# 项目书

1. **项目选题**

**预期完成对哪类数据的爬取**

对链家网南昌市的二手房源各类信息进行爬取，具体内容包括：

1. 爬取基本发布信息：房主发布的售房信息，房源的位置信息与小区名称
2. 爬取房源具体信息：如户型，朝向，装修情况，位于的楼层，楼房类型
3. 爬取热度信息：关注人数，发布时间以及看房方式
4. 爬取价格信息：房源的房价，单价

**为什么要爬取此类数据**

1. 链家网内部的房源信息十分全面，包含了各城市业主对于售房信息的发布，数据源可信度及集成度都比较高，方便进行数据分析及实现选房的目标。
2. 爬取链家网南昌市的二手房源如房价，单价，楼房类型等数据并进行相应分析可以让我们了解到如今二手房的价格信息和业主对于他们持售房源的供求情况以及本地户型分布的一些规律。
3. 爬取链家网二手房源对应关注热度及发布时间可以让我们了解到二手房源的关注情况如何，这有利于我们查看南昌市本地的二手房源需求的部分情况。
4. 爬取链家网的二手房源可以推及到租房信息的爬取，对应的链家网网站还收集了南昌等各个城市的租房信息的渠道，因此对于我们日后租房提供了分析的渠道。

**数据可能来源于哪些网站**

链家网二手房主页 url:https://nc.lianjia.com/ershoufang/

1. **项目组成员及分工**

1、完成主要数据的爬取如完成BeatifulSoup解析网页的功能函数，线程池的多页面爬取功能函数，显示爬取信息并执行等待时间的功能函数及其他一些数据爬取工作。

2、完成部分数据分析以及数据可视化如jieba库分词和制作词云图

3、完成部分数据的爬取如存储数据写入csv文件的功能函数，随机UA伪装防止网站反爬及其他一些数据爬取工作

4、完成部分数据分析以及数据可视化如利用excel工具完成各类数据图表的制作

**三、使用到的技术**

数据爬取部分：

1、Requests

2、Beatifulsoup

3、ThreadPoolExecutor

4、csv

数据分析及可视化部分：

1. Jieba
2. pandas
3. Wordcloud
4. Excel工具

**四、实施过程**

首先，进行数据爬取部分的工作，具体包括：



1、爬取基本发布信息：房主发布的售房信息，房源的位置信息与小区名称

2、爬取房源具体信息：如户型，朝向，装修情况，位于的楼层，楼房类型

3、爬取热度信息：关注人数，发布时间以及看房方式

4、爬取价格信息：房源的房价，单价

其次，在完成上述数据爬取操作之后进行数据分析操作：

1. 对爬取得到的数据进行存储并简单的数据清洗等操作
2. 对于csv文件title列数据初步进行jieba分词提取中文高频词并制作词云图
3. 运用excel工具对于各个类型数据进行深层分析如：
4. 制作户型的词云图或者每种户型的直方图
5. 统计房源面积分布情况，并绘制密度图展示。
6. 利用折线图或柱形图展示关注人数趋势和发布时间分布情况。
7. 展示看房方式的比例，以饼图的形式呈现。
8. 分析房源的朝向、装修情况、楼层、类型等特征，绘制相应的饼图或条形图，显示比例和趋势。

