实践报告：城市租房数据分析 与可视化

学号：2212534

姓名：魏思诚

日期：2024.10.2

南开大学 软件学院

目 录

[1 实验介绍 3](#_Toc7jmh2f)

[1)对于长沙租房数据的分析 3](#_Toclypsj7)

[2 实验目的 4](#_Tocts8p9h)

[1)学习掌握爬虫技术爬取链家数据 4](#_Tock523rk)

[2)增强对数据的处理分析能力 4](#_Tocpmdw9e)

[3 实验流程 5](#_Toc5ihwrt)

[1)使用python爬取链家长沙租房数据30页 5](#_Toc8tbxzn)

[2)对数据进行处理与分析 5](#_Toc5489qp)

[4 实验结果或结论 6](#_Toc2fe5jy)

[1)长沙租金分析 6](#_Tocmqn8po)

[2)长发房源分析 6](#_Tocaflm23)

# 实验介绍

## 对于长沙租房数据的分析

使用python爬取链家上长沙的租房数据(<https://cs.lianjia.com/zufang/>),整理数据获取规范的csv数据文件

对数据进行预处理,

进行数据可视化

分析数据, 获取信息与结论

# 实验目的

## 学习掌握爬虫技术爬取链家数据

使用python爬取链家数据,由于链家具有反爬取机制,在多次访问后会出现人机验证,这部分需要采用分批次爬取来解决,同时处理爬取的数据组成csv文件并保存

## 增强对数据的处理分析能力

使用matplotlib进行数据可视化,筛选出感兴趣的几列分析其中包含的相关性

# 实验流程

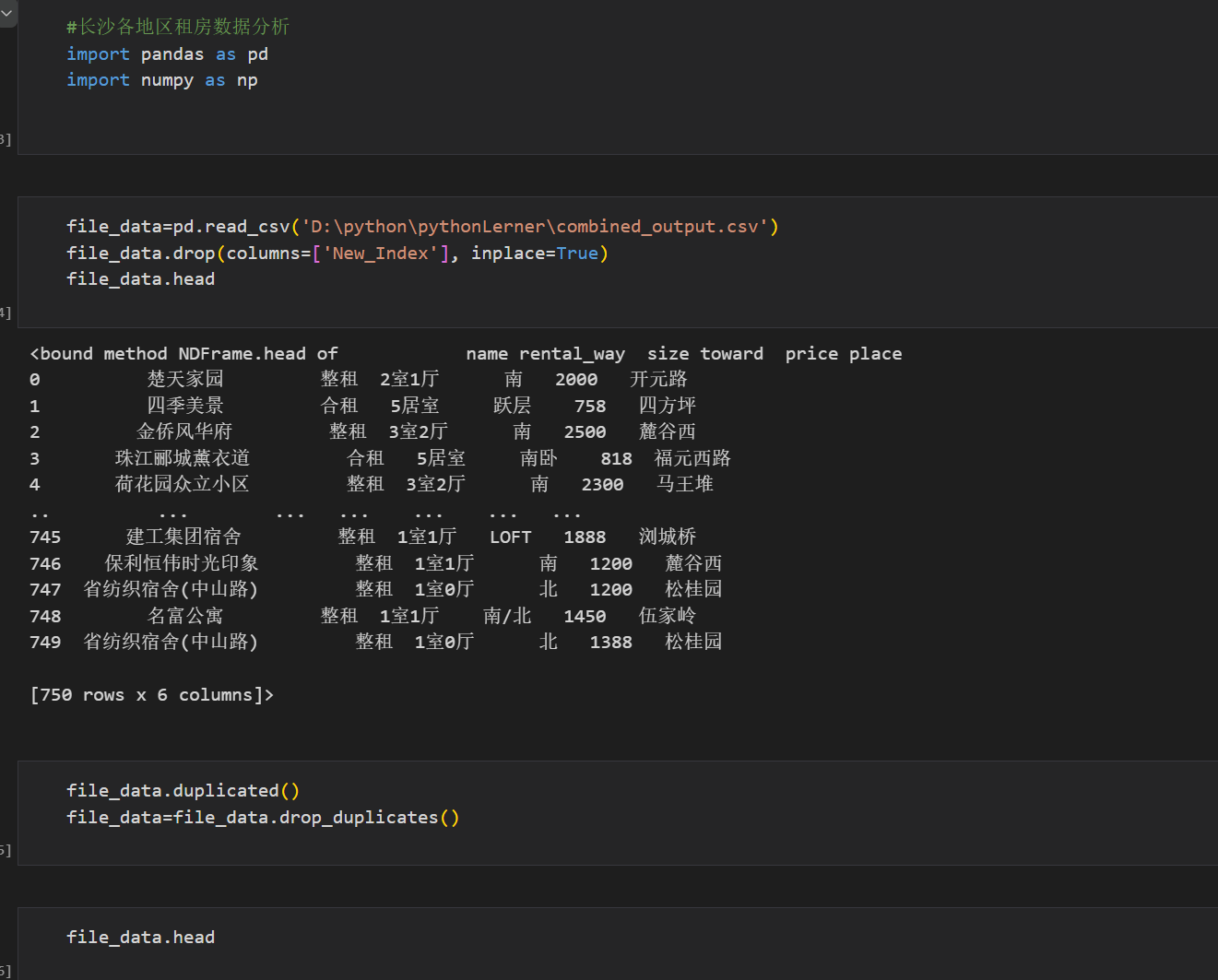
## 使用python爬取链家长沙租房数据30页

from urllib.request import urlopen  
  
import numpy as np  
import pandas as pd  
from lxml import etree  
  
import requests  
from win32api import Sleep  
  
headers = {  
 "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/129.0.0.0 Safari/537.36",  
}  
url = "https://cs.lianjia.com/zufang/"  
i = 1  
final\_data = []  
while i < 30:  
 i += 1  
 url\_ = url + '/pg' + str(i)  
