**Final Project**

天津中德应用技术大学

**软件与通信学院**

****

**基于Python的天津二手房数据爬取及分析**

**班 级**

**课程名称 云计算与大数据实验**

**小组成员**

**小组成员**

**2023年6月27日**

# 正 文 部 分

**1. 项目概述**（概述你的报告选题背景、选题意义、主要内容、数据调研途径及手段）

（宋体，小四，间距为固定值22磅，蓝色说明部分请勿删除）

报告选题背景：

国内房地产快速发展，近几年来，公民越来越关心房价问题，2016年的中央经济工作会议明确提出“房子是用来住的、不是用来炒的”，展现了住房是民生之本。面对攀升的房价，人们买房的压力越来越大，为了缓解住房压力，人们会选择购买二手房。二手房的房价除了受到最基本的区域、总价、单价、建筑面积的影响,更受到了其他社会因素的影响。所以我们有必要对二手房价格背后的影响因素进行研究

选题意义：

在买房之前，人们会在相关网站上搜索自己想要了解的住房信息，但是面对庞大且来源广泛的网络数据，如何缩短网页下载时间，如何分析大量数据并找出有用信息，就需要用到网络爬虫技术。本文通过Python爬虫程序爬取链家网上所有的二手房数据，爬取的字段包括所在区域、面积、朝向、楼层、户型结构、装修情况、总价、单价等，并对采集到的这些数据进行清洗。最后对清洗过的数据进行可视化分析，探索天津二手房建筑面积、位置、户型等对二手房房价的影响情况。

主要内容：

本文利用Requests和Beautifulsoup库进行网络爬虫。Requests是一个基于urllib，并且是用Python语言编写的HTTP库。Requests比urllib更加方便，它可以向需要爬取的网页发送requests请求，然后自动爬取HTML页面并模拟浏览器请求提交，从而得到目标网页资源。beautifulsoup库作为Python的第三方库，可以解析HTML或XML文件中的数据，提取网页数据的Python库[4]，使其更加具结构化和格式化。

在进行数据分析时用到了Numpy、Matplotlib、Pandas这些第三方库。把这些库用在二手房信息爬取及分析中，不但能够节约项目成本，还能提高信息资源的利用效率。Numpy底层使用C语言编写，是Python的基础计算系统，通过提供Python对多维数组对象的支持即ndarray，直接对元素级别和数据进行操作，运算效率远高于纯Python代码。Matplotlib是一种可以生成高质量图标的基于Python的数据可视化模块，帮助用户在各种硬拷贝格式和交互式环境下进行MATLAB风格的绘图。Pandas作为数据分析包，经常在Python做数据分析时用到，能够快速地给Python提供结构化的数据资源。

数据采集途径：主要是通过python爬虫的方式访问链家的天津二手房

**2. 分工情况**（说明每个人的工作分工及完成情况）

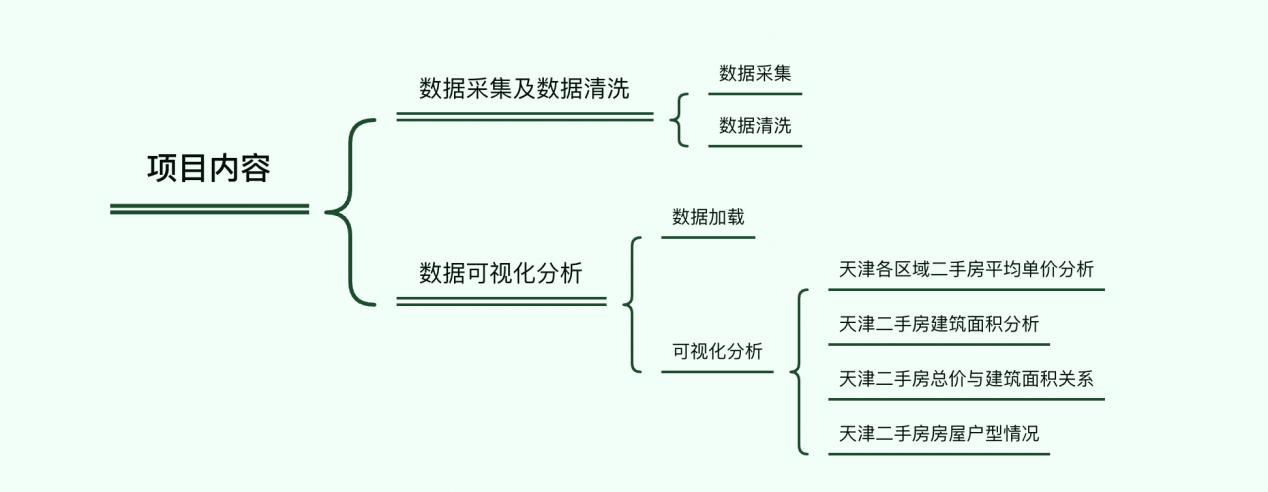
（宋体，小四，间距为固定值22磅，蓝色说明部分请勿删除）

xx负责数据采集，数据清洗

xx负责数据可视化分析

**3. 项目内容**（详细说明项目的内容、完整代码及解释、实现方法及结果显示，结论分析总结）

（宋体，小四，间距为固定值22磅，蓝色说明部分请勿删除）



1.1数据采集及数据清洗

1.1.1数据采集

⑴链家网网站结构分析

链家网(http://www.lianjia.com/)成立于2009年，是链家房地产经纪有限公司为发展主要业务领域——新房、二手房、租房而创建的线上房产交易平台。链家网上的二手房的主界面如图1所示。图1最上方的红色方框展示的是目前天津市内各地区关于二手房房源的位置区域，而中间的红色方框表示在售房源的总数量，图1最下方的红色方框其实是二手房房源信息的缩略图。由于链家网自带有百度地图的定位功能，所以这个缩略图涵盖了二手房房源页面的URL地址标签，能够直接得到小区的经纬度信息。在用户与网站的交互过程中，链家网能对这个过程产生的比如搜索数据、点击数据等大数据进行分析，掌握用户的看房偏好，为其推荐感兴趣的房源并且推荐公司好评优先、熟悉该小区的房屋经纪人对该用户进行服务。