最后，进行数据可视化汇总操作，完成以下下数据的可视化操作：

1. 业主售房公布信息标题的词云图
2. 户型频数直方图
3. 楼房装修情况饼图
4. 发布时间密度图
5. 看房方式汇总条形图
6. 房价区间频数条形图

**五、程序主要流程**

**数据爬取：**

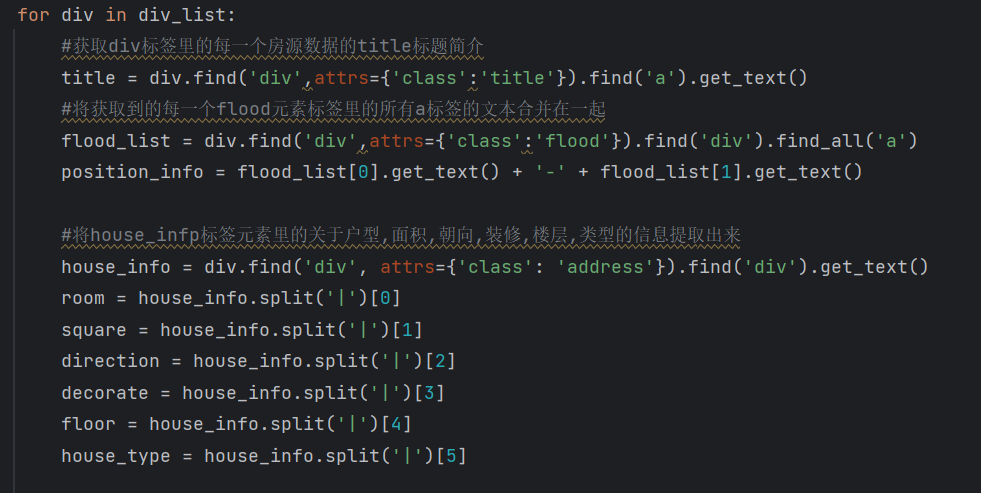
1、完成对链家网南昌市的二手房源各类信息的爬取：

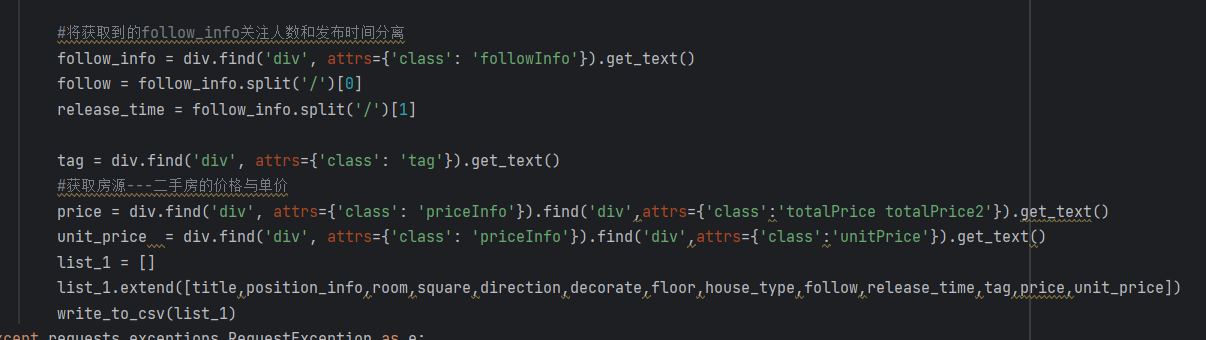
首先，我们定义了一个fun函数用来编写主要的爬取网页源代码数据的程序：设置随机user-agent来传参以及实例化一个BeatifulSoup类对象来爬取指定url的HTML网页元素标签节点内容，并从中提取所有class属性为’info clear’的div标签，因为我们想要获取的信息全在这个div标签中。





其次，再对网页源代码部分进行分析，我们发现关于业主发放售房标题，小区名称, 户型,面积,朝向,装修,楼层,类型, 关注人数, 发布时间, 看房方式，总价,单价等等信息都包含在下一层节点元素标签中，我们只需获取到它的属性(attrs)即可，获取属性之后再对应信息中进行字符串的分割操作，获取到一个个子数据，因此，我们只需对div标签通过调用find\_all,find,select等方法就能获取到关键数据。