 resp = requests.get(url\_, headers=headers)  
 print(resp.text)  
 html = etree.HTML(resp.text)  
 # 定义XPath查询  
 xpath\_query = '//\*[@id="content"]/div/div/div/div/p[1]/a'  
  
 # 使用XPath查询  
 links = html.xpath(xpath\_query)  
 price = html.xpath('//em/text()')  
 place = html.xpath('//\*[@id="content"]/div/div/div/div/p[2]/a[2]/text()')  
 area = html.xpath('//\*[@id="content"]/div/div/div/div/p[2]/text()[4]')  
  
 # 输出结果  
  
 def split\_(links):  
 rental\_way = []  
 size = []  
 toward = []  
 name = []  
 for i in range(len(links)):  
 p = links[i].split('·')  
 rental\_way.append((p[0].replace(" ", ''))[1:])  
 rest = p[1]  
 rest = rest.split()  
 name.append(rest[0])  
 size.append(rest[1])  
 toward.append(rest[2])  
 return name, rental\_way, size, toward  
  
  
 name, rental\_way, size, toward = split\_(links)  
  
 data = np.column\_stack((name, rental\_way, size, toward, price[:-1], place))  
 if not final\_data:  
 final\_data = data  
 else:  
 final\_data = np.vstack((final\_data, data))  
 Sleep(800)  
final\_data = pd.DataFrame(final\_data, columns=['name', 'rental\_way', 'size', 'toward', 'price', 'place'])  
final\_data.to\_csv('final\_data\_1.csv')

