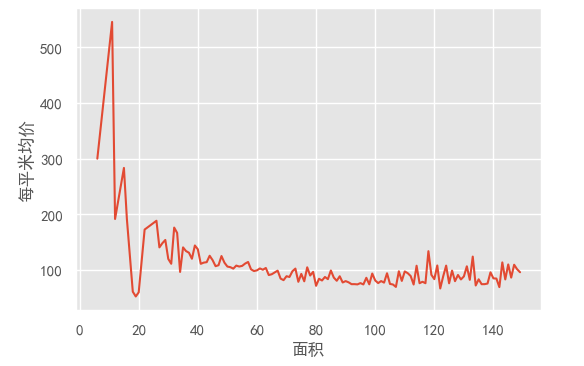
g.set\_xlabel('面积', fontweight='bold')

g.set\_ylabel('每平米均价', fontweight='bold')

return g

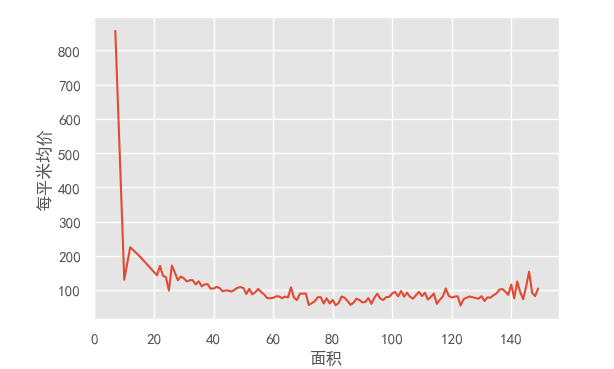
##### 8.6.1 北京

area\_price\_relation('北京')



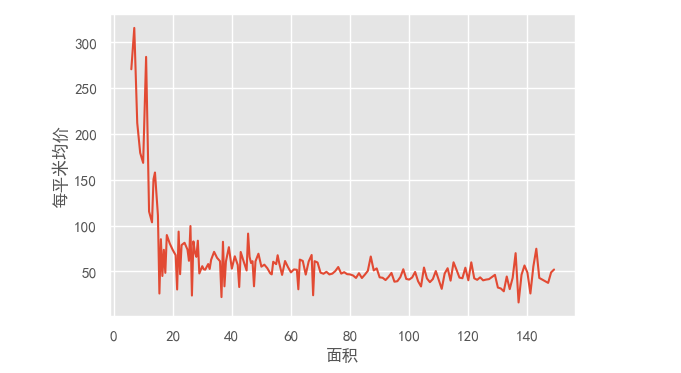
##### 8.6.2上海

area\_price\_relation('上海')



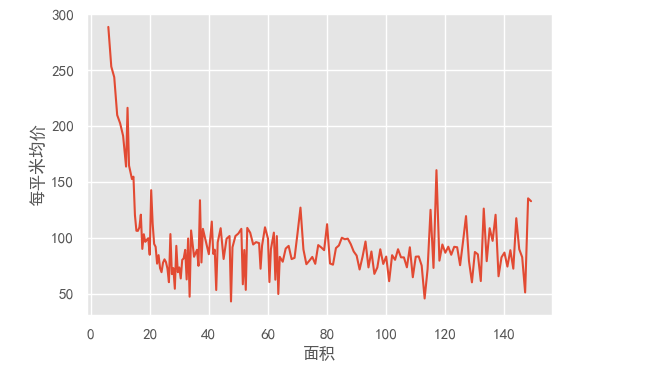
##### 8.6.3 广州

area\_price\_relation('广州')



##### 8.6.4 深圳

area\_price\_relation('深圳')



### 8.7 个人房和公寓房

# 根据house\_title和house\_tag再造一个字段：is\_dep，也就是“是否是公寓”

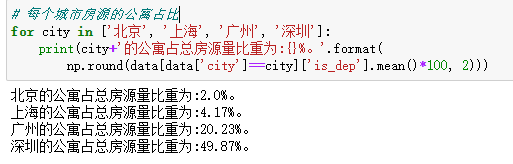
data['is\_dep'] = (data['house\_title'].str.contains('公寓') + data['house\_tag'].str.contains('公寓')) > 0

# 每个城市房源的公寓占比

for city in ['北京', '上海', '广州', '深圳']:

print(city+'的公寓占总房源量比重为:{}%。'.format(

np.round(data[data['city']==city]['is\_dep'].mean()\*100, 2)))



##### 8.7.1 北京公寓房占比

plt.rcParams['font.sans-serif']=['SimHei']

labels = ['公寓房','个人房']

sizes = [2,98]

explode = (0.1,0)

plt.pie(sizes,explode=explode,labels=labels,autopct='%1.1f%%',shadow=False,startangle=150)

plt.title("北京公寓占比")

plt.show()



##### 8.7.2 上海公寓房占比

plt.rcParams['font.sans-serif']=['SimHei']

labels = ['公寓房','个人房']

