一、程序功能

爬取上海链家二手房数据，并进行分析

二、数据来源

数据集下载：http://sh.lianjia.com/ershoufang/

数据含义：爬取的csv文件，每一行代表每一条房源的具体数据，每一列是房源相关的信息，包括：总价，面积，户型等。

三、 分析和代码

分析：爬取上海链家二手房数据，并进行数据分析，包括：

1、根据户型(一室一厅，二室一厅等)绘制户型柱状图；

2、根据房屋面积绘制柱状图；

3、根据房屋所在行政区，利用饼状图绘制房屋分布饼状图；

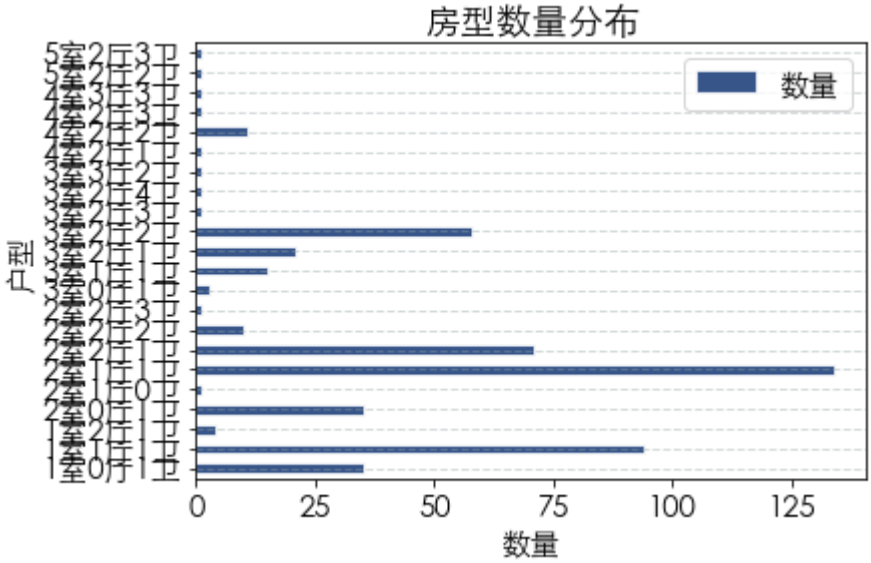
4、根据房价，面积信息进行房源聚类分析。

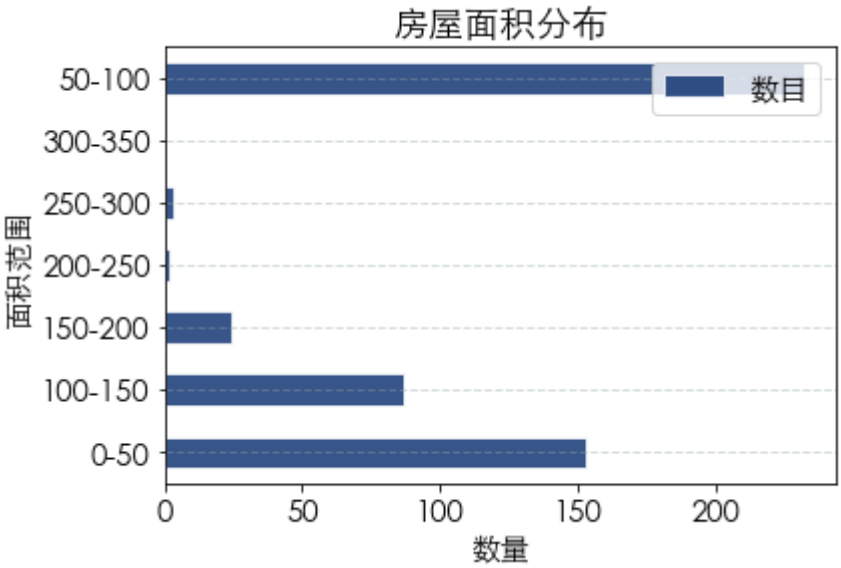
程序：

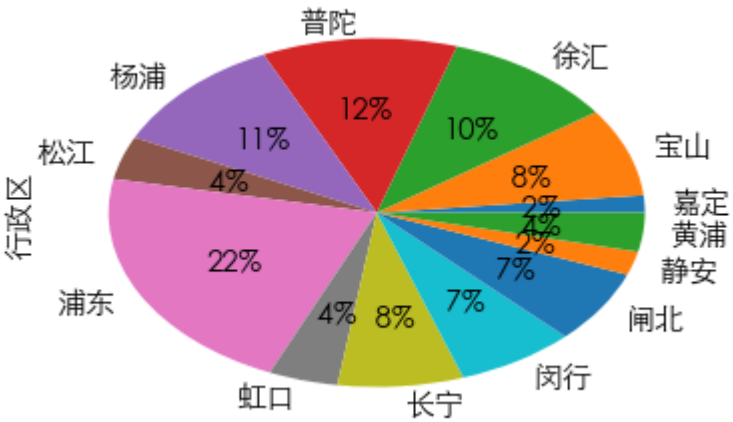
```
# -*- coding: utf-8 -*-
"""
Created on Mon Nov 20 22:42:45 2017

@author: manny

爬取上海链家二手房信息，并进行数据分析，程序主要包括如下内容：
#1爬取数据
#2#户型绘制柱状图
#3面积绘制柱状图
#4房屋分布饼状图
#5 房源聚类分析
"""

import requests
from bs4 import BeautifulSoup
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import time
import re
from sklearn.cluster import KMeans

######数据获取#################################
##############################################
#设置headers信息
headers = {'User-Agent':'Mozilla/5.0 (Windows NT 6.1) \
AppleWebKit/537.11 (KHTML, like Gecko) Chrome/23.0.1271.64 Safari/537.11'}
#主域名
domain='http://sh.lianjia.com'
#存放信息的列表
info_total=[]
#爬取总页数
for i in range(1,20):
    #爬取拼接域名
    res=requests.get('http://sh.lianjia.com/ershoufang/d'+str(i),headers=headers)
    #使用lxml筛选器
    soup = BeautifulSoup(res.text,'lxml')
    #网站每页呈现30条数据，循环爬取
    for j in range(0,30):
        url1=soup.select('.prop-title a')[j]['href']
        #构造子域名
        url=domain+url1
        pi=[]
        a=[soup.select('.info-col a')[1+3*j].get_text()]
        res_sub=requests.get(url,headers=headers)
        soup_sub=BeautifulSoup(res_sub.text,'lxml')
        d=re.findall(re.compile('span class="price-num">(.*?)</span>'),res_sub.text)
        pi=pi+a+d
        #print(pi)
        c=soup_sub.find_all(lambda tag: tag.name=='span' and tag.get('class')==['item-cell'])
        for b in c:

            pi.append(b.get_text(strip=True))