由于反爬取机制,实验时并没有一次性爬取所有数据,采用1-5 ,5-10 ,10-15..方式分批次爬取,

## 对数据进行处理与分析

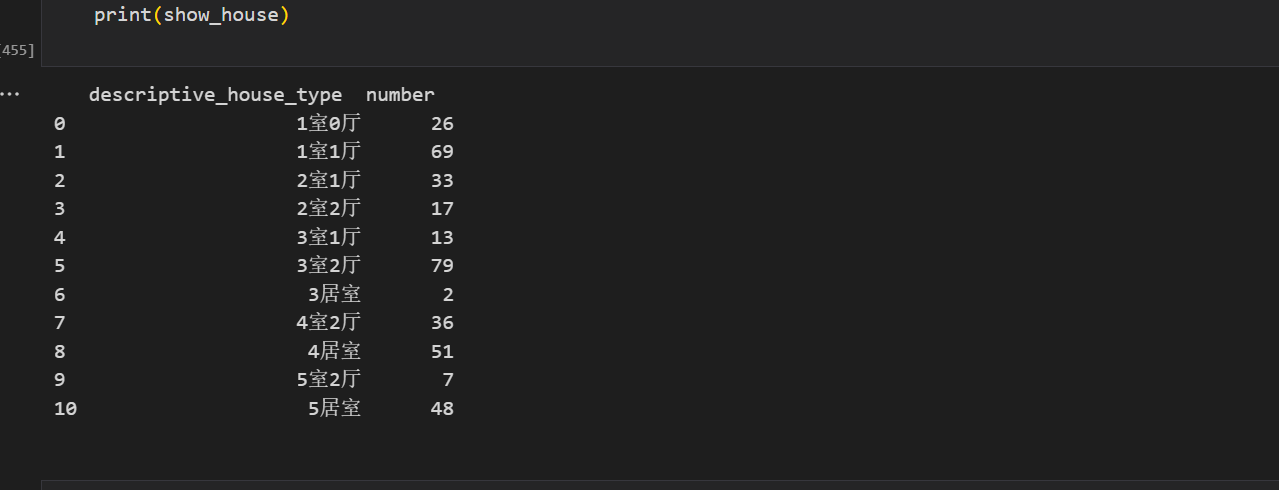
数据读取:



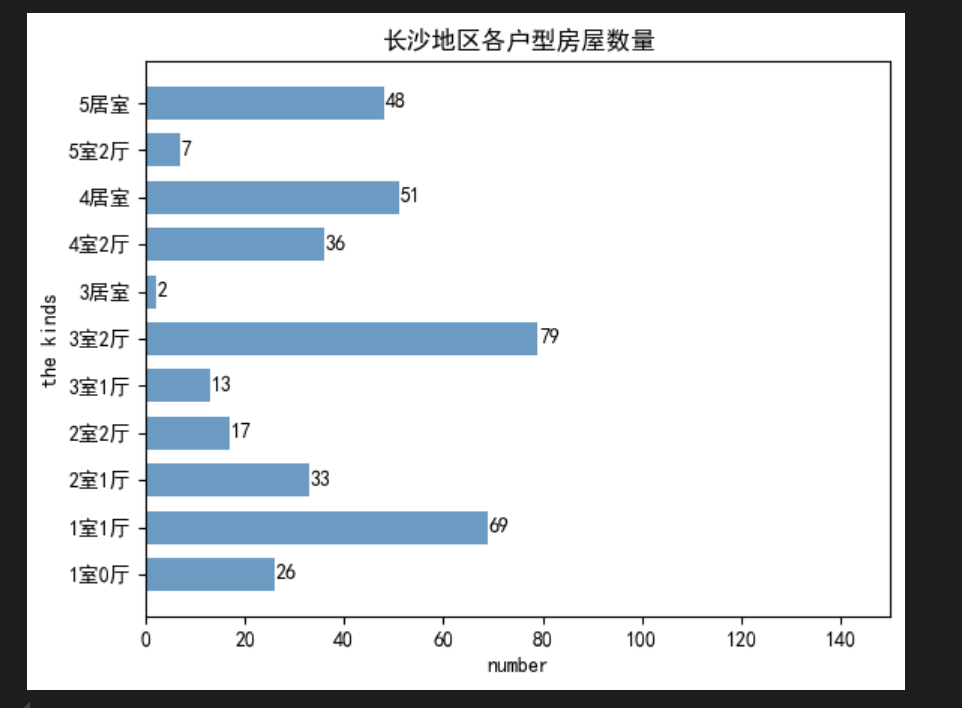
数据预处理获取地区分布信息:



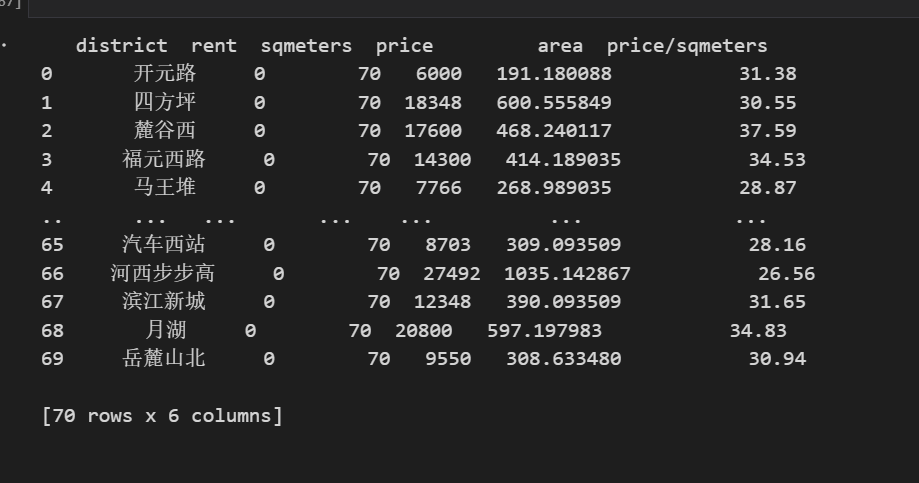
处理数据获取户型信息:

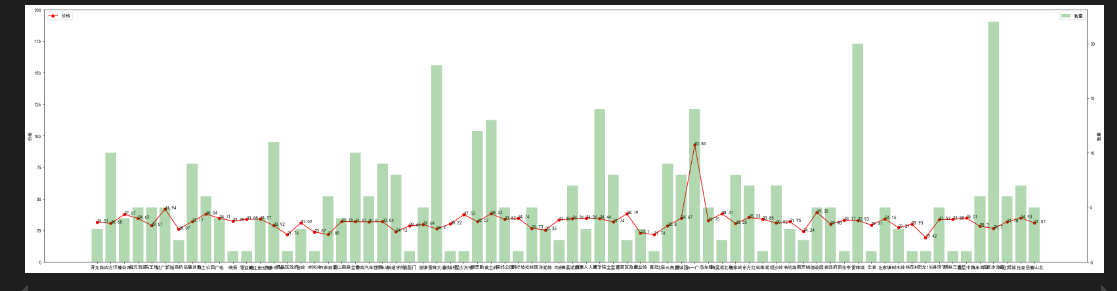


数据可视化:



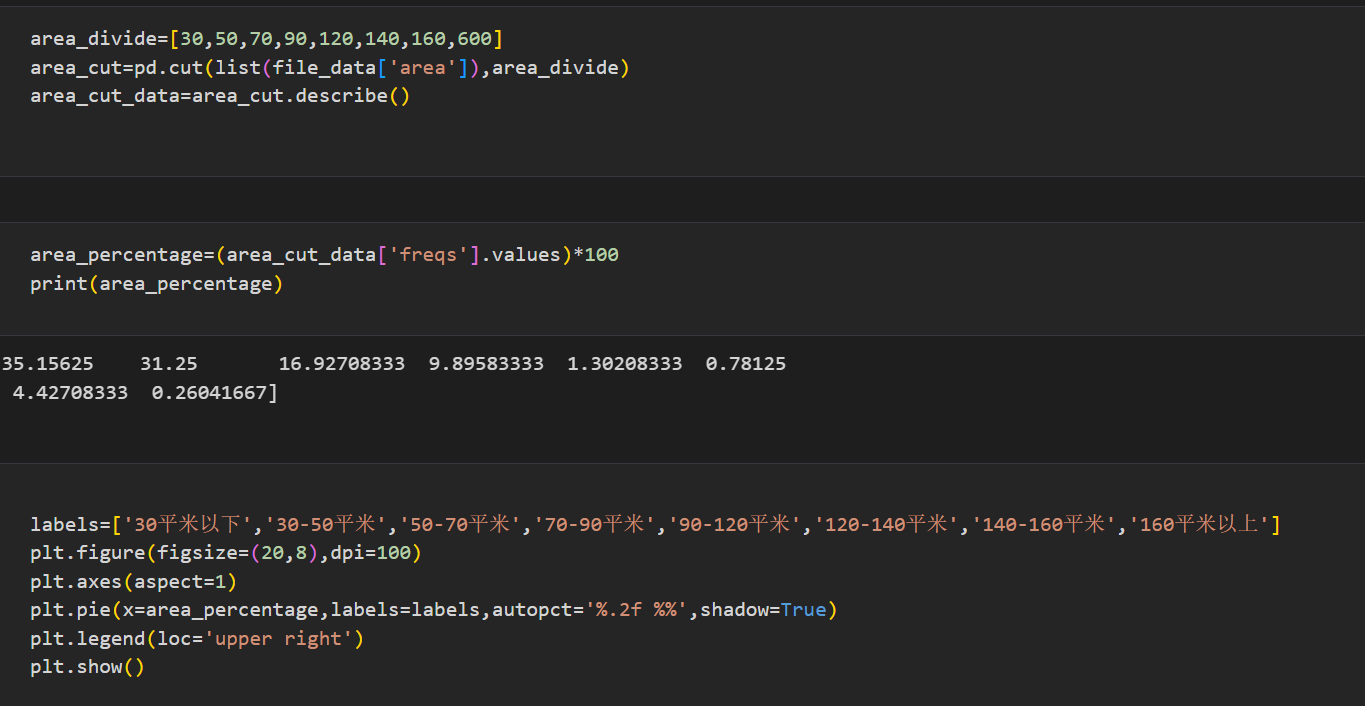
分析价格与面积与房源相关性:

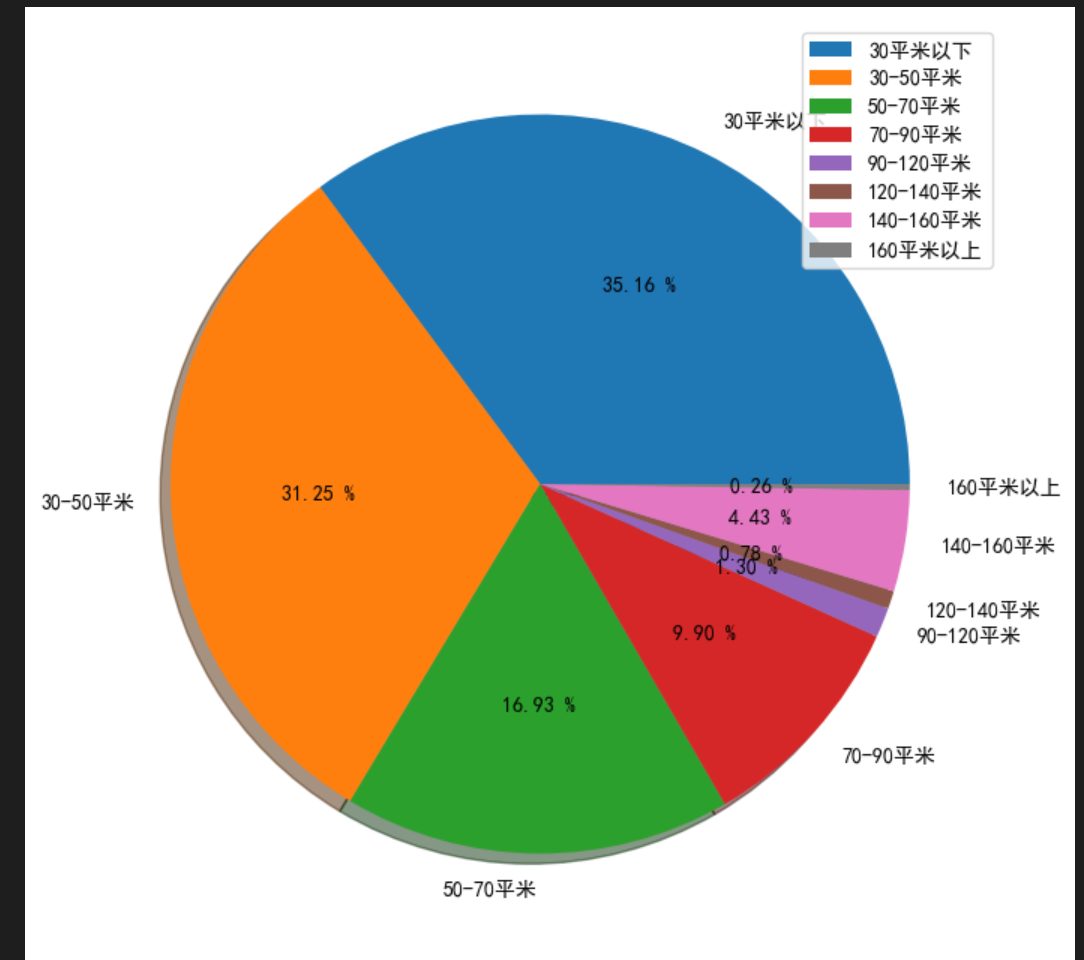


数据可视化: 