（2）需要采集的网站数据

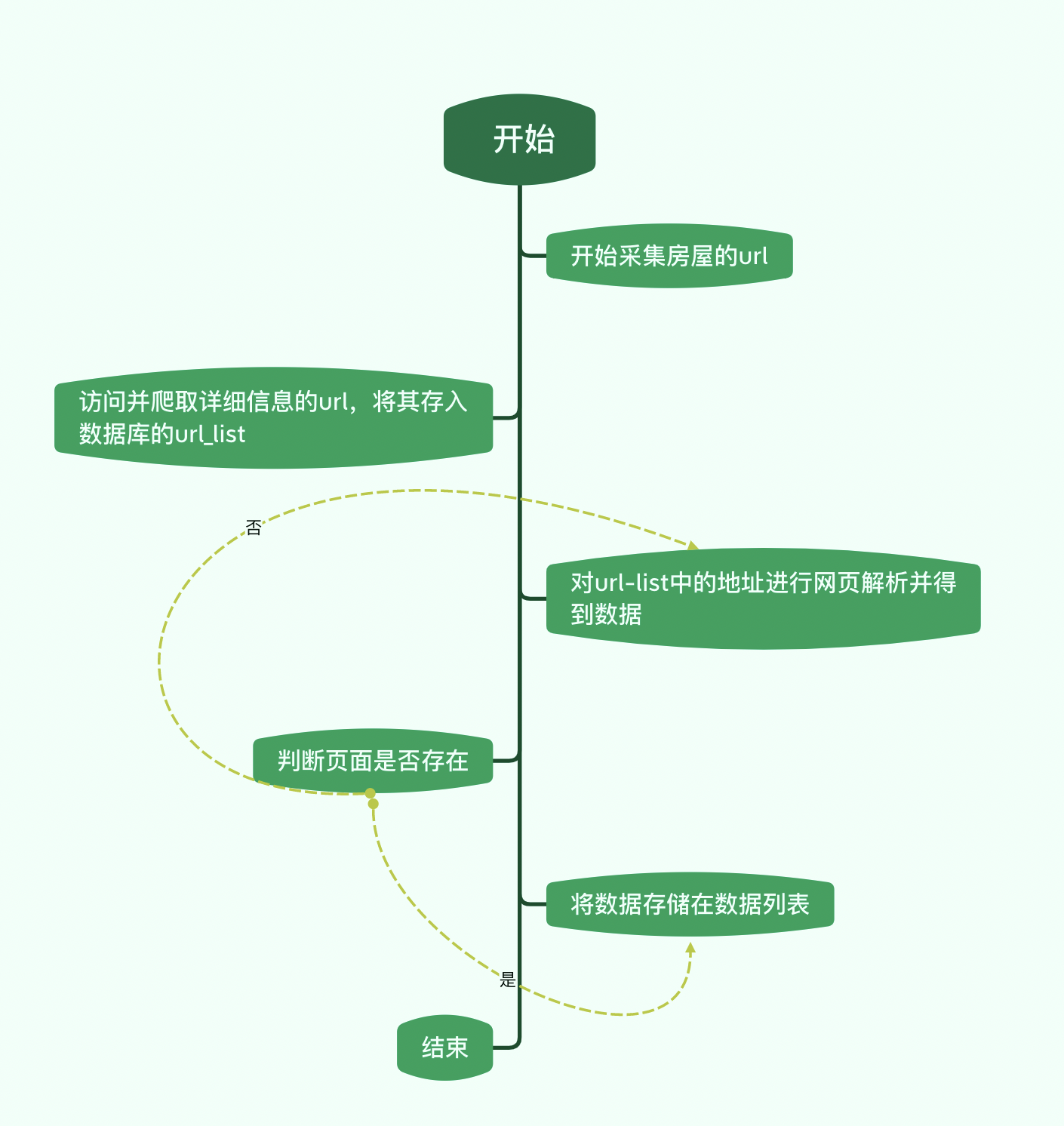
需要采集的房源数据主要包括二手房源的基本信息（总价、单价、所在小区名称以及所在区域）；基本属性（房屋户型、建筑面积、套内面积、房屋朝向、装修情况、配备电梯、所在楼层、户型结构、建筑类型、建筑结构、梯户比例）；交易属性（交易权属、房屋用途、产权所属、房本备件、挂牌时间、上次交易、房屋年限、抵押信息）。