我们可以从上述图片中发现对每一个div标签进行循环遍历，从中获取房源的标题标题简介、地理位置信息、房屋户型、面积、朝向、装修、楼层、类型等详细信息。

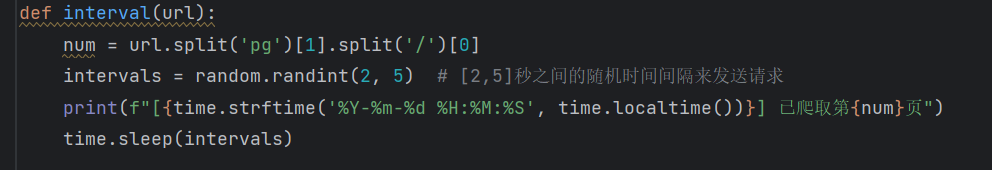
在标题标签的div标签中使用find函数找到第一个a标签来获取标题字符数据，使用flood类的div元素中的a标签列表来获取地理位置信息。这些信息都可以使用get\_text()方法提取。比如对于房源的详细信息，使用address这个class类的div元素来获取房屋信息字符串数据并使用split函数进行分割，再挑选其中的值并进行赋值给相应的变量。之后将爬取到的数据按顺序放置在list\_1列表中，以便后续进行写入csv文件操作。

我们可以在链家网二手房南昌市的网页中发现，一共具有100页的数据，每页有30条售房信息，但是我们以爬取单页面的过程来爬取网页的话效率太慢，因此我们通过线程池模块来进行多线程的操作，用一个循环生成了第1页到第100页的url，并将这些url存入列表变量url\_list中。然后使用线程池ThreadPoolExecutor创建一个数量为10的线程池，遍历url\_list中的每个url，提交到线程池中执行爬取任务main。这样可以同时执行多个网络请求，提高爬取效率。





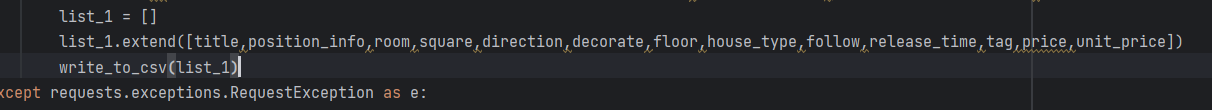
为了解决爬取页面并能够给用户展示爬取页面信息的提示，我们又加了一个显示时间戳和爬取页面提示的函数，这段代码从url中提取出页面页码接下来，生成一个随机的时间间隔intervals（范围在2到5之间），再通过time.sleep()等待这个时间间隔，以模拟人工打开页面的时间。最后，将已爬取的页面信息格式化输出，显示当前时间戳和页码num。