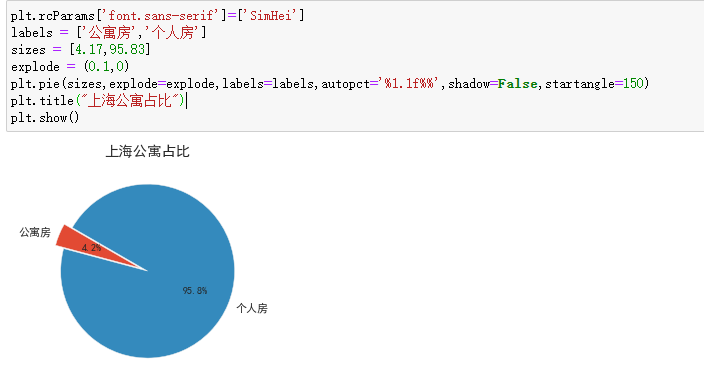
sizes = [4.17,95.83]

explode = (0.1,0)

plt.pie(sizes,explode=explode,labels=labels,autopct='%1.1f%%',shadow=False,startangle=150)

plt.title("上海公寓占比")

plt.show()



##### 8.7.3 广州公寓房占比

plt.rcParams['font.sans-serif']=['SimHei']

labels = ['公寓房','个人房']

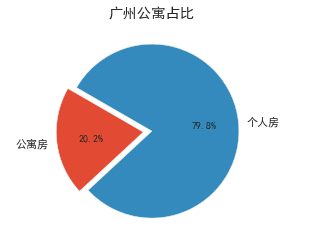
sizes = [20.23,79.77]

explode = (0.1,0)

plt.pie(sizes,explode=explode,labels=labels,autopct='%1.1f%%',shadow=False,startangle=150)

plt.title("广州公寓占比")

plt.show()



##### 8.7.4 深圳公寓房占比

plt.rcParams['font.sans-serif']=['SimHei']

labels = ['公寓房','个人房']

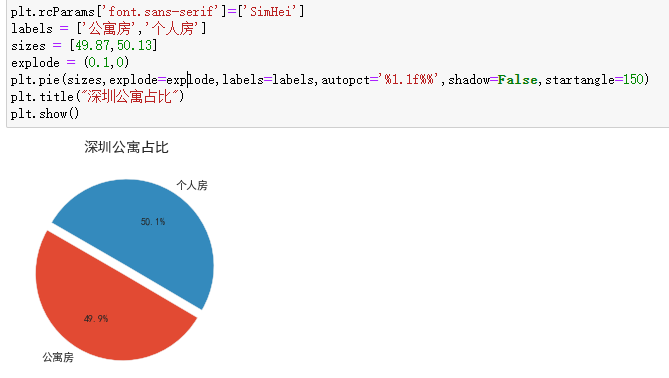
sizes = [49.87,50.13]

explode = (0.1,0)

plt.pie(sizes,explode=explode,labels=labels,autopct='%1.1f%%',shadow=False,startangle=150)

plt.title("深圳公寓占比")

plt.show()



# 数据分析（阮汉成）

# 总结（郭晓东）

通过这次课程设计，我认识到了自己有许多不足之处，上课的基础不是很扎实，许多知识点的应用还不是很熟练，在此次课程设计中也因此遇到很多现实的困难，但是通过去网上查找资料包括问老师相关问题，最终都能够自己解决好，但是也不得不说，暴露了很多问题，平时上课做书上的代码不觉得，但是真正自己应用的时候还是很懵，这也表明了实践是检验真理的标准吧，我们只有在真正做项目的时候，才能够将自己所学的知识真正应用起来，达到融会贯通，让之前不懂的知识也能够得到理解。

此次实验也是收获颇丰，查找并使用了许多方便的绘图工具，遗憾的是pyechart和mysql因为安装的问题无法使用，总的来说对于这次课程设计中我自己的表现我是比较满意的，因为我参与了答辩，可以说是对自己的一种锻炼，希望以后还会有这种机会。

# 参考文献

[1]Python基础与大数据应用.人民邮电出版社，2017年1月

[2]百度文库

[3]CSDN

[4]百度百科

**课程设计成绩评定表**

|  |  |  |  |
| --- | --- | --- | --- |
| 课程设计题目 | 数据分析的学习与实践 | | |
| 课程设计学生答辩或质疑记录：   1. 此次课程设计中用到的绘图工具和你认为最好用的   答：此次课程设计中用到了matplotlib、seaborn、pyecharts等绘图工具，我觉得基于现阶段还是matplotlib比较简单实用，可以绘制各种图形，但是对于pyecharts我觉得是几种绘图工具中相对来说绘制出来的种类比较多而且外观比较好看的，但是同样它存在许多bug，因为版本问题可能存在使用问题。   1. 能说说距离地铁远近对房屋单位面积租金影响是怎么计算的么？   答：房屋单位面积租金是根据房屋租金除以房屋总面积来计算的，然后根据距离地铁远近这个参数进行两个参数的绘图，通过科学计算函数绘制散点图和正态分布图，同样能够得出斜率，比较直观反映距离房屋远近对单位面积租金的影响 | | | |
| **评 分 依 据** | | **分 值** | **评分成绩** |
| 1．团队协作能力 | | 25分 |  |
| 2．python综合运用能力 | | 25分 |  |
| 3．态度认真、刻苦钻研、遵守纪律 | | 10分 |  |
| 4．过程完成、对工具的使用、对github的运用 | | 20分 |  |
| 5．课程设计答辩逻辑清晰，内容正确 | | 10分 |  |
| 6. 课程设计期间的课堂考勤、创新能力 | | 10分 |  |
| 总 分 | | 100分 |  |
| 最终评定等级为：  指导老师签字：  2021 年 1 月 15日 | | | |