图形用户界面, 应用程序, 表格

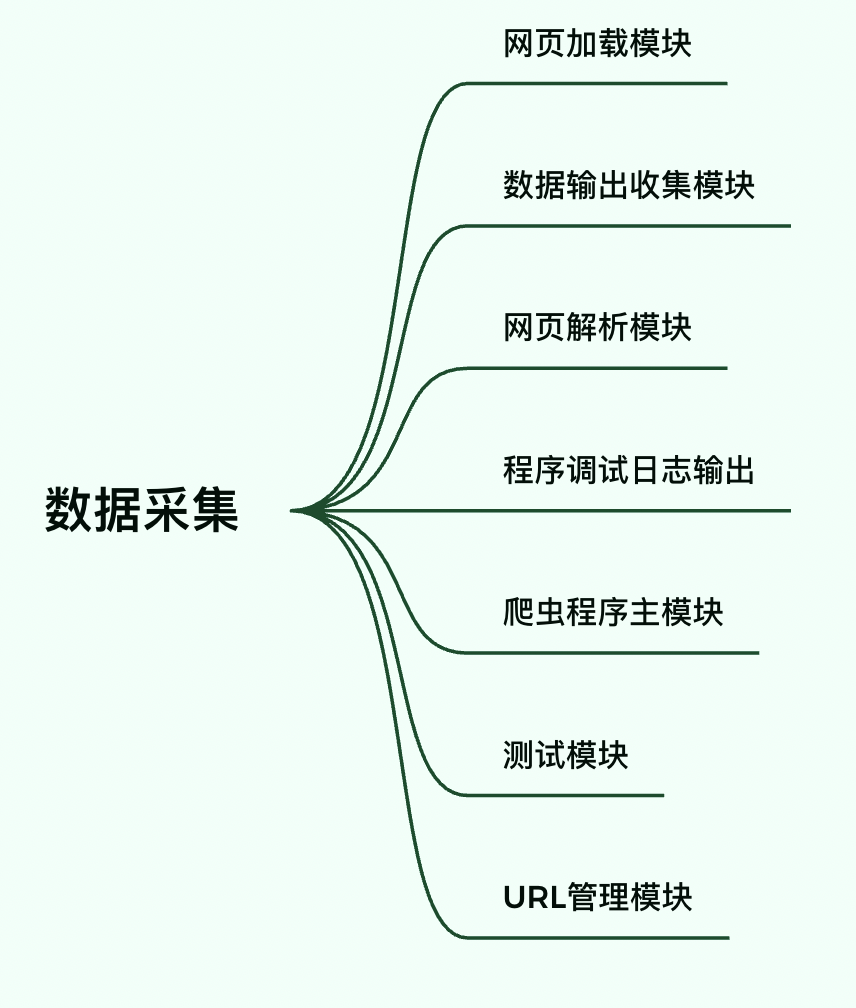
中度可信度描述已自动生成

（3）网站数据爬取流程

爬取链家网中天津市二手房的工作流程如图2所示。



（4）实现数据采集的具体代码模块分布，如图所示



网页加载模块





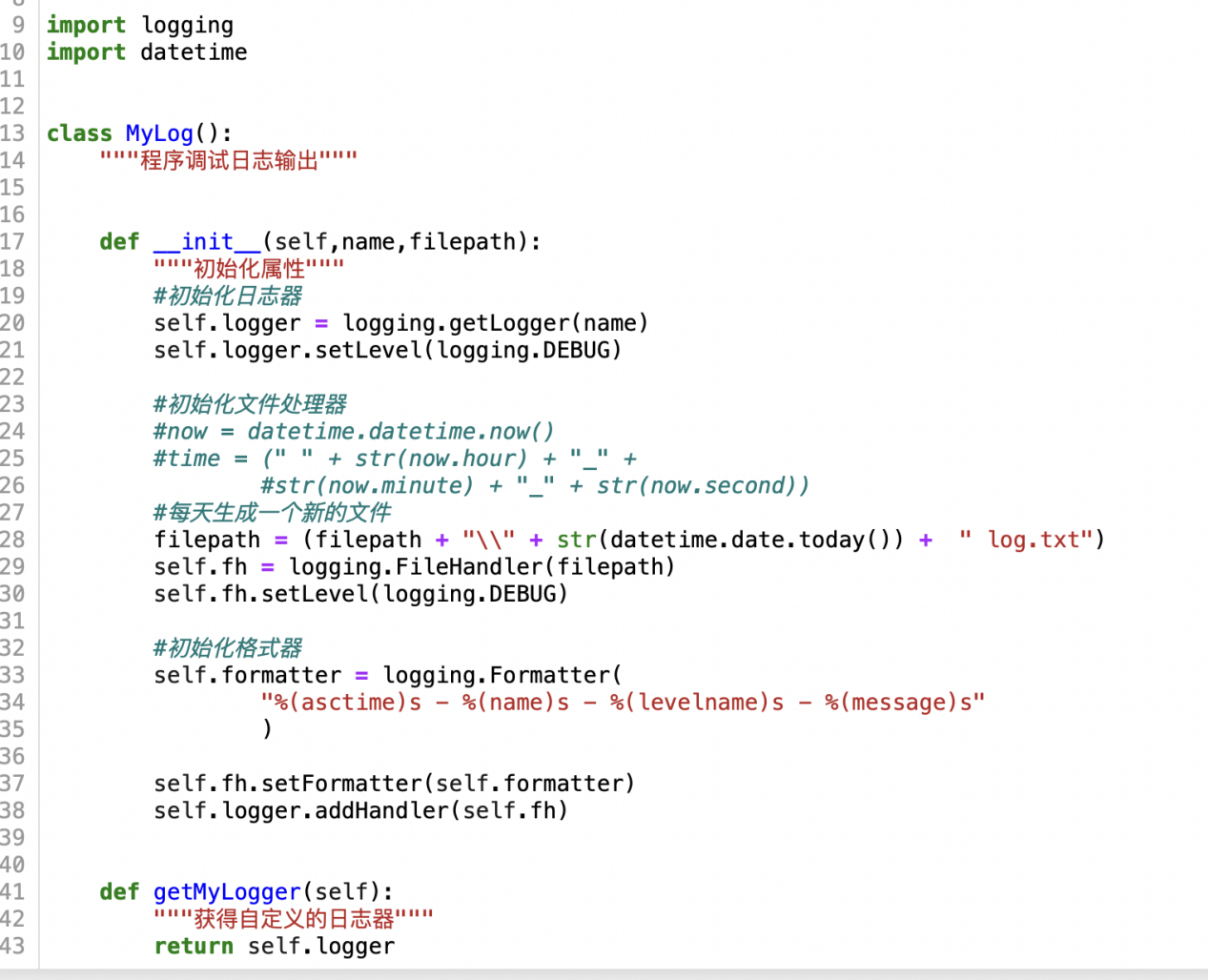
数据输出采集模块



网页解析模块



程序调试日志模块

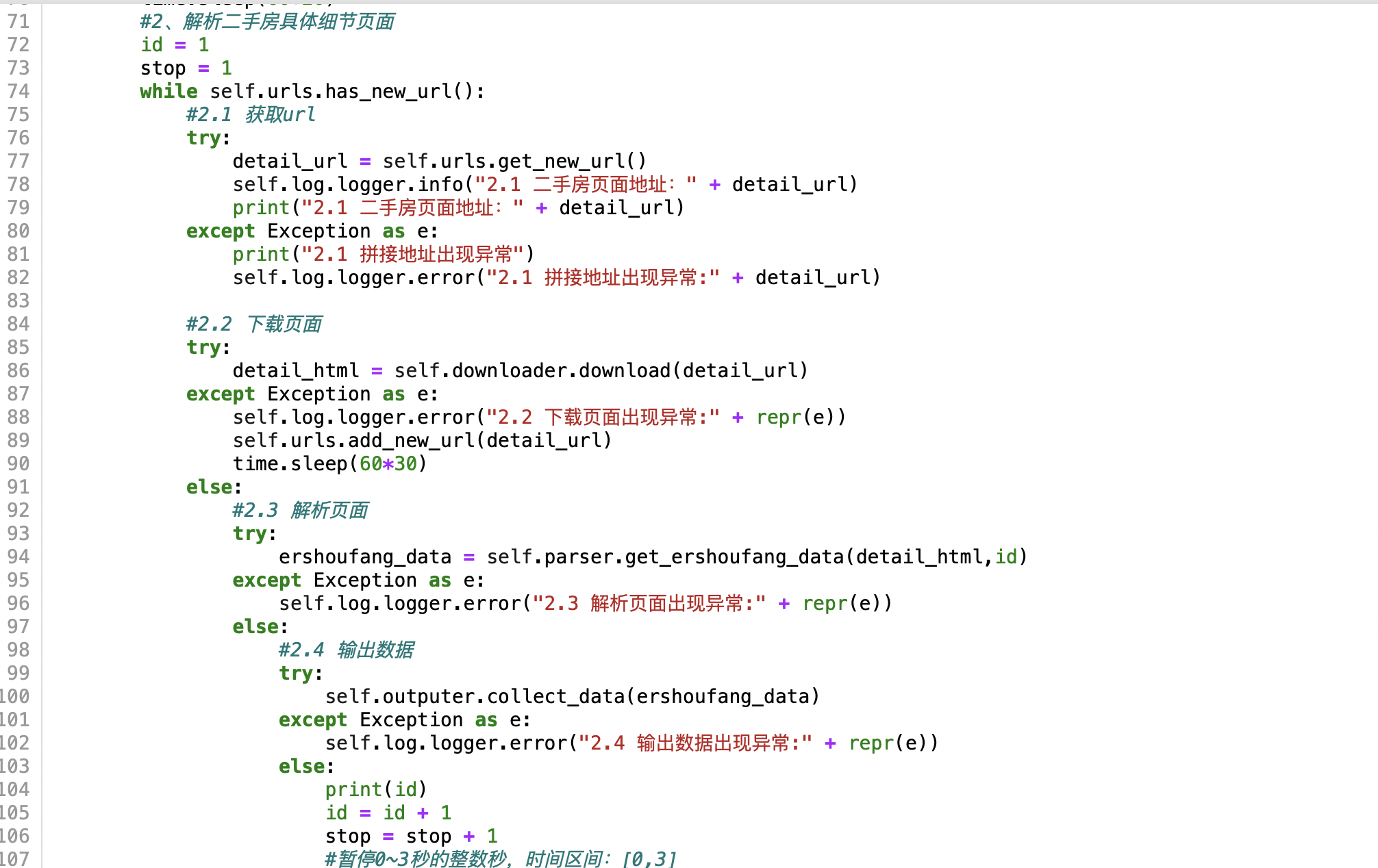


结果显示

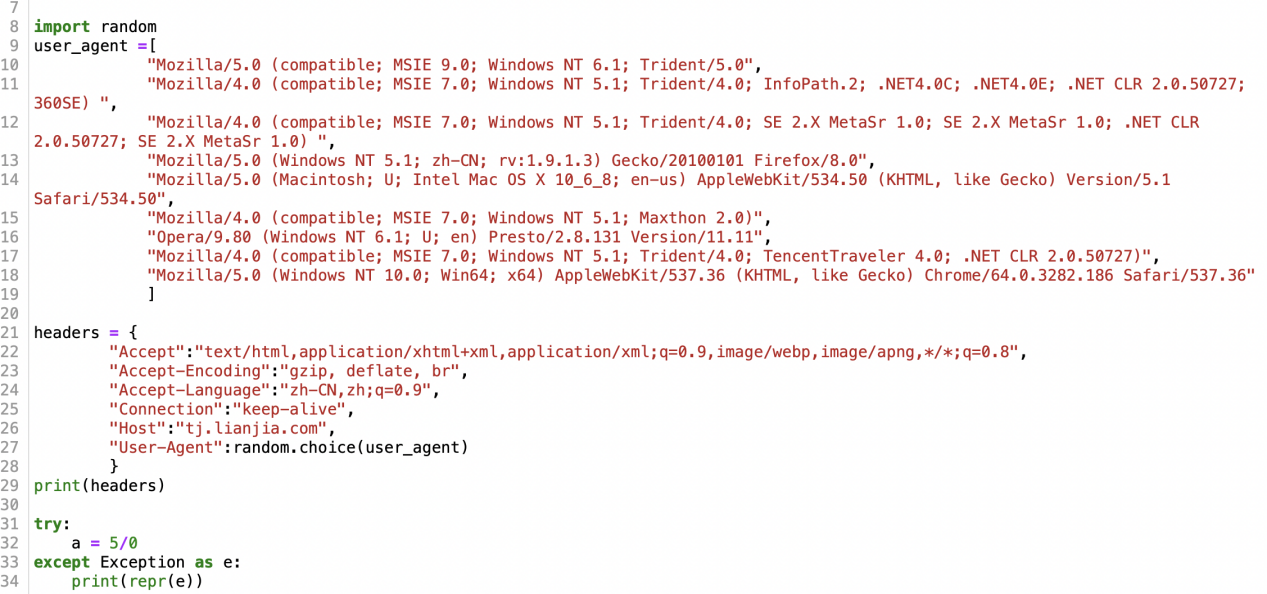


爬虫程序主模块





测试模块



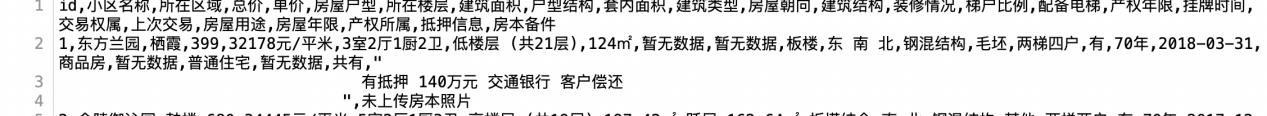
URL管理模块