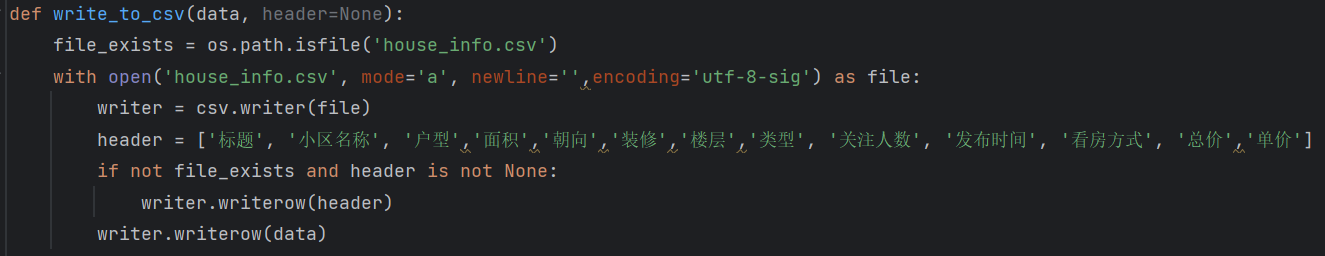
**数据处理：**

1. 对爬取数据进行存储排序：

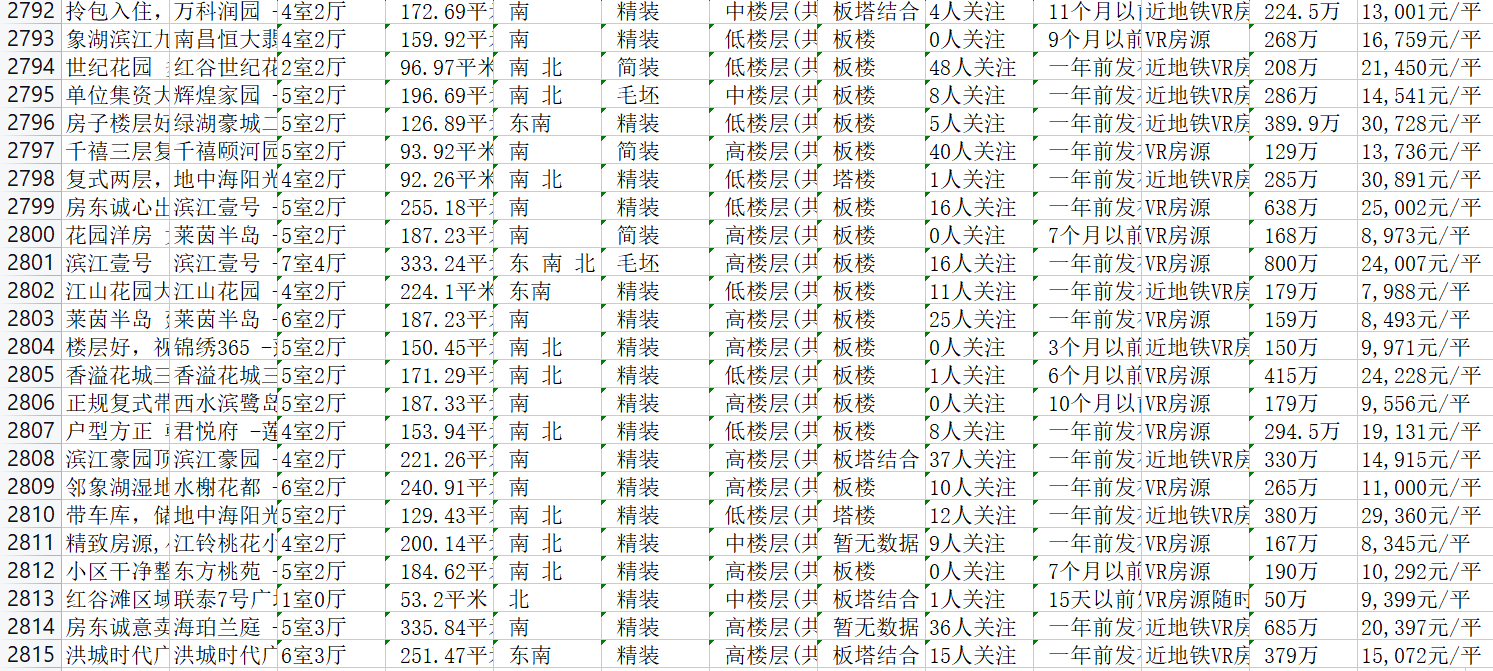
业主发布售房标题简介、地理位置信息、房屋户型、面积、朝向、装修、楼层、类型，房价等详细信息存储在list\_1列表中：