```

```
        if len(pi) > 26:
            for i in range(0,3):
                pi.pop(18)
        info_total.append(pi)
        time.sleep(0.5)
columns=['行政区','price','最低首付','参考月供','环线信息','小区名称','房源编号','房屋户型',\
    '配备电梯','areas','供暖方式','所在楼层','装修情况','房屋朝向','上次交易','房本年限','售房原因','房屋类型',\
    '挂牌均价','建筑年代','物业类型','楼栋总数','房屋总数','物业公司','开发商','挂牌房源']
df=pd.DataFrame(info_total,columns=columns)
#数据清洗
#删除空行
df=df.dropna(axis=0)
#删除无效数据
ex_list1=list(df.areas)
ex_list=[]
for i in ex_list1:
    if(i != "暂无数据"):
        ex_list.append(i)
df=df[df.areas.isin(ex_list)]
#原始数据保存
df.to_csv('data_sh.csv',encoding='utf-8-sig')


######数据分析#################################
################################################
#1#户型绘制柱状图
df=pd.read_csv('data_sh.csv')
#按房源户型类别进行汇总
huxing=df.groupby('房屋户型')['房屋户型'].agg(len)
#房源户型分布绘图
plt.rc('font', family='STXihei', size=15)
huxing.plot(kind='barh', color='#052B6C', alpha=0.8,
            align='center', edgecolor='white')
plt.grid(color='#95a5a6', linestyle='--', linewidth=1, axis='y', alpha=0.4)
plt.xlabel('数量')
plt.ylabel('户型')
plt.title('房型数量分布')
plt.legend(['数量'], loc='upper right')
plt.show()


#2面积绘制柱状图
#对房源面积进行二次分列
#mianji_num_split = pd.DataFrame(df.areas.split('平'),index=df.index,columns=['area','平方米'])
mianji_num_split = pd.DataFrame((x.split('平') for x in df.areas),index=df.index,columns=['area','平方米'])
#将分列后的房源面积拼接回原数据表
df = pd.merge(df,mianji_num_split,right_index=True,left_index=True)
df['area'] = df['area'].map(str.strip)
# 更改mianji_num字段格式为float
df['area'] = df['area'].astype(float)

# 查看所有房源面积的范围值
```

```
# print(house['mianji_num'].min(),house['mianji_num'].max())
# 对房源面积进行分组
bins = [0, 50, 100, 150, 200, 250, 300, 350]
group_mianji = ['0-50', '50-100', '100-150','150-200', '200-250', '250-300', '300-350']
df['group_mianji'] = pd.cut(df['area'], bins, labels=group_mianji)
# 按房源面积分组对房源数量进行汇总
group_mianji = df.groupby('group_mianji')['group_mianji'].agg(len)
#绘制房源面积分布图
plt.rc('font', family='STXihei', size=15)
group_mianji.plot(kind='barh', color='#052B6C', alpha=0.8,align='center', edgecolor='white')
plt.xlabel('数量')
plt.ylabel('面积范围')
plt.title('房屋面积分布')
plt.legend(['数目'], loc='upper right')
plt.grid(color='#95a5a6', linestyle='--', linewidth=1, axis='y', alpha=0.4)
plt.show()

#3房屋分布饼状图
#按行政区进行划分
quyu=df.groupby('行政区')['行政区'].agg(len)
quyu.plot(kind='pie',autopct='%2.0f%%',labeldistance=1.1)
plt.show()


#4 房源聚类分析
# 使用房源总价，面积和关注度三个字段进行聚类
house_type = np.array(df[['price', 'area']])
# 设置质心数量为3
clf = KMeans(n_clusters=3)
# 计算聚类结果
clf = clf.fit(house_type)
# 查看分类结果的中心坐标
center = pd.DataFrame(clf.cluster_centers_, columns=['房价', '面积'])
# 在原数据表中标注所属类别
df['label'] = clf.labels_
print(df.label)


###保存数据
df.to_csv('data_sh_final.csv',encoding='utf-8-sig')
```


房型数量分布
数量
户型
0
25
50
75
100
125
数量


房屋面积分布
数目
面积范围
50-100
300-350
250-300
200-250
150-200
100-150
0-50
0
50
100
150
200
数量


普陀
徐汇
杨浦
宝山
松江
嘉定
黄浦
静安
浦东
闸北
闵行
虹口
长宁
12%
10%
11%
8%
4%
2%
4%
2%
22%
7%
4%
8%
7%
行政区

```
0      1
1      1
2      0
3      0
4      0
5      0
6      1
7      1
8      1
9      2
10     0
11     1
12     1
13     1
14     0
15     1
16     0
17     0
18     0
19     0
20     2
21     0
22     0
23     0
24     0
25     2
26     0
27     1
28     2
29     0
      ..
471    2
472    0
473    0
474    0
475    0
476    0
477    2
478    1
479    1
480    0
481    2
482    0
483    2
484    0
485    0
486    0
487    1
488    0
489    0
490    0
491    0
492    1
```

```
493    0
494    0
495    0
496    0
497    0
498    1
499    0
500    1
Name: label, Length: 501, dtype: int32
```