1.1.2 数据清洗

数据清洗就是对数据文件进行重新监测，识别错误并纠正的过程，对所爬取的数据进行规范化操作，比如删除重复数据、处理无效值和缺失值等。原始数据通常来源于各种业务系统，只是由于主题一致而集合在一起，故难免会有冲突的数据和错误的数据，这些我们并不需要的数据称之为“脏数据”。而消除这些“脏数据”，使过滤后的数据具有一致性的过程就是数据清洗。

原始数据



清洗后的数据

电子计算器

中度可信度描述已自动生成

2.1数据可视化分析

数据可视化技术就是把数据以图形图表的方式呈现出来，它的目的是帮助人们更加方便地分析总结出数据背后隐藏的规律，为决策提供好的依据。可视化技术越来越受大众欢迎，由于它的直观性，这项技术发展迅速并成功取得相应的成就。本文主要对二手房房源信息进行分析，例如房源面积、户型、总价、单价、所在区域等。主要步骤有数据加载、数据转换、数据可视化呈现、数据规律总结等。

2.1.1数据加载

数据准备需要做很多工作，其中大部分来自于数据分析和数据建模，包括转换、清理、加载等。经过数据清洗后，完整的数据依然还保存在CSV文件中，这时候要将这些数据按照一定的要求加载到内存里。为了加载处理清洗过滤后的数据，可以使用Pandas所提供的Data Frame对象，Pandas主动将表格型的数据转换为Data Frame可以读取的函数。数据加载的过程中需要注意关于数据项的行列索引处理，还有对缺失值的处理，以及数据类型的推断和数据转换的问题。

2.1.2数据可视化分析

数据可视化呈现：选择适当的图表或图形类型来展示二手房房源信息。例如，可以使用条形图或饼图显示不同区域的房源数量，使用散点图展示房源面积和总价之间的关系，使用箱线图比较不同户型的房价分布等。可以根据需要选择最合适的图表类型。

详细展示如下

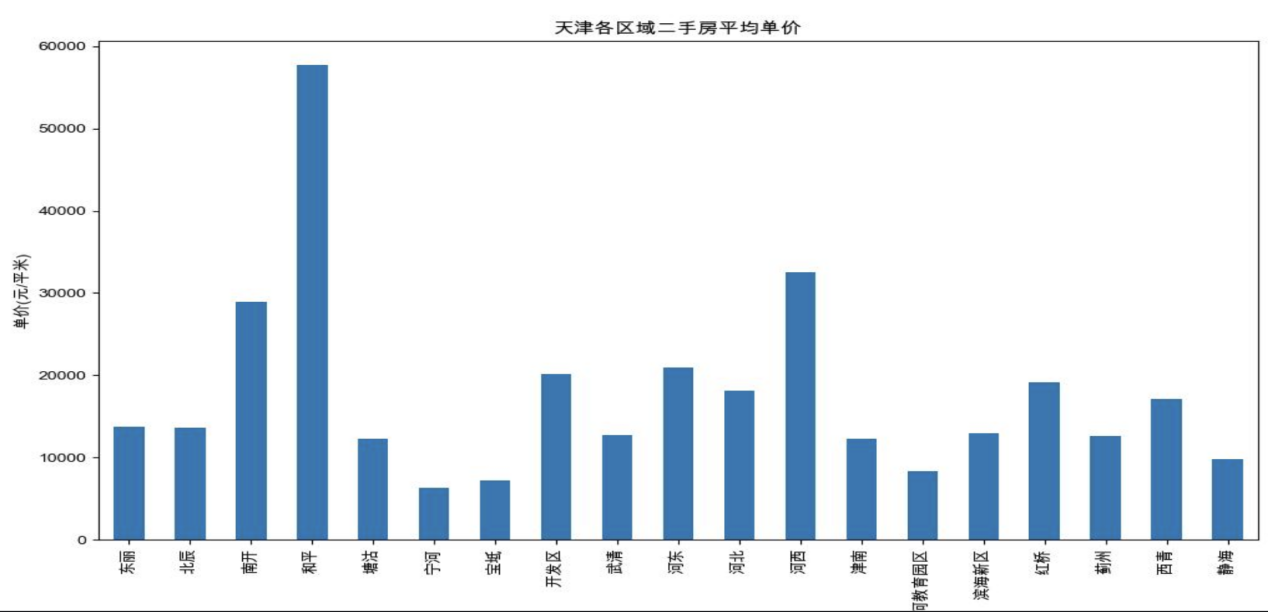
1. 词云图

通过天津二手房的词云图可以发现以下一些规律或趋势：



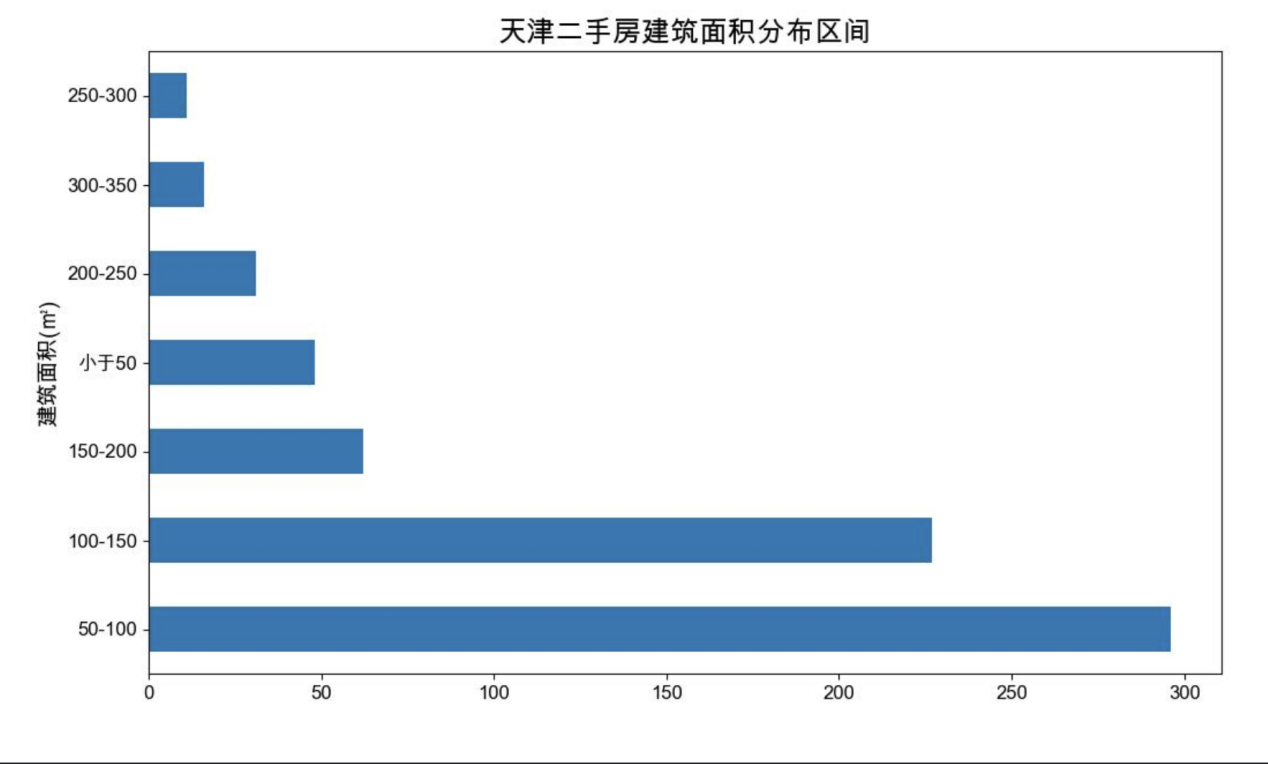
1. 天津二手房平均单价分析

天津市各区域二手房的平均单价柱状图，x轴是天津市各区域名称，y轴是单价（元/平米）。和平区还有河西区的平均单价最高，高达60000元/平米。可能是因为这两个区的地理位置特别好，作为中心城区的和平区，经济发展迅速，是中国东部地区的金融服务中心，房价一直持高，到现在已经是全天津房价最高的区域之一。而河西区是天津市的核心地段，也是国家重要的科技创新中心以及航运物流服务中心，河西区地理环境优越坐拥许多商场，聚集了休闲、餐饮与购物于一体的商业氛围，所以平均单价这几年来也一直位居高位。根据柱状图显示，不难看出天津市各个区域的二手房平均单价都在20000元/平米以上（除去有误差的六合区），还有看起来平均单价偏低的津南区，也比前几年翻了一倍，说明天津这几年的整体房价涨势迅猛。



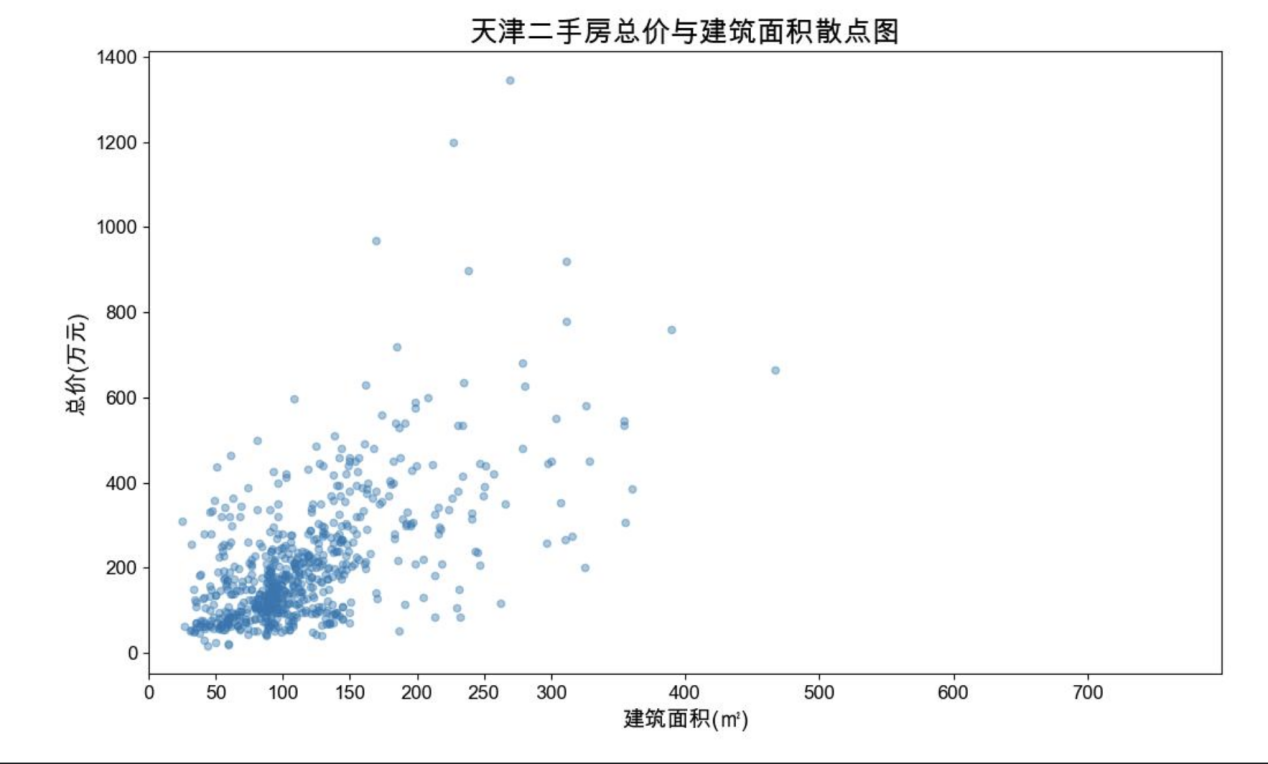
1. 天津二手房建筑面积分析

天津二手房建筑面积分布区间柱状图如图所示，x轴是房源数量，y轴是建筑面积。其中，建筑面积在50-100平米内的房源最多，超过一万多套，100-150平米内的房源位居第二，而250-300平米的房源最少。说明在天津二手房市场中，中小户型占据了主流。



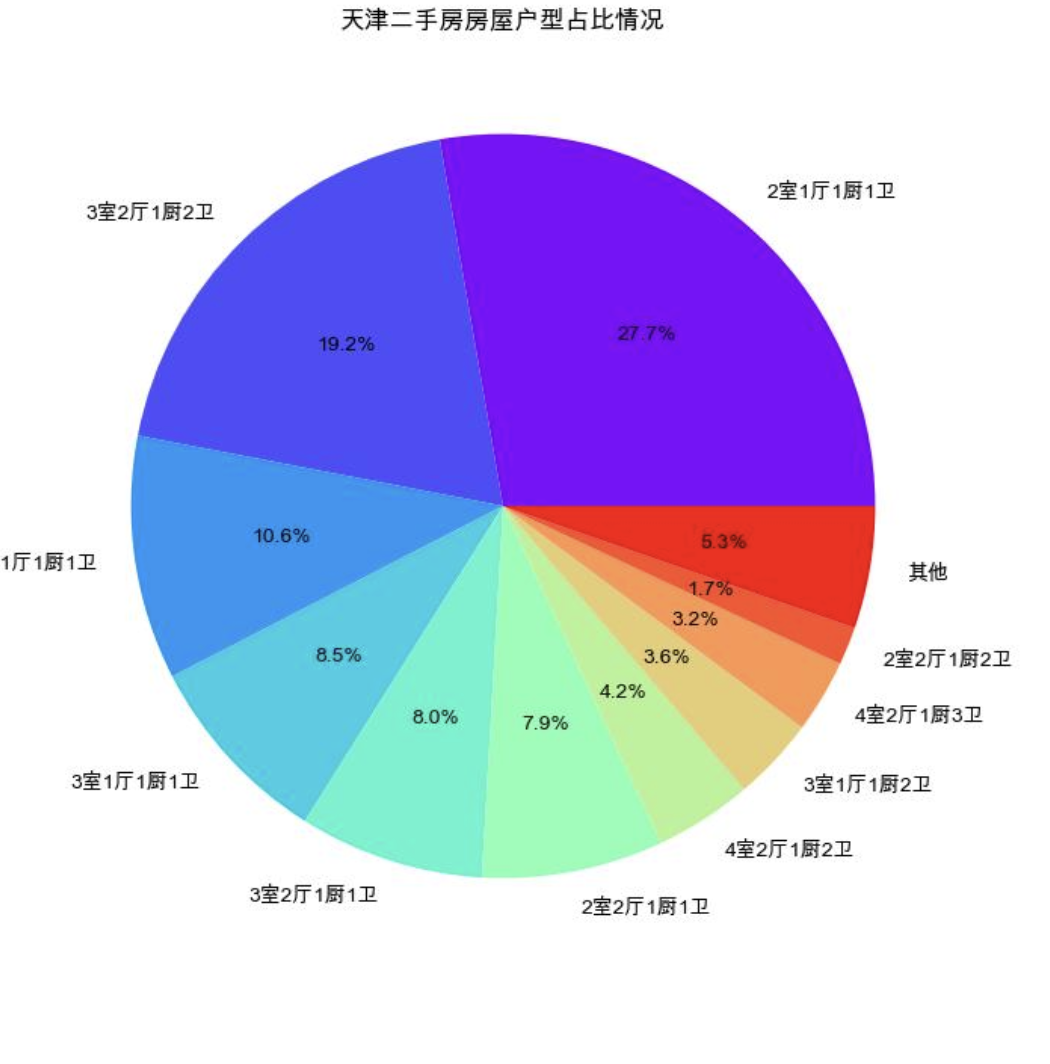
1. 天津二手房总价与建筑面积关系

天津二手房总价与面积关系散点图如图8所示，x轴表示建筑面积，y轴表示总价。我们从图8中了解到总价与建筑面积呈正相关关系。房屋建筑面积越大，价格也就越高，符合基本常识。总价0-200万元与建筑面积0-150平米这个区域内，数据点分布密集。

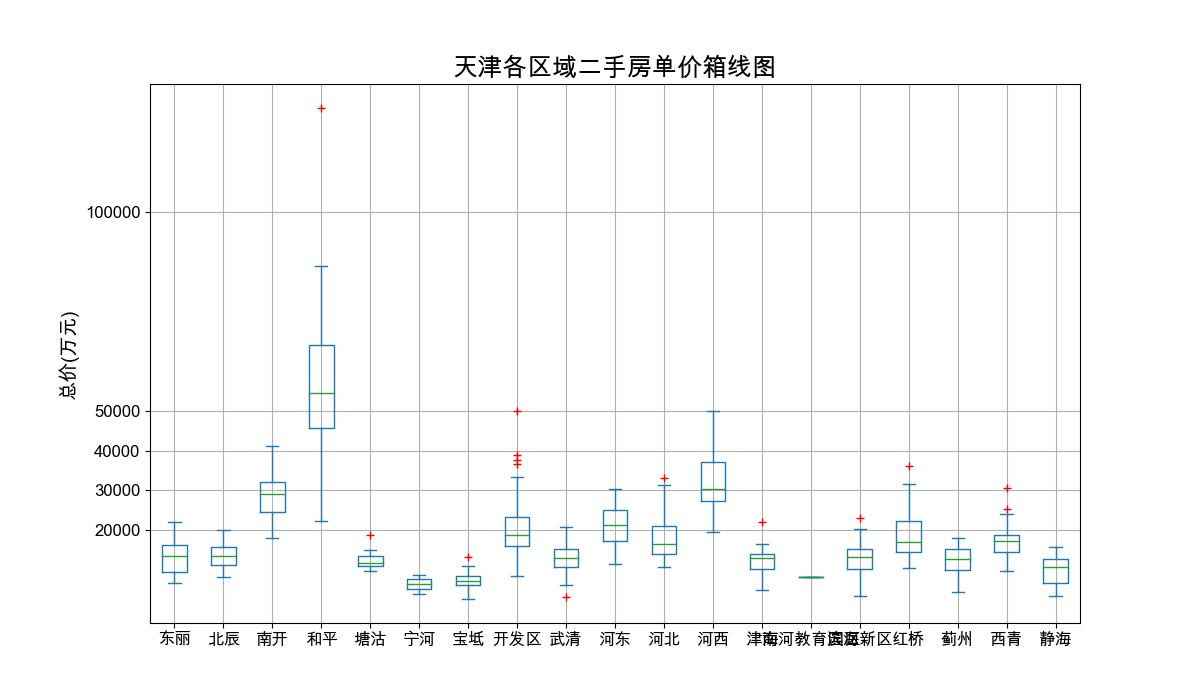


1. 天津二手房房屋户型情况

对天津二手房最常出现的户型进行计数并进行可视化分析如图9所示，天津二手房最常出现的户型为二室一厅，一卫的户型也不少。



1. 天津各区域二手房单价箱线图



根据箱线图的信息所示，我们可以看到，和平区的二手房单价中位数位于首位，数据的分散程度较高，但是在二手房单价中位数较低的一些区中，数据分散程度较低，说明二手房的单价对其价格的分散程度有影响。

**4. 收获与感想**（分别说明对本门课程及本次项目的收获及感想）

（宋体，小四，间距为固定值22磅，蓝色说明部分请勿删除）

在这门课程中我们学习了数据分析的第三方库，例如pandas与numpy等，所以我们用爬虫爬取了链家二手房网站，关于天津二手房的部分数据，对其进行了数据分析。

该论文将通过对天津二手房市场的详细分析，展示市场现状和趋势。数据的收集和处理将确保研究结果的准确性和可靠性。在分析中，将重点关注房价分布、区域差异和影响因素，以全面了解天津二手房市场的情况。最后，通过对未来趋势的展望和相应的建议，为购房者、投资者和政府决策者提供有益的参考，促进天津二手房市场的健康发展。

**5. 参考文献**（如果有，请列出；若无，则注明无）

（宋体，小四，间距为固定值22磅，蓝色说明部分请勿删除）

［1］王晓娟.重庆二手房价格影响因素研究［J］.廊坊师范学院学报：自然科学版，2019.3.

［2］刘航.基于Python的重庆二手房爬取及分析［J］.电脑知识与技术，2019.15（36）:6-7，17

［3］朱俊.二手房数据分析系统的设计与实现［D］.西南交通大学硕士学位论文，2017.

［4］王碧瑶.基于Python的网络爬虫技术研究［J］.数字技术与应用，2017.5:76-76

［5］罗博炜.基于Python的数据可视化［J］.信息记录材料，2019.

［6］姚冲，闭鑫业.基于集成学习的武汉二手房估价模型研究［J］.商讯，2019.

［7］王芳.基于Python的招聘网站信息爬取与数据分析［J］.信息技术与网络安全，2019.8.

［8］赵绿草，饶佳冬.基于python的二手房数据爬取及分析［J］.电脑知识与技术（学术版），2019.7:1-3

[9]戴瑗,郑传行.基于Python的南京二手房数据爬取及分析[J].计算机时代,2021(01):37-40+45.DOI:10.16644/j.cnki.cn33-1094/tp.2021.01.009.