将数据进一步保存在csv文件中并在excel中清洗掉空白行和其它错误数据信息：



在完成数据清洗操作之后，可以看到关于此类售房数据共有2815条，接下来就是进行数据分析及可视化的操作。

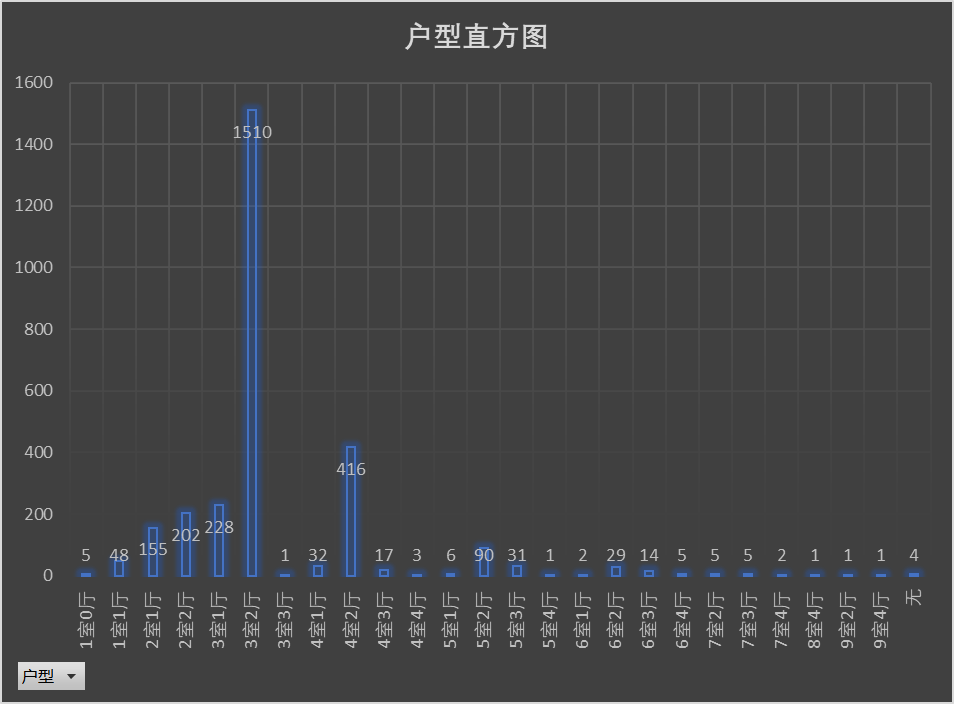


**数据可视化操作：**

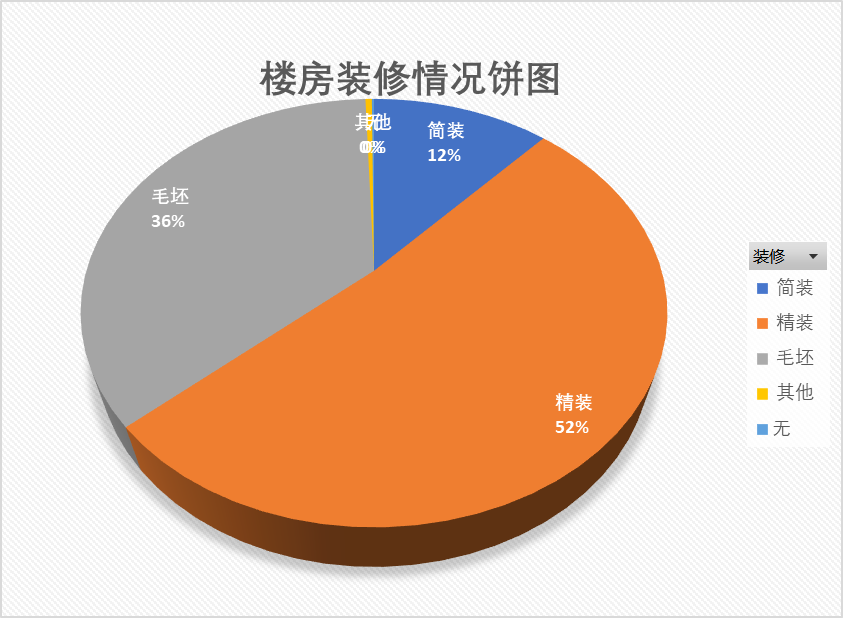
1、业主售房公布信息标题的词云图



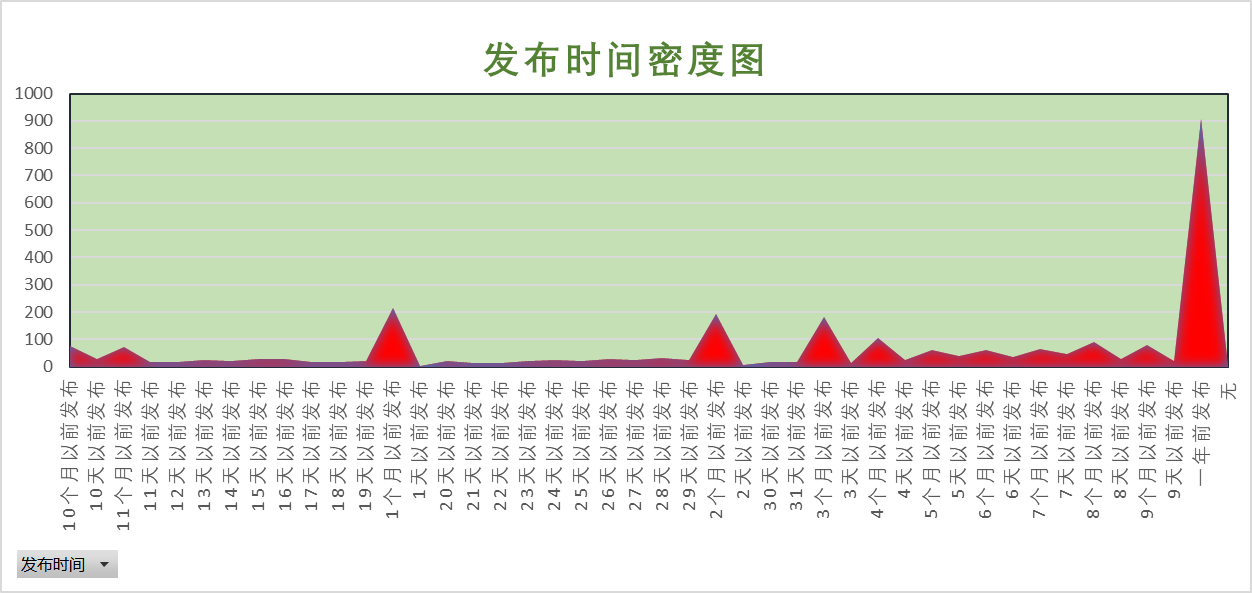
2、户型频数直方图

‘

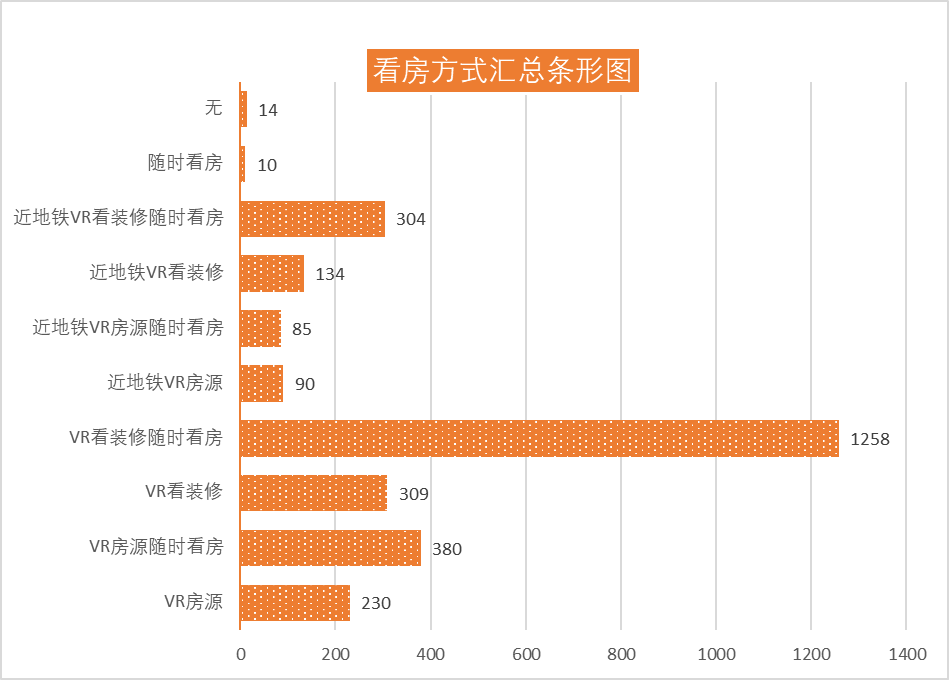
3、楼房装修情况饼图



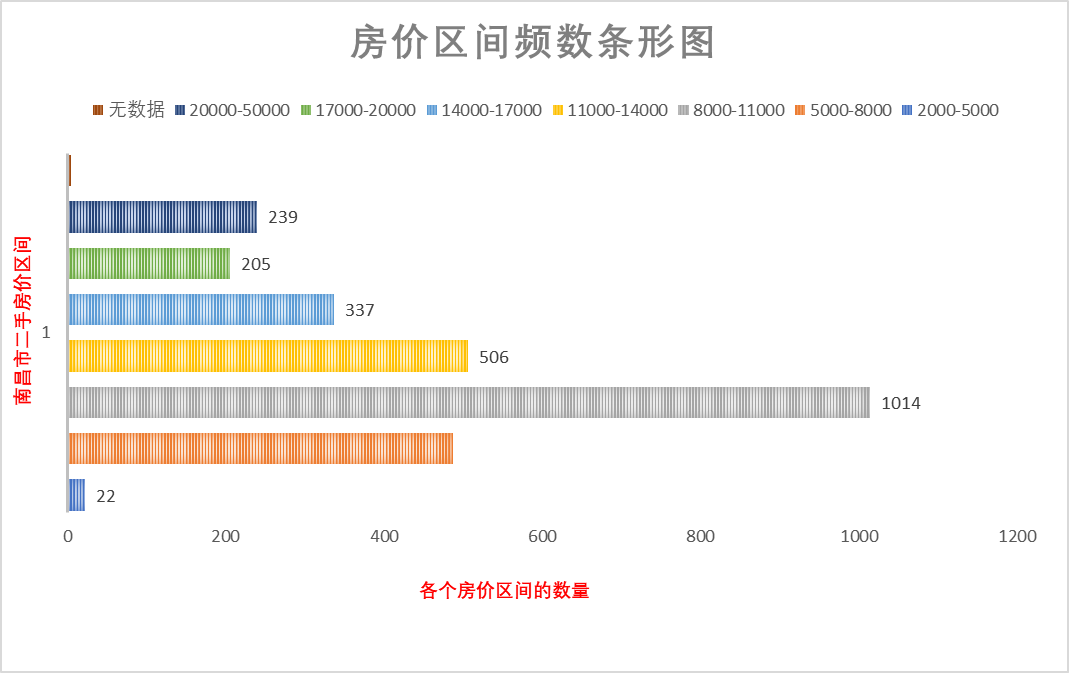
4、发布时间密度图



5、看房方式汇总条形图



6、房价区间频数条形图



1. **项目总结**

本次实验是基于链家网的基本发布信息，房源具体信息，热度信息，价格信息等数据的爬取及可视化分析，通过学习使用requests,bs4等模块完成对页面数据的获取及解析，通过使用Jieba库分词，pandas操作excel文件，Wordcloud等数据分析和绘图功能库完成数据的可视化给我们提供了能够深层次的分析数据的工具。而且我们还在此改进了我们对于链家网项目的爬取功能，比如增加了多线程同时爬取多页面的功能函数，这不仅提高了程序的整体效率，而且也让我们更好地了解了多线程编程的相关知识和技能，此外能够将爬取的页面信息格式化输出，显示当前时间戳和已完成爬取的网页页码，但同时我们的项目其实还存在着一些问题，比如：第一、该程序目前只能抓取链家网上部分的房源信息，因此无法满足一些特殊需求。针对这个问题，我们可以进一步调研链家网的反爬虫机制，以更好地抓取数据。第二，程序的可扩展性目前还不够好，该程序目前难以扩展到其它网页的信息，比如链家网不仅又二手房源的信息，还有租房，小区以及新房等房源信息，我们想要爬取这些信息还得重新编写我们的程序。第三，链家网站可能对爬虫行为进行限制，我们并没有进行测试过，我们设置的随机UA伪装可能反爬机制比较低级，可能会被识别为爬虫程序。

总之，这个链家网的爬取项目，不仅让我们更好地掌握了Python爬虫技术、html，数据可视化等知识，还让我们更好地了解了程序设计、项目管理等方面的知识。通过实践，我们不断发掘问题和解决问题的能力得到提升，也感受到了技术实践的乐趣和挑战。