由于房源比较广泛,导致图较长

房间面积数据处理与可视化:





# 实验结果或结论

## 长沙租金分析

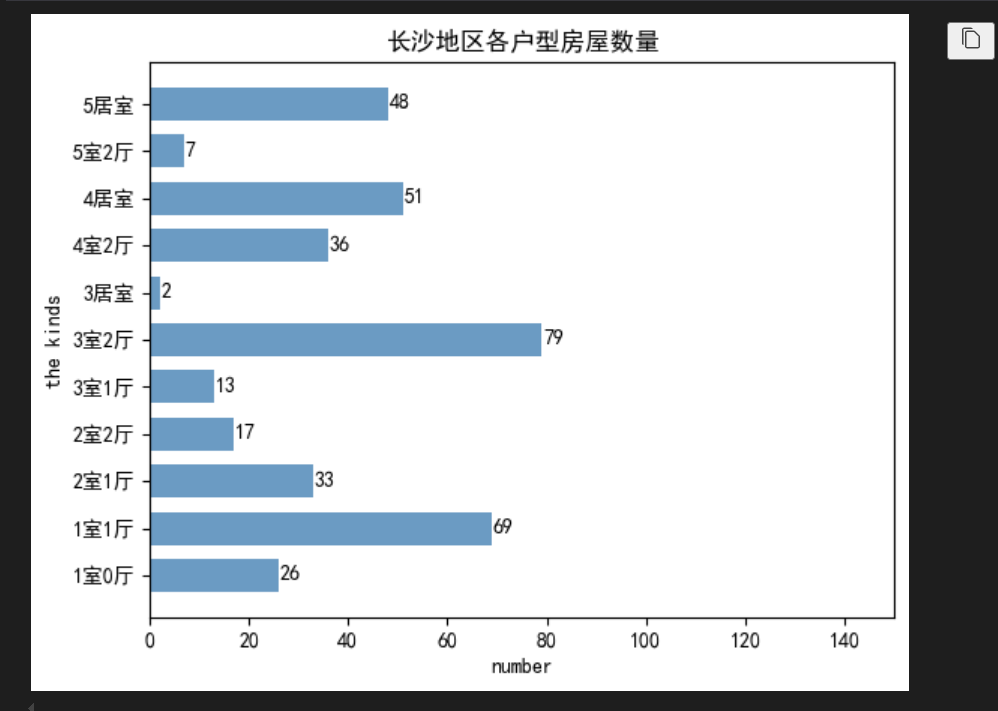
经过对租金可视化图标分析发现:长沙的租金面积比相较于其他一线城市而言,租金普遍较低,最高租金面积比位于五一广场,猜测位于市中心,实际上五一广场是长沙最繁荣的地区。对于房源的面积而言:出租大多数面积位于90一下,猜测应该属于公寓,猜想大多数租房的人群为参与工作的青年群体,有相当一部分(35%)出租面积极小,更偏向于为工作群体租用的房屋,可以推测长沙现在有相当一部分人口是刚参与工作不久的在长沙工作的群体,且大多数出租地区在查看地图后发现都位于较新市区,且一般为小型公寓。

## 长沙房源分析

对于可视化后的图表分析发现,主要出租的地区集中在河西步步高, 雷锋镇, 雷锋大道, 五一广场等区域,其中河西步步高的房源大多数为大面积,猜测应该是商铺出租,五一广场也类似,但五一广场是长沙最繁荣的地区,猜测出租一般面向高消费群体的临时住所与铺面出租。长沙出租的房屋大多数都集中在特定几个区域,其他区域出租房屋极少,且通过对地理位置的分析,发现老城区出租较少或者很少,新城区很多,认为长沙是发展活力很强的城市

## 长沙户型分析:

由于爬取数据的结构不规整,存在居室与室厅之分,如下图,我们可以认为3居室,4居室以及五居室都属于3室2厅及以下规模,分析数据可以发现,其实长沙大多数出租的房屋规模都偏小,但是3室2厅类房屋出租数量很大,小规模3室2厅以下的房源很多,受众趋向于为刚来长沙的工作群体准备,同时也存在部分为家长为孩子在长沙读书而临时租住的房屋需求



数据集以及分析文件:

